Image Captioning with External Knowledge

Published by

LOT
Binnengasthuisstraat 9
1012 ZA Amsterdam
The Netherlands

phone: +31 20 525 2461

e-mail: lot@uva.nl
http://www.lotschool.nl

ISBN: 978-94-6093-426-1
DOI: https://dx.medra.org/10.48273/LOT0641
NUR: 616

Image Captioning with External Knowledge

Automatische Beeldbeschrijving met Externe Kennis

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
woensdag 5 april 2023 te 12.15 uur

door

Sofia Nikiforova

geboren op 28 augustus 1994
te Surgut, Rusland

Promotor:

Prof. dr. Y.S. Vinter Seggev

Copromotoren:

Dr. T. Deoskar

Dr. D. Paperno

Beoordelingscommissie:

Prof. dr. S. Zarrieß

Prof. dr. A.P.J. van den Bosch

Prof. dr. J.E.J.M. Odijk (voorzitter)

Prof. dr. C.J. van Deemter

Dr. A. Gatt

This research was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 742204).

Contents

Acknowledgements

In all the dissertations I have come across, the first thing I read was the "Acknowledgements". Whether I knew the author or not, their words of gratitude, love and relief have always brought me comfort: I, too, can probably make this journey. And now I am happy that I can write my own, saying thanks to those who have helped me on this long path.

First of all, of course, my promotor, Yoad Winter, and my copromotors, Tejaswini Deoskar and Denis Paperno. Yoad, thank you for giving me this amazing opportunity to work in your project and then, freedom to find my place within it. I admire your incredible academic intuition and how you could steer me in the right direction every time when I was lost or at a crossroads. Tejaswini, to me you are an example of tenacity, integrity and hard work. Thank you for the long productive discussions and always finding the time to help improve my countless drafts. Denis, without you all of this simply would not have happened. Thank you for that email a long time ago, which put in my mind the idea of doing a PhD in Europe and ultimately changed my life. And thank you for your never-ending positive attitude, inspiring scientific curiosity and many valuable insights you have offered throughout the years.

I would like to thank the members of my assessment committee: Antal van den Bosch, Kees van Deemter, Albert Gatt, Jan Odijk, Sina Zarrieß. I am honored that you agreed to read and evaluate this dissertation. Thank you, Rick Nouwen and Emiel van Miltenburg, for helping me translate its title into Dutch — only a few words but surprisingly tricky!

My deepest gratitude goes to the members of the ROCKY project: Yoad Winter, Tejaswini Deoskar, Denis Paperno, Joost Zwarts, James Hampton, Lasha Abzianidze, Imke Kruitwagen, Giada Palmieri, Sonya Ros, Myrthe Hemker. With so many different perspectives in the room, our conversations have always been extremely engaging. A special thank you for all the fun times at our ROCKY outings!

Thank you to the members of the NLP Reading Group: Lasha Abzianidze, Tejaswini Deoskar, Denis Paperno, Meaghan Fowlie, Kokos Kogkalidis, Pablo Mosteiro Romero, Aleksei Nazarov, Adriana Duarte Correia. I treasured our meetings, which provided an opportunity to discuss so many interesting papers (much needed especially when we were all stuck at home during the lockdown!) and to receive valuable feedback on my own work.

I am happy to have spent these past four years at the Institute for Language Sciences (formerly UiL-OTS — that is how it will live on in my memory anyway). Maaike Schoorlemmer, spasibo for your continuous support in all matters and for the 'Game of Trans'! Thank you, my dear PhD colleagues, it was a pleasure to get to know you and talk to you in the winding corridors of Trans 10, during lunches, at borrels, and so on and so on: Corentin Bourdeau, Emma Everaert, Alberto Frasson, Rachida Ganga, Mariano González, Na Hu, Shuangshuang Hu, Bambang Kartono, Anika van der Klis, Martijn van der Klis, Marta Massaia, Chou Mo, Victoria Reshetnikova, Nina Sangers, Tijn Schmitz, Iris Selten, Luana Sorgini, Lex Tavenier, Silvia Terenghi, Quy Ngoc Thi Doàn, Joanna Wall, Jan Winkowski, Yuan Xie.

I want to especially thank Jianan Liu — for being the best office mate, Kokos Kogkalidis — for your sense of humour and helping out when I was drowning in 'CUDA out of memory' errors, Sonya Ros — for your optimism, encouragement and our game nights, Giada Palmieri and Imke Kruitwagen — for walking this path alongside me from day one.

Finally, a huge thank you goes to all the people who kept reminding me that there is life outside of the university walls, and it is beautiful. My Expat volleyball teammates, thank you for the laughs and epic memories we have shared. Masha Kozlova, my longest and best friend, thank you for always being there for me, even when we are thousands of kilometers apart. Erwin, Ria en Coen, ik ben dankbaar en gelukkig dat jullie me hebben opgenomen in jullie familie. Bedankt voor alle gezelligheid! Мама, папа, спасибо вам большое за вашу любовь и поддержку. Где бы я ни была, что бы я ни делала, они дают мне силы идти вперед.

Bas, you deserve the most important acknowledgement. Dank je wel voor wie je bent en voor je eindeloze steun. Спасибо за то, что я никогда не сомневаюсь: ты будешь рядом.

CHAPTER 1

Introduction

1.1 Automatic image captioning

This dissertation is dedicated to image captioning, the task of automatically generating a natural language description of a given image. This task is inherently multimodal: an effective caption generator needs to both interpret the visual input (i.e., what is depicted in the image) and, based on that, produce a fluent textual output, the image description. Image captioning ties together vision and language, making it part of the broader study of grounding in Natural Language Processing (NLP), which explores connections between text and non-textual modalities (Chandu et al., 2021).

The structure of most modern image captioning models is an encoder-decoder pipeline, originally adapted from machine translation (Cho et al., 2014). Conceptually, image captioning can be interpreted as a form of “translation” from an image to its description (Vinyals et al., 2015). The basic architecture of an encoder-decoder captioning model is shown in Figure 1.1. First, the encoder deals with visual processing, creating an informative representation of the image. It often utilizes a Convolutional Neural Network (CNN) that is independently pre-trained on Computer Vision tasks, such as object detection or image classification, to extract the most important image features. Then, the decoder carries out the actual caption generation. It generally

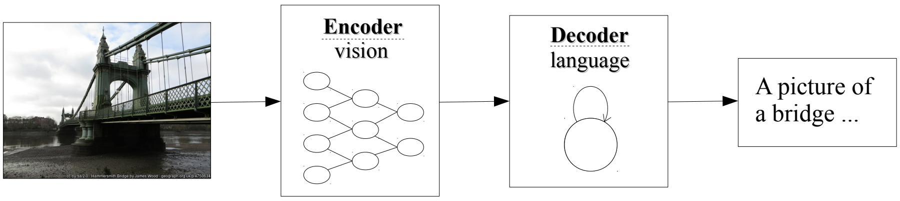


Figure 1.1: The basic encoder-decoder pipeline for image captioning; the illustration is inspired by Vinyals et al. (2015).

involves a language model (in the form of a Recurrent Neural Network (RNN) or a Transformer), which uses the output of the encoder as a prompt to produce the image description.

The training data for image captioning models consists of images with corresponding captions. The most commonly used datasets, such as MSCOCO (Lin et al., 2014) and Flickr30K (Young et al., 2014), contain crowdsourced captions, produced by human annotators who were specifically instructed to provide an exclusively visual description of all the salient objects in a given image. This ensures that images and captions contain the same information, different only in modality — ideal for training a captioning model to "translate" images into texts. In these datasets each image is paired with more than one caption, which allows the model to be trained on a greater variety of linguistic expressions: for example, as shown in Figure 1.2, the same scene can be equally correctly described as "a ginger cat in a green field" and "an orange and white cat in the grass".

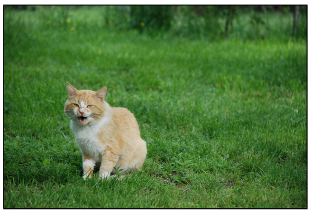
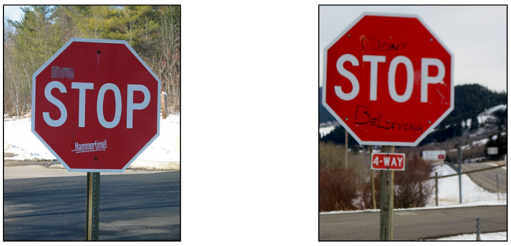


Figure 1.2: An image from the MSCOCO dataset and its five captions.

(1) a cat sitting in an open field meowing.
(2) a ginger cat sitting in a green field and purring
(3) an orange and white cat sitting in the grass meowing.
(4) a tan and white cat sitting in the green grass
(5) an orange and white cat is sitting outside in the grass.

However, captions created without any particular instructions may also include information that cannot be directly seen in the images. Figure 1.3 shows the captions of similar images in two datasets: MSCOCO and YFCC100M (Thomee et al., 2016).

The latter dataset contains captions that were created by the original authors of the photographs, without an intention of using them in Vision and Language applications.



(a) MSCOCO: "A large red stop sign on a pole."

(b) YFCC100M: "Stop sign outside Bozeman MT."

Figure 1.3: Similar images in two captioning datasets.

Both captions mention the primary subject of the photographs, the stop sign, but while the MSCOCO caption focuses on its visual characteristics, such as the size and the color, the YFCC100M caption makes a reference to its location, which cannot be inferred from the image itself.

The general goal of automatic image captioning, as in most machine learning tasks, is to train a model to imitate human behavior, in this case, by generating similar captions to what humans have produced for the same images. Trained on the datasets like MSCOCO, modern captioning models are generally able to produce well-formed and accurate descriptions of what can be seen in the images, but do not extend beyond that.

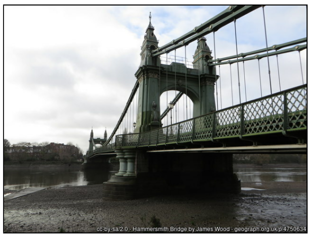

cc-by-sa/2.0 - Hammersmith Bridge by James Wood - geograph.org.uk/p/4750634

Figure 1.4:
Human: "Hammersmith Bridge. Designed by Sir Joseph Bazalgette and opened in 1887."
Automatic: "a picture of a bridge that is in the middle of the day"

For example, in Figure 1.4, a standard captioning model generates a description that refers exclusively to the visual content of the image: a bridge photographed in the daytime. On the other hand, the original human-written caption mentions the name of the bridge, its designer and opening year. To produce such details, information in the image itself is not enough: unless the object is particularly famous, it is hardly possible to determine its name, let alone the specific circumstances of its construction, solely from the image.

In order to automatically generate captions like the human-written ones in Figure 1.3b and Figure 1.4, the model needs to use *image-external world knowledge*[1]. This goes beyond the standard approach to image captioning as an "image-to-text translation" task and ties in with captioning viewed as grounded language generation. This way, the caption is grounded not only by the image, but also by the contextually relevant non-visual data. Exploring ways to enrich image captioning with external knowledge is the main focus of this dissertation.

1.2 Captioning with external knowledge

The goal of this dissertation is to develop a new method of incorporating image-external knowledge into an otherwise standard image captioning pipeline. Such a task involves several challenging steps. The relevant external knowledge needs to be (1) identified and extracted, (2) properly represented, and (3) effectively integrated into the captioning process. Then, once the captions are generated with references to external knowledge, (4) the evaluation procedure needs to assess whether these references are accurate and fitting to the images.

1.2.1 Identification of relevant knowledge

The first step is identifying which part of the world knowledge available in external resources would be useful for captioning a given image. Taking Figure 1.4 as an example, it is fairly obvious that an arbitrary general fact, such as "the Earth is the third planet from the Sun", is not pertinent to the image and should not have an effect on its caption. Even facts about bridges, although one is depicted in the image, are not necessarily useful, e.g., "the oldest existing bridge is the Arkadiko Bridge in Greece" or "types of bridges include, among others, beam bridge, arch bridge and suspension bridge". It is important to identify information that is directly relevant to the specific image, such as, in this case, facts about Hammersmith Bridge in particular.

In our approach, this task is carried out by a **contextualization anchor**. It is an element of image-related data that is used to retrieve real world entities and facts that are likely to be useful for captioning a given image. Importantly, the anchor is separate from the image itself and can be non-visual. In previous research, the

[1] In the examples in this section, for illustration purposes, we used captions influenced by geographic knowledge in particular: information about real world objects located in and around the image. This type of knowledge will also be prominent further throughout this dissertation.

connection to external knowledge was often established through object detection or image classification (Mogadala et al., 2018; Zhou et al., 2019; Huang et al., 2020; Bai et al., 2021), leaving unexplored the potential benefits of utilizing the associated non-visual data. For example, certain elements of image metadata, such as the coordinates of its location or the date and time of its creation, can be used as an anchor, since they provide information about the circumstances in which the image originated and thus can help identify relevant entities and events.

Another important aspect is the selection of an external knowledge resource, and possible domain restrictions on the types of knowledge to be extracted. Depending on the goal and conditions of captioning and which information is considered to be of interest, the extracted knowledge can include, for example, only historical facts ("Hammersmith Bridge was opened in 1887"), or only architectural facts ("Hammersmith Bridge is a suspension bridge"), or the physical characteristics of the objects ("the total length of Hammersmith Bridge is 213 meters"), etc. In this dissertation, we utilize geographic knowledge from OpenStreetMap (`https://www.openstreetmap.org/`) in Chapters 3 and 4, and open-domain encyclopedic facts from DBpedia (Auer et al., 2007) in Chapters 4 and 5.

1.2.2 Knowledge representation

The retrieved knowledge needs to be represented in an efficient and informative way for further use by the captioning model. The specifics of this process depend on the original format and content of the knowledge: different strategies may apply to, for example, free-form text from Wikipedia and structured database entries. In our approach, the knowledge is organized into two data structures: the entity context and the knowledge context.

Real world entities that are retrieved through the contextualization anchor constitute the **entity context**. This is the part of the general image context that consists of the entities relevant to the image, which may or may not be depicted in it. The term 'entities' here does not refer solely to 'named entities', as not every entity is necessarily associated with a proper name. For example, if a map provides information about a certain convenience store in the vicinity of the image location but does not actually name it, this store is still an entity that can be included in the entity context. Instead of a name, it can be associated with another textual label, such as simply "convenience store".

The entities should be encoded in a way that facilitates their correct use in the captions. Unlike regular vocabulary words, entity names (or other labels) are usually not represented well by their distribution patterns in textual corpora. Rare ones might not even be present in the corpora at all, and even if they are, the particular contextual relation linking them to the image would then be lost. Taking geographic entities as an example, for the model to produce "outside Bozeman MT" for Figure 1.3b, the encoding of 'Bozeman' in the entity context should reflect that it is some type of settlement and its relative location with respect to where the photograph was taken.

The **knowledge context** extends the entity context with various facts about the entities in it, providing a wider variety of external data to influence caption generation. For example, if the entity context contains an entity 'Hammersmith Bridge', the knowledge context can include the fact that it carries the A306 road or that it was designed by Joseph Bazalgette. The entity context and the knowledge context together are intended to reflect important aspects of relevant world knowledge that humans might have when they are describing a given image.

1.2.3 Incorporating knowledge into caption generation

A captioning model with a standard architecture as shown in Figure 1.1 only processes the image as input and, subsequently, bases the caption generation only on the image encoding. If the captions in its training data contain references to image-external knowledge, the model learns to produce similar expressions as well, but since this kind of information cannot be inferred from the image alone, the generated facts are likely to be incorrect or irrelevant to the image (the phenomenon commonly referred to as "hallucination" in natural language generation (Ji et al., 2022)). We propose to use external knowledge explicitly as an extra input for the captioning model. The model architecture needs to be modified accordingly, in order to allow the added data to inform the captioning process. The goal of the resulting knowledge-aware captioning model is to generate captions that are influenced by the relevant external knowledge and possibly include explicit references to it.

Thus, we integrate both entity and knowledge contexts into the captioning pipeline as extra sources of information for generating the caption, alongside the image itself. The generation module in the decoder is specifically adapted to be able to produce not only the visual description but also the entity names and facts related to the image.

1.2.4 Evaluation of knowledge-aware captions

In standard image captioning, evaluation is typically based on comparing the automatically generated captions to the ones that were produced for the same images by humans (Stefanini et al., 2021). If they are sufficiently close, it is taken as an indication of the high quality of the captioning model. Evaluation methods that rely on comparison to reference texts are generally common in natural language generation (NLG) (Gatt and Krahmer, 2018). One downside of such methods is that reference texts cannot cover all possible correct variations of the output. In image captioning, for example, there are always multiple different ways to describe an image, and not matching the human-written caption does not necessarily make the model-generated description incorrect (this especially affects datasets with only one caption per image). This and other disadvantages of the widely used reference-based metrics have been extensively discussed in the NLG literature (Novikova et al., 2017; Mathur et al., 2020; Reiter and Belz, 2009; Fabbri et al., 2021; Liu et al., 2016).

Although we do not aim to solve all the problems associated with the standard metrics, this dissertation addresses one particular aspect of evaluation that they are not well-equipped for: assessing the *factual accuracy* of the generated captions. Producing external knowledge references in a caption is only justified if they are actually correct and relevant to the image. Here it is especially important to extend beyond a comparison to the human-written captions. While they are considered to be the ‘ground truth’, meaning that information in them is assumed to be correct, the absence of certain information does not indicate that it is false. For example, if an automatically generated caption for Figure 1.4 stated that the bridge crosses the River Thames, it would not be a mistake, even though it is not mentioned in the original human-written caption. A reliable evaluation procedure focusing on factual accuracy should utilize the available image-external data directly. In this dissertation, we pay special attention to evaluating the veridicality of image-external knowledge in the automatically generated captions.

1.3 Models and datasets

This dissertation presents three image captioning models enriched with external knowledge. These models are developed with varying degrees of specificity in the application of our approach. We start from the domain-specific problem of generating references to the geographic context of the image, then expand to broad encyclopedic knowledge about geographic entities, and finally, to knowledge-aware captioning for diverse images in a more general domain. Table 1.1 provides an overview of the models, including the datasets they are trained on and the high-level descriptions of what constitutes the main components of our approach in each of them.

	Geo-aware (§3)	**Knowledge-aware (§4)**
Dataset	GeoRic	K-GeoRic
Contextualization anchor	Image location metadata	Image location metadata
Entity context	Geographic entities	Geographic entities
Knowledge context	—	DBpedia facts about geo-entities

	News-knowledge-aware (§5)
Dataset	NYTimes800k
Contextualization anchor	News article
Entity context	Named entities in the article text
Knowledge context	DBpedia facts about named entities

Table 1.1: Overview of the models presented in this dissertation.

The first model, introduced in Chapter 3, aims at geo-aware image captioning — an application of our approach to the geographic domain. In this model, the coordinates of the image location are used as a contextualization anchor, which helps retrieve a set of geographic entities around the image. The retrieved entities make up the domain-specific realization of an entity context, a geographic entity context, which is further integrated into the captioning pipeline to inform caption generation. The model is trained on a new GeoRic dataset, which we also present in this chapter. This dataset contains naturally produced captions with many references to the geographic context of the images, similar to "outside Bozeman MT" in Figure 1.3b. We evaluate the trained model with a particular emphasis on the correctness of the generated geographic references.

The second model, introduced in Chapter 4, builds upon the previous one, utilizing the same anchor and entity context, but expands to a wider variety of external data via the knowledge context. This knowledge-aware captioning model is trained on the K-GeoRic dataset, with captions that include different kinds of encyclopedic facts about the geographic entities in the images, such as "designed by Sir Joseph Bazalgette and opened in 1887" in Figure 1.4. We conduct a specialized factual accuracy evaluation, which shows that the knowledge context helps the model generate captions with facts that are both correct and relevant to the image.

The third model, introduced in Chapter 5, extends beyond the geographic domain and thus provides evidence for the generality of our approach. This model is trained on a qualitatively different image captioning dataset, NYTimes800k (Tran et al., 2020), which contains news images with original captions, extracted from New York Times articles. The model architecture is largely the same as in the previous chapter, with additional generalization to an entity context that includes named entities of different types (not only geographic), collected from the news article, which serves as a contextualization anchor. The entity context and the knowledge context of related encyclopedic facts allow the model to produce informative captions with accurate image-external knowledge.

1.4 Contributions

To summarize, the contributions of this dissertation are as follows:

- A new approach to image captioning that incorporates relevant image-external knowledge. Its effectiveness for generating contextualized and informative captions is demonstrated by multiple experiments, using standard captioning metrics as well as custom metrics we introduce to evaluate factual accuracy specifically.

- The GeoRic and K-GeoRic datasets of images paired with naturally created captions and geographic metadata. Both datasets contain captions that are highly contextualized and include many references to external knowledge: purely geographic in GeoRic and diverse encyclopedic facts in K-GeoRic. This is in stark contrast to the commonly used image captioning datasets that avoid references to image-external data by design.

    - The dissertation also includes a novel systematic analysis of a sample of existing datasets (Section 2.2.2.3), showing that naturally created captions have a larger, more diverse vocabulary than the crowdsourced ones. They are also much more likely to contain information that cannot be inferred from the image alone, which motivates the use of such captions in our research.

- Three image captioning models that present different practical applications of our general approach: geo-aware captioning, knowledge-aware captioning in the geographic domain, and news image captioning with external encyclopedic knowledge.

1.5 Dissertation outline

Chapter 2 contains an overview of related work in image captioning and existing approaches to world knowledge integration into language models. In relation to image captioning, we discuss the state-of-the-art of the field and a common limitation of neural-based captioning models, hallucination in the generated captions. We describe the widely used datasets (highlighting the contrast between the ones with crowdsourced captions and naturally created captions) and evaluation metrics. We then review the most relevant previous research on enhancing language models by incorporating world knowledge from external databases.

As discussed in detail in Section 1.3, Chapters 3, 4 and 5 present three models that are developed based on our general approach to image captioning with external knowledge. Chapter 3 introduces a captioning model that focuses on geographic knowledge integration. Chapter 4 describes a model that incorporates diverse encyclopedic facts about geographic entities. Chapter 5 explores an application of our approach to the domain of news image captioning.

Chapter 6 contains a summary of the contributions of this dissertation and a discussion of opportunities for future research.

Appendix A describes the implementation details of our image captioning models.

Appendix B provides links to the software and datasets developed in this work.

CHAPTER 2

Related work

2.1 Introduction

The topic of this dissertation, image captioning with external knowledge, combines two distinct streams of research: first, general image captioning, and second, integration of external world knowledge into diverse (neural network-based) language generation models.

In image captioning in general, the goal of the model is to process the visual content of the image and to produce a well-formed textual description of this content. Current state-of-the-art caption generators are able to achieve impressive results by leveraging large-scale pre-trained image recognition and language models and powerful deep learning algorithms, such as a Transformer network (Vaswani et al., 2017). In this chapter, Section 2.2 provides an overview of modern approaches to image captioning[1], as well as the most commonly used datasets and evaluation metrics. In our discussion of the datasets, we also present a systematic analysis of the captions in several of them, focusing on the vocabulary diversity and the presence and types of named entities. This analysis, which, to the best of our knowledge, has not been done in previous research, demonstrates the difference between two kinds of datasets: ones with crowdsourced

[1] For a comprehensive survey on image captioning, including the history of the field, see Hossain et al. (2019); Stefanini et al. (2021).

captions and naturally created captions. The results of the analysis provide motivation for using the latter kind in this dissertation.

World knowledge integration into machine learning models is made possible by long-standing AI projects that aim to collect human knowledge about the world and store it in structured databases (Lenat and Guha, 1989; Miller, 1995; Suchanek et al., 2008). In the NLP field in particular it has been claimed that in order to achieve human-like ability to generate and interpret language, it is necessary to take into account knowledge from external resources (Poerner et al., 2019; Cao et al., 2021). Section 2.3 reviews the most relevant research on incorporating world knowledge into language models.

2.2 Image captioning

2.2.1 Approaches

2.2.1.1 Encoder-decoder architecture for captioning

Practically all modern image captioning models are end-to-end trainable neural networks with two components: one for the visual processing, the other one for the text generation. The components are organized into an encoder-decoder pipeline, which was initially proposed in the machine translation field (Cho et al., 2014). The encoder takes an input image and transforms ("encodes") it into a feature representation, which is intended to reflect all the information needed to produce a caption. The decoder receives the result of the encoding and, based on it, generates ("decodes") a text description of the image. Both encoder and decoder are jointly optimized end-to-end during training, typically by minimizing negative log-likelihood of the caption given the image.

This neural encoder-decoder architecture for image captioning was first introduced in the Show and Tell model in Vinyals et al. (2015). This model uses a Convolutional Neural Network (CNN) as an encoder to create a dense representation of the image features. The CNN is pre-trained for an image classification task on the ImageNet dataset (Russakovsky et al., 2015). The last hidden layer of the CNN, believed to hold the most information about the image, is passed to a Recurrent Neural Network (RNN), more specifically, a Long Short-Term Memory network (LSTM), which acts as a decoder. The LSTM generates a caption word by word, at each time step taking into account the words produced at the previous steps.

An extension of this approach was proposed in Xu et al. (2015). Their Show,

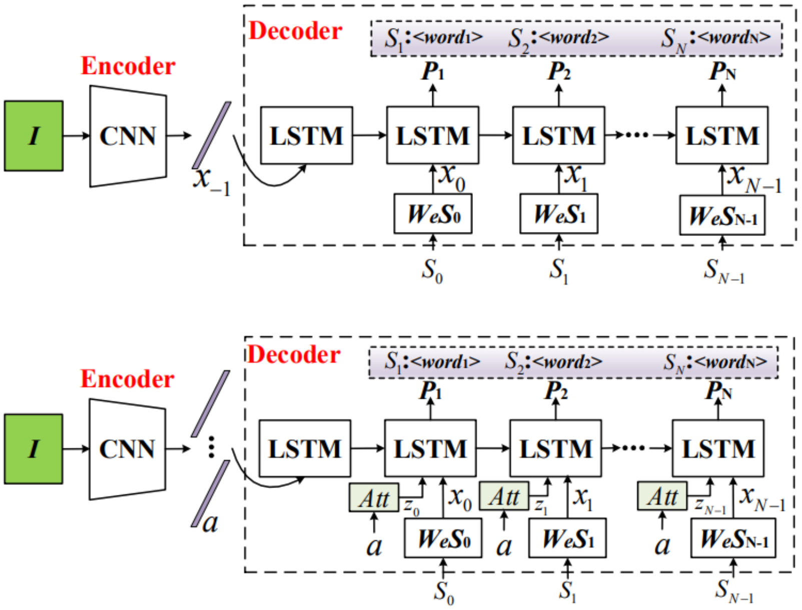


Figure 2.1: Image captioning models with an encoder-decoder structure: Show and Tell (Vinyals et al., 2015), top; Show, Attend and Tell (Xu et al., 2015), bottom. An illustration from Wang et al. (2019).

Attend and Tell model was the first to introduce the mechanism of attention into image captioning (see Figure 2.1 for a schematic comparison of the model architectures in Vinyals et al. (2015) and in Xu et al. (2015)). A model with attention can assign different weights to different parts of the image representation, thereby giving them more or less influence over the caption generation process. In Show, Attend and Tell, a dynamic attention mechanism is added on top of the CNN-LSTM pipeline. At every time step during the decoding, the attention weights are calculated based on the previously generated words, applied to the encoded image, and the result is taken into account when the next word is being selected. Various improvements of the attention mechanism in image captioning were developed in later work (You et al., 2016; Lu et al., 2017a; Zhu et al., 2018b; Yan et al., 2021; Anderson et al., 2018; Jiang et al., 2022).

The image captioning models developed in this dissertation are structured as standard encoder-decoder pipelines. Following the traditional approaches, we use a CNN model, pre-trained for image classification, to create an image representation for the subsequent use in the decoder.

2.2.1.2 Further advances

In recent years, Transformer-based models (Vaswani et al., 2017) have become standard in natural language generation (Topal et al., 2021). The advantages of a Transformer network include the ability to retain contextual information over long token sequences and a much faster generation speed compared to the traditional recurrent architectures. In image captioning, it also became widely used in both the decoder (Zhu et al., 2018a; Li et al., 2019a) and the encoder (Fang et al., 2022; Liu et al., 2021c; Zeng et al., 2022; Cornia et al., 2020). In our models, Transformer networks are used for encoding the additional context (non-visual external knowledge) and generating the caption at the decoding stage.

The latest advancements in image captioning research are brought about by the development of joint vision-language pre-trained models. These models are able to represent both images and texts in the shared feature space and can be fine-tuned for the image captioning task (Li et al., 2020, 2021; Zhang et al., 2021; Zhou et al., 2020; Hu et al., 2022). Alternatively, a pre-trained vision-language model can be used as an encoder in an image captioning pipeline. One of the baselines in this dissertation is provided by the ClipCap caption generator (Mokady et al., 2021), which has the CLIP vision-language model (Radford et al., 2021) as the image encoder and another large-scale pre-trained model GPT-2 (Radford et al., 2019) as the caption decoder.

2.2.1.3 Hallucination

Overall, modern image captioning models are able to generate image descriptions of high quality, and they are continuously improving with the development of better datasets and more sophisticated techniques. One challenge that remains largely unsolved is hallucination in generated captions (Rohrbach et al., 2018). The phenomenon of hallucination is common in all areas of neural natural language generation (Ji et al., 2022). It refers to the generation of text that contradicts or cannot be supported by the input data. In image captioning, the model “hallucinates” when it produces a caption that is unsupported by the image, for example, because it mentions something that is not actually present in the image or provides an inaccurate description of an object.

Rohrbach et al. (2018) find that hallucination is caused mostly by the linguistic component of the captioning pipeline: the errors tend to be driven by language priors. Although various techniques have been proposed for reducing hallucination in image captioning (Biten et al., 2022; Xiao and Wang, 2021; Dai et al., 2022), it is still quite common.

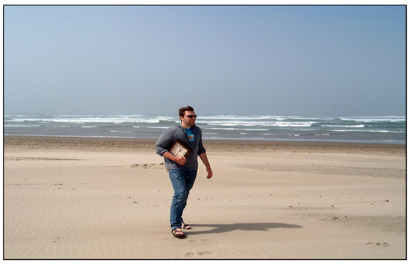


Figure 2.2: From Biten et al. (2022). Two automatic captioning models hallucinate the object that the man is holding in the image:
"A man on a beach with a *surfboard*", "A man standing on a beach holding a *frisbee*"

Strictly speaking, generation of external knowledge in captions, which we explore in this dissertation, can also be viewed as a kind of hallucination, since, by definition, it extends beyond what can be seen in the image. Here we draw a distinction between, on the one hand, generating statements that are not only unsupported by the visual content but are also incorrect or unrelated to the image, and, on the other hand, statements that are factually accurate and relevant to the image. For example, if the captioning model used the metadata of the image in Figure 2.2 to generate a caption that mentions the correct location of the beach or the name of the man, we would not consider it a hallucination and, on the contrary, encourage the valid use of image-external data. However, if the model produced an incorrect name of the beach or the man, it is as much a hallucination as stating that the man is holding a surfboard or a frisbee.

The image-external knowledge present in captions that are used for training a captioning model creates a particular risk of hallucination if the model is not able to utilize any data besides the image itself. This is especially the case for the datasets with naturally created captions, which are described in the next section.

2.2.2 Datasets

Image captioning datasets are large collections of images with the corresponding descriptions, used for training automatic caption generation models. This section describes the most widely used publicly available datasets with English-language captions[2]. We specifically discuss the distinction between the datasets with crowdsourced captions

[2] For a comprehensive survey of image captioning datasets, including multilingual (Elliott et al., 2016; Li et al., 2016; Yoshikawa et al., 2017; Lan et al., 2017; Li et al., 2019b), domain-specific (Radev et al., 2016; Lu et al., 2017b; Chen et al., 2019a) and stylized (Mathews et al., 2016; Shuster et al., 2019; Gan et al., 2017), see Luo et al. (2022).

and naturally created captions. Using descriptive metrics, we show the difference between these two dataset types in vocabulary diversity and specificity, and in the nature of the relationship between the images and the captions.

2.2.2.1 Crowdsourced captions

Several image captioning datasets were created by crowdsourcing captions for the images from the Flickr image hosting platform[3]. Through crowdsourcing, the dataset creators were able to collect more than one caption per image, which provides the models that are trained on these datasets with a variety of possible descriptions for a single scene.

Flickr8K (Hodosh et al., 2013) contains photographs of various day-to-day activities, objects and events from Flickr. Each photograph is associated with 5 descriptions obtained through the Amazon Mechanical Turk crowdsourcing service[4]. **Flickr30K** (Young et al., 2014) extends Flickr8K and adds entity annotations grounded in image regions.

