

MMChat: Multi-Modal Chat Dataset on Social Media

Yinhe Zheng^{♣*}, Guanyi Chen[♣], Xin Liu[♡], and Ke Lin[♡]

[♣] Alibaba Group

[♣]Department of Information and Computing Sciences, Utrecht University

[♡]Samsung Research China - Beijing (SRC-B)

zhengyinhe1@163.com

Abstract

Incorporating multi-modal contexts in conversation is an important step for developing more engaging dialogue systems. In this work, we explore this direction by introducing MMCHAT: a large scale multi-modal dialogue corpus (32.4M raw dialogues and 120.84K filtered dialogues). Unlike previous corpora that are crowd-sourced or collected from fictitious movies, MMCHAT contains image-grounded dialogues collected from real conversations on social media, in which the *sparsity* issue is observed. Specifically, image-initiated dialogues in common communications may deviate to some non-image-grounded topics as the conversation proceeds. We develop a benchmark model to address this issue in dialogue generation tasks by adapting the attention routing mechanism on image features. Experiments demonstrate the usefulness of incorporating image features and the effectiveness in handling the sparsity of image features.

1 Introduction

Attempts to build human-like open-domain dialogue systems generally follow two angles: 1). enriching the dialogue system with textual or structural contexts such as knowledge (Madotto et al., 2018) or personalities (Zhang et al., 2018a; Zheng et al., 2019); and 2). enabling the dialogue systems to perceive multi-modality contexts beyond text, such as vision, voice, or even gesture (Shuster et al., 2020b,c; Liao et al., 2018; Ju et al., 2019). Systems that are built following the second angle are also known as Multi-Modal Dialogue Systems (MMDSs).

To facilitate the development of data-driven MMDSs, a few dialogue datasets containing visual information have been constructed (Mostafazadeh

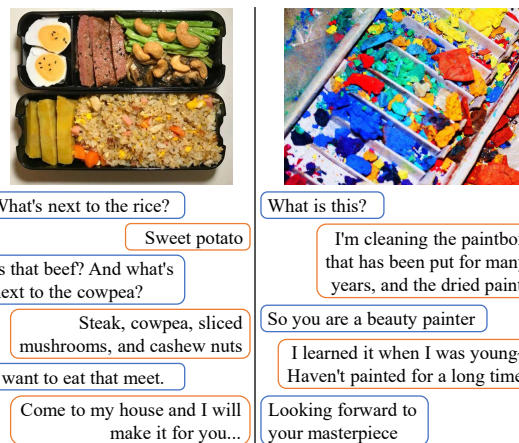


Figure 1: Example dialogues from MMCHAT (translated from Chinese).

et al., 2017; Mogadala et al., 2019; AlAmri et al., 2019; Kottur et al., 2019; Pasunuru and Bansal, 2018). For instance, Shuster et al. (2020a) introduced a crowd-sourced image grounded dialogue corpus IMAGE-CHAT, in which annotators are employed to chat based on given images. Meng et al. (2020) proposed OPENVIDIAL by directly extracting dialogues and their visual contexts from movies and TV series. There are also many works on visual question answering (Das et al., 2017) that focuses on the QA tasks involving image inputs.

One drawback of existing datasets is that they enforce every utterance in a dialogue to be grounded on the given image. Nevertheless, this is not always true in our daily communications. Concretely, the topic triggered by the image may drift in the conversation flow so that not every utterances in a dialogue session are image-grounded. Taking the right dialogue in Figure 1 as an example, it is initialized by the objects shown in the given image (i.e., “paintbox” and “dried paint”) but the following dialogue’s focus move to the speaker’s own experience as a painter, which is not image-related anymore.

* Work done while Yinhe Zheng worked in Samsung Research.

A similar pattern is also observed in the left example of Figure 1. We refer to this phenomenon as the issue of *sparsity* and dialogues that exhibit this phenomenon as the *sparse image-grounded dialogues*.