MSCOCO (Lin et al., 2014) is one of the largest and most commonly used image captioning datasets. It contains images collected by querying various scene types and object categories on Flickr, and each of the images is also annotated with 5 captions via Mechanical Turk.

The annotator instructions for the Flickr8K, Flickr30K and MSCOCO datasets specify that the captions should only describe prominent entities in the image and should not contain any speculations made on the basis of the image content. As a result, captions in these datasets are largely decontextualized and lack image-external knowledge references.

2.2.2.2 Naturally created captions

Datasets with naturally created captions are significantly less popular in automatic image captioning. They present a bigger challenge, since a caption that a human produces in a non-controlled environment is likely to extend beyond a description of what can be directly seen in the image, which makes it harder for the model to imitate it.

Conceptual Captions (Sharma et al., 2018) consists of images harvested from various webpages, with the image descriptions extracted from their Alt-text HTML

[3] `https://www.flickr.com/`

[4] `https://www.mturk.com/`

attributes[5]. The captions in this dataset are naturally created, contextualized and represent a wide variety of styles. They were, however, additionally processed to remove information that was considered to be irrelevant or impossible to infer from the image, such as named entities and event names.

GoodNews (Biten et al., 2019) is a dataset for news image captioning, comprised of captioned photographs from the New York Times newspaper archive. The dataset contains news article texts as well, providing the context for each image-caption pair.

YFCC100M (Thomee et al., 2016) is the largest general-purpose multimedia dataset to date, with 100M media objects, 99.2M of which are photographs (although not all of them have captions). The photographs of YFCC100M are extracted from Flickr and linked to their original captions, if available, and various metadata (owner, camera type, geolocation, etc.).

2.2.2.3 Dataset analysis

This section provides an analysis of the datasets introduced above, using descriptive metrics that are inspired by the quality criteria for Vision and Language datasets suggested in Ferraro et al. (2015) and the metrics proposed in Van Miltenburg et al. (2018) for measuring the diversity of automatically generated captions. This analysis highlights the contrast between the datasets of crowdsourced captions and naturally created captions. The following metrics are used[6]:

- Dataset size. The number of captions in the dataset.

- Average caption length. Calculated in tokens. Longer captions indicate more detailed descriptions of the images.

- Vocabulary size. The number of unique vocabulary words. This metric is affected by the dataset size, so, to mitigate this effect, we report the average vocabulary size per 10k tokens.

- Normalized type-token ratio (TTR). Type-token ratio is calculated as the number of unique n-grams divided by the total number of n-grams. Here we report the normalized type-token ratio, which is an average TTR per 1,000 n-grams, less likely to be affected by extremely rare tokens and the dataset size.

[5] `https://en.wikipedia.org/wiki/Alt_attribute`

[6] These metrics will also be used in the coming chapters to describe the datasets we construct for training the captioning models with external knowledge.

- Top frequent tokens ratio. The ratio of top-20% frequent tokens to all the tokens in the captions. We propose to use this metric as an additional measure of diversity for the dataset captions: the lower the ratio, the longer the tail of the Zipfian distribution of word frequencies in the dataset (Zipf, 1949), and therefore, the higher the estimated diversity.

- Named entity representation. These metrics are calculated using the named entity recognition module in SpaCy (Honnibal and Montani, 2017).
  - The percentage of captions that contain at least one named entity.
  - The percentage of named entity tokens among all tokens.
  - The distribution of named entity types in the captions. Some of the fine-grained types are combined into coarser thematic categories: Geographic (SpaCy's GPE, LOC and FAC) and Numerical (CARDINAL and ORDINAL).

Table 2.1 contains the results of the analysis. The named entity type distribution is given separately in Table 2.2.

Dataset	**Num Cap (k)**	**Avg Cap Len**	**Vocab Size (k), per 10k Tokens**	**Norm TTR**		**Freq Tokens Ratio**	**% Cap with NE**	**% NE in All Tokens**
				1-gram	2-gram			
MSCOCO	593.96	11.32	1.57	39.00	78.08	92.22	16.27	1.58
Flickr30K	154.57	13.42	1.77	39.35	78.54	86.70	27.95	2.48
Conceptual Captions	2364.2	10.68	2.76	54.49	90.43	89.00	13.88	1.46
YFCC100M	326.78	75.87	2.95	45.18	75.06	83.67	77.13	4.23
GoodNews	471.34	21.17	3.92	60.55	93.89	76.66	95.86	15.48

Table 2.1: Exploratory analysis of a sample of existing image captioning datasets.

Dataset	**% Person**	**% Org**	**% Geo**	**% Num**	**% Date**	**% Other**
MSCOCO	2.79	8.23	3.55	71.43	5.33	8.67
Flickr30K	2.31	5.94	3.43	71.69	4.09	12.54
Conceptual Captions	0.03	0.4	0.17	31.05	49.60	18.75
YFCC100M	14.91	20.77	21.82	15.21	12.01	15.28
GoodNews	31.33	16.17	21.87	6.99	14.93	8.71

Table 2.2: Named entity type distribution in the captions of different datasets.

As seen in Table 2.1, naturally created captions are characterized by the higher vocabulary diversity than the crowdsourced ones: their vocabulary size is larger, they mostly have a higher type-token ratio and a lower frequent token ratio. They also contain much more named entities of various types (with the exception of Conceptual Captions, where they were purposefully removed).

Table 2.3 shows some examples of how similar images are captioned in these datasets. The MSCOCO and Flickr30K datasets describe the images listing the directly visible objects and actions, with a special attention to the visual attributes ("blond-haired", "brown and white"). The caption from the Conceptual Captions dataset is mostly descriptive but mentions a "semifinal match", which cannot be inferred from the image itself. The caption from the GoodNews dataset contains extensive references to the image-external knowledge (the name of the person, the location, the time, the background information). The YFCC100M caption presents a humorous interpretation of the image instead of a literal visual description.

Flickr30K	**Conceptual Captions**	**GoodNews**
A blond-haired women tennis player hitting the tennis ball.	tennis player returns the ball to tennis player during their semifinal match	Serena Williams in action at Wimbledon last year. She has five Wimbledon singles titles and five doubles titles.
MSCOCO	**YFCC100M**	
A brown and white cow standing on a grass field.	The cow just looked at me as if to say it wasn't all me you know	

Table 2.3: Similar images in the datasets with crowdsourced and naturally created captions.

The focus of this dissertation is on image-external knowledge influencing caption generation. Our analysis of the existing datasets shows that standard crowdsourced captions provide much less suitable data for this study than naturally created ones, which generally account for the image context and world knowledge in addition to the image itself. All the models we develop are trained and tested on the datasets of naturally created contextualized captions (each is described in detail in the relevant chapter).

2.2.3 Evaluation

The most common approach to evaluating automatically generated captions is comparing them to the human-written ‘ground truth’ captions for the same images. The higher the similarity between the ground truth captions (reference) and the generated ones (candidate), the higher the score the captioning model gets. This creates a quick and objective evaluation procedure that does not require labor- and time-intensive manual annotation. Metrics that implement this approach are widely used in image captioning as well as other language generation tasks, such as machine translation and text summarization, although they have been criticized for a relatively low correlation with human judgements (Sulem et al., 2018; Reiter, 2018; Kilickaya et al., 2017; Elliott and Keller, 2014). In this dissertation, we utilize the following standard metrics:

- **BLEU** (Papineni et al., 2002). This metric counts the number of overlapping n-grams between the candidate and reference captions and is often referred to as BLEU-1, BLEU-2, BLEU-3, BLEU-4 to indicate the maximum n-gram length. It is a precision-based metric, meaning that it measures how much of the candidate caption is found in the reference one.

- **ROUGE** (Lin, 2004). This metric is similar to BLEU but also accounts for recall (how much of the reference caption is found in the candidate one) and calculates the F-score based on the longest common subsequence of tokens between the candidate and reference captions.

- **METEOR** (Denkowski and Lavie, 2014). This metric calculates the overlap between the candidate and reference captions, additionally using synonym matching, paraphrasing and stemming for more flexibility and a more reliable alignment between the captions. The metric score is defined as a harmonic mean of the token precision and n-gram recall, with different weights for content and function words.

- **CIDEr** (Vedantam et al., 2015). This metric is based on the cosine similarity of TF-IDF vectors of n-grams in the candidate and reference captions. It gives the higher weight to the n-grams with higher TF-IDF scores, making more informative words count for more in the evaluation.

Other existing metrics aim to improve the candidate-reference matching quality by utilizing contextualized word embeddings (Zhang et al., 2019a; Lee et al., 2020) and scene graphs (Anderson et al., 2016). Some methods use a classifier that determines if a given caption was produced by a human or a machine and take the probability of the "human" class as the metric score (Sharif et al., 2018; Cui et al., 2018). Finally, several recently developed metrics make use of the pre-trained vision-language models to calculate the similarity between the caption and the image (Hessel et al., 2021; Lee et al., 2021). Notably, none of these metrics attempt to explicitly evaluate the correctness of the generated captions given the images. To a certain extent this is done by the CHAIR metric, which was proposed in Rohrbach et al. (2018) for measuring the amount of hallucination in the captions. However, this metric only deals with generic objects (e.g., if a caption mentions a table, the metric checks if there is a table in the image) and does not apply to situations where the caption contains image-external knowledge references. In this dissertation we develop custom metrics for measuring the factual accuracy of the generated captions, in addition to the standard metrics described above.

2.3 World knowledge-aware language generation

2.3.1 World knowledge in databases

Throughout history, knowledge about the world has been recorded in countless encyclopedias, dictionaries, glossaries, atlases, etc. Since the second half of the 20th century, structured knowledge bases have been widely used in AI and machine learning-related tasks. The different kinds of knowledge they provide include, for example:

- Encyclopedic fact knowledge, e.g., "Paris is a capital of France", "Douglas Adams was a member of the Footlights club": YAGO (Mahdisoltani et al., 2014; Hoffart et al., 2013; Suchanek et al., 2008), DBpedia (Auer et al., 2007), WikiTaxonomy (Ponzetto and Strube, 2008), Freebase (Bollacker et al., 2008), NELL (Carlson et al., 2010), Wikidata (Vrandečić and Krötzsch, 2014), Wikidata5M (Wang et al., 2021b).

- Common sense knowledge, e.g., "Ice makes the road slippery", "A book can be placed in a drawer": Cyc (Lenat, 1995; Lenat and Guha, 1989), ConceptNet (Speer et al., 2017; Liu and Singh, 2004), SenticNet (Cambria et al., 2014, 2010), COGBASE (Olsher, 2014), WebChild (Tandon et al., 2017, 2014), ATOMIC (Sap et al., 2019), ASER (Zhang et al., 2020).

- Semantic relations between words and concepts, e.g., "*violin* is an instance of *string instrument*, other instances of which include *viola*, *cello* and *double bass*": WordNet (Miller, 1995), VerbNet (Schuler, 2005).

- Domain-specific knowledge, e.g., biomedical (SNOMED[7], PDBe-KB[8]), geographic (OpenStreetMap[9], GeoNames[10]), etc.

This data is commonly represented in a knowledge graph — a directed labeled graph with entities $\mathcal{E}$ as nodes and relations $\mathcal{R}$ between the entities as edges. Formally, a knowledge graph is a set of triples $\{(p, r, c) \mid p \in \mathcal{E}, r \in \mathcal{R}, c \in \mathcal{E}\}$, where p is a parent entity with relation r to a child entity c. Figures 2.3 and 2.4 demonstrate snippets of the knowledge stored in DBpedia and ConceptNet. The web interface provides the data in an easily readable format, but it can also be represented as triples like <Paul McCartney, birthPlace, Liverpool> and <bicycle, used for, transportation>.

In this dissertation, we inject two kinds of knowledge into our captioning models: geographic (from OpenStreetMap) and open-domain encyclopedic facts (from DBpedia). Our proposed way to encode factual knowledge is in general applicable to all kinds of data that can be converted to the knowledge graph format.

2.3.2 Language modeling

Language models (LMs) are trained to estimate the probability of a sequence of words $P(w_0, w_1, \ldots, w_n)$. At timestamp t, a language model estimates the probability distribution over all the words in the vocabulary V, given the sequence of words observed before t:

$$P(w_t = w_i \mid w_0 \ldots w_{t-1}),\ \forall w_i \in V \tag{2.1}$$

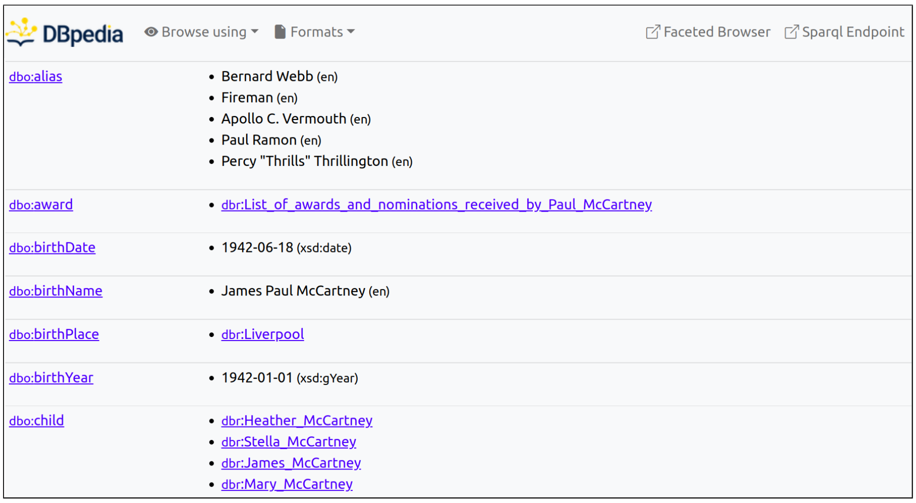
Language models are at the core of most modern language generation pipelines (including image captioning), where they are used to predict the most probable word to be generated next given the context of preceding words.

[7] https://www.snomed.org/

[8] https://www.ebi.ac.uk/pdbe/pdbe-kb/

[9] https://www.openstreetmap.org/

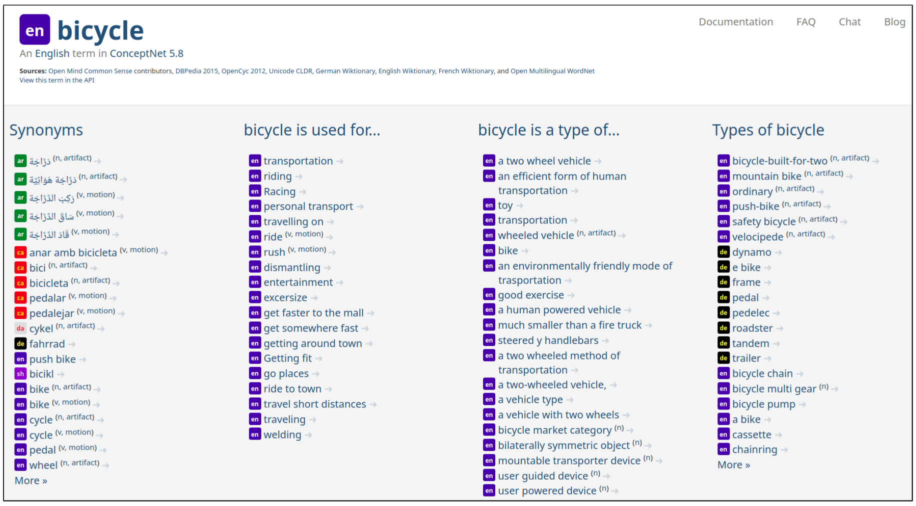
[10] http://www.geonames.org/



DBpedia Browse using ▾ Formats ▾ Faceted Browser Sparql Endpoint

dbo:alias	• Bernard Webb (en) • Fireman (en) • Apollo C. Vermouth (en) • Paul Ramon (en) • Percy "Thrills" Thrillington (en)
dbo:award	• dbr:List_of_awards_and_nominations_received_by_Paul_McCartney
dbo:birthDate	• 1942-06-18 (xsd:date)
dbo:birthName	• James Paul McCartney (en)
dbo:birthPlace	• dbr:Liverpool
dbo:birthYear	• 1942-01-01 (xsd:gYear)
dbo:child	• dbr:Heather_McCartney • dbr:Stella_McCartney • dbr:James_McCartney • dbr:Mary_McCartney

Figure 2.3: A snippet of the ‘Paul McCartney’ entry in DBpedia. `https://dbpedia.org/page/Paul_McCartney`



Documentation FAQ Chat Blog

en **bicycle**

An English term in ConceptNet 5.8

Sources: Open Mind Common Sense contributors, DBPedia 2015, OpenCyc 2012, Unicode CLDR, German Wiktionary, English Wiktionary, French Wiktionary, and Open Multilingual WordNet
View this term in the API

Synonyms

- ar دَرَّاجَة (n, artifact) →
- ar دَرَّاجَة هَوَائِيَّة (n, artifact) →
- ar رَكِبَ الدَّرَّاجَة (v, motion) →
- ar سَاقَ الدَّرَّاجَة (v, motion) →
- ar قَادَ الدَّرَّاجَة (v, motion) →
- ca anar amb bicicleta (v, motion) →
- ca bici (n, artifact) →
- ca bicicleta (n, artifact) →
- ca pedalar (v, motion) →
- ca pedalejar (v, motion) →
- da cykel (n, artifact) →
- de fahrrad →
- en push bike →
- sh bicikl →
- en bike (n, artifact) →
- en bike (v, motion) →
- en cycle (n, artifact) →
- en cycle (v, motion) →
- en pedal (v, motion) →
- en wheel (n, artifact) →

More »

bicycle is used for...

- en transportation →
- en riding →
- en Racing →
- en personal transport →
- en travelling on →
- en ride (v, motion) →
- en rush (v, motion) →
- en dismantling →
- en entertainment →
- en excersize →
- en get faster to the mall →
- en get somewhere fast →
- en getting around town →
- en Getting fit →
- en go places →
- en ride to town →
- en travel short distances →
- en traveling →
- en welding →

bicycle is a type of...

- en a two wheel vehicle →
- en an efficient form of human transportation →
- en toy →
- en transportation →
- en wheeled vehicle (n, artifact) →
- en bike →
- en an environmentally friendly mode of trasportation →
- en good exercise →
- en a human powered vehicle →
- en much smaller than a fire truck →
- en steered y handlebars →
- en a two wheeled method of transportation →
- en a two-wheeled vehicle, →
- en a vehicle type →
- en a vehicle with two wheels →
- en bicycle market category (n) →
- en bilaterally symmetric object (n) →
- en mountable transporter device (n) →
- en user guided device (n) →
- en user powered device (n) →

Types of bicycle

- en bicycle-built-for-two (n, artifact) →
- en mountain bike (n, artifact) →
- en ordinary (n, artifact) →
- en push-bike (n, artifact) →
- en safety bicycle (n, artifact) →
- en velocipede (n, artifact) →
- de dynamo →
- de e bike →
- de frame →
- de pedal →
- de pedelec →
- de roadster →
- de tandem →
- de trailer →
- en bicycle chain →
- en bicycle multi gear (n) →
- en bicycle pump →
- en a bike →
- en cassette →
- en chainring →

More »

Figure 2.4: A snippet of the ‘bicycle’ entry in ConceptNet. `https://conceptnet.io/c/en/bicycle`

2.3.2.1 Knowledge from pre-training

Language models learn to predict the most probable word sequences, and thus, to generate texts, by pre-training on large text corpora. Datasets like CommonCrawl[11]

[11] `https://commoncrawl.org/`

provide the models with billions of tokens worth of human-written text, which allows them to generalize linguistic rules and meanings. The most advanced LMs, such as GPT-3, are reported to produce texts that are so fluent and semantically coherent that it is difficult to distinguish them from the ones that were written by humans (Brown et al., 2020). It has also been shown that large-scale LMs acquire a significant amount of world knowledge during pre-training (Petroni et al., 2019; Trinh and Le, 2019; Roberts et al., 2020; Heinzerling and Inui, 2021). From enough exposure in the corpora, LMs are able to learn, for example, that the correct way to continue "Dante was born in..." is "Florence", without the need to look up this fact in a database.

2.3.2.2 Limitations

However, multiple studies have claimed that LMs in their current state cannot replace explicit knowledge bases in knowledge-intensive tasks (Poerner et al., 2019; Razniewski et al., 2021; Cao et al., 2021; AlKhamissi et al., 2022). First of all, they are limited by the knowledge present in the training data, and even the largest corpus of natural texts will not represent all the information stored in dedicated resources. Language models do not generalize well to rare or unseen entities, so, if prompted by an entity they do not "recognize", they are likely to generate sentences that are uninformative or incorrect, suffering from hallucination (Ji et al., 2022). Moreover, there is no easy way to introduce new facts or update the existing ones without re-training the model.

These limitations of the standard text-only LMs motivated the development of *knowledge-aware* language models that incorporate structured fact and entity knowledge from external resources[12]. They are reported to achieve lower perplexity than their knowledge-agnostic counterparts when tested on texts that contain a significant amount of factual knowledge and named entities, e.g., texts from Wikipedia. In addition to that, they perform better in downstream tasks, including LAMA unsupervised question answering benchmark (Petroni et al., 2019), relation classification, entity linking and fact completion.

2.3.3 Knowledge-aware LM approaches

We identify three main groups of approaches to incorporating world knowledge into language models, discussed in turn below. The method developed in this dissertation

[12] Since the focus of this dissertation is on external knowledge integration, we do not review knowledge- or entity-aware language models that do not explicitly utilize world knowledge from external resources, such as, among others, Rosset et al. (2020); Liu et al. (2022); Févry et al. (2020). For a survey that also describes this line of research, see Wei et al. (2021); Safavi and Koutra (2021).

most resembles the approaches from the third group, which involves generating some words directly from an external data source (Section 2.3.3.3).

2.3.3.1 Modeling entity names using KBs

A prominent group of approaches focuses on injecting factual information about real world entities into the language model during pre-training. This typically involves two steps: first, identifying named entities in the texts, and second, modifying the text encoding procedure so that it takes into account the information about these entities retrieved from an external knowledge base.

For example, for training the ERNIE model, Zhang et al. (2019b) link named entities in texts to their corresponding entries in Wikidata, then use the entity-related knowledge subgraphs, embedded with the TransE algorithm (Bordes et al., 2013), as entity representations. KEPLER (Wang et al., 2021b) embeds entities and relations using their textual descriptions in knowledge graphs, e.g., Wikidata5M and WordNet. DKPLM (Zhang et al., 2022) represents the long-tail entities through the related knowledge triples from the Wikidata5M database.

Some models build the factually enhanced entity representations on top of the already pre-trained LMs. For example, KnowBert (Peters et al., 2019) inserts a special knowledge integration mechanism between two layers of the pre-trained BERT (Devlin et al., 2019). This mechanism identifies entity spans in the input text, links them to an external database (such as WordNet or Wikipedia) and uses the extracted knowledge to update the original BERT subword representations. Similarly, E-BERT (Poerner et al., 2020) combines BERT's subword vectors for entity names with the embeddings created for the same entities by Wikipedia2Vec (Yamada et al., 2018).

2.3.3.2 Appending KB knowledge

Various approaches aim to provide the language model with the factual knowledge from a database in combination with the pre-training texts.

The K-Adapter framework (Wang et al., 2021a) adds a plug-in factual adapter (a compact neural network) on top of the pre-trained RoBERTa model (Liu et al., 2019b). The adapter is trained separately on the relation classification task using Wikipedia and Wikidata, and the knowledge it acquires is used as an additional input for downstream tasks. CoLAKE (Sun et al., 2020) and K-BERT (Liu et al., 2020) merge the graph representation of the input text with the knowledge subgraph related to the entities mentioned in this text. Similarly, LM-CORE (Kaur et al., 2022) concatenates the encoded text with the relevant knowledge triples extracted from an external database. The resulting structure incorporates extra factual information besides what is already given in the text.

2.3.3.3 Generation from KBs

Another group of approaches lets the language model choose between generating a word from the standard vocabulary or copying a word directly from an external knowledge source.

The NKLM model developed in Ahn et al. (2016) makes a distinction between two types of words in a text: regular ones and ones that represent a certain fact from the Freebase database. The model takes into account the preceding words and facts and learns to either generate a word from the regular vocabulary or to copy a word from a Freebase fact. A similar approach is implemented with a special gating mechanism in the KALM model (Liu et al., 2019a). In Logan et al. (2019), the KGLM model maintains a dynamically growing local knowledge graph with Wikidata facts about the entities mentioned in the text. At each time step the model decides, based on the previous context, whether to generate a regular word from the vocabulary, to copy an entity from the local knowledge graph, or to produce a new entity from Wikidata and add facts about it to the local graph.

In this dissertation, the language models trained to generate captions follow an approach that is most similar to this category. The external knowledge sources act as additional vocabularies of tokens that refer to real world entities and facts. As is common for all the approaches discussed above, these tokens are represented through the information available about the entities and facts in the knowledge base, instead of their distributional patterns in the pre-training texts.

2.4 Discussion

This dissertation presents an approach to knowledge-aware image captioning, a combination of general image captioning and knowledge-aware language generation. It is inspired by existing works that recognize the limitations of language models in knowledge-intensive tasks, and incorporate structured information from knowledge bases into the training and generation process of LMs. Standard image captioning models have a similar limitation — the inability to utilize the data beyond the images and captions in the training datasets. In many cases this does not affect the perceived quality of the models, since the commonly used datasets contain captions that account only for what can be directly seen in the images. However, naturally created captions often provide a description of the image that is informed by relevant world knowledge. In this dissertation, the captioning models are trained on such captions, which include abundant references to domain-specific (geographic) and general encyclopedic knowledge. The base of the models is a simple traditional captioning pipeline; the main focus is placed on the integration of external data.

CHAPTER 3

Geo-aware image captioning

3.1 Introduction

This chapter presents our approach to image captioning with external geographic knowledge integration[1]. It describes the development of a captioning model that uses image location as a contextualization anchor. This "geographic" anchor provides natural grounding to the captioning process: objects in the image are no longer generic, but, linked to the specific location, become concrete and often identifiable from external sources.

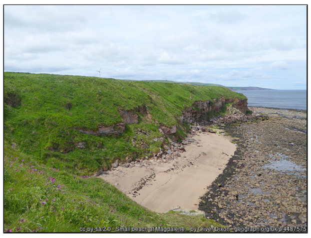

cc-by-sa/2.0 - Small beach at Magdalene... by Oliver Dixon - geograph.org.uk/p/4487575

Figure 3.1: "Small beach at Magdalene Fields. At the foot of the cliffs, north of Berwick-upon-Tweed." © Oliver Dixon.

For example, a non-expert is unlikely to identify the beach in Figure 3.1 from the

[1] A preliminary version of our work on this topic has been published in Nikiforova et al. (2020); this chapter contains a further development of the ideas in the paper, with a refined model architecture and substantially improved results.

image alone, but, if the coordinates of the image location are known, it becomes a much easier task. For an automatic captioning model, it increases the probability of imitating the human-written caption for this image, which makes multiple references to its geographic context.

The ‘geo-aware’ captioning model presented in this chapter utilizes external geographic knowledge, which is retrieved based on the image location, as an extra input for the encoder and an extra vocabulary for the decoder. The model is trained on a new dataset that we compile from naturally created image-caption data similar to the example in Figure 3.1. The model’s evaluation includes an assessment of the accuracy of generated geographic references, carried out with a custom metric that measures the ability of the model to correctly use spatial expressions.

Section 3.2 provides an overview of the use of location metadata for contextualization in various NLP and Computer Vision tasks. Section 3.3 presents the geo-aware image-caption data that is used for training and testing our captioning model. Section 3.4 discusses the concept and implementation of the image-specific geo-entity context that contains information about the real world geographic entities relevant to the image. Section 3.5 describes the procedure of encoding the captions that takes into account the difference between regular words and geographic entity names. Section 3.6 describes the architecture of the image captioning model we develop, and Section 3.7 reports the results of its evaluation in comparison to several decontextualized baseline models, with a special focus on the accuracy of geographic references. Finally, Section 3.8 contains a discussion of the benefits and limitations of the developed geo-aware captioning model.

3.2 Contextualization with location metadata

Geographic metadata, the latitude and longitude coordinates of the image location, is easily available for many real-life photographs due to the built-in GPS in modern cameras and phones. Thus, using it as a contextualization anchor to access external information for captioning would not necessarily require any additional annotation. Nevertheless, to the best of our knowledge, there has been no research yet that utilizes location metadata for contextualization in image captioning. At the same time the benefits of using geographic information have been attested in various other tasks in Computer Vision and, to a lesser degree, in NLP.

In image classification, geolocation has been reported to significantly improve the quality of the models, providing additional information when the visual distinction between classes is minimal (see Figure 3.2). Different approaches have emerged regarding how the image location is used to improve the model performance. The most straightforward way is to integrate (add, concatenate, fuse with a multilayer perceptron) raw or normalized location coordinates with the visual features of the image extracted by a CNN (Tang et al., 2015; Chu et al., 2019; Yang et al., 2022a; Arbinger et al., 2022; Zou et al., 2022). Another method uses the distribution of labels per geolocation in the training data as a classification prior (Chu et al., 2019; Skreta et al., 2020) or as the basis for filtering the predicted labels at the post-processing stage (Chu et al., 2019; Zou et al., 2022). In Ayush et al. (2021), geographic metadata is used to improve self-supervised learning of image representations by implementing geolocation prediction as an auxiliary objective. Finally, a large group of approaches focuses on the additional information that the image location can help retrieve. For example, Tang et al. (2015) extract a multitude of data from external resources, including the area's elevation, precipitation, vegetation and other geographic indicators, as well as various demographic data (statistics of age, sex, race, income, education, etc.) of the relevant zip code district. Arbinger et al. (2022) utilize the satellite images of the location,

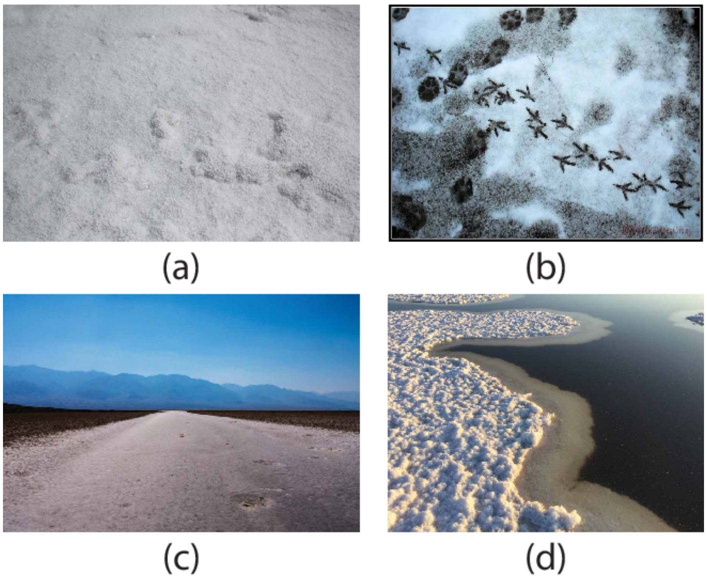


Figure 3.2: From Tang et al. (2015). "Which of these are images of snow? Just by looking at the images, it may be difficult to tell. However, what if we knew that (a) was taken at the Bonneville Salt Flats in Utah, (b) was taken in New Hamsphire, (c) was taken in Death Valley, California and (d) was taken near Palo Alto, California?"

addresses of the nearby places and user tags of the photographs taken in the area. The user tags of the "neighbor" photographs are also used as additional input in Liao et al. (2015), Tang et al. (2015) and Nitta et al. (2020). Our approach to geo-aware image captioning would fall loosely into the last category, which is characterized by extracting relevant data from external resources based on the geolocation of the image.

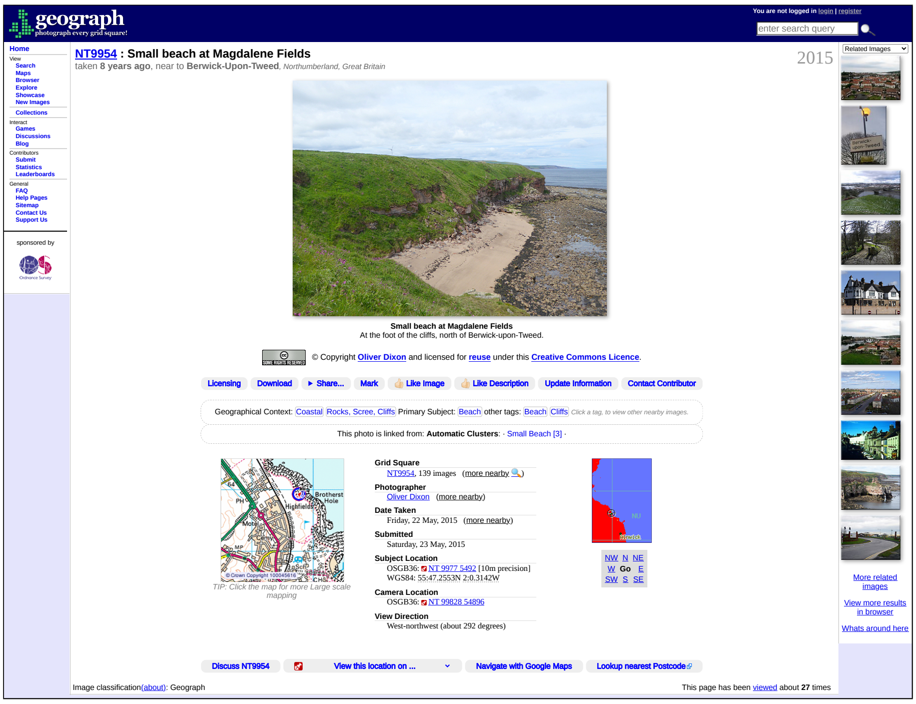
The use of geographic data is less common in NLP. Naturally, the challenges of jointly processing linguistic and geospatial data are addressed in multidisciplinary studies bordering geographic sciences, such as geographic question answering (Mai et al., 2021, 2020) and geographic information retrieval (Purves et al., 2018). Sporadically, geolocation of the text has been utilized to contextualize word embeddings (Cocos and Callison-Burch, 2017) and to improve named entity disambiguation (Srinivasan and Rafiei, 2021). Overall, the potential of enriching NLP models with geographic information is currently under-researched. To the best of our knowledge, this chapter presents the first model that includes automatic text generation (in the form of image captions) where geolocation is used for contextualization and for producing accurate references to geographic data.

3.3 Data

This section presents the image-caption data that is used to train and test the geo-aware captioning model in this chapter. Section 3.3.1 describes the original source of the data, the Geograph project. Section 3.3.2 introduces GeoRic, the dataset that we compile from the Geograph data and use in our experiments.

3.3.1 The Geograph project

Geograph (`https://www.geograph.org.uk/`) is an on-going project that aims to collect photographs of every square kilometer of Great Britain and Ireland. At the time of writing, the website of this project contained more than 7 million photographs. Each photograph is accompanied by extensive metadata: the title and caption, the name of the photographer, the dates when the photograph was taken and when it was submitted to the website, the location of the photograph subject and the camera, view direction and various semantic tags concerning the content of the image (e.g., "bridge", "village sign", "M74"), photograph characteristics (e.g., "wideangle", "closeup"), geographic location (e.g., "near Blackpool"). An example of a page on the Geograph website, with a captioned photograph and all the related metadata, is given in Figure 3.3.



NT9954 : Small beach at Magdalene Fields
taken **8 years ago**, near to **Berwick-Upon-Tweed**, *Northumberland, Great Britain*

Small beach at Magdalene Fields
At the foot of the cliffs, north of Berwick-upon-Tweed.

© Copyright Oliver Dixon and licensed for reuse under this Creative Commons Licence.

Geographical Context: Coastal Rocks, Scree, Cliffs Primary Subject: Beach other tags: Beach Cliffs *Click a tag, to view other nearby images.*

This photo is linked from: **Automatic Clusters**: · Small Beach [3] ·

TIP: Click the map for more Large scale mapping

Grid Square
NT9954, 139 images (more nearby)

Photographer
Oliver Dixon (more nearby)

Date Taken
Friday, 22 May, 2015 (more nearby)

Submitted
Saturday, 23 May, 2015

Subject Location
OSGB36: NT 9977 5492 [10m precision]
WGS84: 55:47.2553N 2:0.3142W

Camera Location
OSGB36: NT 99828 54896

View Direction
West-northwest (about 292 degrees)

Image classification(about): Geograph

This page has been viewed about **27** times

Figure 3.3: A page on the Geograph project website:
`https://www.geograph.org.uk/photo/4487575`

The data on the Geograph project website is licensed for reuse under the Creative Commons BY-SA 2.0 license[2], which allows sharing and adapting the data as long as appropriate credit is given, a link to the license is provided, there is an indication if any changes were made, and the modified data is distributed under the same license as the original.

3.3.2 The GeoRic dataset

The GeoRic (Geo-aware Rocky Image Captioning) dataset[3] consists of the data collected from the Geograph project website. Unlike the standard datasets, such as

[2] `https://creativecommons.org/licenses/by-sa/2.0/`

[3] The first version of the GeoRic dataset, GeoRic v1.0, was developed for Nikiforova et al. (2020) and is available at `https://rocky.sites.uu.nl/datasets/#georic-dataset`. The dataset described here is the second version, GeoRic v2.0, which features an updated collection of images and captions, selected with a less restrictive procedure (e.g., GeoRic v1.0 included only single sentence captions, while GeoRic v2.0 does not have this restriction), and with more related textual data included (titles and captions from Geograph instead of only captions).

MSCOCO or Flick30k, GeoRic contains 'geo-aware' data: images tagged with the coordinates of their location and captions produced with the geographic context in mind. A single entry in the dataset includes the image, its title and caption[4], the latitude and longitude coordinates of the image location. Each image is accompanied by its original URL that acts as a unique identifier as well as an author attribution. Table 3.1 demonstrates a few sample dataset entries.

Image	**URL**	**Title**	**Caption**	**Latitude**	**Longitude**
	`https://www.geograph.org.uk/photo/4487575`	Small beach at Magdalene Fields	At the foot of the cliffs, north of Berwick-upon-Tweed.	55.7876	-2.00524
	`https://www.geograph.org.uk/photo/5325648`	Our Lady of Good Counsel Catholic Church	The red brick Catholic church in Broughty Ferry from across Westfield Road.	56.4673	-2.88119

Table 3.1: Sample entries in the GeoRic dataset.

3.3.2.1 Train-validation-test data split

The GeoRic dataset contains overall 25,987 entries[5], split between separate sets for training (19,490, 75%), validation (3,248, 12.5%) and testing (3,249, 12.5%). Importantly, in order to avoid assigning different photographs of the same place to both train and validation/test sets, we base the split on the latitude of the image location instead of splitting the dataset randomly. The photographs that were taken to the north of the 54.9768° latitude are assigned to the test set, between the 53.7065° and the 54.9768° latitude to the validation set, and the rest to the train set. With the latitude-based split, we ensure testing on previously unseen data, which helps to detect possible overfitting. Figure 3.4 provides a visualization of the GeoRic data points, plotted according to the image location.

[4] The title and caption are separate fields on the Geograph website, which is why they are kept separate in GeoRic as well. In practice we use a combination of them as a caption for the captioning model, since generally both of them contain relevant descriptions of the image and its context.

[5] With the vast and continuously growing number of images on the Geograph website, it is possible to extend the GeoRic dataset with more data in future work.

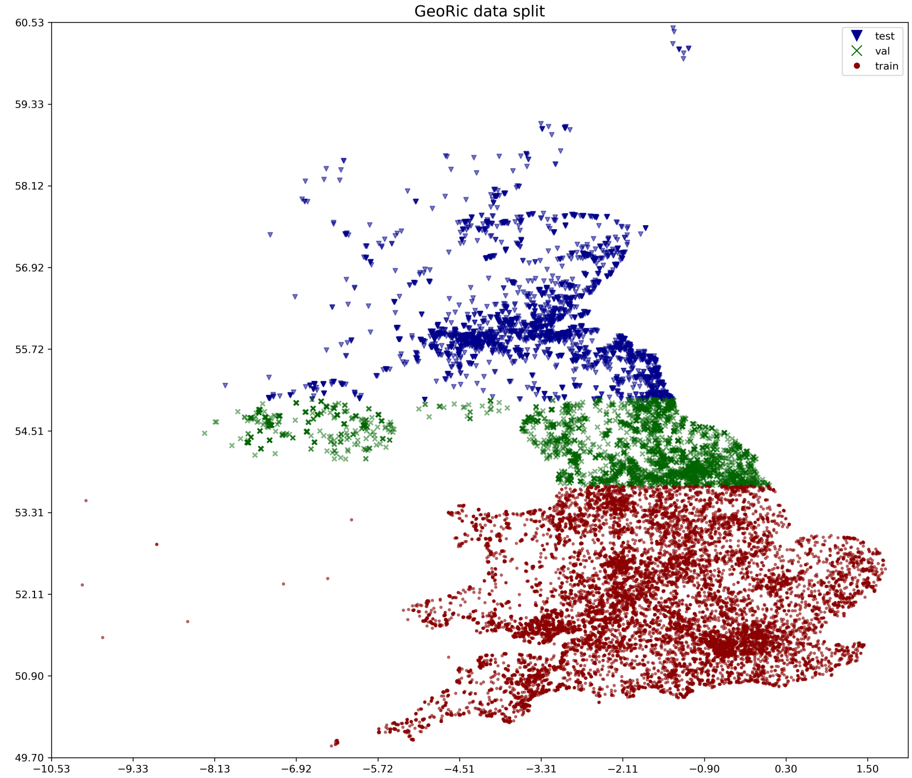


Figure 3.4: Locations of GeoRic images, split into train, validation and test sets.

3.3.2.2 Selection criteria

The images in the dataset were selected randomly from those that satisfied the following requirements:

(i) The length of the caption (i.e., the combined length of the title and the caption on the Geograph website) is less than 100 tokens. This limit is introduced for practical reasons, since training the captioning model to produce extremely long texts is both computationally expensive and undesirable for our research goals: the majority of such captions contain copied excerpts from Wikipedia.

(ii) The caption contains at least one reference to the geographic context of the image. This requirement ensures that there is enough material for the captioning model to learn to generate appropriate geographic references.

3.3.2.3 Statistics

Table 3.2 shows quantitative statistics of the captions in the GeoRic dataset, next to the statistics of the MSCOCO dataset for comparison. As is evident from the table, although GeoRic is a substantially smaller dataset, its captions are on average longer and their vocabulary is much more diverse. Importantly, unlike the decontextualized MSCOCO captions, almost all of the GeoRic ones contain named entities, with about 48% of them identified as names of geographic objects (see Table 3.3).

	Num Cap	**Avg Cap Len**	**Vocab Size (k), per 10k Tokens**	**Norm TTR**		**Freq Tokens Ratio**	**% Cap with NE**	**% NE in All Tokens**
				1-gram	2-gram			
GeoRic	25,987	18.05	2.68	44.66	84.57	65.86	99.63	17.7
MSCOCO	593,968	11.32	1.57	39.00	78.08	92.22	16.27	1.58

Table 3.2: Statistics of the GeoRic dataset, compared to MSCOCO. The metrics are defined in Section 2.2.2.3.

	% Person	**% Org**	**% Geo**	**% Num**	**% Date**	**% Other**
GeoRic	20.48	24.06	48.03	2.26	1.87	3.3
MSCOCO	2.79	8.23	3.55	71.43	5.33	8.67

Table 3.3: Named entity type distribution in the GeoRic captions, compared to MSCOCO.

Geographic references in GeoRic captions can take various forms; most frequently they are either directly naming an object in the image ("Our Lady of Good Counsel Catholic Church. ..."), or indicating where the photograph was taken relative to other places ("...north of Berwick-upon-Tweed"). The latter is often realized through spatial expressions, such as "near [X]", "(to the) north of [X]". The most frequent ones in the dataset are shown in Table 3.4.

	Near	**In**	**Along**	**Across**	**North of**	**South of**	**East of**	**West of**
# Caps	8,003	6,940	5,312	3,428	1,849	1,791	1,166	1,082

Table 3.4: Number of GeoRic captions per spatial expression.

3.3.2.4 Normalization

Captions in GeoRic undergo minimal normalization. They are converted to lower case and split into tokens with the standard NLTK Tokenizer package[6]; formatting artifacts are removed and some tokens are replaced with others for unification purposes (e.g., “&” is replaced with “and”, “saint” is replaced with “st”).

3.4 The (geographic) entity context of an image

The entity context of an image is, generally speaking, a collection of real world entities relevant to the image and useful for its description. Here, given the special focus on geographic knowledge, we interpret the entity context as a *geographic* entity context (or simply geo-entity context): relevant geographic entities around the image location, represented through their geographic properties.

3.4.1 Geographic data resources

A pair of latitude and longitude coordinates of a certain place can provide access to an abundance of data about geographic objects nearby, which is available in open-domain web databases (e.g., in the geocoded portion of Wikipedia) as well as in dedicated resources, Geographic Information Systems (GIS). GIS databases are particularly useful as they specifically accumulate geographic data, extending beyond landmarks and points-of-interest, and often also provide tools for data management, analysis and visualization.

In this dissertation we utilize one of the largest publicly available GIS resources, OpenStreetMap (OSM, `https://www.openstreetmap.org/`). It is a collaborative online project that at the time of writing had over 8.5 million registered users and over 8.7 billion described geographic entities. Each entity in OSM is linked to its latitude and longitude coordinates and can have an arbitrary number of annotations, which are called “tags”. Tags are key-value pairs that describe various aspects and features of an entity, for example, *building=house*, *name=Main Road*. Although the amount of available data in OSM is vast (more than 135 million unique tags and more than 88,000 unique keys), we opt to create a simple approximation of the geo-entity context, representing the entities in it with a few basic features.

[6] `https://www.nltk.org/api/nltk.tokenize.html`

3.4.2 Building the geo-entity context

To build the geo-entity context for a given image, first, we retrieve all the available geographic entities from OSM that are located around the image location. Based on preliminary experiments, this area is limited to the radius of 1 kilometer, as anything further away is likely to be less relevant to the image description.

Second, we extract features to represent the retrieved entities. The **distance *d*** between the entity and the image location and the **azimuth *a*** of the angle between them are calculated based on their coordinates. The **size** *s* of the entity is estimated as the area of the polygon of its shape (with the value of 0.0 if the entity is given as a single point in OSM). The last feature is the **type *t*** of the entity, as provided by OSM. This set of features is intended to reflect the salience of the entities (especially through distance and size) and to ensure valid usage of the most frequent spatial expressions, e.g., the distance value is crucial for generating appropriate entities after prepositions "near" and "in", the azimuth value — for "north of", "south of", etc., and prepositions "across" and "along" are compatible only with entities of certain types. A fragment of a geographic entity context, with entities represented through these features, is shown in Figure 3.5.

e_1 – Magdalene Fields	(d_1= 0.127, a_1= -29.31, s_1= 0.001, t_1= golf)
e_2 – Berwick-upon-Tweed	(d_2= 0.723, a_2= -168.19, s_2= 0.0081, t_2= city)
…	
e_n – St. George's Road	(d_n= 0.641, a_n= 131.49, s_n= 0.0045, t_n= residential)

Figure 3.5: Sample fragment of a geographic entity context.

Finally, the entities are ranked by the estimated probability of them being mentioned in an image caption. To achieve that, we train a *ranking model* on a subset of 10,000 additional Geograph captions (not overlapping with the ones in GeoRic). This model is a simple logistic regression classifier with the standard implementation from scikit-learn (Pedregosa et al., 2011) that predicts whether the entity will be mentioned in a caption based on the following features: the distance between the image location and the entity, the entity's size, type and the visual features extracted from the image by the scene detection model pre-trained on the Places365 database (Zhou et al., 2017) (the model's checkpoint is available at `https://github.com/GKalliatakis/Keras-VGG16-places365`).

Top 300 entities of the ranked list constitute the final geo-entity context *G*. This number, as well as the radius of 1 km, were selected after training the ranking model.

For at least 95% of the images in the test set of the ranking model, top 300 of the image's ranked entities, all of which are located no further than 1 km from the image location, contain at least one that appears in this image's original caption.

3.4.3 Geographic embedding function

We introduce an embedding function that transforms raw entity features, as seen in Figure 3.5, into a "geographic embedding" vector:

$$\textsc{GeoEmb}(e_i) = \text{Concat}[d_i,\ norm(a_i),\ s_i,\ Emb_t(t_i)] \tag{3.1}$$

where Emb_t is a specialized embedding layer for the entities' types, the distance and the size parameters are used without any normalization, and the azimuth is normalized as follows:

$$\begin{aligned} &norm(a_i) = \text{Concat}[a_{north};\ a_{east}], \text{ where} \\ &a_{north} = |a_i|/180 \\ &a_{east} = \begin{cases} |90 - a_i|/180, a_i >= -90 \\ (90 + |a_i + 180|)/180, \text{otherwise} \end{cases} \end{aligned} \tag{3.2}$$

This ensures that the azimuth values that are close to each other "on the compass" (e.g., -5° and 5°, -175° and 175°, 85° and 95°, but not -90° and 90°, etc.) become minimally different after normalization. To be processed further in the captioning pipeline, each of the entities in the geo-entity context is embedded, as shown in Equation 3.3.

$$EmbG = (\textsc{GeoEmb}(e_1)\dots\textsc{GeoEmb}(e_n)), e_i \in G \tag{3.3}$$

3.5 Encoding captions

Encoding the captions for further processing by the captioning model involves mapping each token to a number given the vocabulary compiled from the dataset train set. Here we distinguish two types of tokens: regular vocabulary words and geographic entity names. We believe this distinction to be critical for training a geo-aware captioning model. While the model can learn to produce regular vocabulary words according to their distribution patterns in the training corpus, no corpus could ever be large enough to accurately and unambiguously represent the important characteristics of the real world entities that the model needs to take into account to correctly use their names in a caption. For example, to say that a given image depicts something "to the north of Blackpool", the model needs to consider the location of Blackpool relative to where the photograph was taken, which is not reflected in any textual corpus.

Thus, regular words are mapped to their indices from the vocabulary, which is shared among all the captions, and geographic names are mapped to the indices of the corresponding geographic entities in the geo-entity context specific to a given image. For example, consider the following fictitious caption:

> "Sacred Heart Church. An impressive Gothic building, located near University Hospital."

Let the geo-entity context G of the corresponding image contain the following entries:

$$\begin{bmatrix} \text{Sacred Heart Church } \{\mathrm{d}_0, \mathrm{a}_0, \mathrm{s}_0, \mathrm{t}_0\} \text{ — } 0 \\ \dots \\ \text{University Hospital } \{\mathrm{d}_{12}, \mathrm{a}_{12}, \mathrm{s}_{12}, \mathrm{t}_{12}\} \text{ — } 12 \\ \dots \end{bmatrix}$$

Let the vocabulary V, common for all the captions, contain the following entries:

$$\begin{bmatrix} \text{<pad> — 0} \\ \text{. — 1} \\ \text{an — 2} \\ \text{impressive — 3} \\ \text{gothic — 4} \\ \text{building — 5} \\ \text{, — 6} \\ \text{located — 7} \\ \text{near — 8} \\ \dots \\ \text{<unk> — 9047} \\ \text{<start> — 9048} \\ \text{<end> — 9049} \end{bmatrix}$$

With simple string matching against the geo-entity context, "Sacred Heart Church" and "University Hospital" are recognized as names of geographic entities in the caption. The rest of the caption tokens are assumed to be regular vocabulary words. The encoding follows the following mapping rule:

$$w_i \to \begin{cases} \text{len}(V) + G[w_i], & \text{if } w_i \in G \\ V[w_i], & \text{otherwise} \end{cases} \tag{3.4}$$

where $X[y]$ means getting an index of y in X. Therefore, the original caption is mapped to the following sequence of numbers:

Sacred Heart Church ⟶ 9050

. ⟶ 1

An ⟶ 2

impressive ⟶ 3

Gothic ⟶ 4

building ⟶ 5

, ⟶ 6

located ⟶ 7

near ⟶ 8

University Hospital ⟶ 9062

. ⟶ 1

We also store a binary "mask" vector, which indicates the type of each token: 0 for a vocabulary word and 1 for a geographic entity name. The mask is passed to the captioning model along with the encoded caption and used to embed the tokens according to their type.

3.6 Model architecture

Our image captioning model has a general encoder-decoder structure, but, in contrast to the standard models, both encoder and decoder incorporate information from the geo-entity context in order to generate geo-aware captions. Figure 3.6 shows the overall architecture of the model; its separate components are discussed in detail below.

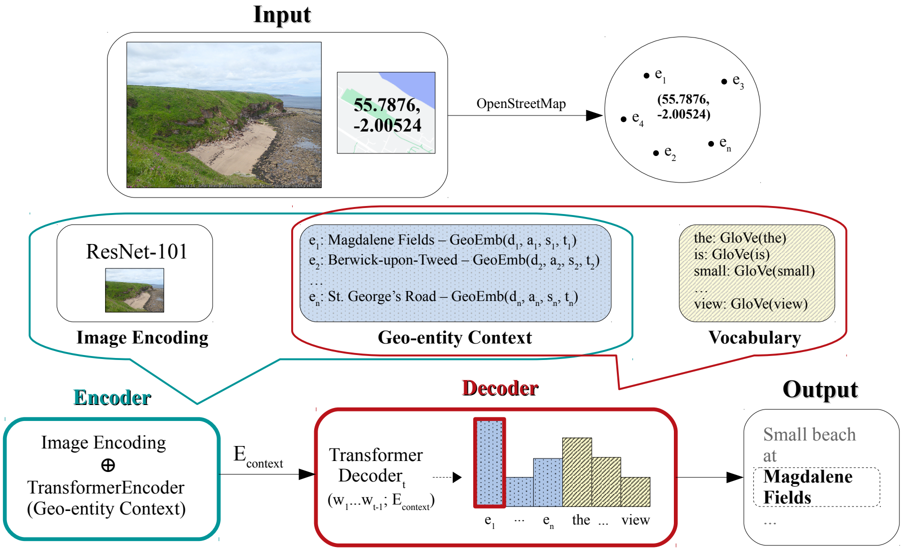


Figure 3.6: Overview of the geo-aware captioning model architecture.

3.6.1 Encoder

The encoder in a standard encoder-decoder captioning pipeline is responsible for creating the initial representation of an input image, which is passed on to the next stage, the decoder. Generally speaking, the purpose of the encoder is to capture all the important information needed to produce the caption later on. In a standard setting, this refers only to the visual features of the image. In the geo-aware captioning model, it is extended to include both the image and the external knowledge from the geographic entity context.

The images are encoded with the ResNet-101 CNN (He et al., 2016) model, pre-trained for the image classification task on the ImageNet dataset (Russakovsky et al., 2015). As is standard for captioning, the last two (Linear and Pool) layers of the model are removed, since what is needed is not the actual classification labels but the informative visual features that would have been used for classification.

For encoding the geo-entity context *EmbG*, we utilize a Transformer encoder with the standard structure proposed in Vaswani et al. (2017). The outcome is concatenated with the image encoding, resulting in a combined representation of the extended

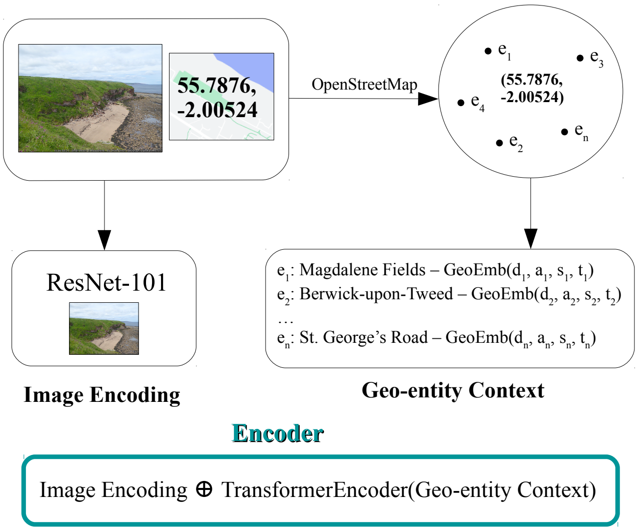

55.7876,
-2.00524
OpenStreetMap
• e_1
• e_3
(55.7876,
-2.00524)
• e_4
• e_n
• e_2
ResNet-101
e_1: Magdalene Fields – GeoEmb(d_1, a_1, s_1, t_1)
e_2: Berwick-upon-Tweed – GeoEmb(d_2, a_2, s_2, t_2)
…
e_n: St. George’s Road – GeoEmb(d_n, a_n, s_n, t_n)
Image Encoding
Geo-entity Context
Encoder
Image Encoding ⊕ TransformerEncoder(Geo-entity Context)

Figure 3.7: The geo-aware encoder.

multimodal (visual and geographic) context for generating the caption.

$$E_{context} = \text{Concat}[\,\textsc{ResNet-101}(\text{Img}),\ \textsc{TransformerEncoder}(\text{EmbG})] \tag{3.5}$$

3.6.2 Decoder

The decoder receives the context representation from the encoder as input and, based on it, generates the output caption. The challenge of a decoder in a geo-aware captioning model is to produce a caption that combines a description of the image itself, composed of regular vocabulary words, and appropriate geographic referencing, with the names from the geo-entity context.

Vocabulary tokens and geographic names are embedded with different embedding functions. For the former, we make use of pre-trained GloVe word embeddings (Pennington et al., 2014), which capture their semantics through the distribution patterns in a large corpus. However, for the geographic names this would not be enough. For the decoder to correctly use them in the generated captions, in phrases like “near [X]” or “to the south of [X]”, it needs to have access to the geographic characteristics of the real world objects they represent. Therefore, we use the geographic embedding function introduced in Equation 3.1, to convey information about their distance, azimuth, size and type.

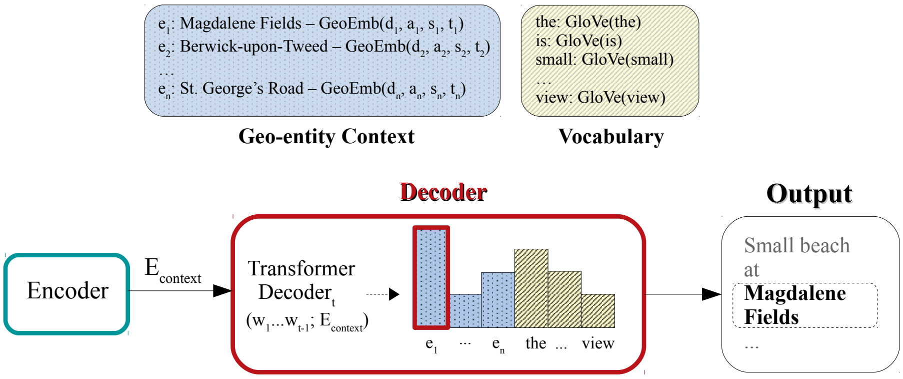


Figure 3.8: The geo-aware decoder.

$$Emb(w_i) = \begin{cases} \textsc{GeoEmb}(w_i), \text{if } w_i \in G \\ \textsc{GloVe}(w_i), \text{otherwise} \end{cases} \tag{3.6}$$

Each token is represented as a sum of its embedding (GloVe or geographic, depending on the token type) and the encoding of its position in the caption.

$$PosEmb(w_i) = Emb(w_i) + Pos(w_i) \tag{3.7}$$

The decoder in the model is a Transformer decoder with a standard structure. At each time step t it takes into account the context representation from the encoder $E_{context}$ and the previously produced tokens $w_1 \dots w_{t-1}$.

$$h_t = \textsc{TransformerDecoder}(PosEmb(w_{1\dots t-1}); E_{context}) \tag{3.8}$$

The output of the decoder h_t is generally used to estimate the probability distribution over the vocabulary, to select the most likely token to appear next in the sequence. In this case, there are two probability distributions to consider, as shown in Figure 3.8: the one over the vocabulary, and the one over the entities in the image-specific geo-entity context. So, we calculate the scores for the vocabulary items $v_1 ... v_l \in V$ and the scores for the geographic entity names $e_1 ... e_n \in G$:

$$\begin{aligned} y_{v_1} ... y_{v_l} &= h_t \, W_{vocab} \\ y_{e_1} ... y_{e_n} &= (EmbG \ \textsc{diag}(h_t)) \, \vec{w}_{geo} \end{aligned} \tag{3.9}$$

where W_{vocab} and $\vec{w}_{geo}$ are trainable linear transformation matrix/vector (bias terms are omitted for simplicity), and $\textsc{diag}(h_t)$ denotes a diagonal matrix with the h_t vector in the main diagonal.

The two sets of scores are then concatenated and passed through a softmax layer, and the token of either type with the highest score is generated at position t.

$$\begin{aligned} w_t &= \arg\max_{w_i} P(w_i), w_i \in V \cup G \\ \text{where } P(w_i) &= \sigma_i(\text{ Concat}[y_{v_1}...y_{v_l}, y_{e_1}...y_{e_n}]) \end{aligned} \tag{3.10}$$

3.7 Evaluation

For any image captioning model, the goal of evaluation is to assess how well the generated captions describe their corresponding images. Traditionally this is done by comparing the automatically produced captions to the ground truth (human-written) captions for the same images. In this section we present the results of this standard caption quality evaluation with commonly used metrics (BLEU, ROUGE, METEOR, CIDEr).

Contextualized image captioning adds another dimension to evaluation: estimating how accurately the context is represented in the generated captions. In geo-aware captioning in particular, the goal is not only to produce captions with geographic references, but to ensure that these references are in fact correct and appropriate for a given image. This is why we also conduct specialized geographic accuracy evaluation, which is based on the comparison between the model's use of spatial expressions (phrases like "near [X]", "to the north of [X]") and how they are used in human-written captions.

3.7.1 Baselines

We introduce the following baselines for comparison with the geo-aware model.

ClipCap-MSCOCO/Conceptual A standard pre-trained captioning model ClipCap (Mokady et al., 2021), with two implementations: trained on the MSCOCO dataset (ClipCap-MSCOCO) and on the Conceptual Captions dataset (ClipCap-Conceptual). Running this model on the out-of-domain GeoRic images puts it at a considerable disadvantage; the purpose of this baseline is not to provide a competitive alternative but rather to demonstrate how a model that achieves strong results on two other datasets would fare on the geo-aware captions of GeoRic.

No-context A model that we train on the GeoRic dataset, similar in architecture to the geo-aware one, but with no contextualization component. This model is structured as a standard encoder-decoder captioning pipeline, where the input to the encoder is only the image itself, and the decoder produces the caption only from the regular vocabulary. This baseline is intended to demonstrate the effect that the geographic contextualization component has on the performance of a caption generator.

Random geo-entity This baseline is used in geographic accuracy evaluation. Its goal is to test whether the geo-aware captioning model has actually learned to select the specific geographic entities that are appropriate in a given context. We expect that if this is in fact the case, then its performance should be higher than if the entities are selected at random. Here we do not train a new model; instead, we take the captions generated by the geo-aware model and replace each generated geographic name with a randomly selected one from the same geo-entity context. This creates a strong baseline, since the geo-entity context was specifically constructed to include entities which are highly likely to be relevant to the image description.

3.7.2 Results

3.7.2.1 Quantitative evaluation

We measure how close the generated captions are to the ground truth, human-written ones using the standard captioning metrics: BLEU, ROUGE, METEOR and CIDEr (see Section 2.2.3 for the metric descriptions). While it is well known that BLEU-type metrics are suboptimal in terms of correlation with human judgements (Reiter, 2018; Sulem et al., 2018; Kilickaya et al., 2017), especially if there is only one reference ground truth caption per image (and not multiple, like in MSCOCO or Flickr30k), they still provide a reliable way to capture the differences between the alternative models.

		BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
ClipCap	MSCOCO	6.39	0.10	7.44	4.56	3.17
	Conceptual	4.17	0.22	11.38	4.91	3.29
No-context		21.09	1.23	20.22	5.57	3.28
Geo-aware		**27.77**	**4.89**	**25.84**	**10.61**	**45.90**

Table 3.5: Standard metric scores of the models, measured on the test set of GeoRic.

As shown in Table 3.5, the two standard pre-trained ClipCap models get extremely low metric scores, since they produce very different captions from the ones in the GeoRic dataset. This is expected, because both training datasets, MSCOCO and Conceptual Captions, contain no specific references to external geographic knowledge (MSCOCO was created with the focus on purely visual descriptions, and the captions in Conceptual Captions were processed to remove named entities).

The no-context baseline model achieves higher metric scores than ClipCap, although the difference varies from major (e.g., in BLEU-1) to minimal (e.g., in METEOR), and the CIDEr score is even slightly lower than that of the ClipCap-Conceptual model. The no-context model was trained on the in-domain data, so it is not surprising that it is able to better imitate the ground truth captions.

The best performance is demonstrated by the geo-aware model. Its metric scores are significantly higher than those of the alternative models (two-sample t-test, p <0.001), showing the benefits of the additional contextualization component. The ability to produce correct geographic names leads to a particularly big jump in CIDEr, which weighs the tokens according to the TF-IDF score and therefore makes a match in named entities especially valuable.