To tackle the above issue, we introduce MMCHAT: a large-scale dataset that contains sparse image-grounded dialogues in Chinese. The dialogues in MMCHAT originate from real conversations collected from Weibo (A large Chinese social media). Specifically, MMCHAT contains 120.84K sessions of open-domain dialogues (filtered out of 32.4M sessions of raw dialogues) and 204.32K corresponding images. Two example dialogues are shown in Figure 1. Each image in MMCHAT is also released with a set of object labels and a descriptive caption that is predicted by a state-of-the-art pre-trained image caption model. Besides, we also provide a strong benchmark model to tackle the image-sparsity issue in open-domain dialogue generation tasks based on the attention routing mechanism. Evaluation results suggest that incorporating visual contexts contributes positively to dialogue modeling, and the approach used in our benchmark model helps alleviate the sparsity issue.

Besides enlightening new models for realistic multi-modal conversations, MMCHAT is also built to help understand how Chinese multi-modal communications are conducted from the aspect of social science (Jovanovic and Van Leeuwen, 2018). As being indicated in the Broader Considerations, due to privacy issues, please contact the corresponding author for accessing the dataset.

2 Dataset Construction: MMCHAT

2.1 Data Collection

MMCHAT originates from Weibo, on which users can share their daily lives through images and texts. In this study, a two-phase pipeline is used to construct MMCHAT:

The first phase aims to collect seed users who are active on Weibo. Specifically, we start this phase with a few hand-collected mass media Weibo accounts. These accounts are maintained by professionals and are committed to post daily news on broad topics. The users who comment under these news are collected as our seed users.

The second phase starts from the seed users collected above. Specifically, the images posted by these seed users are obtained, and the comments under these images are collected. Dialogues along

these images are constructed by restoring the reply relationship between these comments.

The two-phase data collection approach used in our study effectively avoids spammers’ noises since most spammers will not bother to follow and reply to daily news. Moreover, we also filter out seed users that are not active to make the data collection process more effective. Finally, we collect about 32.4M sessions of raw dialogues.

2.2 Data Filtering and Processing

To improve the quality of MMCHAT, a set of rules are carefully designed to filter out low-quality images and dialogues from the raw corpus collected in Section 2.1. Specifically, images with extremely low resolution (< 500 pixels) or high aspect ratios are abandoned, and dialogues that contain utterances longer than 50 tokens are filtered. The offensive contents are also filtered using an offensive word list and a pre-trained offensive content classifier (Wang et al., 2020).

To ensure the dialogue contents in MMCHAT are related to the corresponding images in the first few turns of the conversation, we filter out images that are not uploaded through the direct-share mode of Weibo. This mode allows users to share images without providing textual content. We argue that the initial few turns of the dialogues following these image-only weibo posts are triggered by the visual information because there are no previous textual contexts except for the uploaded images.

Note that eliminating weibo posts that are not uploaded through the direct-share mode filters out a vast majority of collected raw dialogues. However, this rule is adopted not because these filtered dialogues are of low quality but because we only have limited computation resources. We want our model to focus on dialogues that can more significantly capture the sparsity issue. We believe these filtered raw dialogues are useful in building larger-scale multi-modal dialogue models. We will release all the collected raw dialogues and the data processing scripts to facilitate further studies in this direction.

To protect the data privacy, MMCHAT is released under strict terms for academic users only. Further, we only release the URL links to these dialogues and images. Scripts are provided to help user collect MMCHAT from Weibo’s server.

2.3 Data Analysis

The statistics of the resulting MMCHAT dataset is shown in Table 1. Each dialogue session is associ-

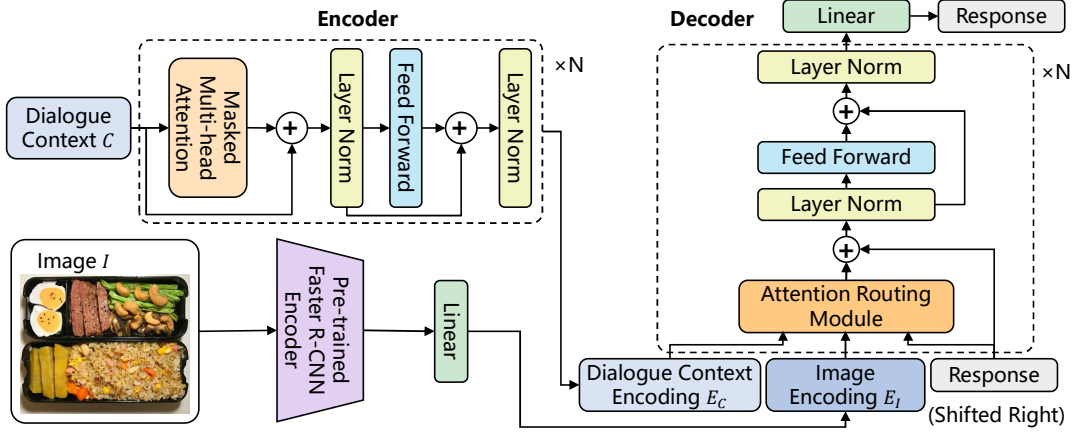


Figure 2: Overview of multi-modal dialogue generation model.

ated with at least one image (9 images maximum), and a considerable amount of sessions (more than 17K) in MMCHAT contain at least 4 utterances (i.e., 2 turns). Note that different dialogues may share the same weibo post (i.e., the same set of images).

We also analyzed the contents of the images. Specifically, a Faster R-CNN model (Ren et al., 2015; Lu et al., 2019) trained with attribute head on the Visual Genome (Krishna et al., 2017) dataset is used to detect objects in each image. The regions where any class detection confidence exceeds a threshold (0.2 in our case following (Anderson et al., 2018)) are selected to further detect the specific object labels. We follow the work of Anderson et al. (2018) to use an object and attribute vocabulary with the size of 1600 and 400, respectively. Finally, an average count of 11.42 objects is detected in each image. This indicates that images in our dataset contain rich semantic information and thus being informative. To facilitate future studies, we also release these detected object labels and the spatial location of each bounding box. Moreover, a descriptive caption for each image in MMCHAT is further predicted using an image caption model pre-trained on the MSCOCO dataset (Lin et al., 2014). We release all these predicted English captions and their Chinese translations, which can serve as further indicators of the image contents and can be utilized in further studies.

3 Dialogue Generation on MMCHAT

3.1 Task Definition

The task of MMDS is to learn a function f that can map textual contexts \mathcal{C} (e.g., dialogue histories) and multi-modal contexts \mathcal{I} (e.g., images) into dialogue responses Y , i.e., learn $f : \{\mathcal{C}, \mathcal{I}\} \mapsto Y$. In this

#(Dialogues)	120.84K
#(Total Images)	204.32K
#(Total Utterances)	314.13K
#(Dialogue Sessions) Longer than 4	17.32K
#(Image) per Dialogue	2.91
#(Character) per Utterance	8.52
#(Detected Object) per Image	11.42
#(Raw Dialogues)	32.4M

Table 1: Statistics of MMCHAT.

study, we focus on the image modality in \mathcal{I} , i.e., \mathcal{I} is composed of a set of images $\{I_n\}_{n=1}^N$.

3.2 Dialogue Generation Model

The Seq2Seq architecture is used as our backbone to build multi-modal dialogue generation model. As shown in Figure 2, two encoders are used to respectively encode the textual context \mathcal{C} and image context \mathcal{I} into encoded representations E_C and E_I . An attention routing module is utilized to merge E_C and E_I in the decoder, and the response Y is decoded auto-regressively.

3.2.1 Encoder

The encoder for the textual context \mathcal{C} and the dialogue decoder are both parameterized with the Transformer architecture (Vaswani et al., 2017) (12 layers, 12 attention heads, and 768 hidden states) and share the same set of weights. To further improve the generation quality, we initialize their weights using a pre-trained GPT model (Radford et al., 2018). Utterances in the dialogue history are concatenated using a special token “[SEP]”, and E_C is obtained by feeding the concatenated token sequence into the textual encoder.

Model	BLEU-2,3,4			Dist-1	Dist-2	Ent-1	Ent-2
Seq2Seq	2.830	1.376	0.805	2.63	33.92	6.00	9.47
Seq2Seq+PIMG	2.928	1.469	0.888	2.73	34.34	6.01	9.45
	(+3.46%)	(+6.76%)	(+10.31%)	(+3.80%)	(+1.24%)	(+0.17%)	(-0.21%)
Seq2Seq+IMG	3.001	1.588	1.006	2.82	35.38	6.07	9.52
	(+6.04%)	(+15.41%)	(+24.97%)	(+7.22%)	(+4.30%)	(+1.17%)	(+0.53%)
Human Reference	N/A	N/A	N/A	9.09	48.77	6.69	9.64

Table 2: Evaluation Results. Relative improvements comparing to the Seq2Seq baseline is shown in parentheses.