The metric scores in Table 3.5 are overall much lower than those of the state-of-the-art captioning models that are trained and tested on standard datasets (e.g., in Li et al. (2020), OSCAR achieves BLEU-4 of 41.7 and CIDEr of 140.0 on the MSCOCO dataset). However, it is harder to generate diverse, contextualized captions such as the ones in GeoRic. It is also harder to achieve high metric scores with only one reference caption per image (MSCOCO provides five, and a match with any of them counts towards the score). The geo-aware model's results are however comparable with those achieved on average by other contextualized image captioning models that deal with non-crowdsourced, naturally produced captions (Hu et al., 2020; Tran et al., 2020; Zhao et al., 2021; Bai et al., 2021). Although a direct comparison between these models and the ones developed in this dissertation is not possible due to the differences in the datasets and the task specifics, the metric scores reported in these works are close to ours, with BLEU-4 ranging from 1.71 to 8.8 and CIDEr ranging from 9.1 to 54.47.

3.7.2.2 Qualitative evaluation

Table 3.6 provides examples of captions generated by the geo-aware and baseline models for the sample captions from the test set.

Both ClipCap models generate straightforward descriptions of what can be seen in

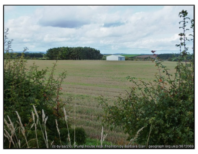
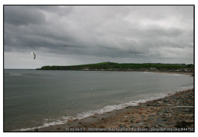
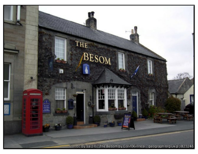
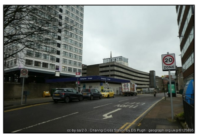
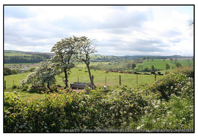
(a)		Ground truth: Pump house near Thornton. Looking northwesterward over farmland to the east of Thornton. ClipCap-MSCOCO: A field with a herd of sheep grazing on it. ClipCap-Conceptual: a view of the fields. No-context: farmland, *chaddleworth*. a crop of wheat to the east of forest Geo-aware: farmland **near thornton**. looking across farmland **to the east of thornton**
(b)		Ground truth: Stonehaven Bay. From the promenade looking across Stonehaven Bay ClipCap-MSCOCO: A large body of water with a boat in it. ClipCap-Conceptual: the beach on the coast. No-context: the *river thames*. looking downstream *from the thames near pangbourne*. Geo-aware: **stonehaven bay**. looking **across stonehaven bay** from the beach.
(c)		Ground truth: The Besom. The Besom public house in Coldstream ClipCap-MSCOCO: A bar called the dockside bar and grill. ClipCap-Conceptual: the pub in the village. No-context: the *white lion*. public house *along white lion street in geddington*. Geo-aware: the **besom inn**. pub **in coldstream**.
(d)		Ground truth: Charing Cross Station. Looking along Elmbank Crescent towards the railway station. ClipCap-MSCOCO: A city street with a bus and cars. ClipCap-Conceptual: the bus station is seen. No-context: the old railway line. the railway line *near chatham station*. Geo-aware: **charing cross station**. looking *along newton street* **towards charing cross**.
(e)		Ground truth: Valley of Wooler Water. From the minor road to Earle, south of Wooler. ClipCap-MSCOCO: A herd of sheep grazing on a lush green field. ClipCap-Conceptual: view from the top of the hill. No-context: *mid devon*: countryside scenery. looking across the countryside *near uffculme*. Geo-aware: viewpoint **south of wooler**. looking across the valley floor *to the west of wooler*.

Table 3.6: Examples of the captions generated for the GeoRic images from the test set. Correct geographic references are given in **bold**; incorrect ones are given in *italics*.

the images, without any names or other context-specific references. The descriptions themselves are of good quality overall, although the MSCOCO-trained model is prone to some hallucination (e.g., “sheep” in image (a), “boat” in image (b)).

Captions produced by the no-context baseline model are similar to the contextualized ground truth ones, with many geographic references included. Importantly though, all these references are incorrect, which is explained by the fact that they were generated from the regular vocabulary, without taking into account any external data. This shows that it is hardly possible to identify non-famous entities, such as a specific pub or a station, based on the image alone.

Finally, captions produced by the geo-aware model show that it can successfully use the geo-entity context to generate accurate geographic references without compromising the quality of image description. However, not all generated references are correct. For example, “Wooler” in caption (e) is mentioned twice, and only one of these times correctly, specifically when it is used in a spatial expression with “south of” and not with “west of”. In caption (d), interestingly, the model was right to choose an object of the type *road* to follow the preposition “along”, which satisfies this preposition’s semantic requirements. But this road is in fact located a little further away from the image location, so the reference is overall incorrect. In the next section we conduct a systematic evaluation of geographic accuracy in the generated captions, with a custom metric designed specifically for this task.

3.7.2.3 Geographic accuracy evaluation

Comparing the generated captions to the ground truth ones is not enough to evaluate whether or not the captioning model has acquired the ability to process geographic data and to produce correct geographic references. Just like there are multiple ways to describe what can be seen in a given image, there are multiple ways to indicate its location. The methods employed by the standard metrics to mitigate this issue, such as using synonyms and paraphrases in METEOR, would not, for instance, help identify the generated “near Rheims Way” as correct if the ground truth caption describes the image location as “in Kent”. Therefore, we develop a custom metric that aims to specifically assess the geographic “competence” of the captioning models.

We assume that a captioning model is competent in handling geographic knowledge if it is able to correctly use spatial expressions. In particular, we focus on the most frequent ones in the GeoRic dataset (see Table 3.4): phrases with “near”, “in”, “along”, “across”, “north of”, “south of”, “east of” and “west of”. Because of their prevalence in the training data, a captioning model can be expected to generate them often enough for a reliable automatic evaluation.

We find that even though there is some intuitive understanding of how these expressions are generally used, there are no strict definitions which we could utilize in an automatic quantitative evaluation. The word “near” is a classic example of vague language (Mark and Frank, 1989; Altman, 1994; Bennett, 2010): “near my house” and “near France” correspond to very different possible distance values. In fact, all the spatial expressions listed above are vague as well. For example, although there is an absolute north on the compass, there are no hard constraints to where “north of” begins and ends. Similarly, there is a certain inherent rule that in a phrase “along [X]” where X is a geographic object, X has to be elongated (e.g., “along the road/river”, but not “along the statue”), but there is no definitive list of object types that are compatible with “along”. In the scope of this work, we do not aim to come up with definitions or exhaustive constraints of using the aforementioned spatial expressions. Since any captioning model is only as good as its training data, we turn to it for the “gold standards” of how these expressions should be used with geographic entity names.

We assume that the following features are the most critical for determining compatibility with the prepositions in question, based on the theoretical and empirical research concerning these prepositions (Robinson, 1990; Gahegan, 1995; Garrod et al., 1999; Takemura et al., 2005):

- “X is **near** Y”: distance between X and Y
- “X is **in** Y”: distance between X and Y, type of Y
- “(looking from) X **along/across** Y”: distance between X and Y, type of Y
- “X is **to the north/south/east/west of** Y”: azimuth of the angle between X and Y

In all of the cases, X is the location of the image and Y is a geographic entity. Figure 3.9 presents the probability distributions of relevant feature values[7] in the training set of GeoRic: these distributions show how people have used the spatial expressions we focus on. To collect the feature values, we identified the locations and types of the geographic entities in these expressions with the Nominatim[8] geocoding service that utilizes OpenStreetMap data (in case of ambiguous entity names that can refer to multiple different objects, the closest one to the image location was selected).

As seen in Figure 3.9, the distance values for “near”, “in”, “along” and “across” present an overall similar picture of the probability decreasing as the distance increases, with the highest peak at less than 100 meters. The rate with which the probability decreases is lower for “near” than the other prepositions, apparently showing that it can combine more freely with entities that are located somewhat further away. The

[7] Continuous distance and azimuth variables are grouped into bins; for the sake of clarity, type values are plotted only if their probability is greater than 0.01.

[8] `https://nominatim.org/`

little peaks at the 2 kilometer mark for “along” and “across” are explained by their common co-occurrence with *road* objects (see the type probability distributions in the same figure). The locations of the roads in OpenStreetMap are sometimes only given as a single ‘point’ instead of a more realistic ‘line’, and the provided point can be situated further from the image location than the closest point of the road actually is. The azimuth values in Figure 3.9 are distributed according to the expected prototypical “north” (highest probability around 0°), “south” (two peaks around -180° and 180°), “east” (highest probability around 90°) and “west” (highest probability around -90°). Most frequent entity types that follow “in” include different kinds of settlements (*village*, *town*) and roads (*residential*). Both “along” and “across” are most often used in combination with roads (*residential*, *unclassified*, *primary*, *secondary*, *tertiary*), while “across” also prominently co-occurs with objects related to water (*bay*, *water*, *stream*, *river*, *beach*).

Then, we compute the distributions of the same feature values, now based on the captions generated by the automatic captioning model. The core idea is that if it has indeed learned how a certain spatial expression should be used, the distribution of its relevant feature values in the generated captions will be similar to the one observed in the training data. To quantify the similarity between the two distributions, we use the Jensen-Shannon (JS) distance metric (Endres and Schindelin, 2003). The JS distance is the square root of the JS divergence (Lin, 1991), which is based on the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). The KL divergence of the probability distribution P from the reference probability distribution Q is the expected excess surprise from using Q as a model when the actual distribution is P. The JS divergence is a symmetrized and smoothed version of the KL divergence, with values ranging between 0 (distributions are identical) and 1 (distributions are maximally different). The JS distance metric is calculated as follows:

$$\begin{gathered}\mathrm{JS}(P \,||\, Q) = \sqrt{\frac{1}{2}\mathrm{KL}(P \,||\, M) + \frac{1}{2}\mathrm{KL}(Q \,||\, M)}, \quad M = \frac{1}{2}(P+Q) \\ \text{where } \mathrm{KL}(P \,||\, Q) = \sum_{x \in \mathcal{X}} P(x) \, \log_2\left(\frac{P(x)}{Q(x)}\right)\end{gathered} \tag{3.11}$$

We take P to be the probability distribution of the feature values $x \in \mathcal{X}$ in the training data, and Q to be the probability distribution of the feature values in the generated captions.

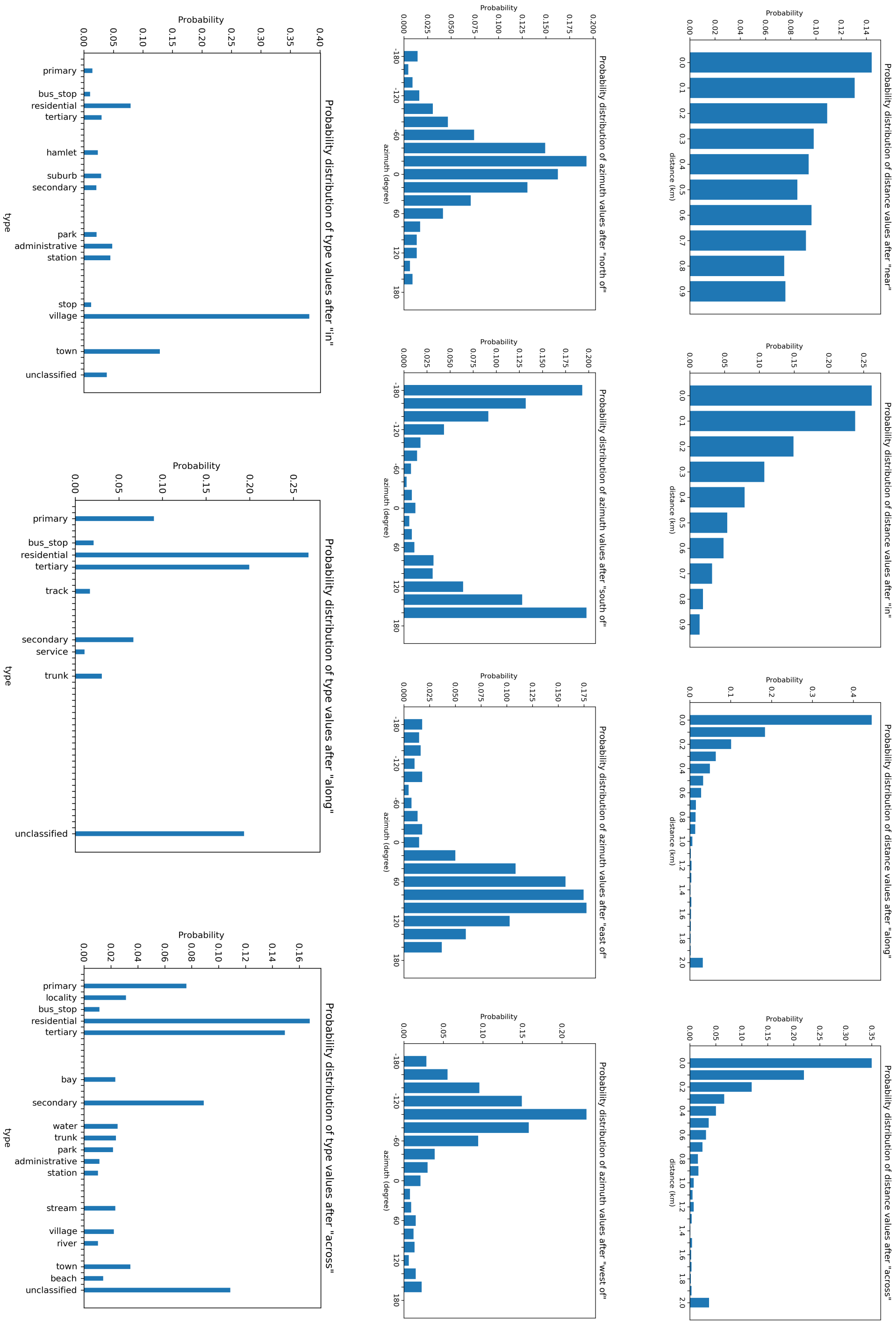


Figure 3.9: Probability distributions of expression-specific feature values in the training set of GeoRic.

		Near	**In**		**Along**		**Across**	
			dist	**type**	**dist**	**type**	**dist**	**type**
No-context	Count	1,238	702		168		108	
	JS	1.0	1.0	0.657	0.948	0.622	0.941	0.595
Random geo-entity	Count	1,625	378		458		348	
	JS	0.217	0.277	0.711	0.465	0.580	0.443	0.541
Geo-aware	Count	1,625	378		458		348	
	JS	**0.061**	**0.240**	**0.436**	**0.454**	**0.311**	**0.338**	**0.349**

		North of	**South of**	**East of**	**West of**
No-context	Count	73	381	376	80
	JS	0.667	0.661	0.682	0.688
Random geo-entity	Count	200	121	138	145
	JS	0.427	0.546	0.443	0.474
Geo-aware	Count	200	121	138	145
	JS	**0.316**	**0.316**	**0.266**	**0.300**

Table 3.7: Jensen-Shannon metric scores of the models. The lower the score, the better the estimated ability of the model to produce valid geographic references.

Table 3.7 shows the JS metric scores of the geo-aware captioning model, the no-context and the random geo-entity baseline models. The standard pre-trained models are not included here, since they do not produce spatial expressions with geographic names. We also report the number of times the expressions were generated, showing that there is a sufficient amount of data for a reliable evaluation of each of them.

The JS metric scores confirm that the geo-aware model has indeed learned the semantic requirements of the most frequent spatial prepositions, confidently outperforming the strong random geo-entity baseline model and especially the no-context one. Figure 3.10 shows the probability distributions of the feature values in the captions generated by the geo-aware model, which look a lot like the ones in Figure 3.9.

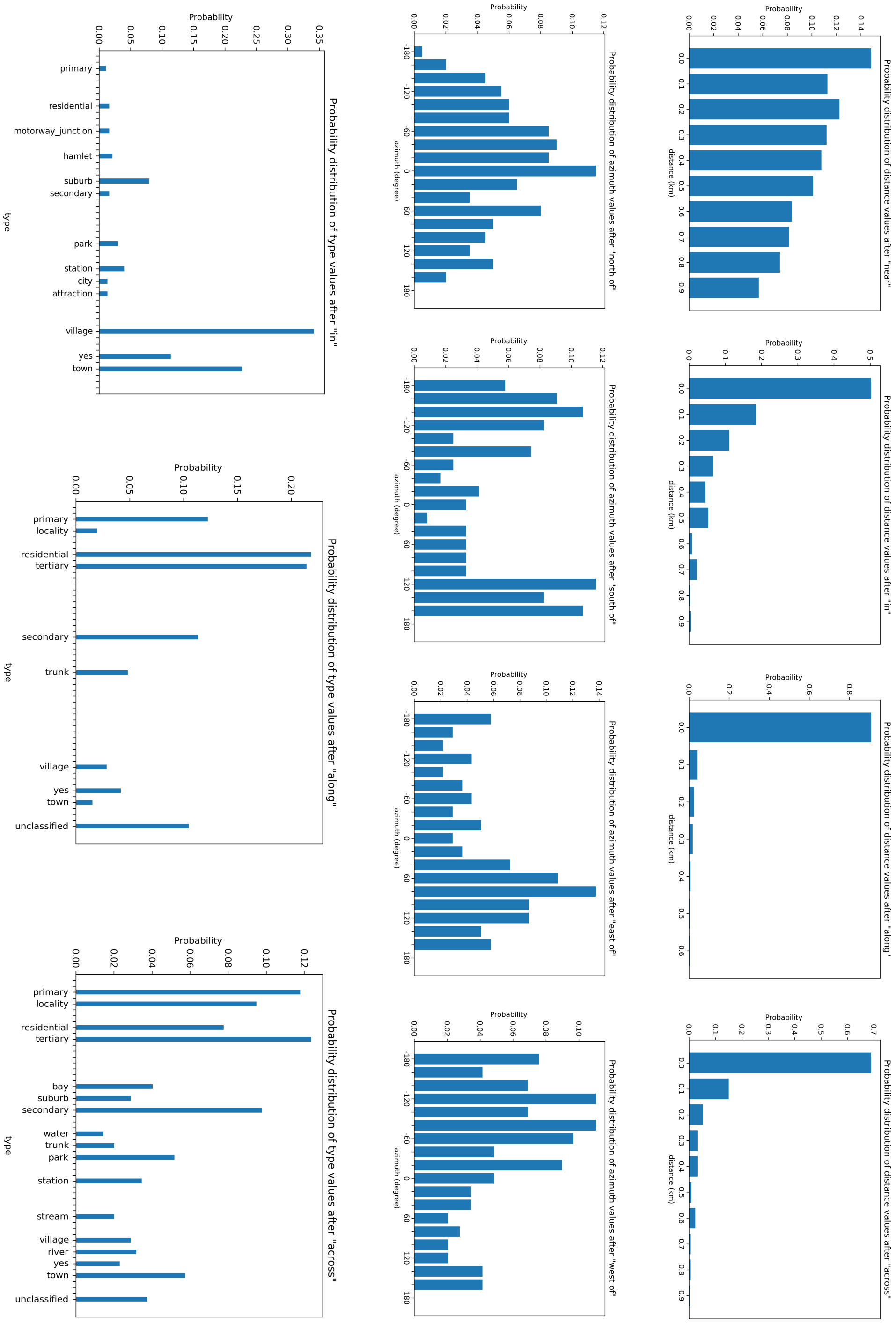


Figure 3.10: Probability distributions of expression-specific feature values in the captions generated by the geo-aware model.

3.8 Discussion

This chapter has demonstrated the effectiveness of our approach for producing contextualized and informative image captions. Within this approach, the captioning model describes an image while drawing information not only from the image itself but also from the general world knowledge (in this chapter, geographic knowledge in particular) bounded by the specifics of the image context. A piece of image-related data, such as the coordinates of its location, serves as an anchor, with which all the available external knowledge is reduced to the small subset related to the specific image, which is subsequently integrated into the captioning process. The approach proves to be effective for generating captions with relevant and accurate references to image-external data, without the need for extensive annotation efforts or large-scale pre-training.

However, the model described in this chapter also has obvious limitations. Consider the following caption from the GeoRic dataset: “Whithaugh Bridge. The new bridge crossing the river Liddel in Newcastleton opened in 2014”. The geo-aware model can utilize information about the geographic entities around the image location and is likely to produce a caption that correctly mentions Whithaugh Bridge in Newcastleton. But additional encyclopedic knowledge about this bridge, such as when it was opened, is not available to the model. Geographic knowledge is only a part of the general world knowledge that can potentially inform the caption generation process. The next chapter will describe an extension of the approach to more diverse external data and introduce a new component into the captioning pipeline, the knowledge context, which contains open-domain encyclopedic facts about the entities relevant to the image.

CHAPTER 4

Encyclopedic knowledge-aware image captioning

4.1 Introduction

This chapter presents a further development of our approach to incorporating various kinds of external knowledge into an image captioning pipeline[1]. The previous chapter demonstrated how using image location as a contextualization anchor allows a captioning model to gather information about relevant geographic entities around the image (a geo-entity context) and produce accurate references to them in the caption.

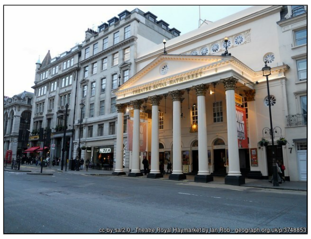

cc-by-sa/2.0 - Theatre Royal Haymarket by Ian Rob - geograph.org.uk/p/3748853

Figure 4.1: “Theatre Royal Haymarket. Dating back to 1720.” © Ian Rob.

Here, in addition to the geographic data, the model makes use of a wide range of open-domain encyclopedic facts that constitute a knowledge context. It is integrated into the pipeline alongside the geo-entity context at both the encoding and decoding

[1] The research presented in this chapter was previously described in Nikiforova et al. (2022).

stages, creating a 'knowledge-aware' captioning model. Such a model is able to produce informative captions with facts about the objects in the image, which are not inferrable from the image itself, such as, for example, "dating back to 1720" in Figure 4.1.

Section 4.2 reviews previous research on image captioning enriched with external encyclopedic data. Section 4.3 describes the types of knowledge found in the captions of the Geograph project, and a new image captioning dataset that we compile for training our knowledge-aware captioning model. Section 4.4 provides a detailed account of the knowledge context and its implementation in the captioning model. Section 4.5 describes the process of encoding the captions using two kinds of external data, geographic and encyclopedic, while Section 4.6 describes the architecture of the captioning model that incorporates the geographic and encyclopedic knowledge components. Section 4.7 presents its quantitative and qualitative evaluation, including a focused evaluation of the factual accuracy of the generated captions. Finally, Section 4.8 discusses the advantages of applying our approach to knowledge-aware image captioning and outlines the remaining issues.

4.2 Enhancing caption generation with encyclopedic data

Integrating external encyclopedic data into image captioning has not been the focus of much prior research, although the few existing works (Mogadala et al., 2018; Zhou et al., 2019; Huang et al., 2020; Bai et al., 2021) show its potential for improving informativeness and overall quality of the generated captions.

In contrast to the method developed in this dissertation, some of the existing approaches do not attempt to recognize specific entities in an image but instead utilize external knowledge about common objects. Mogadala et al. (2018) extract labels of the common objects in the image using custom classifiers trained on the MSCOCO dataset and retrieve DBpedia subgraphs related to the detected objects. These subgraphs, embedded with the RDF2Vec algorithm (Ristoski and Paulheim, 2016), are used as additional context for caption generation. Similarly, Zhou et al. (2019) and Huang et al. (2020) start with applying a pre-trained model to detect objects in the image (e.g., "stop sign", "book", "dog"). The resulting object labels are used to query the ConceptNet knowledge base to retrieve a set of related terms (e.g., "stop sign" ⟶ "bus", "bus station", etc.). In Zhou et al. (2019) the embeddings of the related terms, combined with the image features, initialize the caption generation module[2]; in Huang et al. (2020) the probability of generating the related terms is increased at decoding time. To test the efficacy of this methodology on our data and to compare it to the approach proposed in this dissertation, we develop a baseline model that involves common object and scene detection in the image, extracting information about the related concepts from an external database and introducing it into the captioning process.

Bai et al. (2021) develop a captioning model for fine art paintings. They train specialized classifiers to determine various specific characteristics of the painting: its author, style, time period and school, as well as employ a standard object detection model to identify what is depicted in the image. Based on the output of both custom classifiers and object detection, they extract relevant information from Wikipedia and use it in the captioning model, which increases the specificity and informativeness of the generated captions.

In general, all of these approaches rely exclusively on image processing to access relevant external knowledge. Thus, the result is always only as good as the image recognition algorithm, and the potential benefit of utilizing additional non-visual related data is left unexplored. We present an approach that uses the image location to unambiguously identify relevant specific entities and encyclopedic knowledge about them in external resources, followed by incorporating the extracted knowledge into the captioning pipeline.

4.3 Data

4.3.1 Knowledge types in Geograph captions

Naturally produced captions on the Geograph project website not only describe what can be directly seen in the image, but also contain various references to image-external knowledge. We distinguish the following broad categories of knowledge most commonly found in Geograph captions:

- **Geographic**: information about the image location and its relation to other geographic entities. For example, "Rough pasture to the north of Cogry Road", "This is a park between Catford and Lewisham.". This type of knowledge was

[2] A similar method is also explored in application to the related task of visual question answering in Wu et al. (2017), where short descriptions of the detected objects are extracted from DBpedia, encoded as vectors by Doc2Vec (Le and Mikolov, 2014) and added to the representations of the image and the question to produce an answer.

the focus of the previous chapter, in which we presented a geo-aware image captioning model.

- **Personal**: individual experiences and opinions of the person who produced the caption. For example, “Golders Green Road at dusk. This is where I have shopped all my life.”, “The churchyard at St Clement Eastcheap. my favourite city church”. Such remarks are strictly personal and subjective, and therefore, we do not intend to reproduce them in the automatically generated captions.

- **Situational**: temporary circumstances of image-related entities. For example, “Currently for sale - a conifer plantation southwest of Lauder.”, “The Thames Path is temporarily diverted due to building.”. This reflects the situation-specific knowledge of current events.

- **Encyclopedic**: ‘static’ information about image-related entities. For example, “St John’s Church, Bethnal Green. Designed by Sir John Soane.”, “Herne Hill Velodrome, London. Built in 1891”. These are facts about events or circumstances that are, if not permanent, stable for long enough to be documented in general databases. For instance, facts mentioned in the examples above are found in DBpedia: <St John on Bethnal Green, architect, John Soane> and <Herne Hill Velodrome, built, 1891>.

Our method relies on the relevant information being available in external resources (e.g., a database), which is typically not the case with situational and personal knowledge. Therefore, in this chapter we focus exclusively on ‘static’ encyclopedic facts about the entities related to the image. Even though the knowledge we introduce into the captioning model is only encyclopedic, for the sake of brevity, we call the model simply ‘knowledge-aware’ instead of ‘encyclopedic knowledge-aware’. Expanding the model to include situational or even personal knowledge is also possible, provided that the data is available in a suitable format (which will be discussed further in Section 4.4.1); this direction is, however, not explored in the present dissertation.

4.3.2 The K-GeoRic dataset

For training and testing the knowledge-aware image captioning model, we compile the K-GeoRic dataset (K for “knowledge”), following mostly the same procedure as for GeoRic (Section 3.3.2). Like GeoRic, this dataset contains images from the Geograph project website, with titles, captions and location metadata.

Image	URL	Title	Caption	Latitude	Longitude
	`https://www.geograph.org.uk/photo/3748853`	Theatre Royal Haymarket	Dating back to 1720.	51.50845	-0.13177
	`https://www.geograph.org.uk/photo/4750634`	Hammersmith Bridge	Designed by Sir Joseph Bazalgette and opened in 1887	51.48875	-0.22966

Table 4.1: Sample entries in the K-GeoRic dataset.

4.3.2.1 Train-validation-test data split

The train-validation-test split is again based on the latitude of the image location, which ensures testing on photographs of previously unseen objects. Out of 7,128 images in the dataset overall, 891 (12.5%) are assigned to the test set (taken to the north of the 54.8975° latitude), 891 (12.5%) to the validation set (taken between the 53.5706° and the 54.8975° latitude), and the rest, 5,346 (75%) to the train set.

4.3.2.2 Selection criteria

Data selection for K-GeoRic involves the same caption requirements as for GeoRic: the maximum length of 100 tokens and the presence of at least one geographic reference. In addition, we select the image only if its caption contains at least one encyclopedic fact about a relevant geographic entity. This requirement is enforced so that there is enough data for the captioning model to learn to incorporate external knowledge into caption generation. The resulting dataset has no overlap in images with GeoRic.

4.3.2.3 Statistics

Table 4.2 presents the statistics of the size and vocabulary diversity of the K-GeoRic dataset, compared to MSCOCO. Table 4.3 shows the named entity type distribution in K-GeoRic compared to GeoRic. K-GeoRic is very small by the general standards of image captioning datasets. However, its captions are almost 5 times longer than the MSCOCO ones and build up a vocabulary that is more diverse, as shown by multiple metrics. All of the K-GeoRic captions contain named entities, and there is a more even distribution in their types as compared to GeoRic, with the most substantial increase in the share of the DATE type, presumably due to a large presence of encyclopedic facts like “built in 1619”.

	Num Cap	**Avg Cap Len**	**Vocab Size (k), per 10k Tokens**	**Norm TTR**		**Freq Tokens Ratio**	**% Cap with NE**	**% NE in All Tokens**
				1-gram	2-gram			
K-GeoRic	7,128	53.14	2.15	40.78	80.85	73.28	100.0	14.95
MSCOCO	593,968	11.32	1.57	39.00	78.08	92.22	16.27	1.58

Table 4.2: Statistics of the K-GeoRic dataset, compared to MSCOCO. The metrics are defined in Section 2.2.2.3.

	% Person	**% Org**	**% Geo**	**% Num**	**% Date**	**% Other**
K-GeoRic	15.44	23.09	29.44	4.82	22.04	5.17
GeoRic	20.48	24.06	48.03	2.26	1.87	3.3

Table 4.3: Named entity type distribution in the K-GeoRic captions, compared to GeoRic.

4.3.2.4 Normalization

The normalization procedure for the captions remains the same as in GeoRic (see Section 3.3.2.4) and includes converting them to lower case, tokenization, minimal formatting clean-up and unification of commonly occurring synonyms.

4.4 The knowledge context of an image

Encyclopedic data is incorporated into the captioning model by adding a new component to the pipeline, the knowledge context: a set of facts about the entities in the entity context. Expanding the model described in Chapter 3, we focus on the entities from the geographic entity context *G* in particular.

4.4.1 Building the knowledge context

As the source of encyclopedic data, we use DBpedia, an open database that stores information available in Wikipedia in a structured format of <subject, predicate, object> triples[3]. Given the entities $(e_1 \dots e_n) \in G$, we extract DBpedia triples where the entities are the subjects, resulting in a collection of facts $(f_1 \dots f_m) \in K$.

[3] The terms ‘subject’ and ‘object’ here are defined technically as DBpedia entities, where the subject is the topic of a given DBpedia webpage and the object is an entity connected to it by any relation, as shown

f_1 – < Theatre Royal, built_in, 1720 >
f_2 – < Theatre Royal, architect, John Nash >
. . .
f_m – < Charing Cross, opened_in, 1864 >

Figure 4.2: Sample fragment of a knowledge context.

4.4.1.1 Number of facts

The number of facts extracted for different images can vary drastically: in urban, and especially central, historic areas, geographic entities are generally much better described in DBpedia than in less developed locations. In the K-GeoRic dataset, the initial number of facts in the knowledge contexts for different images ranges from 1 to 39,448. For the sake of computational efficiency, we limit the number of facts per knowledge context to a fixed value, but, in order to retain the most relevant ones, they are first ranked according to how probable they are to appear in a caption. For that, we train a ranking model, similar to the ranking model for the geo-entity context (Section 3.4.2): a logistic regression classifier that predicts whether a given fact will be mentioned in the caption or not. This model is trained on a subset of 1,000 Geograph captions, taking into account the fact's predicate, the ranking of the fact's subject in the geo-entity context and its geographic features (distance, size and type). Based on the experiments with the trained ranking model, the maximum size of the knowledge context for an image is set to 50 encyclopedic facts: for at least 95% of the images in the ranking model's test set, the top 50 ranked facts include at least one that appears in the corresponding caption.

Not every entity from G corresponds to a single fact in the knowledge context K: some are not featured in DBpedia, some, on the other hand, have more than one fact associated with them. Table 4.4 presents the statistics of the average number of facts and unique predicates per image and entity in the knowledge context.

on the subject's webpage. Thus, the page dedicated to Christopher Wren contains facts with the subject "Christopher Wren", including, for example, one with the object "St. Paul's Cathedral", connected to the subject by the relation "architect": <Christopher Wren, architect, St. Paul's Cathedral>. Conversely, on the page about St. Paul's Cathedral, there is a fact <St. Paul's Cathedral, architect, Christopher Wren>, where the subject is St. Paul's Cathedral and the object is Christopher Wren.

	Per image	Per entity
Average number of facts	36.49	7.34
Average number of predicates	24.96	5.82

Table 4.4: Average number of facts and predicates in the knowledge context.

4.4.2 Fact features in geographic embeddings

The geographic embedding function from Chapter 3 (Equation 3.1) is modified to include information about the number of facts associated with a given entity. Two new features are introduced, both of which are intended to reflect the salience of the geo-entity through the amount of data available about it in a knowledge base. The first one is the **presence of knowledge** $\exists \boldsymbol{f}$ — a binary indicator that shows whether or not the entity corresponds to any facts in the knowledge context, and the second one is the **number of facts** $\#\boldsymbol{f}$ that correspond to the entity in the knowledge context. The new geographic embedding function is given by Equation 4.1 and is used to update the geo-entity context embedding, Equation 4.2.

$$\textsc{GeoEmb}^k(e_i) = \text{Concat}[d_i,\ norm(a_i),\ s_i,\ \exists f_i,\ \#f_i, Emb_t(t_i)] \tag{4.1}$$

$$EmbG^k = (\textsc{GeoEmb}^k(e_1) \ldots \textsc{GeoEmb}^k(e_n)), e_i \in G \tag{4.2}$$

4.4.3 Identifying facts in captions

The collected knowledge context contains facts that are likely to be relevant for captioning, but it is much less straightforward to isolate a reference to a certain fact in a caption than, for example, a reference to a geographic entity. Each geo-entity is associated with a name, the presence of which is easy to detect in the caption. A fact can be expressed in various ways, e.g., <Theatre Royal, built, 1720> as “Theatre Royal was built in 1720”, “the building year of Theatre Royal is 1720” or “Theatre Royal dates back to 1720”. It can be noted, however, that the biggest source of this variability is generally the fact’s predicate. Indeed, in the example above, “Theatre Royal” and “1720” are always the same, and it is the predicate *built* that is expressed differently. Since we are not restricting the possible predicates in the knowledge context, it is infeasible to attempt to list all the lexical realizations of all the DBpedia predicates in order to match them in the captions. Instead, we match only the least variable parts of the fact: the subject and the object, and assume that if the caption contains the subject and the object of the fact f_i, then it contains a reference to the fact f_i.

In practice, since the subjects of all the facts in the knowledge context are geographic entities from *G*, we treat their occurrences in the captions as geo-entity names, and the occurrences of the fact objects are treated as the proxy for the facts. For example, we locate the fact <Theatre Royal, built, 1720> in a caption "Theatre Royal Haymarket. Dating back to 1720" because it contains the fact's subject "Theatre Royal" and object "1720". During caption encoding though, "Theatre Royal" is treated as a geo-entity name supplied by *G*, and only the token "1720" is considered to come directly from the knowledge context (more details to follow in Section 4.5).

4.4.4 Fact embedding

In the captioning process each fact is encoded as a linear combination of its subject and predicate embeddings, see Equation 4.3. The object of a fact is considered to be a function of the subject and predicate combined: e.g., the meaning of "1720" is simply '[when] Theatre Royal [was] built'. In a way, a fact object is a mere label of the fact itself, representing it in the caption text, similarly to how the names of geographic entities are also just labels of the real world entities from the geo-entity context. In a sentence "Theatre Royal dates back to 1720", the name "Theatre Royal" is a label of an entity from *G*, which has a certain set of geographic and physical characteristics (distance from the image, size, etc.); the token "1720" is similarly a label of the fact about '[when] Theatre Royal [was] built'. Thus, the "fact embedding" is computed as shown in Equation 4.3.

$$\textsc{FactEmb}(f_i) = \textsc{GeoEmb}^k(e_i) + Emb_p(p_i) \tag{4.3}$$

where e_i is the subject of the fact f_i and p_i is its predicate. Since the subjects of the facts are entities from *G*, we can use the geographic embedding function $\textsc{GeoEmb}^k$ for their representation. Emb_p is a separate embedding function for the predicates; predicate embeddings are initialized randomly and updated in an end-to-end fashion during the training of the captioning model.

The function is applied to each element of the knowledge context to embed it for further use by the captioning model:

$$EmbK = (\textsc{FactEmb}(f_1) \dots \textsc{FactEmb}(f_m)), f_i \in K \tag{4.4}$$

With fact objects interpreted as labels that represent facts in captions, the knowledge context now looks as shown in Figure 4.3.

f_1 – 1720	< Theatre Royal, built_in >
f_2 – John Nash	< Theatre Royal, architect >
…	
f_m – 1864	< Charing Cross, opened_in >

Figure 4.3: Knowledge context with fact objects as labels.

4.4.5 Knowledge contexts in K-GeoRic

Knowledge contexts of all the images in the K-GeoRic dataset contain altogether facts with 1,542 unique predicates, 336 of which appear at least once in the captions (a caption in the dataset contains on average 2.2 fact references).

Figure 4.4 shows the distribution of the top 50 frequent predicates to appear in the captions. Predicates of the form "years_[NUM]" ("years_0", "years_1", etc.) have all originated from the same ambiguous DBpedia predicate "years" that vaguely means 'an important year' and can denote the year of building, opening, closing, etc. On a given DBpedia page, it usually occurs in multiple different facts with the same subject, and there are certain tendencies connecting the order of the occurrence and the predicate's concrete meaning, e.g., the first occurrence in most cases refers to the opening year. So we alleviate the ambiguity of this predicate by combining the predicate itself with its sequence number.

Many of the predicates displayed in Figure 4.4 are in fact synonymous, i.e., reflect the same relation between the subject and object entities. For example, "opened" and "openingyear" are both used to link a geographic entity (the subject) and the year when it was opened (the object). Ideally, every unique relation between entities should be represented by a single predicate. On this account, we merged frequent synonymous predicates into a single predicate that represents their common meaning, e.g., both "opened" and "openingyear" are merged into "opened"[4]. The distribution of the top 50 most frequent predicates after merging the synonymous ones are shown in Figure 4.5.

[4] We take a cautious approach determining which predicates to merge. For example, even though the predicate "years_1" is very often synonymous with "closed", there is a significant number of cases when it denotes the year of (re)opening an entity, which is why we do not consider it fully synonymous with "closed" and do not merge them into one. Future work could explore this issue more systematically and develop a general way to identify and merge synonymous predicates in any domain.

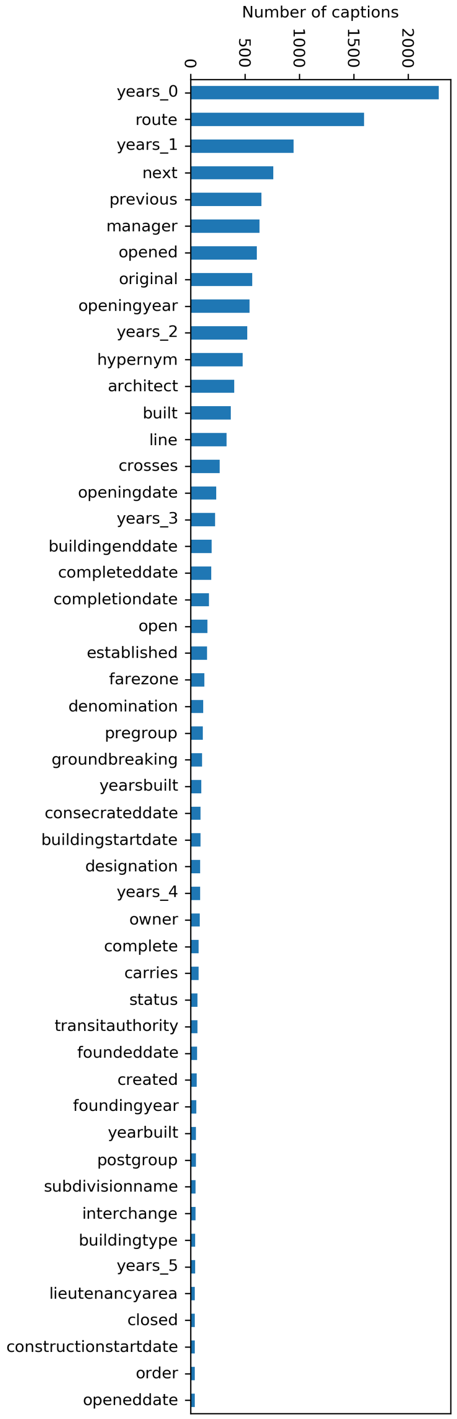


Figure 4.4: Distribution of the top 50 frequent original predicates of facts detected in the captions of the K-GeoRic dataset.

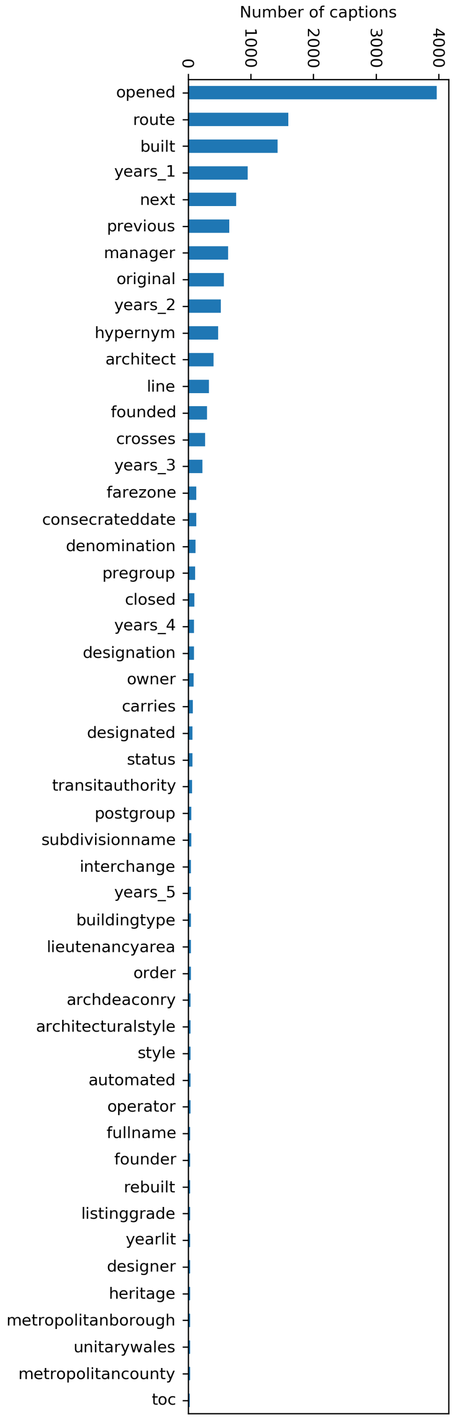


Figure 4.5: Distribution of the top 50 frequent predicates after merging the synonymous ones.

4.5 Encoding captions

The caption encoding procedure is similar to the one employed in the geo-aware model (Section 3.5). A significant difference is the introduction of a third type of tokens, in addition to geographic entity names and regular vocabulary words: fact objects — caption tokens that refer to the objects of the facts from the knowledge context. The mask vector, which indicates the type of each token and is used during caption embedding, now has three possible values: 0 (regular vocabulary), 1 (geographic entity) and 2 (fact object).

Consider the previous example from Section 3.5, where a sample fictitious caption “Sacred Heart Church. An impressive Gothic building, located near University Hospital.” was encoded based on the geo-entity context *G* and the vocabulary *V* with the mapping function in Equation 3.4. The resulting series of numbers, [9050, 1, 2, 3, 4, 5, 6, 7, 8, 9062, 1], represented “Sacred Heart Church” and “University Hospital” as geographic entity names and the rest of the tokens as regular vocabulary words. We now consider the additional knowledge context *K* and update the mapping function to include the third option for encoding the tokens that are present as objects in *K*:

$$\begin{bmatrix} \text{1575} & < \text{Sacred Heart Church, built} > \text{— 0} \\ \text{gothic} & < \text{Sacred Heart Church, style} > \text{— 1} \\ \dots & \\ \text{general} & < \text{University Hospital, type} > \text{— 28} \\ \dots & \end{bmatrix}$$

$$w_i \rightarrow \begin{cases} \text{len}(V) + \text{len}(G) + K[w_i], & \text{if } w_i \in K \\ \text{len}(V) + G[w_i], & \text{if } w_i \in G \\ V[w_i], & \text{otherwise} \end{cases} \tag{4.5}$$

Thus, “gothic” is now recognized as a fact object from the knowledge context, which updates its encoding from 4 (the index of the word “gothic” in the vocabulary) to 9351 (taking the length of *V* to be 9050 and the length of *G* to be 300). The resulting sequence of numbers that encodes this caption’s tokens is [9050, 1, 2, 3, 9351, 5, 6, 7, 8, 9062, 1].

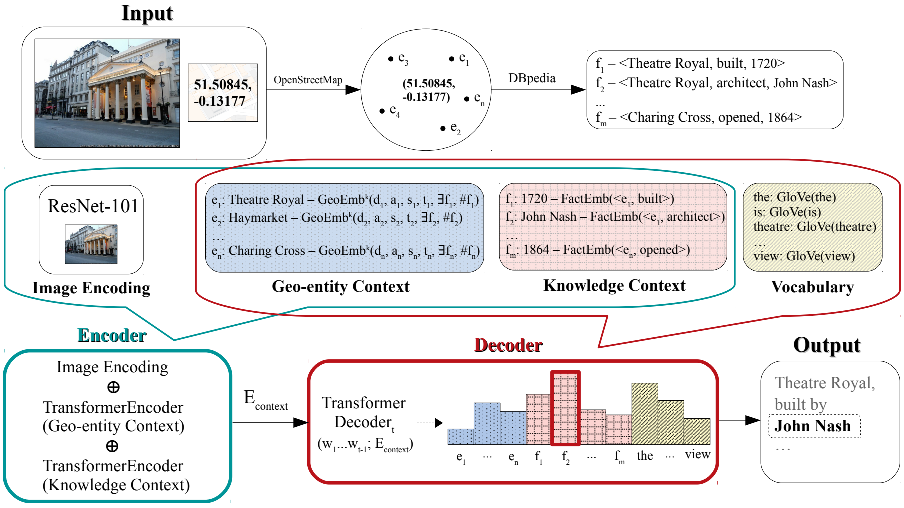


Figure 4.6: Overview of the knowledge-aware captioning model architecture.

4.6 Model architecture

The knowledge-aware captioning model extends the architecture of the geo-aware model introduced in the previous chapter. All the stages and components of the model integrate the knowledge context with relevant encyclopedic facts, which influence the captioning process, including being directly generated in the captions. The general overview of the model's architecture is shown in Figure 4.6.

4.6.1 Encoder

The encoder of the captioning model is further enhanced with the addition of the knowledge context, to be processed along with the visual features of the image and the relevant geographic entities in the geo-entity context. The embedded knowledge context is passed through a Transformer encoder with a standard structure and concatenated with the rest of the context components (visual and geographic), the encoding of which is the same as described in Section 3.6.1.

$$\begin{aligned} E^k_{context} = \text{Concat}[\,&\textsc{ResNet-101}(\text{Img}),\ \textsc{TransformerEncoder}(\text{EmbG}^k), \\ &\textsc{TransformerEncoder}(\text{EmbK})] \end{aligned} \tag{4.6}$$

The resulting representation comprises a diverse, informative context for caption

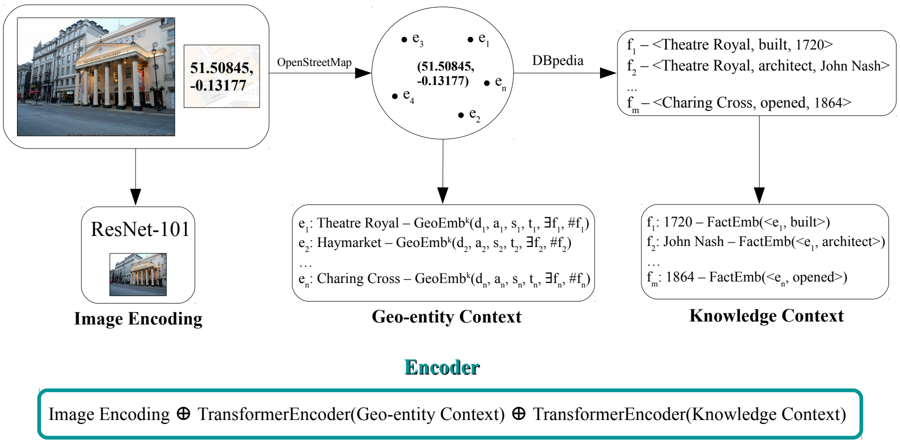

51.50845,
-0.13177
OpenStreetMap
e3
e1
(51.50845,
-0.13177)
en
e4
e2
DBpedia
f1 – <Theatre Royal, built, 1720>
f2 – <Theatre Royal, architect, John Nash>
...
fm – <Charing Cross, opened, 1864>
ResNet-101
Image Encoding
e1: Theatre Royal – GeoEmbk(d1, a1, s1, t1, ∃f1, #f1)
e2: Haymarket – GeoEmbk(d2, a2, s2, t2, ∃f2, #f2)
…
en: Charing Cross – GeoEmbk(dn, an, sn, tn, ∃fn, #fn)
Geo-entity Context
f1: 1720 – FactEmb(<e1, built>)
f2: John Nash – FactEmb(<e1, architect>)
…
fm: 1864 – FactEmb(<en, opened>)
Knowledge Context
Encoder
Image Encoding ⊕ TransformerEncoder(Geo-entity Context) ⊕ TransformerEncoder(Knowledge Context)

Figure 4.7: The knowledge-aware encoder.

generation that allows the decoder to take into account the visual, geographic and encyclopedic data.

4.6.2 Decoder

During caption generation, the decoder also attends to the knowledge context besides the vocabulary and the geographic entities. The process resembles decoding in the geo-aware model, with several important updates to account for the additional data.

The tokens are embedded according to which of the three types they represent: either with the fact embedding function, or the geographic embedding function, or using the pre-trained GloVe embeddings for regular vocabulary words (Equation 4.7). The embeddings are combined with the encoding of the token's position in the sentence (Equation 4.8). The output of the encoder (the combined representation of the visual, geo-entity and knowledge contexts) and the positional embeddings of the preceding tokens are passed through a standard Transformer decoder (Equation 4.9).

$$Emb^k(w_i) = \begin{cases} \textsc{FactEmb}(w_i), \text{if } w_i \in K \\ \textsc{GeoEmb}^k(w_i), \text{if } w_i \in G \\ \textsc{GloVe}(w_i), \text{otherwise} \end{cases} \tag{4.7}$$

$$PosEmb^k(w_i) = Emb^k(w_i) + Pos(w_i) \tag{4.8}$$

$$h_t^k = \textsc{TransformerDecoder}(PosEmb^k(w_{1...t-1}); E_{context}^k) \tag{4.9}$$

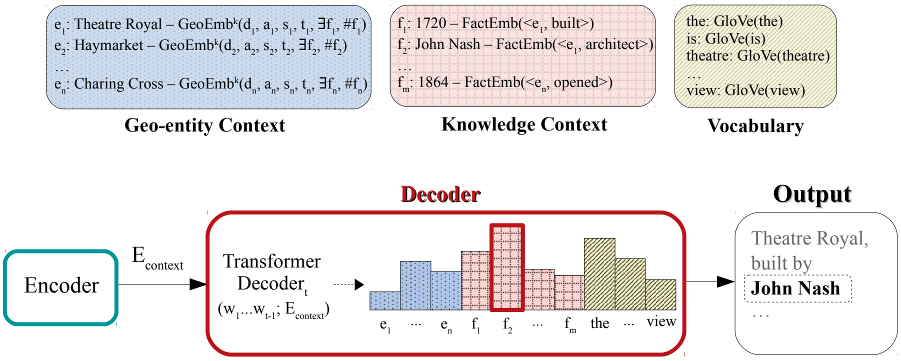

e$_1$: Theatre Royal – GeoEmb^k(d$_1$, a$_1$, s$_1$, t$_1$, ∃f$_1$, #f$_1$)
e$_2$: Haymarket – GeoEmb^k(d$_2$, a$_2$, s$_2$, t$_2$, ∃f$_2$, #f$_2$)
…
e$_n$: Charing Cross – GeoEmb^k(d$_n$, a$_n$, s$_n$, t$_n$, ∃f$_n$, #f$_n$)
Geo-entity Context
f$_1$: 1720 – FactEmb(<e$_1$, built>)
f$_2$: John Nash – FactEmb(<e$_1$, architect>)
…
f$_m$: 1864 – FactEmb(<e$_n$, opened>)
Knowledge Context
the: GloVe(the)
is: GloVe(is)
theatre: GloVe(theatre)
…
view: GloVe(view)
Vocabulary
Decoder
Output
Encoder
E$_{context}$
Transformer Decoder$_t$ (w$_1$...w$_{t-1}$; E$_{context}$)
e$_1$
…
e$_n$
f$_1$
f$_2$
…
f$_m$
the
…
view
Theatre Royal, built by
John Nash
…

Figure 4.8: The knowledge-aware decoder.

Then, given h_t^k, three sets of scores are calculated: probability distributions over the vocabulary, the geo-entity context and the facts in the knowledge context.

$$\begin{aligned} y_{v_1}...y_{v_l} &= (h_t^k \circ (p_{ind}\, W_{pred} + \beta))\, W_{vocab} \\ y_{e_1}...y_{e_n} &= (EmbG^k \ \textsc{diag}(h_t^k))\, \vec{w}_{geo} \\ y_{f_1}...y_{f_m} &= ((EmbK \circ g_{ind})\ \textsc{diag}(h_t^k))\, \vec{w}_{fact} \end{aligned} \tag{4.10}$$

where W_{pred}, W_{vocab}, $\vec{w}_{geo}$ and $\vec{w}_{fact}$ are trainable linear transformation matrices/vectors, β is a bias vector (the rest of them are omitted for simplicity), $\circ$ stands for a Hadamard product, p_{ind} and g_{ind} are described in turn below. This calculation is similar to Equation 3.9 in Chapter 3, with two significant modifications: computing the additional scores for the knowledge context and introducing p_{ind} into the calculation of vocabulary scores.

Finally, the scores for the vocabulary words, geographic entities and facts are concatenated and fed to a softmax layer, which produces an overall probability distribution over the tokens of all types, as shown in Figure 4.8. The one with the highest probability is then generated at position t.

$$\begin{aligned} & w_t = \arg\max_{w_i} P(w_i), w_i \in V \cup G \cup K \\ & \text{where } P(w_i) = \sigma_i(\text{ Concat}[y_{v_1}...y_{v_l}, y_{e_1}...y_{e_n}, y_{f_1}...y_{f_m}]) \end{aligned} \tag{4.11}$$

Predicate indicator p_{ind} stands for ‘predicate indicator’, which is used for extra contextualization in generating regular vocabulary words. It is a binary vector that reflects for each of the known DBpedia predicates whether the knowledge context of a given image contains a fact with this predicate *and* the subject that is already present in the caption.

For example, suppose that the first three slots of the vector are dedicated to the predicates “built”, “opened” and “designer”. Suppose also that the previously generated tokens include the name “Theatre Royal” and no other geographic names, and the knowledge context contains only the following facts: <Theatre Royal, built, 1720> and <Charing Cross, opened, 1864>. Then the first three slots of the p_{ind} vector would be [1, 0, 0]: there is a fact in *K* that contains the predicate “built” and the subject that is already mentioned in the caption (“Theatre Royal”), there is no fact in *K* with the predicate “opened” and the subject mentioned in the caption (since “Charing Cross” is not in the caption), and there is no fact in *K* with the predicate “designer”.

Introducing this indicator into the calculation of vocabulary scores is intended to reduce the risk of producing phrases that would require a specific type of fact being generated afterwards if no such facts are available in the knowledge context. In the example above, it should reduce the probability of generating “designed by” after “Theatre Royal”, since there is no fact in *K* that would provide the correct name of the designer to follow that.

Geo-entity indicator The calculation of the scores for the knowledge context facts includes g_{ind}, the geo-entity indicator. It specifies for each fact in *K* whether its subject has already been mentioned in the caption at a given time step (1 for ‘has been mentioned’, 0 for ‘has not been mentioned’). Facts with the subjects that are already in the caption would generally get a higher generation probability assigned to them by the model.

4.7 Evaluation

Evaluation of the model is carried out using both standard and custom captioning metrics, with its performance compared to a range of baselines. Just like in the previous chapter, the goal is not only to compare the generated captions to the ground truth ones but also to assess how accurately they represent the provided context. In this case we focus on the knowledge context and the correctness of the verifiable factual statements produced in the captions.

4.7.1 Baselines

Some of the baselines in this chapter are the same that were used in comparison with the geo-aware model (Section 3.7.1). Several new ones provide a focused contrast with the knowledge-aware model that incorporates encyclopedic facts into image captioning.

ClipCap-MSCOCO/Conceptual These two baselines demonstrate the performance level of a standard pre-trained captioning model ClipCap (Mokady et al., 2021) on the K-GeoRic test set. Both implementations were trained on the out-of-domain data: the MSCOCO (ClipCap-MSCOCO) and Conceptual Captions (ClipCap-Conceptual) datasets.

No-context This model shares the general structure with the knowledge-aware one but both geographic and knowledge contextualization components have been removed; the remaining encoder-decoder pipeline was trained on the K-GeoRic dataset. Its performance represents the level that can be achieved by a standard image caption generator, trained on the in-domain data with no additional contextualization.

Related-concepts The goal of this baseline is to compare the approach proposed in this dissertation to another one that is based on the existing research (Huang et al., 2020; Zhou et al., 2019, see also Section 4.2). Here, external knowledge that informs caption generation is extracted based solely on the image[5]. Similarly to the previous works, we use a pre-trained object recognition model to detect generic objects and scenes in the image (e.g., "beach"). Specifically, we employ the VGG16 CNN model trained on the Places365 database (Zhou et al., 2017) for scene recognition (`https://github.com/GKalliatakis/Keras-VGG16-places365`). Then we query the ConceptNet knowledge base to retrieve the terms related to these objects (e.g., "coast", "driftwood", "ocean"). Top 5 objects are identified in every image and top 10 closest ConceptNet terms are extracted for each of them. Their GloVe embeddings are passed through a Transformer encoder, and the result is concatenated with the encoding of the image to construct a combined context representation for caption generation. The decoder of this captioning model is a standard Transformer decoder with no specific modifications.

Geo-aware This baseline is the model we introduced in Chapter 3, trained on the K-GeoRic dataset. It utilizes the original geographic embeddings without the fea-

[5] The same baseline was called K-(IMAGE-ONLY) in Nikiforova et al. (2022).

tures based on the number/presence of related facts, the encoder processes only the visual image features and the geo-entity context, and the decoder generates the caption picking the tokens only from the vocabulary and the geographic entity names. The difference between the performance level of the geo-aware and knowledge-aware models demonstrates the specific impact that the knowledge context has on the generated captions.

The remaining three baselines are only used in factual accuracy evaluation, where they present the biggest contrast with the knowledge-aware model. In terms of the standard metrics that compare the generated captions to the ground truth ones, these models do not differ much from the knowledge-aware one.

No predicate indicator This baseline presents an alternative setup without the predicate indicator p_{ind}. The scores calculation from Equation 4.10 is modified for the vocabulary words $v_1...v_l \in V$ as follows:

$$y_{v_1}...y_{v_l} = h_t^k \, W_{vocab} \tag{4.12}$$

No geo-entity indicator This baseline presents an alternative setup without the geo-entity indicator g_{ind}. Here, the scores calculation from Equation 4.10 is modified for the facts $f_1...f_m \in K$ as follows:

$$y_{f_1}...y_{f_m} = (EmbK \ \textsc{diag}(h_t^k)) \ \vec{w}_{fact} \tag{4.13}$$

Random fact object This baseline is similar in its purpose and implementation to the random geo-entity baseline from Section 3.7.1. It emulates a captioning model that has access to the relevant knowledge context for each image but is only able to pick the fact object to produce in the caption at random. Here we do not train a new model but take the captions generated by the trained knowledge-aware model for the images in the test set and replace the fact objects in each caption with the objects randomly picked from the knowledge context, preserving the type (e.g., replace generated years with randomly picked years, generated people's names with randomly picked people's names, etc.). With this restriction, the median number of unique objects to choose from is only 3.0, making it an especially strong baseline.

4.7.2 Results

4.7.2.1 Quantitative evaluation

Table 4.5 contains the results of the evaluation with the standard captioning metrics that compare the generated captions to the ground truth ones.

		BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
ClipCap	MSCOCO	0.33	0.0	5.62	2.45	0.14
	Conceptual	0.04	0.0	8.45	2.96	0.10
No-context		16.29	2.25	23.14	7.24	1.75
Related-concepts		19.71	2.4	23.48	7.76	1.63
Geo-aware		21.91	5.32	30.19	10.98	16.59
Knowledge-aware		**26.84**	**9.5**	**33.47**	**13.52**	**32.41**

Table 4.5: Standard metric scores of the models, measured on the test set of K-GeoRic.

Both ClipCap models produce captions that directly describe the visual content of the images, which is different from the contextualized and knowledge-rich K-GeoRic captions, yielding low scores in the metrics. The no-context and related-concepts models achieve higher scores, the latter with a marginal improvement over the former in most of the metrics. This is expected since they were trained on the in-domain images of the K-GeoRic dataset.

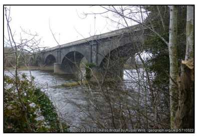
The geo-aware model is a further improvement over all the previous ones. An increase in all the metrics, with a particular surge in CIDEr (which is sensitive to rare and informative words), is the same phenomenon that we have previously observed in Section 3.7.2.1. It is explained by the model's access to image-specific geo-entity contexts, which enable it to produce contextually appropriate geographic names, instead of having to pick them from the regular vocabulary.

The knowledge-aware model's performance is the highest among all, with statistically significant gains in all the metrics (two-sample t-test, p <0.001), and especially in CIDEr. The improvement over the geo-aware model in particular shows the benefits of the added knowledge context component.

4.7.2.2 Qualitative evaluation

Table 4.6 shows the captions generated for the sample images by all the models, along with the original human-produced ones.

(a)



Ground truth: Kelso Bridge. Below the confluence of the Rivers Tweed and Teviot. John Rennie engineered the bridge, which was built between 1800 and 1803.

ClipCap-MSCOCO: A river with a bridge and a train on it.

ClipCap-Conceptual: the bridge over river in the village.

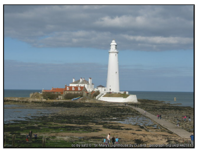
No-context: the *river dee*. *farndon bridge* was opened in *1339* by *monks* from *farndon bridge*.

Related-concepts: *chertsey bridge*. *chertsey bridge* dates from *1785*.

Geo-aware: the **river tweed**. the bridge is a *grade ii listed* building. the bridge was built in *1826* and completed in *1805*.

Knowledge-aware: **kelso bridge**. the bridge over the **river tweed** was built in **1800**, and was designed by **john rennie the elder**.

(b)



Ground truth: St. Mary's Lighthouse. On St Mary's Island, just north of Whitley Bay. Completed in 1898 the lighthouse remained operational until 1984 when it was superseded by modern navigational techniques.

ClipCap-MSCOCO: A lighthouse with a white tower on top.

ClipCap-Conceptual: lighthouse on the coast of person.

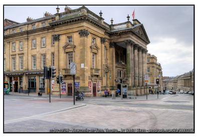
No-context: the *naze tower, walton on naze*. the *naze tower* was built in *1720* as a navigational aid to *1720* as a navigational aid to aid

Related-concepts: *portland bill car park*. *portland bill* is a popular destination for visitors to the lighthouse, built in *1906*.

Geo-aware: **st marys lighthouse**. the lighthouse was built in *1833*, and is **grade ii listed**.

Knowledge-aware: **st marys lighthouse**. the lighthouse was built in **1898** and is now a **grade ii listed** building.

(c)



Ground truth: Newcastle-on-Tyne, Theatre Royal. The Theatre Royal is a Grade I listed building (English Heritage List Entry Number: 1145840 Link) situated on Grey Street. It was designed by local architects John and Benjamin Green as part of Richard Grainger's grand design for the centre of Newcastle, and was opened on 20 February 1837. The interior of the building was destroyed by a huge fire in 1899 (Link The Guardian). It reopened in 1901 with the interior redesigned. Externally, the building is exactly as it was when it was first built.

ClipCap-MSCOCO: A street corner with a building and a horse statue.

ClipCap-Conceptual: the building in the centre.

No-context: the *old shades public house*. the house was built in *whitehall*, completed in *1874*, designed by *charles driver*. it is a *grade ii* listed building.

Related-concepts: the *angel, islington*. the portico of the roman numerals on the street. the building was built in *1906* to the designs of architect *edwin maxwell fry*. the building is *grade ii* listed.

Geo-aware: the **theatre royal**. the **theatre royal** is a *grade ii* listed building. the theatre was built in *1890*.