The encoder for the image context \mathcal{I} is implemented as the Faster R-CNN model with ResNet-101 backbone. The weights of this encoder are pre-trained on the Visual Genome dataset and fixed in the training process. Specifically, a feature vector with the size of 2048 is extracted from each image region. The top-50 high confidence regions are used to produce $E_{\mathcal{I}}$ with a linear layer to adjust the feature-length, i.e., the resulting $E_{\mathcal{I}}$ contains 50 features each has a length of 768.

3.2.2 Decoder

To tackle the sparsity issue, we equip the dialogue decoder with the attention routing mechanism (Zheng et al., 2020) to balance the contribution of each region feature. Specifically, given the encoding of the dialogue context E_C , image context $E_{\mathcal{I}}$, and previous decoded tokens E_{pre} , three attention routes are computed as:

$$O_C = \text{MHA}(E_{\text{pre}}, E_C, E_C), \quad (1)$$

$$O_{\mathcal{I}} = \text{MHA}(E_{\text{pre}}, \gamma E_{\mathcal{I}}, \gamma E_{\mathcal{I}}), \quad (2)$$

$$O_{\text{pre}} = \text{MMHA}(E_{\text{pre}}, E_{\text{pre}}, E_{\text{pre}}), \quad (3)$$

where E_{pre} is the query, $\gamma \in [0, 1]$ is a hyper-parameter to re-scale $E_{\mathcal{I}}$, MHA and MMHA represents masked and unmasked multi-head attention, respectively. The results of each attention operation are averaged before proceeding to the next sub-module:

$$O_{\text{merge}} = \frac{O_C + O_{\mathcal{I}} + O_{\text{pre}}}{3} + E_{\text{pre}}. \quad (4)$$

Note that the attention route on image features (i.e., Eq. 2) assigns different weights to different image regions. This facilitates more flexible control over image features in the decoding process and thus helps ease the sparsity issue.

4 Experiments

4.1 Implementation Details

In our proposed dialogue model, the encoder and decoder are 12-layer transformers with 768-dimensional hidden states and 12 attention heads. For the position-wise feed-forward networks, 3,072-dimensional inner states are used. The Adam optimizer is used to train our model with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. The maximum learning rate is set to $1.0e-4$. The training starts with a *warmupstep* of 1,000, and the learning rate is annealed proportionally to the inverse square root of the step number. Batch size is set to 360, and the training iterates 60 epochs. A character-level vocabulary of size 13,084 is used. Other settings of our Transformer model follow the work of Radford et al. (2018).

We share the weights of the encoder and decoder in the dialogue model and initialize these weights using a pre-trained GPT model (Wang et al., 2020). The pretraining corpus contains about 0.5 billion tokens, and the pretraining process lasts for a week on 8 GTX1080Ti GPUs. In the inference phase, we use the top-K ($K = 20$) decoding scheme with beam search for the dialogue model. The beam size and length penalty is set to 4, and 2.0, respectively. The training of our model last for about 24 hours on 2 NVIDIA Tesla V100 GPUs. The number of parameters for our model (textual encoder and decoder, not including the image encoder) are 191.01M. Note that due to the large computation loads to utilize the pre-training based method in our model and our baselines, we inherit most of the hyper-parameter settings from the previous studies, such as Radford et al. (2018), and skip the hyper-parameter tuning process. Moreover, for fair comparisons, we use a fixed set of the hyper-parameters in all our experiments (including all the baselines).