Knowledge-aware: **theatre royal, newcastle upon tyne**. *grade ii* listed building dates from **1837**.

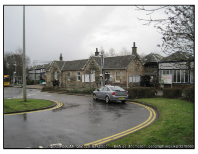
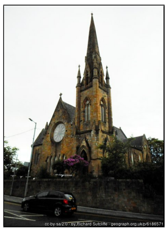
(d)		Ground truth: Milngavie railway station, East Dunbartonshire. Opened in 1863 by the Glasgow & Milngavie Junction Railway. View south west at forecourt. ClipCap-MSCOCO: A car driving down a street past a building. ClipCap-Conceptual: the main road to the village. No-context: *st johns wood church, threapwood*. the church was built in *1848* by the architect *john soane*. *grade ii* listed building (english heritage building id: link). Related-concepts: *staines station*. *staines station* is a station on the *waterloo to reading line, and the junction of windsor*. the station was opened in *1848*. Geo-aware: **milngavie station**. the station is located on the **milngavie branch**. the station is located on the site of the station, and the station was opened in **1863**. Knowledge-aware: **milngavie station**. the station was opened in **1863**. it is now part of the **abellio scotrail**.
(e)		Ground truth: Lenzie Old Parish Church. A Category C listed church [Link], dating from 1874. ClipCap-MSCOCO: A church with a clock tower and a cross on top. ClipCap-Conceptual: the church in the centre. No-context: *st peters church, colwyn bay*. *st asaph*. *st pauls church* was consecrated in *1887*. Related-concepts: *christ church, heaton norris*. *christ church* was built in *1846*. a developed *early english style*. the church was built in *1846*. the church was designed by *sir christopher wren*. Geo-aware: **lenzie old parish, lenzie**. built in *1846*, the church was built in *1846*. Knowledge-aware: **lenzie old parish church , lenzie**. **lenzie old parish church** was built in **1874** and designed by **clarke and bell**. it is now a *grade ii* listed.

Table 4.6: Examples of the captions generated for the K-GeoRic images from the test set. Correct geographic references and facts are given in **bold**; incorrect ones are given in *italics*. Facts about entities that are unrelated to the image are also given in italics, even if they are technically correct on their own.

The ClipCap models produce mostly accurate descriptive captions, although hallucination remains an issue (e.g., “train” in image (a), “cross” in image (e)). There are no concrete references to any image-external data. Interestingly, the caption generated by ClipCap-Conceptual for image (b) specifies the location of the lighthouse as “the coast of person”. This is most likely due to the pre-processing procedure employed by the creators of the Conceptual Captions dataset: named entities were replaced with tokens that indicate their type. For example, phrases like “the coast of Saint Helena” could be replaced by “the coast of PERSON”, which was then learned by the model from the training data.

The no-context and related-concepts models produce captions of similar quality to each other, indicating that the additional external knowledge from ConceptNet does not have a major effect on caption generation for K-GeoRic images. Both models imitate the contextualized ground truth captions by attempting to name the geographic objects in the images and producing specific encyclopedic facts about them. The facts can even be technically correct but they are never related to the images that they are supposed to describe. For example, in (a), Farndon Bridge does indeed cross the River Dee and was opened in 1339, and Chertsey Bridge does date from 1785. But these facts are not relevant for the description of this particular image, since it depicts Kelso Bridge. Information about the other bridges was evidently “memorized” from the training data of K-GeoRic. For these two captioning models, the training data is the only source of knowledge about specific real world entities. Thus, the risk of hallucination is very high for the images of objects that were not seen during training. All of the K-GeoRic test images present this challenge due to the latitude-based train-validation-test split (Section 4.3.2.1).

The geo-aware model makes use of the geo-entity context to generate captions with correct references to relevant geographic entities. However, facts about them are generated from the regular vocabulary and, as a result, are mostly incorrect.

Finally, the knowledge-aware model is able to utilize the knowledge context in addition to the geo-entity context, and produce correct encyclopedic facts as well as accurate geographic references. Out of all the factual statements produced by the knowledge-aware model in Table 4.6, only two are incorrect: “grade ii” in captions (c) and (e). Both of them were actually generated as regular vocabulary words and not drawn from the knowledge context. This means that “grade ii” was selected because it occurred frequently in similar sentences in the training data, and not because it was related to a specific relevant entity, as it would have been if it was generated based on the external knowledge.

4.7.2.3 Factual accuracy evaluation

Measuring the accuracy of factual statements in the captions is in essence a task of fact verification in the presence of relevant background data (the knowledge context). Fact verification is a challenging problem on its own, with dedicated datasets (Thorne et al., 2018a; Chen et al., 2019b; Jiang et al., 2020; Mishra et al., 2022), shared tasks (Thorne et al., 2018b, 2019; Aly et al., 2021), and a wide application in specific domains, such as dialogue generation (Santhanam et al., 2021; Honovich et al., 2021; Dziri et al., 2022, 2021). To make the task manageable and at the same time more focused, we target only those factual statements that correspond to some of the most frequent predicates that appear in the ground truth captions of the K-GeoRic dataset.

We create a rule-based metric that searches for common predicate-specific key phrases in the caption (e.g., *designer* — "designed by", *opened* — "opened in"). It then verifies that the geographic name generated in the caption in connection to a given phrase is actually related to the image; if it is not, the fact is omitted from the subsequent calculations, since the correctness of facts that are unrelated to the image is of less importance in this evaluation. Once the predicate and the geo-entity name (the fact's subject) are identified, we extract the corresponding fact objects from the knowledge context and check if they are present in the caption. A fact is considered correct if its object is found in the caption and incorrect otherwise. The metric's final score is the percentage of correct facts among all generated facts.

The additional complexity in the fact accuracy measurement comes from the ambiguity of certain predicate-related phrases. For example, if the generated caption contains a phrase "[X], dating from 1710", it is not clear to which of the facts in the knowledge context it maps to: <[X], founded, 1709>, <[X], built, 1710> or <[X], opened, 1711>. Since it is impossible to disambiguate the phrase "dating from" in this context, all three options are considered correct, so, the metric treats the generated fact as accurate if it includes any of these three years: "[X], dating from 1709/1710/1711".

Another situation when there is more than one possible 'correct answer' for the metric is when the knowledge context contains several facts with the same subject and predicate but different objects. This can happen, for example, when a single construction is linked to multiple architects, or, due to closing and then reopening, there have been multiple opening years for a single building. In such cases the metric considers any of these different fact objects to be correct, if they are generated in the caption with the corresponding subject and a predicate-specific phrase.

Table 4.7 shows the accuracy scores of the knowledge-aware model and the three baseline models. The standard pre-trained (ClipCap), no-context and related-concepts models are not shown in the table, since they did not produce any facts that would be included in the accuracy measurement (standard models trained on the out-of-domain data do not produce factual statements at all, and the no-context and related-concepts ones do not produce facts about the entities related to the image).

	Accuracy, %
Geo-aware	6.56
Random fact object	53.35
No geo-entity indicator	76.86
No predicate indicator	80.74
Knowledge-aware	**86.08**

Table 4.7: Factual accuracy scores of the models.

The geo-aware model's encyclopedic facts are mostly drawn from the regular vocabulary, making it unlikely that they will be correct. The few that are actually correct are either coincidental guesses (such as "opened in 1863" in caption (d) in Table 4.6) or result from using a relevant geographic name in the factual statement (such as "located on the milngavie branch" in caption (d), where "milngavie" is actually a name from the geo-entity context). All the other baseline models show significant improvements over the metric score of the geo-aware model. The random fact object model chooses the fact objects randomly, but only from the highly relevant ones in the knowledge context, which makes its metric score quite competitive[6].

The architecture without the geo-entity indicator scores almost 10 percentage points lower than with it. When the geo-entity indicator is not included, the model is more prone to generating facts, the subjects of which have not been mentioned in the caption. This often results in a mistake (a hallucination) when the model selects a fact with an appropriate predicate but an incorrect subject, e.g., generating "neidpath viaduct, built in 1263" with the following two facts in the knowledge context: <Neidpath Viaduct, built, 1863> and <Neidpath Castle, built, 1263> (the predicate "built" is fitting to the generated "built in" but the year is taken from the fact about an entity that is not present in the caption, Neidpath Castle).

In the setup without the predicate indicator, the influence of the knowledge context on vocabulary word selection is not as strong. This leads to a decrease in the generated fact accuracy by almost 6 percentage points compared to when it is present. A typical

[6] The score is further increased because of the situations when there are multiple options accepted as correct by the fact accuracy metric, as described above. These situations, however, lead to the increased metric scores for all the models, which is why the comparison between them is not compromised.

mistake of this model is generating a phrase from the vocabulary that warrants a certain type of fact that is not available in the knowledge context; for example, generating "elie house, founded in 1697" when there is no fact in the corresponding knowledge context with the subject "Elie House" and the predicate "founded" (though there is a fact <Elie House, built, 1697>). Because there is no appropriate fact, generating "founded in" after "elie house" would very likely lead to producing an incorrect statement. The predicate indicator's purpose is to account for what is available in the knowledge context while generating vocabulary words, in order to avoid this kind of mistakes.

The knowledge-aware model significantly outperforms all the alternatives. Its score in general is quite high — over 86% of the generated facts are correct, showing that the approach it is based on is indeed effective for incorporating structured encyclopedic knowledge into the generated captions. The two models without the predicate and geo-entity indicators perform much worse, showing the importance of these components in reducing hallucinations and improving factual accuracy.

4.8 Discussion

This chapter has presented an extension of the geo-aware model from Chapter 3, incorporating open-domain encyclopedic knowledge that is related to the image but not directly inferrable from it. As shown by standard and custom metrics, integrating this knowledge into the captioning process facilitates generation of informative captions with correct facts about entities relevant to the image.

The principal difficulty in generating knowledge-rich captions is that there is no cue for the knowledge in the image itself. It is hardly possible to identify, for example, the name of a bridge just by looking at its photograph, let alone the year when this bridge was opened. Standard captioning models can produce this information only based on a caption of a similar photograph in the training data, but the risk of hallucination is high and generalization to unseen objects cannot be achieved. Our approach solves this issue. The captioning model is able to access external world knowledge in the form of the geo-entity context and the knowledge context. The former provides information about the nearby objects (including, for example, the name of a relevant bridge), and the latter provides various facts about these objects (including the bridge's opening year). During caption generation, the model switches between describing the visual content of the image and, when necessary, copying names and facts from image-external knowledge sources.

An important question that remains is whether our approach can be generalized to other domains. All the experiments described so far have utilized images and captions of the Geograph project. Naturally produced and highly contextualized, they provided a good testbed for our research questions. However, different types of images and captions, as well as a different contextualization anchor, might present new challenges. The next chapter will describe an application of this approach to a qualitatively different dataset of news images and discuss related issues and necessary enhancements of the model.

CHAPTER 5

Towards generalization: news image captioning data

5.1 Introduction

In the previous chapters, our image captioning approach was shown to be effective in application to the images and captions of the Geograph project, with image location used as a contextualization anchor. This chapter presents evidence for the generality of this approach by applying it to the qualitatively different dataset of images from news articles. This case study will demonstrate what benefits and limitations our approach exhibits in a new domain, in which factual accuracy of captions is highly important.

News image captioning is a challenging problem in contextualized caption generation that has recently attracted considerable attention (Biten et al., 2019; Tran et al., 2020; Liu et al., 2021a; Hu et al., 2020; Yang et al., 2021). Human-written captions for news images generally provide much more information than what is shown in the image, drawing heavily from the article text and overall world knowledge. A standard captioning model trained on such data is prone to hallucination, and this is especially undesirable in the news domain, where accuracy is critical. But unlike standard captioning methods, our approach is specifically designed to ensure generation of correct image-external knowledge. In this chapter we train our knowledge-aware model on the news image-caption data and examine its performance in comparison with a specialized model for news image captioning.

Section 5.2 provides a high-level overview of the existing research on news image captioning, with a particular focus on Tran et al. (2020), which presents a large news image-caption dataset and a successful contextualized captioning method. Section 5.3 describes the application of our approach to this dataset. Section 5.4 demonstrates the results of the trained model evaluation, including a focused factual accuracy analysis. Section 5.5 discusses the implications of this case study and potential directions for future work.

5.2 News image captioning

Captions of images that illustrate news articles practically always contain information beyond what can be directly seen in the image. Figure 5.1 shows an example of a photograph from a New York Times article, with its original caption.

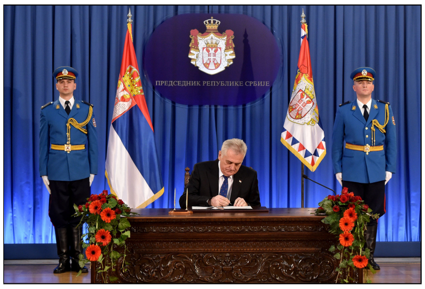


Figure 5.1: "President Tomislav Nikolic of Serbia signed the measure to dissolve Parliament in Belgrade on Friday."
An image and caption from the NYTimes800k dataset (Tran et al., 2020).

This caption includes multiple details that standard captioning models would not be able to produce based on the image alone: the name and title of the person, the precise nature of the action, its time and place. Generating such a caption requires utilizing extensive image-external knowledge. Naturally, much of the relevant knowledge for describing a given news image is provided in the corresponding article. Thus, existing approaches to news image captioning always use it as an external knowledge source. Some of them start by generating a caption template with placeholder slots for named entities, e.g., "PERSON from ORG in GPE". The placeholders are later replaced with the names extracted from the article (Biten et al., 2019; Jing et al., 2020). This straightforward fill-in-the-slot method can, however, be problematic if none of the available entities fits the already generated slot. Another common method of integrating knowledge from the article into the captioning process is to modify the model decoder to let it choose at each step between generating a word from the vocabulary or copying a word from the article text (Whitehead et al., 2018; Chen and Zhuge, 2020).

Our approach can similarly use the article as the source of knowledge about real world entities relevant for the image description. But, differently from most news image captioning methods, it can also utilize other external resources, retrieving information that is not necessarily present in the article. For example, the original caption for Figure 5.1 refers to Belgrade as the site of the event. However, only Serbia and not Belgrade is actually mentioned in the article that this image illustrates[1]. The fact that Belgrade is the capital of Serbia is available in external databases, such as DBpedia, which makes it possible for a model based on our approach to produce it in the caption.

Next we describe the particular news image-caption dataset that is used in this chapter for training and testing our model, and the dedicated captioning method originally developed by the dataset creators.

5.2.1 The Transform and Tell model and NYTimes800k

We make use of one of the largest datasets for news image captioning to date — NYTimes800k (Tran et al., 2020). This dataset consists of 793K images with captions extracted from New York Times articles spanning 14 years. The captions generally contain abundant references to real world entities and events related to the article content (there is at least one named entity in 96% of the captions). In addition, the dataset includes various metadata, such as the publication date and the position of the image within the article text.

The original paper presents the Transform and Tell model for entity-aware image captioning, trained on the NYTimes800k dataset. The overview of its architecture is shown in Figure 5.2. The model consists of a Transformer decoder and four encoders that produce high-level multimodal context representations to inform caption generation.

The first encoder, standard for captioning, creates an overall image representation using the pre-trained ResNet-152 model (He et al., 2016). The second encoder, which has also been successfully used for caption generation (Yang et al., 2017; Yao et al., 2018; Herdade et al., 2019), outputs the ResNet-152 encoding of the objects detected in

[1] `https://www.nytimes.com/2016/03/05/world/europe/serbia-dissolves-parliament-and-calls-early-elections.html`

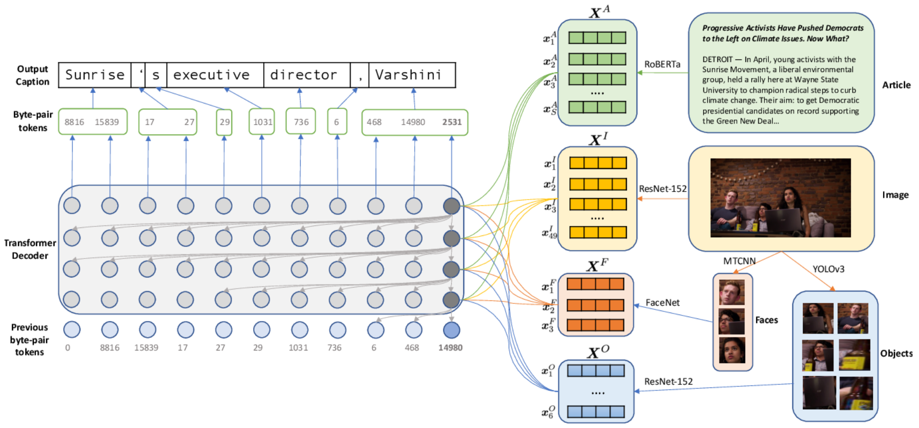

Output Caption
Sunrise
‘
s
executive
director
,
Varshini
Byte-pair tokens
8816 15839
17 27
29
1031
736
6
468 14980 2531
Transformer Decoder
Previous byte-pair tokens
0
8816
15839
17
27
29
1031
736
6
468
14980
X^A
x_1^A
x_2^A
x_3^A
....
x_S^A
RoBERTa
Progressive Activists Have Pushed Democrats to the Left on Climate Issues. Now What?
DETROIT — In April, young activists with the Sunrise Movement, a liberal environmental group, held a rally here at Wayne State University to champion radical steps to curb climate change. Their aim: to get Democratic presidential candidates on record supporting the Green New Deal...
Article
X^I
x_1^I
x_2^I
x_3^I
....
x_{49}^I
ResNet-152
Image
MTCNN
YOLOv3
X^F
x_1^F
x_2^F
x_3^F
FaceNet
Faces
Objects
X^O
x_1^O
....
x_6^O
ResNet-152

Figure 5.2: From Tran et al. (2020). Overview of the Transform and Tell model.

the image with YOLOv3 (Redmon and Farhadi, 2018). The third encoder, specifically useful for news images, detects faces in the image with MTCNN (Zhang et al., 2016) and encodes them with the pre-trained FaceNet model (Schroff et al., 2015). This encoder aims to improve the quality of people's names generation, which is particularly important given that, as reported in Tran et al. (2020), 71% of the training images in the NYTimes800k dataset contain at least one face and 68% of the training captions mention at least one person's name. The fourth encoder creates a representation of the news article text: a weighted sum of the outputs of the pre-trained RoBERTa layers (Liu et al., 2019b). Importantly, this encoder processes only the specific part of the article that surrounds the image, which is assumed to contain the most relevant information.

The decoder consists of four Transformer blocks that use dynamic convolutions to condition on the previously generated tokens (Wu et al., 2019), and multi-head attention over the four encoder outputs. The decoder uses byte-pair encoding (Sennrich et al., 2016), which creates an unlimited vocabulary and thus enables the model to generate words unseen in the training data (crucial for producing rare entity names). The output byte-pair tokens are combined to form whole words and punctuations.

Tran et al. (2020) also present several variations of this model as ablation studies (for example, with GloVe embeddings instead of RoBERTa, without the face or object recognition, etc.). All of the alternatives do not reach the performance level of the full Transform and Tell model.

In this chapter, we apply our approach to the NYTimes800k dataset with the goal of testing its viability and studying its performance on the data that is considerably different from what was used in the previous chapters (specific differences concerning the amount and types of image-external knowledge in captions will be discussed throughout the chapter). The first question that this case study aims to answer is whether our model is able to utilize both the article and an external database to produce informative captions with accurate knowledge. The second question is what benefits our approach provides for generating captions in this domain, compared to a sophisticated news image captioning method, such as Transform and Tell.

5.3 Our approach applied to news data

5.3.1 Contextualization anchor

The first step in applying our approach to the new type of data is to establish the contextualization anchor, the purpose of which is to connect a given image to relevant information in external resources. In the case of the NYTimes800k dataset (as well as any news image captioning dataset), a natural anchor is the article that the image illustrates. The entities mentioned in the article are assumed to be relevant to the image; therefore, using them during caption generation is expected to make the resulting captions more contextualized and informative. Accordingly, we use the news article as the contextualization anchor in this chapter.

5.3.2 Entity context

5.3.2.1 Collecting the entities

For each image we define the entity context as a set of named entities[2] in the corresponding news article, extracted with the SpaCy named entity recognition module (Honnibal and Montani, 2017). Since there are no restrictions on the named entity type, this creates a more general context than in the previous chapters, where the entities were exclusively geographic. On average, a caption in the NYTimes800k dataset contains 3.25 named entities and an article (including the headline) contains 65.15 named entities. Figure 5.3 shows the distribution of the types of named entities in captions and articles in this dataset.

[2] Unlike in the previous chapters, here the distinction between an 'entity' and a 'named entity' is minimal. In general, named entities are a subset of entities; when in Chapters 3 and 4 the entities were collected from OpenStreetMap, they could be in principle denoted by a common phrase (e.g., "convenience store") and therefore not be considered 'named'. Here, all the entities we retrieve from the article text are named entities.

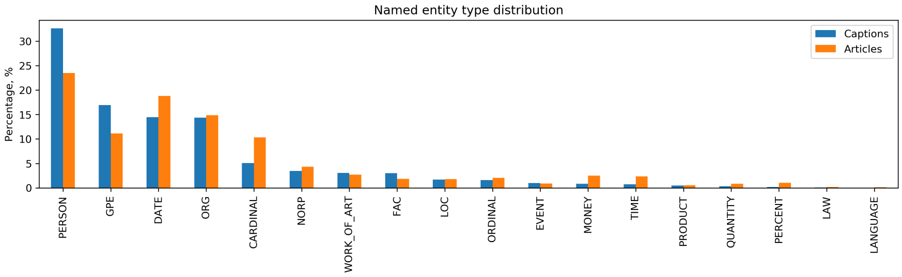


Figure 5.3: Distribution of the types of named entities extracted from the captions and articles of the NYTimes800k dataset.

Only 73.84% of the named entities detected in captions also appear in the corresponding articles. In some cases this is due to SpaCy's imperfect named entity recognition algorithm, which can, for example, misidentify a regular token or phrase as a named entity, especially if it is capitalized (e.g., "last week", "beauty salon", "parliament", "md."). In other cases an entity in a caption is related to the entities in the article while not being directly named in it. For example, a caption may mention Rio de Janeiro while the corresponding article talks about Brazil without naming any specific cities, or a caption may mention ISIS while in the article the group is referred to only as Islamic State.

5.3.2.2 Ranking the entities

The items in the entity context are arranged according to the probability of them being mentioned in a caption. The probability is estimated with a ranking model, similar to the ones described in Sections 3.4.2 and 4.4.1.1, — a logistic regression classifier that we train on a randomly selected subset of 10,000 captions from the training set of NYTimes800k. The classifier predicts whether a given entity will be mentioned in a caption, given the count of the entity in the article, its presence in the headline and the first paragraph, and the entity's type. We set the fixed length of the entity context to 100[3]; so, the top 100 entities of the ranked list constitute the entity context for a given image.

[3] For at least 95% of the images in the ranking model's test set, the top 100 ranked entities include at least one that appears in the corresponding caption.

5.3.2.3 Encoding the entities

In the previous chapters the entity context included only geographic entities, and their geographic characteristics (distance from the image location, size, etc.) were used to represent them. Here the entity context consists of named entities of various types extracted from the article text, which demands a different encoding method. We make use of the following features:

(i) $count$: how many times the entity is found in the article

(ii) $headline$: a binary indicator of the entity's presence in the headline

(iii) $first\_par$: a binary indicator of the entity's presence in the first paragraph of the article

Features (i)-(iii) are intended to reflect the entity's prominence in the article. Thus, they have a similar function to the *distance* and *size* features in the geographic embedding in Section 3.4.3, which reflected the entity's salience in the geographic context of the image.

(iv) $\exists f$: whether there are any facts about the entity in the knowledge context[4]

(v) $\# f$: the number of facts about the entity in the knowledge context

Features (iv)-(v) are also correlated with how salient the entity is: the more facts about the entity there are in an external knowledge source, the more significant and noteworthy the entity is likely to be. The same features were used in the encoding of geo-entities in Section 4.4.2.

(vi) $type$: the type of the entity (e.g., PERSON, ORG)

The type of the entity is an important indicator of its standard usage contexts. For example, in a phrase "X arrived in Y on Z", the usual types of named entities X, Y and Z are PERSON, GPE and DATE, respectively.

In addition, the entity's encoding includes an average of the GloVe embeddings of the words in its name (or randomly initialized embeddings if the words are out-of-vocabulary). Not only does it serve as an additional source of information about the

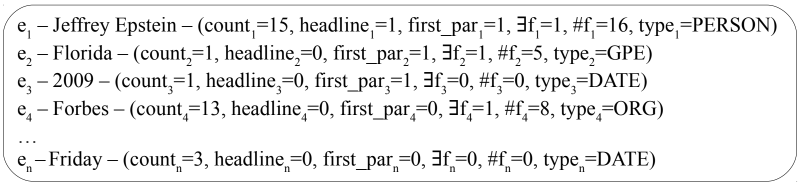
[4] To compute this and the following feature, we use the knowledge context (Section 5.3.3) after it has been constructed but before the facts are encoded. The construction of the knowledge context does not require the full embedding of the entities.

entity's distribution patterns, but it also has the added benefit of making the entity's representation unique, which is necessary for its use in caption generation.

Thus, the embedding for an entity in the entity context is computed as follows ($\circ$ denoting a Hadamard product):

$$\begin{aligned} \textsc{EntEmb}(e_i) &= \textsc{features}(e_i) \circ \textsc{name}(e_i), \text{ where} \\ \textsc{features}(e_i) &= \text{Concat}[count_i,\ headline_i,\ first\_par_i, \\ &\qquad\qquad \exists f_i,\ \#f_i, Emb_e(type_i)] \\ \textsc{name}(e_i) &= \text{Avg}[\textsc{GloVe}(w_j),\ w_j \in e_i] \end{aligned} \tag{5.1}$$

Figure 5.4 shows a snippet of an entity context for a sample image.



e$_1$ – Jeffrey Epstein – (count$_1$=15, headline$_1$=1, first_par$_1$=1, ∃f$_1$=1, #f$_1$=16, type$_1$=PERSON)
e$_2$ – Florida – (count$_2$=1, headline$_2$=0, first_par$_2$=1, ∃f$_2$=1, #f$_2$=5, type$_2$=GPE)
e$_3$ – 2009 – (count$_3$=1, headline$_3$=0, first_par$_3$=1, ∃f$_3$=0, #f$_3$=0, type$_3$=DATE)
e$_4$ – Forbes – (count$_4$=13, headline$_4$=0, first_par$_4$=0, ∃f$_4$=1, #f$_4$=8, type$_4$=ORG)
…
e$_n$ – Friday – (count$_n$=3, headline$_n$=0, first_par$_n$=0, ∃f$_n$=0, #f$_n$=0, type$_n$=DATE)

Figure 5.4: Sample fragment of an entity context for a news image.

5.3.3 Knowledge context

5.3.3.1 Building the knowledge context

The procedure of knowledge context construction is largely the same as the one described in Section 4.4.1. For every entity in the entity context we extract facts from the DBpedia knowledge base where this entity is the subject. As a result of collecting a broader variety of (named) entity types than in the previous chapters, the facts are also more diverse. Some of their predicates are not compatible with geographic entities and therefore were not featured in the knowledge context in the last chapter (e.g., *occupation*, *spouse*, *starring*). The facts are ranked by another logistic regression model that takes into account the subject's position in the entity context and the fact's predicate, and estimates the likelihood of the fact appearing in a caption. This model was trained on the same data subset as the ranking model for the entity context (Section 5.3.2.2). The size of the knowledge context is set to 300 facts. Figure 5.5 shows a fragment of a knowledge context for the same image as the entity context in Figure 5.4.

f_1 – financier	<Jeffrey Epstein, occupation>
f_2 – magazine	<Forbes, hypernym>
…	
f_m – United States	<Florida, country>

Figure 5.5: Sample fragment of a knowledge context for a news image.

5.3.3.2 Encoding the facts

The principle behind encoding the facts in the knowledge context is the same as in Section 4.4.4 (Equation 4.3): each fact is encoded as a linear combination of the embeddings of the fact's subject and predicate. The entity embedding (Equation 5.1) is used to encode the subject, and the predicates' embeddings are initialized randomly and updated during the training of the captioning model. Fact objects in this setup function similarly to the entity names in the entity context: they do not participate in the encoding but rather act as fact "labels" that represent them in the captions.

$$\textsc{FactEmb}^{news}(f_i) = \textsc{EntEmb}(e_i) + Emb_p(p_i) \tag{5.2}$$

5.3.3.3 Knowledge context facts in NYTimes800k

Even though the NYTimes800k captions include many references to image-external knowledge in general, only 11.76% of them contain at least one fact from the knowledge context. A possible reason is that, using the terminology established in Section 4.3.1, DBpedia generally contains *encyclopedic* knowledge about a given entity (e.g., <Joe Biden, party, Democratic Party (United States)>), while the knowledge in news articles and the corresponding images and captions is often *situational*, i.e., related to current events and therefore less likely to be documented in a knowledge base (e.g., "Joe Biden on Saturday at a campaign stop in South Carolina"). Being exposed to enough knowledge context facts in the training captions is necessary for the captioning model to learn to produce them. Thus, the relatively low occurrence of such facts in the NYTimes800k captions is likely to result in a similarly low number of facts being generated from the knowledge context by our model.

Captions in the NYTimes800k dataset contain altogether facts with 1,901 unique predicates[5]. Figure 5.6 presents 50 most frequent ones by the number of facts in captions. Most of the facts in captions are attributed to entities of the types GPE (countries, cities, states), PERSON and ORG, as shown in Figure 5.7.

[5] As compared to 336 unique predicates in the captions of K-GeoRic, see Section 4.4.5

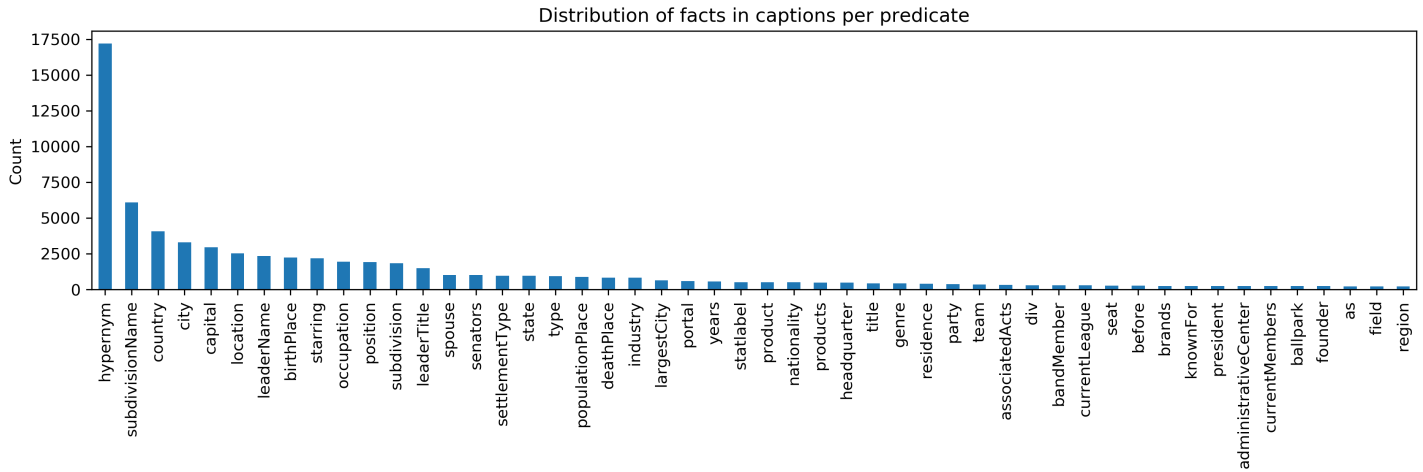


Figure 5.6: Distribution of facts detected in the captions of the NYTimes800k dataset per predicate.