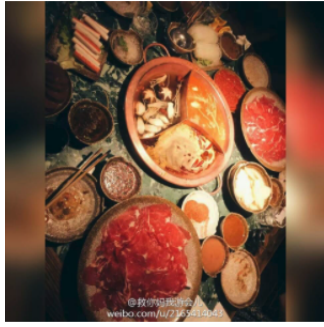

Input Image	Dialogue Contexts and Generated Responses
	<p>Context: “在拜托啦学长那里看到你，欢迎来我们福建玩，坐标福建漳州。不过现在在外上学” (“I saw you in the channel of ”Please! Seniors.”. Welcome to Fujian! I am from Zhangzhou, Fujian, but studying out of Fujian.”)</p> <p>Human: 厦门。 (Xiamen.)</p> <p>Seq2Seq: 啊，那还真的好啊，我已经回来了。 (Aha! That’s great! I have been back.)</p> <p>Seq2Seq+PIMG: 我是漳州的学妹。 (I am your younger female schoolmate from Zhangzhou.)</p> <p>Seq2Seq+IMG: 原来是这样，下次有机再来福建找你玩哈。 (I see. I’ll look you up, if I ever have chance to visit Fujian again.)</p>
	<p>Context: “你又睡那么晚”，“美甲！”“你有空了就撒” (“Why you go to sleep so late again?”, “Just back from a nail salon.”, “Shall we meet if you have spare time?”)</p> <p>Human: 好。 (Okey.)</p> <p>Seq2Seq: 我就在四小时以左右开始睡不着觉！其他时候都是两三点。 (I found it is hard to fall in sleep at 4 o’clock this morning. This used to be 2 or 3 o’clock.)</p> <p>Seq2Seq+PIMG: 上完课就去上，没时间了。 (I’ll go there after class. There is no time.)</p> <p>Seq2Seq+IMG: 好滴！国庆之后就能见到了。 (Okey. I suppose to meet you soon after the national day.)</p>

Table 3: Example outputs (with English translations) generated by our multi-modal dialogue systems.

4.2 Baselines

We also implement two baselines to validate our dataset and model: **Seq2Seq**: a vanilla Seq2Seq model is built with only textual inputs; **Seq2Seq+PIMG**: an image-grounded dialogue model is built with a single pooled image representation. Specifically, an max-pooling operation is applied to E_I , and the pooled vector is added to each representation vector in E_C . The attention route to E_I (i.e., Eq. 2) is not applied. Note that the first baseline does not use image contexts, and the second baseline does not model the sparsity phenomenon.

For fair comparisons, all baselines employ the same architecture, hyper-parameter setting and initialization scheme with our model **Seq2Seq+IMG**.

4.3 Metrics

We use the following metrics: **BLEU** (Papineni et al., 2002) measures the n-gram (n=2,3,4) overlaps between generated and reference responses; **Distinct (Dist)** (Li et al., 2016) measures the proportion of unique n-gram in the generated response (n=1,2); **Entropy (Ent)** (Zhang et al., 2018b) measures how even the empirical n-gram (n=1,2) distribution is.

4.4 Results

As Table 2 shows, our model outperforms all the baseline on all metrics. We can further observe that: 1). Incorporating image contexts in dialogue models generally helps to produce better responses. Specifically, our model’s maximum relative improvement compared to the text-only baseline Seq2Seq is 24.97% (BLUE-4), and the Dist-1 measure achieves a relative improvement of 7.22%. This validates our motivation to incorporate multi-modal features in the dialogue generation model and proves that MMCHAT can be used to build image-grounded dialogue models. 2). Our model, Seq2Seq+IMG, can better capture sparse grounded image features in the response generation task. This indicates that explicitly modeling the sparsity phenomenon helps to further improve the dialogue generation performance, and MMCHAT facilitates the study of such a phenomenon. Table 3 lists a number of example outputs generated by our dialogue generation models and baselines.

5 Conclusion

We introduce MMCHAT, a large scale multi-modal dialogue corpus that reveals the image-sparsity phenomenon in real conversations. Our dataset

contains 120.84K dialogue sessions filtered from 56.9M sessions of raw multi-modal dialogues. A dialogue model is proposed to tackle the image-sparsity issue utilizing MMCHAT. Experiment results indicate that MMCHAT helps to develop image-grounded dialogue systems and facilitates further study of the image-sparsity issue. Besides the filtered dialogues in MMCHAT, we will also release all the raw dialogues obtained in the data collection process to facilitate further studies.