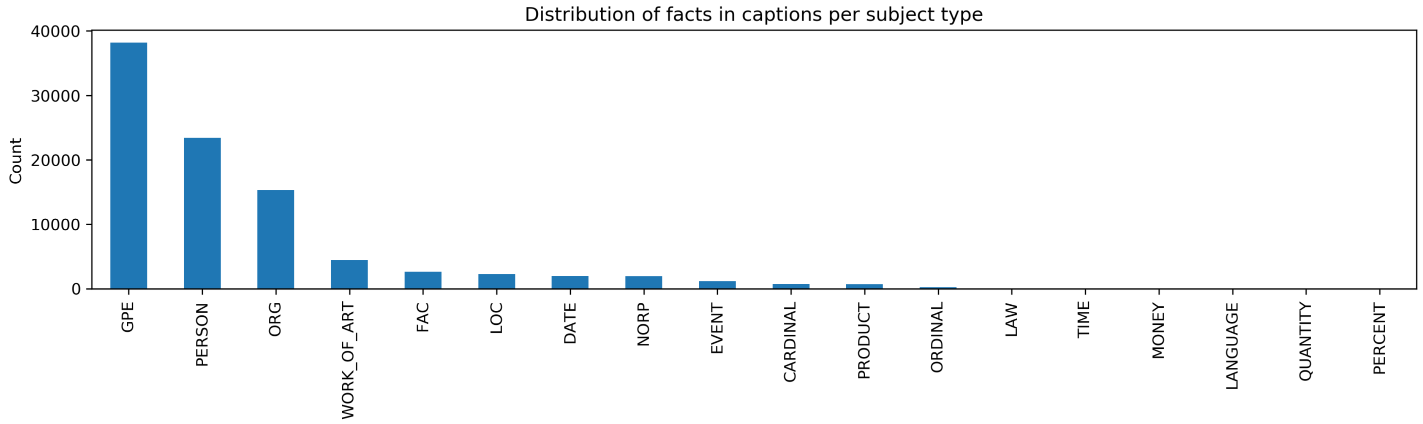


Figure 5.7: Distribution of facts detected in the captions of the NYTimes800k dataset per subject type.

5.3.4 Model architecture

The architecture of the captioning model is the same as the one developed in the previous chapter; the only difference is the updated entity and fact embeddings: the entity embedding $\textsc{EntEmb}$ (Equation 5.1) replaces the geographic embedding $\textsc{GeoEmb}^k$ (Equation 4.1), and the fact embedding $\textsc{FactEmb}^{news}$ (Equation 5.2) replaces $\textsc{FactEmb}$ (Equation 4.3). First, captions are encoded according to the procedure described in Section 4.5, with a distinction made between three types of tokens: regular vocabulary words, entity names and fact objects. The encoder, same as in Section 4.6.1, produces a combined representation of the context for caption generation. It concatenates the encoding of the image produced by the pre-trained ResNet-101 CNN, and the encodings of the entity and knowledge contexts produced by two Transformer networks. Finally, the Transformer decoder generates caption tokens one by one, at each step taking into account the output of the encoder and the previously generated text. As described in Section 4.6.2, the decoder can produce not only words from the regular vocabulary but also names from the entity context and fact objects from the knowledge context. The knowledge context is additionally used to influence the generation of vocabulary words, with the goal of decreasing the risk of hallucination. Overall, the captioning pipeline described in Chapter 4 is applied to the NYTimes800k dataset without any additional modifications.

5.4 Evaluation

5.4.1 Train-validation-test data split

We use the original train-validation-test data split proposed by Tran et al. (2020), which is based on the time of the article publication. With this splitting strategy, the dataset creators aim to avoid overfitting and to test the performance of the model on previously unseen entities and events. These goals are similar to those of the latitude-based splits of GeoRic (Section 3.3.2.1) and K-GeoRic (Section 4.3.2.1).

	Training	Validation	Test
Number of articles	433,561	2,978	8,375
Number of images	763,217	7,777	21,977
Start month	Mar 15	May 19	Jun 19
End month	Apr 19	May 19	Aug 19

Table 5.1: From Tran et al. (2020). Training, validation and test splits in NYTimes800k.

5.4.2 Quantitative evaluation results

The trained model is evaluated on the test set of NYTimes800k using the same metrics as in Tran et al. (2020): BLEU-4, ROUGE, CIDEr, and the precision and recall of the generated named entities as compared to the ground truth captions for the same images. In the original paper, precision and recall are calculated using full string matching (e.g., "Trump" in the generated caption is *not* a match for "Donald Trump" in the ground truth caption); we additionally provide scores calculated with partial matching.

As in the previous chapters, two ClipCap models, pre-trained on MSCOCO and Conceptual Captions (**ClipCap-MSCOCO** and **ClipCap-Conceptual** respectively), are used as baselines to demonstrate the performance level of a standard decontextualized caption generator (these models are not expected to be competitive against the contextualized ones; we include them in the analysis for the sake of completeness).

A more direct comparison with our approach is provided by the two variations of the Transform and Tell model from Tran et al. (2020), which we denote **T&T-GloVe** and **T&T-Full**. T&T-GloVe uses a Transformer decoder for caption generation based on the image representation by ResNet-152 and the article text encoded with GloVe embeddings. Thus, it is the closest to our model, although the underlying methods are still different, e.g., in our approach the article text itself is not embedded but is used only to construct the entity context. T&T-Full is the best model of the original paper, in which generation is based not only on the image as a whole but also on the detected faces and objects in the image, and on the weighted RoBERTa embeddings of the specific part of the article that surrounds the image.

		BLEU-4	ROUGE	CIDEr	NE-Exact		NE-Partial	
					P	R	P	R
ClipCap	MSCOCO	0.08	6.73	1.32	—	—	—	—
	Conceptual	0.18	8.12	2.34	—	—	—	—
T&T	GloVe	2.75	15.9	20.3	13.2	10.8	27.44	22.41
	Full	**6.30**	**21.7**	**54.4**	**24.6**	**22.2**	**39.05**	**35.1**
Ours		1.45	17.24	6.46	11.57	9.85	20.4	16.79

Table 5.2: Metric scores of the models, measured on the NYTimes800k test set.

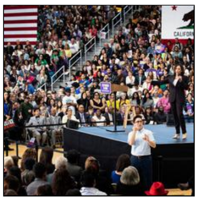
As shown in Table 5.2, our model achieves higher metric scores than the decontextualized ClipCap models but is outperformed by both Transform and Tell ones. The scores of the T&T-Full model are much higher than any of the alternatives. This demonstrates the importance of the task-specific T&T components, which are missing from the current implementation of our approach (e.g., a face recognition module, conditioning only on the most important part of the article, etc.). Importantly, our model is competitive against T&T-GloVe, even though it is not trained to condition the generation on the article text, a big disadvantage as compared to T&T.

Our model is nevertheless able to produce informative captions utilizing the entity and knowledge contexts. 98.6% of the captions generated by our model contain at least one entity from the entity context and 22.88% contain at least one fact from the knowledge context (the latter number is relatively low because knowledge context facts are quite sparse in the training captions, as mentioned in Section 5.3.3.3). Each caption includes on average 2.63 named entities, although only 1.21 of them are generated from either entity or knowledge context. This means that there is a substantial number of named entities generated from the regular vocabulary, which creates a higher risk of hallucination (as supported by the error analysis, see Section 5.4.3.3).

Next, we present a manual qualitative evaluation of the generated captions. As in the previous chapters, a particular emphasis is placed on the factual accuracy of image-external knowledge references. The analysis is conducted on a limited-size sample from the test set, which allows for a focused exploration of the model performance and a better understanding of its advantages and limitations than what is provided by the automatic metrics above.

5.4.3 Qualitative evaluation results

Table 5.3 presents captions produced by all the models for the images from the test set. We intentionally chose images that depict a variety of scenes: cultural, political, sports-related, etc. Since our primary goal is to examine the performance of our model, we also selected a sample showcasing the types of errors it makes (excluding the ones related to fluency, such as ungrammatical phrases or generating the same text in a loop, which are not particularly interesting), and its ability to produce factual information from the entity and knowledge contexts.

(a)		Ground truth: Angelina Jolie in the new trailer for the sequel. ClipCap-MSCOCO: A woman with green hair and a green tie. ClipCap-Conceptual: film character from the movie. T&T-GloVe: **Angelina Jolie** in the trailer for *"The Crown."* T&T-Full: **Angelina Jolie** in a scene from the trailer for **"Maleficent: Mist of Evil."** Ours: **angelina jolie** in the **maleficent**, directed by **joachim ronning**.
(b)		Ground truth: Senator Kamala Harris during a rally in Los Angeles in May. ClipCap-MSCOCO: A crowd of people standing around each other. ClipCap-Conceptual: politician speaks to a crowd of supporters. T&T-GloVe: **Senator** *Elizabeth Warren* at a **rally** in *Oakland*, **Calif.**, on *Saturday*. T&T-Full: **Senator Kamala Harris** of **California**, *center*, at a **rally** in **Los Angeles** *last month*. She is among the **Democratic candidates** who will be the first **Democratic presidential candidate** in **2020**. Ours: the **democratic** *national convention* in **los angeles** in **southern california**.
(c)		Ground truth: Young Simba in the new "Lion King." ClipCap-MSCOCO: A small kitten is standing on a big toy. ClipCap-Conceptual: the lion cub is seen in a still from the film. T&T-GloVe: A scene from **"The Lion King,"** a film directed by *Adam Pangor*. T&T-Full: **Simba**, the new **"Lion King"** star, in the new film. Ours: **simba** in **lion king**.
(d)		Ground truth: Prince Harry and Meghan, Duchess of Sussex, with their son. ClipCap-MSCOCO: A man and a woman in formal wear holding a baby. ClipCap-Conceptual: the couple welcomed their first child, a boy, in july. T&T-GloVe: **Prince Harry** and **Meghan Markle** at the *wedding of their son Harry and Meghan Markle* in **London** in *2015* T&T-Full: **Prince Harry** and **Meghan, Duchess of Sussex**, in **London** in *April* Ours: **prince harry**, *left*, and **meghan** in **london**, **united kingdom**, on wednesday.
(e)		Ground truth: At least 5,000 African migrants are being housed in detention centers across northwestern Libya. ClipCap-MSCOCO: A group of people sitting around a truck. ClipCap-Conceptual: people gather around a tent to buy food. T&T-GloVe: **Migrants** *from Libya* waited for food in the town of *Surt*, **Libya**, on Wednesday. T&T-Full: *A man was carried away from the site of bombing* in Tripoli on Wednesday. Ours: the **migrants at detention center** in tajoura, **libya**, on wednesday.

(f)		Ground truth: Maria Sharapova, currently ranked 87th in the world, has not defeated Williams since 2004. ClipCap-MSCOCO: A woman in a tennis dress holding a tennis racket. ClipCap-Conceptual: tennis player celebrates winning her women's singles final match against tennis player during day. T&T-GloVe: *Serena Williams* after her *victory over Maria Sharapova* in the *final* of the **U.S. Open** on **Monday**. T&T-Full: **Maria Sharapova** after her *victory over Maria Sharapova*. Ours: *serena williams*, who has *won* the **open** *since* *2004*.
(g)		Ground truth: Visitors now throng the Forbidden City in Beijing. ClipCap-MSCOCO: A crowd of people standing in front of a building. ClipCap-Conceptual: tourists visit chinese structure during the festival. T&T-GloVe: The *National Palace Museum* in **Beijing** T&T-Full: The **Forbidden City** in **Beijing**. The complex is *the first to open in the last century*. Ours: the **forbidden city**, in **dongcheng district**, *n.y.*
(h)		Ground truth: Chico MacMurtrie's "Organic Arches (Time Traveller)," 2014/2017, at the Queens Museum. ClipCap-MSCOCO: A large mesh mesh mesh mesh object in a room. ClipCap-Conceptual: the installation is a work of art. T&T-GloVe: *"The Great White,"* by *Alex Rivera*. T&T-Full: *Alex Rivera's 2008 feature film, "Sleep Deal," is a migrant-worker "Blade Runner" in which people in Mexico operate robots in the United States by remote control*. Ours: the **queens museum** in **queens**, where *mundos alternos* was *killed* in *2008*.

Table 5.3: Examples of the captions generated for the NYTimes800k images from the test set. Correct entities and facts are given in **bold**, incorrect ones are given in *italics*; ones that are impossible to fact-check are not marked (for example, for image (e) it is impossible to infer from the article or any other sources on which day and in which specific city in northwestern Libya the photograph was taken). Tokens generated from the entity context are underlined with a straight line, the ones from the knowledge context — with a wavy line.

5.4.3.1 Performance of the decontextualized models

Captions generated by ClipCap-MSCOCO provide generic descriptions of what can be seen in the images, mostly accurate with occasional hallucinations (e.g., “green tie” in caption (a), “big toy” in caption (c)). The ClipCap-Conceptual captions are more specific, for example, people in image (b) are correctly described as a “crowd of supporters” of a politician instead of just a crowd of people, and images (a) and (c) are recognized as movie scenes. This reflects that the model was trained on a dataset of diverse, naturally created captions, which are often similar to the captions in NYTimes800k, but with named entities removed. However, captions generated by ClipCap-Conceptual tend to include false information, such as the month of July in caption (d), the intention to buy food in caption (e), winning the match in caption (f). These hallucinated details result from the lack of image context incorporation in the ClipCap model: “July” was produced in (d) not because it was associated with that specific photograph or the corresponding article, but because it occurred in similar captions in the training set of Conceptual Captions.

5.4.3.2 Image-external information in captions

Two Transform and Tell implementations and our model were trained on the in-domain NYTimes800k data and, by design, incorporate image-external knowledge. As a result, their captions include many references to concrete entities and events, which cannot be inferred from the image alone. Unlike the T&T models, ours is also able to utilize information that is not present in the related news article. For example, caption (g) specifies that the Forbidden City is located in the Dongcheng District, which is correct but not mentioned in the article the image is associated with[6]. This shows one of the potential benefits of using our approach for news image captioning, given that human-produced captions also sometimes include named entities that are related to the article text but not directly mentioned in it (see the “Belgrade”—“Serbia” example in Figure 5.1).

Table 5.3 also shows that, although precision and recall of named entities are useful for the general understanding of the quality of entity selection, they cannot provide a precise indication of how factually accurate the generated captions are. For example, in the captions for image (a), T&T-Full and our model mention the correct name of the film and T&T-GloVe mentions an incorrect one, but since the ground truth caption does

[6] https://www.nytimes.com/2019/08/03/world/asia/beijing-forbidden-city-museum.html

not name the film at all, all three models would equally receive a penalty for a false positive. In addition, the correct name of the director produced by our model would also count as a false positive. On the other hand, in (b), the mistake of calling a rally a “national convention” is not picked up by these metrics, since neither “rally” nor “national convention” are named entities. This shows that the news article itself and not only the ground truth caption should be involved in evaluating the generated captions’ accuracy. Moreover, in order to assess the correctness of the caption generated by our model for image (g), it is necessary to refer to an external database, since the location of the Forbidden City in the Dongcheng District is not reflected in the associated article text. Therefore, a comprehensive metric should take into account the knowledge in the ground truth caption, the article and other external data sources. Given the wide diversity of topics and fact types that would need to be considered, developing such a metric would be extremely challenging, and we leave it for future work.

5.4.3.3 Error analysis

Factual inaccuracies in the captions generated by our model are often a result of producing a named entity from the regular vocabulary instead of drawing it from the entity or knowledge context (for example, “N.Y.” in caption (g)). In order to reduce the risk of such errors, all the named entities in the training captions should be found in either of the contexts, which would help the model learn that particular constructions, such as “in GPE, [...]” in caption (g), call for a token from the available external data and not the vocabulary. Otherwise, if the training set contains a frequent phrase with a named entity that is being treated as a regular vocabulary word, the model is likely to learn to generate this phrase whether or not it is fitting to the context of the image. An extreme example of this is the performance of the no-context baselines in the previous chapters, which produced all the names and facts from the vocabulary, resulting in severe hallucinations. Facts that our model did generate from the knowledge context are mostly related to locations (with predicates such as *capital*, *country*, *subdivisionName*, *location*, etc.). These types of facts are also among the most frequently occurring in the NYTimes800k dataset, as shown in Figure 5.6. This suggests that, given training data that includes captions with more diverse encyclopedic facts, the model could be expected to generate more facts of different types.

Another kind of errors produced by our model is false statements related to situational knowledge, such as, for example, the Queens Museum being described as the place where “mundos alternos was killed” in caption (h), whereas actually the Queens

Museum hosted the “Mundos Alternos” exhibition. Unlike encyclopedic facts that can be retrieved from the knowledge context, situational knowledge cannot be generated correctly unless it is extracted from the news article itself. We believe that this issue could be mitigated by expanding the knowledge context to include situation-specific facts retrieved from the article text.

Finally, the generated captions are sometimes fitting to the article but not to the image. For example, image (f) depicts Maria Sharapova who, as reported in the corresponding article and the ground truth caption, has not defeated Serena Williams in the U.S. Open since 2004. Our model has erroneously identified the woman as Serena Williams, rendering most of the caption incorrect with respect to the image, even though it also mentions winning and 2004, which would be at least partially correct if the image had indeed featured Williams. The T&T-GloVe model also misidentifies the person, but not T&T-Full. The latter includes a dedicated face recognition module in the encoder, which can reduce the risk of such errors. Furthermore, its generation is conditioned only on the specific part of the article that surrounds the image. This way the external knowledge that is produced in the caption is more likely to be directly connected to the image. Introducing a variation of either technique into our model (for example, using only a part of the article for the entity context construction) could also make the generated captions more pertinent to the images.

5.4.4 Accuracy of generated facts

		Sample analysis	
Predicate	Total number of captions	Accurate facts (%)	Facts relevant to the image (%)
hypernym	3,111	100	70
subdivisionName	974	100	80
country	697	90	60
capital	166	100	30
location	110	100	50
director	40	100	50
locationCity	17	100	50
birthPlace	17	100	0
seat	16	100	0
state	13	90	20
Total	5,161	98	41

Table 5.4: Accuracy of facts generated from the knowledge context.

In addition, we conduct a separate evaluation of the accuracy of facts generated from the knowledge context. The goal is to measure the consistency with which our model produces factually accurate information using the available image-external data. We randomly sample 10 captions for each of 10 predicates that are generated most frequently and manually verify the correctness of the produced facts. The results of this analysis are shown in Table 5.4, along with the total number of captions generated with each of the predicates (the relative frequency of the predicates here is very similar to the distribution of the same predicates in the ground truth captions, as seen in Figure 5.6).

Nearly all of the generated facts in the sample are true, which indicates that our method of incorporating external knowledge via the knowledge context is indeed effective for producing highly accurate factual information. However, the generated facts are in many cases not directly related to the image (in 59% of cases overall, and particularly often with the less frequent predicates). Table 5.5 demonstrates examples of generated facts with different predicates. For each predicate, one example contains a fact irrelevant for the image description, and the other contains a relevant one, for comparison.

Predicate: *director*

(a)

IRRELEVANT

Ground truth: Baykali Ganambarr, left, and Aisling Franciosi in “The Nightingale.”

Ours: the midsommar, directed by ari aster.

(b)

RELEVANT

Ground truth: Hatidze Muratova, tending to her bees, in “Honeyland.”

Ours: honeyland, directed by tamara kotevska.

Predicate: *locationCity*

(c)

IRRELEVANT

Ground truth: Six police officers said they were asked to leave a Starbucks in Tempe, Ariz., after a customer said their presence made him feel uncomfortable.

Ours: the starbucks in the company’s headquarters in seattle.

(d)

RELEVANT

Ground truth: The Equifax offices in Atlanta. About 147 million people were affected by Equifax’s breach in 2017.

Ours: the equifax in the united states, in atlanta, calif., on wednesday.

Predicate: *subdivisionName*		
(e)		IRRELEVANT Ground truth: Taxidermy fills the walls at Big Dick's Buckhorn Inn in Spooner, Wis. Ours: the logging museum in minocqua, oneida county, on the 10th and <unk>.
(f)		RELEVANT Ground truth: The front-runner to lead the Conservative Party, Boris Johnson, left, and Foreign Secretary Jeremy Hunt during a BBC television leadership debate in London on Tuesday. Ours: boris johnson in the first conservative and chief executive of the conservative party, in london, in united kingdom, on thursday.

Table 5.5: Examples of the generated facts. The tokens generated from the knowledge context (fact objects) are marked with a wavy line, and the entities they describe (fact subjects) are marked with a straight line. All of the facts are true, but some are irrelevant to the images.

Image (a) depicts a scene from the film “The Nightingale”, but our model incorrectly associates it with another film discussed in the article, “Midsommar”. As a result, the accurate fact that “Midsommar” was directed by Ari Aster is not relevant for the description of this particular image. On the other hand, image (b) is correctly identified as the scene from the film “Honeyland”, so the generated fact about its director is both correct and relevant. In the captions for images (c) and (d), the model produces facts with the predicate *locationCity*, which is used to link a company to the city of its headquarters, and for images (e) and (f) with the predicate *subdivisionName*, which denotes a relation between a settlement and its encompassing geographic or administrative region. The fact that the headquarters of Starbucks is located in Seattle is true, but image (c) shows a concrete branch in Tempe, Arizona. Similarly, in caption (e) it is true that the town of Minocqua is located in Oneida County, but the photograph was taken in Spooner, Washburn County. The facts with the same predicates in captions (d), Equifax in Atlanta, and (f), London in the United Kingdom, are both correct and directly related to the images.

Examples (a), (c) and (e) once again demonstrate the main sources of errors made by our model. First, underutilization of situational knowledge provided by the article text: here it is exemplified by caption (c) and it was previously discussed in relation to the Mundos Alternos example in Table 5.3, image (h). Second, the limited ability to identify the specific entities relevant to the image: captions (a) and (e) here and image (f) in Table 5.3 (the Maria Sharapova/Serena Williams example). We believe that this last limitation is due to the fact that the contextualization anchor adopted in this chapter, the news article, is too broad and loosely related to each given image, and therefore, the entity context is likely to include too many irrelevant entities. Since our model consistently produces accurate factual information, the main challenge is to ensure that it is pertinent to the image, which requires a precise anchor, as image-specific as the location coordinates in the previous chapters.

5.5 Discussion

This chapter has examined the application of our knowledge-aware captioning approach to the challenging domain of news images. The architecture of the captioning model presented here is largely the same as that of the model developed in the previous chapter. The data collection and encoding procedures are adapted to the new conditions: the contextualization anchor is different (the news article instead of the image location), and the entity context is generalized to include named entities of diverse types, as is suitable for a broader domain like news. The resulting model is shown to produce contextualized captions with external knowledge drawn from the article and a general database.

Unlike most methods developed specifically for news image captioning, our approach provides a possibility of generating encyclopedic information that is not given in the article text. This makes the captions more informative and reflects the relevant world knowledge that a human might also have when they describe an image. In addition, a manual evaluation has confirmed that our approach ensures generation of true facts, which is essential for a captioning model in the news domain.

Our analysis also identified the main shortcomings in the current implementation of our general approach. The whole article as a contextualization anchor creates an entity context with too many entities unrelated to a particular image, and the knowledge context does not include important situation-specific facts reported in the article. We believe that these issues can be improved by adjusting how the entity and knowledge contexts are constructed. For example, inspired by the Transform and Tell technique, the entity context can be limited to include only named entities from the part of the article adjacent to the image, which would likely reduce the number of irrelevant ones. The knowledge context can be, conversely, expanded by adding facts retrieved from the article text with a relation extraction model.

Thus, the case study in this chapter has shown that in applying our approach to different domains, its components need to be carefully designed to account for the specifics of the data. It is crucial to identify an appropriate contextualization anchor, which would provide the most relevant information for the entity context. Similarly, the knowledge context needs to be constructed from the most suitable domain-specific knowledge sources. If these components of the overall approach are chosen well, it can be easily and successfully applied to caption generation in a new domain.

CHAPTER 6

Conclusions and future work

This dissertation explores image captioning enriched with contextually relevant external knowledge. To this end, we presented an approach to incorporating image-external information into the automatic caption generation process. Its basic components — contextualization anchor, entity and knowledge contexts — can be applied to different domains, with the specifics of their realization depending on the available and most suitable data.

Chapter 3 presents the approach with a focus on geographic knowledge. The captioning model developed in this chapter uses image location metadata as a contextualization anchor to identify specific geographic entities in and around the image. These entities constitute the geo-entity context, in which they are encoded through their geographic features and incorporated into an otherwise standard encoder-decoder captioning pipeline. The geo-entity context provides extra input for the encoder and an additional vocabulary for the decoder, allowing it to generate entity names in the captions. The model is trained on the GeoRic dataset that contains naturally produced captions with abundant geographic references. The evaluation shows a substantial improvement over the standard baseline models, and specifically the ability of our model to correctly produce spatial expressions like "to the north of [X]" or "near [X]".

Chapter 4 describes a further development of the approach. The image captioning model introduced in this chapter extends the model from Chapter 3, incorporating encyclopedic facts about the geographic entities related to the image. These facts comprise the knowledge context, which is integrated into the captioning process alongside the geo-entity context. The knowledge context is added as another input to the encoder, and in the decoder it provides extra contextualization for the generation of regular words and another vocabulary for generating fact-related tokens. For training this model, we compiled another dataset, K-GeoRic, in which captions contain, besides the geographic references, a broad variety of encyclopedic facts about the relevant geo-entities. On the test set of K-GeoRic, our model confidently outperforms a range of baseline models in standard captioning metrics and, importantly, in the accuracy of the generated facts.

Chapter 5 applies our approach to the qualitatively different data from the news image captioning domain. The model in this chapter replicates the architecture of the model from Chapter 4, with several generalizations that result from using the news article as a contextualization anchor. The entity context is constructed from the named entities in the article text, and the knowledge context includes encyclopedic facts about these entities. Both contexts are integrated into the captioning pipeline in the same way as in Chapter 4. The resulting model is able to generate contextualized captions that incorporate information from both the article and an external knowledge base.

All the models developed in this dissertation implement our approach within a standard captioning pipeline. However, as shown in Chapter 5, image captioning models created for a particular task or domain often complement the standard architecture with specialized techniques that allow them to better adapt to the specifics of the data. We believe that integrating our approach with domain-specific captioning architectures is a promising direction for future work.

On the one hand, our method could be used in an existing domain-specific system to create an additional ‘external knowledge’ module. For example, in the Transform and Tell model (Tran et al., 2020), a new encoder could be added to the existing ones for processing the entity and knowledge contexts, which would allow the decoder to attend to them while generating a caption and to produce relevant information from an external database. On the other hand, a captioning model built on the basis of our approach could utilize domain-specific techniques to improve the realization of its components. In the same example of news image captioning, the entity context could be constructed only from the part of the article directly adjacent to the image, inspired by the way the article is processed in Transform and Tell.

We believe that using our approach for domain-specific caption generation would be beneficial in different areas, especially where humans also tend to use external knowledge for the meaningful interpretation of images. For example, human-written captions of medical images often contain information beyond a purely visual description, with references to the context of the image (e.g., the age, gender, underlying conditions of the patient, etc.) and statements made on the basis of general medical knowledge (Huang et al., 2019; Liu et al., 2021b). Thus, in order to automatically generate an informative and accurate caption for such an image, a model should also utilize relevant image-external data. In this highly knowledge-intensive field, images are usually accompanied by extensive metadata, including patient records, and there are many resources that store structured medical knowledge. This suggests the potential for a profitable application of our approach in this domain.

Another important area for future work is the development of a comprehensive evaluation procedure focusing on the correctness of the generated captions. In this dissertation we proposed custom metrics to assess the accuracy of the generated spatial expressions and various facts about geographic entities. With the restriction to the geographic domain, we were able to conduct a focused automatic accuracy evaluation. However, creating a general metric that could verify the correctness of any given fact produced in a caption is a much more difficult task. In Chapter 5 we argued that such a metric would have to take into account the ground truth caption, the available image-related data (such as the article in news image captioning) and external knowledge (general-purpose and domain-specific databases). More research is needed to achieve this goal. Comprehensive factual accuracy evaluation that does not rely on manual annotation is an important and yet unsolved problem in NLG in general. Promising results have been recently reported for summarization and dialogue generation models (Honovich et al., 2021, 2022; Durmus et al., 2020; Laban et al., 2022; Dziri et al., 2021, 2022; Yang et al., 2022b), and, to the best of our knowledge, this work is among the first to focus on this problem in the area of image captioning.

This dissertation has addressed one of the biggest challenges in modern automatic image captioning: moving from the straightforward "image-to-text translation" to a complex contextualized caption generation process that accounts for relevant image-external knowledge. We believe that our research presents substantial steps in this direction, with an effective captioning method and a contribution to the important issue of factual accuracy evaluation.

APPENDIX A

Implementation details

All of our models are implemented in the PyTorch framework (Paszke et al., 2019), version 1.9.0, with CUDA version 11.0. The versions of all the packages we used are provided in the requirements files that are distributed with the code (see Appendix B.2). Unless specified otherwise, all the implementation details reported in this section are common for all the models we developed (geo-aware in Chapter 3, knowledge-aware in Chapter 4 and news-knowledge-aware in Chapter 5).

Encoder All the images are resized to 256x256 pixels to go through the pre-trained CNN image recognition module, which we do not fine-tune. The output of the image encoder has a size 14x14 with 2048 color channels. The Transformer Encoder networks used for encoding the (geo-)entity and knowledge contexts follow the standard PyTorch implementation, with 3 layers and 10 heads in a layer, with a dropout value of 0.5 and a feedforward network with a dimension size 512.

Decoder The decoder follows the standard PyTorch implementation of a Transformer Decoder. Similarly to the encoder, the decoder has 3 layers and 10 heads in a layer, the dimension of the feedforward network is 512, the dropout value is 0.5. The embeddings of the vocabulary words are initialized from the GloVe embeddings of size 300, pre-trained on the Common Crawl data and fine-tuned during training. The size of the geographic embedding, the entity embedding and the fact embedding is 300.

Training The models are trained with the Adam optimizer with the learning rate of 4e-4. During backpropagation, the decoder gradients are clipped to the absolute value of 5.0. We use cross entropy loss; the early stopping is enabled after 20 consecutive epochs without a loss decrease. The training was carried out with a 4GB NVIDIA Quadro P1000 GPU. The total number of trainable parameters in the network is 14,727,363 for the geo-aware model, 16,121,781 for the knowledge-aware model, 124,082,897 for the news-knowledge-aware model.

The geo-aware model, the results of which are reported in this dissertation, was trained for 108 epochs with the batch size 4, the knowledge-aware model — for 21 epochs with the batch size 4, and the news-knowledge-aware model — for 10 epochs with the batch size 3 (due to the size of the dataset, the computational complexity and our infrastructure, a single epoch took approximately 23 hours, which is why we interrupted the training before early stopping was triggered).

Evaluation At inference time, we use beam size 1, and the maximum caption length of 30 for the geo-aware model and 40 for the knowledge-aware and news-knowledge-aware models (if generation is not stopped by the model when this limit is reached, the process is interrupted automatically). The implementation of the standard captioning metrics (BLEU, ROUGE, METEOR, CIDEr) is adopted from `https://github.com/tylin/coco-caption`.

APPENDIX B

Data and software

B.1 License

This dissertation presents GeoRic and K-GeoRic, two datasets of images, captions and geographic metadata collected from the Geograph project website. All the original data is licensed for reuse under the Creative Commons BY-SA 2.0[1] license. In compliance with the license terms, our datasets are distributed under the more recent version of the same license, Creative Commons BY-SA 4.0[2].

The code of the models developed in this dissertation is made available under the Creative Commons BY-SA 4.0 license.

B.2 Code

The code of our models is available at `https://github.com/sonniki/image-captioning-with-external-knowledge`. The repository contains separate folders for the geo-aware (Chapter 3), knowledge-aware (Chapter 4) and news-knowledge-aware (Chapter 5) models. The corresponding README

[1] `https://creativecommons.org/licenses/by-sa/2.0/`

[2] `https://creativecommons.org/licenses/by-sa/4.0/`

files provide comprehensive descriptions of the contents of the folders and instructions for training and evaluating the models.

B.3 Data

The GeoRic (Chapter 3) and K-GeoRic (Chapter 4) datasets are available at `https://drive.google.com/drive/folders/1vFHvkwV4_3jCg3FAtfghtyl3YReP678V`. The datasets are provided with the train-validation-test splits used in this dissertation.

Bibliography

Ahn, S., Choi, H., Pärnamaa, T., and Bengio, Y. (2016). A neural knowledge language model. arXiv preprint arXiv:1608.00318.

AlKhamissi, B., Li, M., Celikyilmaz, A., Diab, M., and Ghazvininejad, M. (2022). A review on language models as knowledge bases. arXiv preprint arXiv:2204.06031.

Altman, D. (1994). Fuzzy set theoretic approaches for handling imprecision in spatial analysis. International journal of geographical information systems, 8(3):271–289.

Aly, R., Guo, Z., Schlichtkrull, M. S., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., and Mittal, A. (2021). The Fact Extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), pages 1–13, Dominican Republic. Association for Computational Linguistics.

Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). SPICE: Semantic propositional image caption evaluation. In European Conference on Computer Vision, pages 382–398. Springer.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 6077–6086.

Arbinger, C., Bullin, M., and Henrich, A. (2022). Exploiting geodata to improve image recognition with deep learning. WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In The Semantic Web, pages 722–735. Springer.

Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., and Ermon, S. (2021). Geography-aware self-supervised learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10181–10190.

Bai, Z., Nakashima, Y., and Garcia, N. (2021). Explain me the painting: Multi-topic knowledgeable art description generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5422–5432.

Bennett, B. (2010). Spatial vagueness. In Methods for Handling Imperfect Spatial Information, pages 15–47. Springer.

Biten, A. F., Gómez, L., and Karatzas, D. (2022). Let there be a clock on the beach: Reducing object hallucination in image captioning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1381–1390.

Biten, A. F., Gomez, L., Rusiñol, M., and Karatzas, D. (2019). Good News, everyone! Context driven entity-aware captioning for news images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12466–12475.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 1247–1250.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In Advances in neural information processing systems, pages 2787–2795.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In Advances in neural information processing systems, pages 1877–1901.

Cambria, E., Olsher, D., and Rajagopal, D. (2014). SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In Proceedings of the twenty-eighth AAAI conference on artificial intelligence, pages 1515–1521.

Cambria, E., Speer, R., Havasi, C., and Hussain, A. (2010). SenticNet: A publicly available semantic resource for opinion mining. In 2010 AAAI Fall Symposium, pages 14–18. AAAI Press.

Cao, B., Lin, H., Han, X., Sun, L., Yan, L., Liao, M., Xue, T., and Xu, J. (2021). Knowledgeable or educated guess? revisiting language models as knowledge bases. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1860–1874.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E. R., and Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In Proceedings of the 24th AAAI Conference on Artificial Intelligence, pages 1306–1313.

Chandu, K. R., Bisk, Y., and Black, A. W. (2021). Grounding 'Grounding' in NLP. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4283–4305.

Chen, C., Zhang, R., Koh, E., Kim, S., Cohen, S., Yu, T., Rossi, R., and Bunescu, R. (2019a). Figure captioning with reasoning and sequence-level training. arXiv preprint arXiv:1906.02850.

Chen, J. and Zhuge, H. (2020). A news image captioning approach based on multimodal pointer-generator network. Concurrency and Computation: Practice and Experience, page e5721.

Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., Zhou, X., and Wang, W. Y. (2019b). TabFact: A large-scale dataset for table-based fact verification. arXiv preprint arXiv:1909.02164.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734.

Chu, G., Potetz, B., Wang, W., Howard, A., Song, Y., Brucher, F., Leung, T., and Adam, H. (2019). Geo-aware networks for fine-grained recognition. Proceedings of the IEEE International Conference on Computer Vision Workshops.

Cocos, A. and Callison-Burch, C. (2017). The language of place: Semantic value from geospatial context. 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference, 2:99–104.

Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10578–10587.

Cui, Y., Yang, G., Veit, A., Huang, X., and Belongie, S. (2018). Learning to evaluate image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5804–5812.

Dai, W., Liu, Z., Ji, Z., Su, D., and Fung, P. (2022). Plausible may not be faithful: Probing object hallucination in vision-language pre-training. arXiv preprint arXiv:2210.07688.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. Proceedings of the ninth workshop on statistical machine translation, pages 376–380.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pages 4171–4186.

Durmus, E., He, H., and Diab, M. (2020). FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5055–5070.

Dziri, N., Kamalloo, E., Milton, S., Zaiane, O., Yu, M., Ponti, E. M., and Reddy, S. (2022). FaithDial: A faithful benchmark for information-seeking dialogue. arXiv preprint arXiv:2204.10757.

Dziri, N., Rashkin, H., Linzen, T., and Reitter, D. (2021). Evaluating groundedness in dialogue systems: The BEGIN benchmark. arXiv preprint arXiv:2105.00071.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30K: Multilingual English-German image descriptions. In Proceedings of the 5th Workshop on Vision and Language, pages 70–74.

Elliott, D. and Keller, F. (2014). Comparing automatic evaluation measures for image description. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 452–457.

Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. IEEE Transactions on Information theory, 49(7):1858–1860.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. (2021). SummEval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9:391–409.

Fang, Z., Wang, J., Hu, X., Liang, L., Gan, Z., Wang, L., Yang, Y., and Liu, Z. (2022). Injecting semantic concepts into end-to-end image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18009–18019.

Ferraro, F., Mostafazadeh, N., Vanderwende, L., Devlin, J., Galley, M., Mitchell, M., et al. (2015). A survey of current datasets for vision and language research. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 207–213.

Févry, T., Soares, L. B., FitzGerald, N., Choi, E., and Kwiatkowski, T. (2020). Entities as experts: Sparse memory access with entity supervision. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4937–4951.

Gahegan, M. (1995). Proximity operators for qualitative spatial reasoning. In International Conference on Spatial Information Theory, pages 31–44. Springer.

Gan, C., Gan, Z., He, X., Gao, J., and Deng, L. (2017). StyleNet: Generating attractive visual captions with styles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3137–3146.

Garrod, S., Ferrier, G., and Campbell, S. (1999). In and on: investigating the functional geometry of spatial prepositions. Cognition, 72(2):167–189.

Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research, 61:65–170.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.

Heinzerling, B. and Inui, K. (2021). Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1772–1791.

Herdade, S., Kappeler, A., Boakye, K., and Soares, J. (2019). Image captioning: Transforming objects into words. Advances in Neural Information Processing Systems, 32.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. (2021). CLIPScore: A reference-free evaluation metric for image captioning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7514–7528.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47:853–899.

Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence, 194:28–61.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A., and Matias, Y. (2022). TRUE: Re-evaluating factual consistency evaluation. In Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 161–175.

Honovich, O., Choshen, L., Aharoni, R., Neeman, E., Szpektor, I., and Abend, O. (2021). Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7856–7870.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR), 51(6):1–36.

Hu, A., Chen, S., and Jin, Q. (2020). ICECAP: Information concentrated entity-aware image captioning. In Proceedings of the 28th ACM International Conference on Multimedia, pages 4217–4225.

Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., and Wang, L. (2022). Scaling up vision-language pre-training for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17980–17989.

Huang, F., Li, Z., Wei, H., Zhang, C., and Ma, H. (2020). Boost image captioning with knowledge reasoning. Machine Learning, 109(12):2313–2332.

Huang, X., Yan, F., Xu, W., and Li, M. (2019). Multi-attention and incorporating background information model for chest X-ray image report generation. IEEE Access, 7:154808–154817.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2022). Survey of hallucination in natural language generation. ACM Computing Surveys.

Jiang, W., Li, Q., Zhan, K., Fang, Y., and Shen, F. (2022). Hybrid attention network for image captioning. Displays, page 102238.

Jiang, Y., Bordia, S., Zhong, Z., Dognin, C., Singh, M., and Bansal, M. (2020). HoVer: A dataset for many-hop fact extraction and claim verification. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3441–3460.

Jing, Y., Zhiwei, X., and Guanglai, G. (2020). Context-driven image caption with global semantic relations of the named entities. IEEE Access, 8:143584–143594.

Kaur, J. N., Bhatia, S., Aggarwal, M., Bansal, R., and Krishnamurthy, B. (2022). LM-CORE: Language models with contextually relevant external knowledge. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 750–769.

Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., and Erdem, E. (2017). Re-evaluating automatic metrics for image captioning. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 199–209.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. The annals of mathematical statistics, 22(1):79–86.

Laban, P., Schnabel, T., Bennett, P. N., and Hearst, M. A. (2022). SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. In Transactions of the Association for Computational Linguistics, volume 10, pages 163–177.

Lan, W., Li, X., and Dong, J. (2017). Fluency-guided cross-lingual image captioning. In Proceedings of the 25th ACM international conference on Multimedia, pages 1549–1557.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In International conference on machine learning, pages 1188–1196. PMLR.

Lee, H., Yoon, S., Dernoncourt, F., Bui, T., and Jung, K. (2021). UMIC: An unreferenced metric for image captioning via contrastive learning. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 220–226.

Lee, H., Yoon, S., Dernoncourt, F., Kim, D. S., Bui, T., and Jung, K. (2020). ViLBERTscore: Evaluating image caption using vision-and-language BERT. In Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, pages 34–39.

Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11):33–38.

Lenat, D. B. and Guha, R. V. (1989). Building large knowledge-based systems; representation and inference in the CYC project. Addison-Wesley Longman Publishing Co., Inc.

Li, G., Zhu, L., Liu, P., and Yang, Y. (2019a). Entangled transformer for image captioning. In Proceedings of the IEEE International Conference on Computer Vision, pages 8928–8937.

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. (2021). UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2592–2607.

Li, X., Lan, W., Dong, J., and Liu, H. (2016). Adding Chinese captions to images. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pages 271–275.

Li, X., Xu, C., Wang, X., Lan, W., Jia, Z., Yang, G., and Xu, J. (2019b). COCO-CN for cross-lingual image tagging, captioning, and retrieval. IEEE Transactions on Multimedia, 21(9):2347–2360.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. (2020). OSCAR: Object-semantics aligned pre-training for vision-language tasks. In European Conference on Computer Vision, pages 121–137. Springer.

Liao, S., Li, X., Shen, H. T., Yang, Y., and Du, X. (2015). Tag features for geo-aware image classification. IEEE transactions on multimedia, 17(7):1058–1067.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81.

Lin, J. (1991). Divergence measures based on the Shannon entropy. IEEE Transactions on Information theory, 37(1):145–151.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 740–755.

Liu, A., Du, J., and Stoyanov, V. (2019a). Knowledge-augmented language model and its application to unsupervised named-entity recognition. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1142–1150.

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2122–2132.

Liu, F., Wang, Y., Wang, T., and Ordonez, V. (2021a). Visual News: Benchmark and challenges in news image captioning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6761–6771.

Liu, F., Wu, X., Ge, S., Fan, W., and Zou, Y. (2021b). Exploring and distilling posterior and prior knowledge for radiology report generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13753–13762.

Liu, H. and Singh, P. (2004). ConceptNet:a practical commonsense reasoning tool-kit. BT technology journal, 22(4):211–226.

Liu, Q., Yogatama, D., and Blunsom, P. (2022). Relational memory-augmented language models. Transactions of the Association for Computational Linguistics, 10:555–572.

Liu, W., Chen, S., Guo, L., Zhu, X., and Liu, J. (2021c). CPTR: Full transformer network for image captioning. arXiv preprint arXiv:2101.10804.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., and Wang, P. (2020). K-BERT: Enabling language representation with knowledge graph. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 2901–2908.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Logan, R., Liu, N. F., Peters, M. E., Gardner, M., and Singh, S. (2019). Barack's wife Hillary: Using knowledge graphs for fact-aware language modeling. In Annual Meeting of the Association for Computational Linguistics (ACL).

Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017a). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. Proceedings of the IEEE conference on computer vision and pattern recognition, pages 375–383.

Lu, X., Wang, B., Zheng, X., and Li, X. (2017b). Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing, 56(4):2183–2195.

Luo, G., Cheng, L., Jing, C., Zhao, C., and Song, G. (2022). A thorough review of models, evaluation metrics, and datasets on image captioning. IET Image Processing, 16(2):311–332.

Mahdisoltani, F., Biega, J., and Suchanek, F. (2014). YAGO3: A knowledge base from multilingual Wikipedias. In 7th biennial conference on innovative data systems research. CIDR Conference.

Mai, G., Janowicz, K., Cai, L., Zhu, R., Regalia, B., Yan, B., Shi, M., and Lao, N. (2020). SE-KGE: A location-aware knowledge graph embedding model for geographic question answering and spatial semantic lifting. Transactions in GIS, 24(3):623–655.

Mai, G., Janowicz, K., Zhu, R., Cai, L., and Lao, N. (2021). Geographic question answering: challenges, uniqueness, classification, and future directions. AGILE: GIScience Series, 2:1–21.

Mark, D. M. and Frank, A. U. (1989). Concepts of space and spatial language. In Proceedings, Ninth International Symposium on Computer-Assisted Cartography (Auto-Carto 9), Baltimore, Maryland, pages 538–556.

Mathews, A., Xie, L., and He, X. (2016). SentiCap: Generating image descriptions with sentiments. In Proceedings of the AAAI conference on artificial intelligence, volume 30.

Mathur, N., Baldwin, T., and Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4984–4997.

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11):39–41.

Mishra, S., Suryavardan, S., Bhaskar, A., Chopra, P., Reganti, A., Patwa, P., Das, A., Chakraborty, T., Sheth, A., Ekbal, A., et al. (2022). FACTIFY: A multi-modal fact verification dataset. In Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY).

Mogadala, A., Bista, U., Xie, L., and Rettinger, A. (2018). Knowledge guided attention and inference for describing images containing unseen objects. European Semantic Web Conference, pages 415–429.

Mokady, R., Hertz, A., and Bermano, A. H. (2021). ClipCap: CLIP prefix for image captioning. arXiv preprint arXiv:2111.09734.

Nikiforova, S., Deoskar, T., Paperno, D., and Winter, Y. (2020). Geo-aware image caption generation. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3143–3156, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nikiforova, S., Deoskar, T., Paperno, D., and Winter, Y. (2022). Generating image captions with external encyclopedic knowledge. arXiv preprint arXiv:2210.04806.

Nitta, N., Nakamura, K., and Babaguchi, N. (2020). Constructing geospatial concept graphs from tagged images for geo-aware fine-grained image recognition. ISPRS International Journal of Geo-Information, 9(6):354.

Novikova, J., Dušek, O., Curry, A. C., and Rieser, V. (2017). Why we need new evaluation metrics for NLG. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2241–2252.

Olsher, D. (2014). Semantically-based priors and nuanced knowledge core for Big Data, Social AI, and language understanding. Neural Networks, 58:131–147.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, pages 8024–8035.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.

Peters, M. E., Neumann, M., Logan, R., Schwartz, R., Joshi, V., Singh, S., and Smith, N. A. (2019). Knowledge enhanced contextual word representations. In Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473.

Poerner, N., Waltinger, U., and Schütze, H. (2019). BERT is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised QA. arXiv preprint arXiv:1911.03681.

Poerner, N., Waltinger, U., and Schütze, H. (2020). E-BERT: Efficient-yet-effective entity embeddings for BERT. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 803–818.

Ponzetto, S. P. and Strube, M. (2008). WikiTaxonomy: A large scale knowledge resource. In ECAI 2008, pages 751–752. IOS Press.

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., Murdock, V., et al. (2018). Geographic information retrieval: Progress and challenges in spatial search of text. Foundations and Trends® in Information Retrieval, 12(2-3):164–318.

Radev, D. R., Stent, A., Tetreault, J., Pappu, A., Iliakopoulou, A., Chanfreau, A., de Juan, P., Vallmitjana, J., Larrarte, A. J., Jha, R., et al. (2016). Humor in collective discourse: unsupervised funniness detection in the New Yorker cartoon caption contest. In 10th conference on International Language Resources and Evaluation (LREC'16), pages 475–479. European Language Resources Association.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763. PMLR.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Razniewski, S., Yates, A., Kassner, N., and Weikum, G. (2021). Language models as or for knowledge bases. arXiv preprint arXiv:2110.04888.

Redmon, J. and Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Reiter, E. (2018). A structured review of the validity of BLEU. Computational Linguistics, 44(3):393–401.

Reiter, E. and Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Computational Linguistics, 35(4):529–558.

Ristoski, P. and Paulheim, H. (2016). RDF2Vec: RDF graph embeddings for data mining. In International Semantic Web Conference, pages 498–514. Springer.

Roberts, A., Raffel, C., and Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5418–5426.

Robinson, V. B. (1990). Interactive machine acquisition of a fuzzy spatial relation. Computers & Geosciences, 16(6):857–872.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. (2018). Object hallucination in image captioning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4035–4045.

Rosset, C., Xiong, C., Phan, M., Song, X., Bennett, P., and Tiwary, S. (2020). Knowledge-aware language model pretraining. arXiv preprint arXiv:2007.00655.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252.

Safavi, T. and Koutra, D. (2021). Relational world knowledge representation in contextual language models: A review. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1053–1067.

Santhanam, S., Hedayatnia, B., Gella, S., Padmakumar, A., Kim, S., Liu, Y., and Hakkani-Tur, D. (2021). Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. arXiv preprint arXiv:2110.05456.

Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., and Choi, Y. (2019). ATOMIC: An atlas of machine commonsense for if-then reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 3027–3035.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823.

Schuler, K. K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. University of Pennsylvania.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In 54th Annual Meeting of the Association for Computational Linguistics, pages 1715–1725. Association for Computational Linguistics (ACL).

Sharif, N., White, L., Bennamoun, M., and Afaq Ali Shah, S. (2018). NNEval: Neural network based evaluation metric for image captioning. In Proceedings of the European Conference on Computer Vision (ECCV), pages 37–53.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2556–2565.

Shuster, K., Humeau, S., Hu, H., Bordes, A., and Weston, J. (2019). Engaging image captioning via personality. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12516–12526.

Skreta, M., Luccioni, A., and Rolnick, D. (2020). Spatiotemporal features improve fine-grained butterfly image classification. In Conference on Neural Information Processing Systems.

Speer, R., Chin, J., and Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. Thirty-First AAAI Conference on Artificial Intelligence.

Srinivasan, M. and Rafiei, D. (2021). Location-aware named entity disambiguation. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pages 3433–3438.

Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., and Cucchiara, R. (2021). From show to tell: A survey on image captioning. arXiv preprint arXiv:2107.06912.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). YAGO: A large ontology from Wikipedia and WordNet. Journal of Web Semantics, 6(3):203–217.

Sulem, E., Abend, O., and Rappoport, A. (2018). BLEU is not suitable for the evaluation of text simplification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 738–744.

Sun, T., Shao, Y., Qiu, X., Guo, Q., Hu, Y., Huang, X.-J., and Zhang, Z. (2020). CoLAKE: Contextualized language and knowledge embedding. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3660–3670.

Takemura, C. M., Cesar, R., and Bloch, I. (2005). Fuzzy modeling and evaluation of the spatial relation “along”. In Iberoamerican Congress on Pattern Recognition, pages 837–848. Springer.

Tandon, N., De Melo, G., Suchanek, F., and Weikum, G. (2014). WebChild: Harvesting and organizing commonsense knowledge from the web. In Proceedings of the 7th ACM international conference on Web search and data mining, pages 523–532.

Tandon, N., De Melo, G., and Weikum, G. (2017). WebChild 2.0: Fine-grained commonsense knowledge distillation. In Proceedings of ACL 2017, System Demonstrations, pages 115–120.

Tang, K., Paluri, M., Fei-Fei, L., Fergus, R., and Bourdev, L. (2015). Improving image classification with location context. Proceedings of the IEEE International Conference on Computer Vision, pages 1008–1016.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. (2016). YFCC100M: The new data in multimedia research. Communications of the ACM, 59(2):64–73.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018a). FEVER: a large-scale dataset for Fact Extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 809–819.

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2018b). The Fact Extraction and VERification (FEVER) shared task. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 1–9.

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2019). The FEVER2.0 shared task. In Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), pages 1–6.

Topal, M. O., Bas, A., and van Heerden, I. (2021). Exploring transformers in natural language generation: GPT, BERT, and XLNet. arXiv preprint arXiv:2102.08036.

Tran, A., Mathews, A., and Xie, L. (2020). Transform and Tell: Entity-aware news image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13035–13045.

Trinh, T. H. and Le, Q. V. (2019). Do language models have common sense? https://openreview.net/forum?id=rkgfWh0qKX.

Van Miltenburg, E., Elliott, D., and Vossen, P. (2018). Measuring the diversity of automatic image descriptions. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1730–1741.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4566–4575.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and Tell: A neural image caption generator. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 3156–3164.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledge base. Communications of the ACM, 57(10):78–85.

Wang, R., Tang, D., Duan, N., Wei, Z., Huang, X.-J., Ji, J., Cao, G., Jiang, D., and Zhou, M. (2021a). K-Adapter: Infusing knowledge into pre-trained models with adapters. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1405–1418.

Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., and Tang, J. (2021b). KEPLER: A unified model for knowledge embedding and pre-trained language representation. Transactions of the Association for Computational Linguistics, 9:176–194.

Wang, Y., Xu, J., Sun, Y., and He, B. (2019). Image captioning based on deep learning methods: A survey. arXiv preprint arXiv:1905.08110.

Wei, X., Wang, S., Zhang, D., Bhatia, P., and Arnold, A. (2021). Knowledge enhanced pretrained language models: A compreshensive survey. arXiv preprint arXiv:2110.08455.

Whitehead, S., Ji, H., Bansal, M., Chang, S.-F., and Voss, C. (2018). Incorporating background knowledge into video description generation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3992–4001.

Wu, F., Fan, A., Baevski, A., Dauphin, Y. N., and Auli, M. (2019). Pay less attention with lightweight and dynamic convolutions. arXiv preprint arXiv:1901.10430.

Wu, Q., Shen, C., Wang, P., Dick, A., and Van Den Hengel, A. (2017). Image captioning and visual question answering based on attributes and external knowledge. IEEE transactions on pattern analysis and machine intelligence, 40(6):1367–1381.

Xiao, Y. and Wang, W. Y. (2021). On hallucination and predictive uncertainty in conditional language generation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2734–2744.

Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. 32nd International Conference on Machine Learning, ICML 2015, pages 2048–2057.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., and Matsumoto, Y. (2018). Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. arXiv preprint arXiv:1812.06280.

Yan, C., Hao, Y., Li, L., Yin, J., Liu, A., Mao, Z., Chen, Z., and Gao, X. (2021). Task-adaptive attention for image captioning. IEEE Transactions on Circuits and Systems for Video technology, 32(1):43–51.

Yang, L., Li, X., Song, R., Zhao, B., Tao, J., Zhou, S., Liang, J., and Yang, J. (2022a). Dynamic MLP for fine-grained image classification by leveraging geographical and temporal information. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10945–10954.

Yang, X., Karaman, S., Tetreault, J., and Jaimes, A. (2021). Journalistic guidelines aware news image captioning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5162–5175.

Yang, Y., Gao, Y., Li, J., and Huang, H. (2022b). G^2: Enhance knowledge grounded dialogue via ground graph. arXiv preprint arXiv:2204.12681.

Yang, Z., Zhang, Y.-J., Huang, Y., et al. (2017). Image captioning with object detection and localization. In International conference on image and graphics, pages 109–118. Springer.

Yao, T., Pan, Y., Li, Y., and Mei, T. (2018). Exploring visual relationship for image captioning. In Proceedings of the European conference on computer vision (ECCV), pages 684–699.

Yoshikawa, Y., Shigeto, Y., and Takeuchi, A. (2017). STAIR captions: Constructing a large-scale Japanese image caption dataset. arXiv preprint arXiv:1705.00823.

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4651–4659.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations. Transactions of the Association of Computational Linguistics, 2(1):67–78.

Zeng, P., Zhang, H., Song, J., and Gao, L. (2022). S2 transformer for image captioning. In Proceedings of the International Joint Conferences on Artificial Intelligence, volume 5.

Zhang, H., Liu, X., Pan, H., Song, Y., and Leung, C. W.-K. (2020). ASER: A large-scale eventuality knowledge graph. In Proceedings of The Web Conference 2020, pages 201–211.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10):1499–1503.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. (2021). VinVL: Revisiting visual representations in vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5579–5588.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019a). BERTScore: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675.

Zhang, T., Wang, C., Hu, N., Qiu, M., Tang, C., He, X., and Huang, J. (2022). DKPLM: Decomposable knowledge-enhanced pre-trained language model for natural language understanding. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 11703–11711.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2019b). ERNIE: Enhanced language representation with informative entities. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1441–1451.

Zhao, W., Hu, Y., Wang, H., Wu, X., and Luo, J. (2021). Boosting entity-aware image captioning with multi-modal knowledge graph. arXiv preprint arXiv:2107.11970.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6):1452–1464.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., and Gao, J. (2020). Unified vision-language pre-training for image captioning and VQA. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 13041–13049.

Zhou, Y., Sun, Y., and Honavar, V. (2019). Improving image captioning by leveraging knowledge graphs. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 283–293.

Zhu, X., Li, L., Liu, J., Peng, H., and Niu, X. (2018a). Captioning transformer with stacked attention modules. Applied Sciences, 8(5):739.

Zhu, Z., Xue, Z., and Yuan, Z. (2018b). Topic-guided attention for image captioning. In 2018 25th IEEE International Conference on Image Processing (ICIP), pages 2615–2619. IEEE.

Zipf, G. K. (1949). Human behavior and the principle of least effort. Addison-Wesley Press.

Zou, C., Xu, F., Wang, M., Li, W., and Cheng, Y. (2022). Solutions for fine-grained and long-tailed snake species recognition in SnakeCLEF 2022. arXiv preprint arXiv:2207.01216.

Samenvatting in het Nederlands

Dit proefschrift is gewijd aan automatische beeldbeschrijving (*image captioning*), de taak om automatisch een natuurlijke taalbeschrijving van een bepaalde afbeelding te genereren. Moderne automatische beeldbeschrijvingssystemen zijn behoorlijk geavanceerd en presteren goed op standaardbenchmarks. Ze vereenvoudigen echter vaak de realiteit die ze moeten modelleren door de focus alleen te leggen op enkele van de informatiedimensies die mensen gebruiken. De meeste bestaande systemen zijn getraind om een visuele beschrijving te geven van wat direct in de afbeelding te zien is. Een door mensen geschreven beschrijving kan daarentegen ook informatie bevatten die niet uit de afbeelding zelf kan worden afgeleid — verwijzingen naar kennis extern aan de afbeelding. Het verkennen van manieren om automatische beeldbeschrijving te verrijken met contextueel relevante externe kennis is de belangrijkste focus van dit proefschrift.

Hoofdstuk 1 beschrijft de motivatie en het doel van dit werk, gerelateerde uitdagingen en onze bijdragen. Het hoofddoel van dit proefschrift is het ontwikkelen van een nieuwe methode voor het integreren van externe kennis in het automatische beeldbeschrijvingsproces. Zo'n taak omvat verschillende uitdagende stappen. De relevante externe kennis moet (1) worden geïdentificeerd en geëxtraheerd, (2) effectief worden gecodeerd en (3) worden geïntegreerd in de beschrijvingspijplijn. Vervolgens, zodra de beschrijvingen zijn gegenereerd met verwijzingen naar externe kennis, (4) moet de evaluatieprocedure beoordelen of deze verwijzingen nauwkeurig zijn en passen bij de afbeeldingen.

In onze aanpak wordt identificatie van relevante kennis uitgevoerd door een *contextualiseringsanker*. Dat is een element van afbeeldingsgerelateerde gegevens (bijv. metadata) dat wordt gebruikt om te bepalen welk deel van de wereldkennis die beschikbaar is in externe bronnen nuttig zou zijn voor het beschrijven van een bepaalde afbeelding. De opgehaalde kennis wordt vervolgens georganiseerd in twee datastructuren: de *entiteitscontext* en de *kenniscontext*. Entiteiten uit de echte wereld die worden opgehaald via het contextualisatieanker vormen de entiteitscontext. Dit is het deel van de algemene context van de afbeelding dat bestaat uit de voor de afbeelding relevante entiteiten die er al dan niet op zijn afgebeeld. De kenniscontext breidt de entiteitscontext uit met verschillende feiten over de entiteiten. De entiteitscontext en de kenniscontext samen zijn bedoeld om belangrijke aspecten van relevante wereldkennis weer te geven die mensen zouden kunnen hebben wanneer ze een bepaalde afbeelding beschrijven. We integreren zowel de entiteits- als de kenniscontext in een voor de rest standaard beeldbeschrijvingssysteem als extra informatiebronnen voor het genereren van een beschrijving, naast de afbeelding zelf. Het doel van het resulterende "kennisbewuste" systeem is om beeldbeschrijvingen te genereren die worden beïnvloed door de relevante externe kennis en die mogelijk expliciete verwijzingen daarnaar bevatten. Tijdens de evaluatie besteden we speciale aandacht aan het meten van feitelijke nauwkeurigheid, d.w.z. de waarheidsgetrouwheid van kennis extern aan de afbeelding in de gegenereerde beschrijvingen.

Hoofdstuk 2 bevat een overzicht van gerelateerd werk op het gebied van automatische beeldbeschrijving en bestaande benaderingen voor de integratie van wereldkennis in taalmodellen. Met betrekking tot beeldbeschrijving bespreken we de state-of-the-art in het veld en een veelvoorkomende beperking van moderne systemen, namelijk "hallucinatie" in de gegenereerde beschrijvingen. We beschrijven de veelgebruikte datasets (waarbij het contrast wordt benadrukt tussen de datasets met beeldbeschrijvingen die zijn verzameld via crowdsourcing en natuurlijk geproduceerde beeldbeschrijvingen) en evaluatiemetrieken. Vervolgens bekijken we het meest relevante eerdere onderzoek naar het verbeteren van taalmodellen door gebruik te maken van wereldkennis uit externe databases.

Hoofdstuk 3 presenteert onze aanpak met een focus op geografische kennis. Het in dit hoofdstuk ontwikkelde beeldbeschrijvingssysteem maakt gebruik van de locatie van de afbeelding (coördinaten van de lengte- en breedtegraad) als contextualiseringsanker om concrete geografische entiteiten in en rond de afbeelding te identificeren. Deze entiteiten vormen de domeinspecifieke realisatie van een entiteitscontext, een geografische entiteitscontext, waarin ze zijn gecodeerd via hun geografische kenmerken. De geografische entiteitscontext is verder geïntegreerd in een standaard encoder-decoder pijplijn voor beeldbeschrijving, waar het extra input levert voor de encoder en een extra vocabulaire voor de decoder, waardoor het entiteitsnamen in de beschrijvingen kan genereren. Het systeem is getraind op een nieuwe GeoRic dataset, die we ook in dit hoofdstuk presenteren. Deze dataset bevat natuurlijk geproduceerde beschrijvingen met veel verwijzingen naar de geografische context van de afbeeldingen, bijv. “to the north of London [ten noorden van Londen]”, “near High Street [nabij High Street]”. De evaluatie laat een substantiële verbetering zien ten opzichte van de standaard systemen, en met name het vermogen van ons systeem om specifieke verwijzingen naar geografische kennis correct te produceren.

Hoofdstuk 4 beschrijft een verdere ontwikkeling van onze aanpak. Het in dit hoofdstuk gepresenteerde beeldbeschrijvingssysteem breidt het systeem uit Hoofdstuk 3 uit met de introductie van de kenniscontext. De kenniscontext van relevante encyclopedische feiten is geïntegreerd in het beeldbeschrijvingsproces naast de geografische entiteitscontext. Het wordt toegevoegd als een andere input aan de encoder, en in de decoder biedt het extra contextualisatie voor het genereren van reguliere woorden en een ander vocabulaire voor het genereren van feitgerelateerde tokens. Om dit systeem te trainen, hebben we een andere dataset samengesteld, K-GeoRic, waarin natuurlijk geproduceerde beeldbeschrijvingen diverse encyclopedische feiten over de geografische entiteiten bevatten. Op de testset van K-GeoRic presteert ons systeem aanzienlijk beter dan verschillende andere systemen in standaard evaluatiemetrieken en, belangrijker nog, in de nauwkeurigheid van de gegenereerde feiten.

Hoofdstuk 5 past onze aanpak toe op de kwalitatief verschillende gegevens uit het nieuwsdomein, namelijk afbeeldingen en hun beschrijvingen uit de artikelen van de New York Times. Het systeem in dit hoofdstuk repliceert de architectuur van het systeem uit Hoofdstuk 4, met enkele generalisaties die het resultaat zijn van het gebruik van de nieuwsartikelen als contextualiseringsanker. De entiteitscontext is opgebouwd uit de bij naam genoemde entiteiten in de artikeltekst (niet alleen geografische) en de kenniscontext bevat encyclopedische feiten over deze entiteiten. Beide contexten zijn op dezelfde manier in het beeldbeschrijvingsproces geïntegreerd als in Hoofdstuk 4. Het resulterende systeem kan gecontextualiseerde beeldbeschrijvingen genereren die informatie uit zowel het artikel als externe kennisbronnen bevatten.

Hoofdstuk 6 bevat een samenvatting van de bijdragen van dit proefschrift en een bespreking van mogelijkheden voor toekomstig onderzoek.

Bijlage A beschrijft de implementatiedetails van onze beeldbeschrijvingssystemen.

Bijlage B bevat links naar de software en datasets die in dit werk zijn ontwikkeld.

Curriculum Vitae

Sofia (Sonya) Nikiforova was born in 1994 in Surgut, Russian Federation. She graduated *cum laude* from both her BA in Theoretical and Applied Linguistics (Moscow State University, Moscow, 2016) and MA in Computational Linguistics (HSE University, Moscow, 2018). During her studies, Sonya developed a keen interest in language theory and typology (having participated in more than ten linguistic field trips), as well as in programming and building computational models of natural language. In parallel with her education, she has been employed as Lead NLP Developer in several high-tech startups. In 2018, Sonya started her PhD in the ERC-funded ROCKY project at Utrecht University. The results of her research on image captioning with external knowledge are presented in this dissertation. Currently, Sonya continues to work on knowledge-aware image captioning as a postdoctoral researcher in the ERC Proof of Concept project at Utrecht University.