Broader Considerations

Our dataset MMCHAT originate from the social media. The dataset collection and release protocols are carefully designed to avoid violating the privacy of Weibo users. Specifically, each user’s permission setting on Weibo is strictly respected so that only contents that are publicly visible are collected. Rules are designed to filter out dialogues that may potentially expose user’s private information, such as phone numbers or emails. Moreover, we will not host the images and dialogues in MMCHAT on our own server. Only the URLs to these contents will be released along with the download scripts. This means that if any Weibo user want to remove their dialogues or images from MMCHAT, they can just set the corresponding weibo post’s visibility to a higher privacy level.

To further enforce the data privacy, MMCHAT is released under strict terms for academic uses only. Users are required to sign a *Letter of Guarantee*, in which they promise no abuse of MMCHAT besides academic purposes.

In addition to the privacy issues, there might also toxic or biased texts exist in MMCHAT or be generated by MMDs trained on MMCHAT. On the one hand, although we take the responsible to remove toxic texts (using a offensive word list and a offensive content classifier), we cannot guarantee that there is no offensive contents left. As offensive and abusive content recognition is a rapid developing area (Vidgen et al., 2019), we would deploy more advanced filters once the new state-of-the-art offensive and abusive classifiers being proposed in future. Regarding the potential biases, excepting those from the dataset itself (Henderson et al., 2018) (which always exists in dialogue datasets), biases might be introduced by the pre-trained language model (Bender et al., 2021) and the pre-trained image encoder (Steed and Caliskan, 2020) used in this work. In future, we head to

apply and develop corresponding mitigation techniques (following works such as Dinan et al. (2020) and Liu et al. (2020)).

Besides, we also note that the goal of our work is to facilitate further works on multi-modal dialogue systems. Although the model used in this work is still far from realistic, our dataset can be regarded as an initial step towards the sparsity issue in real world conversations.

References

- Huda AlAmri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. 2019. [Audio visual scene-aware dialog](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7558–7567. Computer Vision Foundation / IEEE.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. [Bottom-up and top-down attention for image captioning and visual question answering](#). In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big. *Proceedings of FAccT*.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. [Visual dialog](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1080–1089. IEEE Computer Society.
- Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. [Queens are powerful too: Mitigating gender bias in dialogue generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.
- Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129.
- Danica Jovanovic and Theo Van Leeuwen. 2018. Multimodal dialogue on social media. *Social Semiotics*, 28(5):683–699.

- Da Ju, Kurt Shuster, Y-Lan Boureau, and Jason Weston. 2019. All-in-one image-grounded conversational agents. *arXiv preprint arXiv:1912.12394*.
- Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. **CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog**. 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 582–595, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. **A diversity-promoting objective function for neural conversation models**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 801–809.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zita Liu, and Jiliang Tang. 2020. **Mitigating gender bias for neural dialogue generation with adversarial learning**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903, Online. Association for Computational Linguistics.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. **Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks**. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.
- Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. **Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478, Melbourne, Australia. Association for Computational Linguistics.
- Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. Openvidual: A large-scale open-domain dialogue dataset with visual contexts. *arXiv preprint arXiv:2012.15015*.
- Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2019. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *arXiv preprint arXiv:1907.09358*.
- Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. **Image-grounded conversations: Multimodal context for natural question and response generation**. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Ramakanth Pasunuru and Mohit Bansal. 2018. **Game-based video-context dialogue**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 125–136, Brussels, Belgium. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. **Faster R-CNN: towards real-time object detection with region proposal networks**. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.
- Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020a. **Image-chat: Engaging grounded conversations**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online. Association for Computational Linguistics.
- Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020b. **The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2453–2470, Online. Association for Computational Linguistics.

- Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2020c. Multi-modal open-domain dialogue. *arXiv preprint arXiv:2010.01082*.
- Ryan Steed and Aylin Caliskan. 2020. Image representations learned with unsupervised pre-training contain human-like biases. *arXiv preprint arXiv:2010.15052*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. [Challenges and frontiers in abusive content detection](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.
- Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. [A large-scale chinese short-text conversation dataset](#). In *Natural Language Processing and Chinese Computing*.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.
- Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. [Generating informative and diverse conversational responses via adversarial information maximization](#). In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1815–1825.
- Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.
- Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *AAAI*, pages 9693–9700.