ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Enhancing brain decoding using attention augmented deep neural networks

Ismail Alaoui Abdellaoui, Jesús García Fernández,
Caner Sahinli and Siamak Mehrkanoon

Department of Data Science and Knowledge Engineering
Maastricht University, The Netherlands

Abstract. Neuroimaging techniques have shown to be valuable when studying brain activity. This paper uses Magnetoencephalography (MEG) data, provided by the Human Connectome Project (HCP), and different deep learning models to perform brain decoding. Specifically, we investigate to which extent one can infer the task performed by a subject based on its MEG data. In order to capture the most relevant features of the signals, self and global attention are incorporated into our models. The obtained results show that the inclusion of attention improves the performance and generalization of the models across subjects.

1 Introduction

Magnetoencephalography (MEG) is a non-invasive technique for investigating neuronal activity. It allows neuroscientists to study properties of the working human brain. Recent interest in machine learning techniques has led to the development of data-driven models to learn its underlying patterns. They have been successfully employed to reveal neurological patterns for many diseases and disorders [4, 2]. While there has been extensive research into human activity classification using EEG data [5, 12, 8, 7], there are relatively fewer studies performed for MEG data. Compared to EEG data, MEG signals are more expensive and complex to obtain, yet, they are more precise due to their higher spatio-temporal resolution. The authors in [11] predicted the direction of subjects' hand-movement using Regularized Linear Discriminant Analysis method and EEG/MEG signals. Friston et al. [3] studied how to integrate multiple modalities, restrictions and subjects to achieve higher reproducibility of cortical response across subjects, using MEG, EEG and fMRI data. Zhang et al. [12] explored intention recognition based on EEG by combining recurrent and convolutional components. Lawhern et al. [5] propose an EGG-based architecture that can extract interpretable features from neurophysiological phenomena. In this paper, we augment two deep learning architectures by incorporating attention mechanisms for the purpose of brain decoding using MEG data.

2 Preliminaries

Attention mechanisms allow the models to capture long-range dependencies, and highlight/suppress relevant/irrelevant parts of the input. The models used in this paper are equipped with two types of attention: self and global.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

2.1 Convolutional Multi Head Self-Attention

This mechanism is introduced in [1], and combines convolutions and multi-head self-attention [10]. The *key* (K), *value* (V) and *query* (Q) are extracted from the input (X), and the output of the h^{th} head, i.e. O_h, is calculated as follows:

$$\mathrm{O}_h = \mathrm{Softmax}(\frac{QK^T}{\sqrt{d_k^h}})V, \quad Q = XW_q, \quad K = XW_k, \quad \text{and} \quad V = XW_v. \tag{1}$$

Here d_k^h is the dimension of the queries per attention head. The outputs of the heads are combined and concatenated with a convolution as follows:

$$\mathrm{Heads(X)} = \mathrm{Concat}[\mathrm{O}_1, ..., \mathrm{O}_n]W_o, \quad \mathrm{Output} = \mathrm{Concat}[\mathrm{Conv}(X), \mathrm{Heads}(X)], \tag{2}$$

where W_q, W_k, W_v and W_o are linear transformations learned during training.

2.2 Global Attention

The used Luong's style of global attention is introduced in [6]. Given all source hidden states $\bar{h}_s$ (input attention layer) and the target hidden state h_t (last element input), one obtains the score between the two entities using $score(\bar{h}_s, h_t)$= $h_t^T W_a \bar{h}_s$. An attention score vector a_t is obtained through the dot product between h_t and $score(\bar{h}_s, h_t)$, followed by a softmax function. A context vector c_t is derived as the weighted sum between a_t and $\bar{h}_s$. Then it calculates the output $\tilde{h}$ as follows: $\tilde{h} = tanh(W_c[c_t : h_t]$, where W_a and W_c are learnable parameters.

3 Proposed Models

Here, we use previously studied architectures used to analyze EEG signals as core models. These core models mainly rely on convolutions, which operate in a local scope and thus cannot explicitly capture global dependencies. To address this locality issue, we extend these models by equipping them with different attention mechanisms. Furthermore, the models are adapted to suit MEG data analysis. In the subsequent sections the applied adaptation will be discussed in detail. The models and data used in our experiments can be found on Github by IsmailAlaouiAbdellaoui here.

3.1 Attention Augmented EEGNet (AA-EEGNet)

The model proposed in [5] serves as the basis for our first architecture. AA-EEGNet is a CNN mainly composed of three convolutional layers. First, a $1 \times K$ convolution acts as a band-pass filter, separating ranges of frequency and extracting temporal features from them. Next, a $C \times 1$ depth-wise convolution extracts spatial features for each frequency range. These two convolutions act in the same way as the Filter Bank Common Spatial Pattern (FBCSP). Lastly, the model includes a separable convolution, composed of a $1 \times D$ depth-wise convolution, which summarizes features, followed by a 1×1 point-wise convolution, which combines features. After this sequence, a softmax classifier classifies the

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

extracted features. We augment the first convolutional layer with multi-head self-attention and add global attention before the softmax classifier to enhance its performance. Contrary to [5], we use 16, 2 and 32 filters for the three convolutional layers, respectively, as well as two attention heads. K and D are set to 128 and 16, respectively. C is set to the number of channels in the data. This was empirically found to be the optimal configuration.

3.2 Attention Augmented Cascade Network (AA-CascadeNet)

The architecture in [12] is used as the backbone of our model. Motivated by the approach in [12], we adopt a 2D representation of the input to take into account the spatial information given by the different sensors. In particular, we convert the 1D MEG recordings into a 2D mesh in which the sensors' locations are captured. This mesh $M_t \in \mathbb{R}^{N \times L}$ depicts a top-down view of the human scalp at each time step t, as shown below. A similar input mesh representation has been used in [12].

$M_t =$

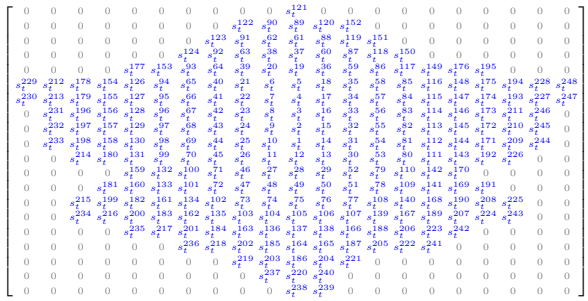

Fig. 1: Mesh input representation of the AA-CascadeNet.

Here, each non-zero element s_t^j corresponds to a specific sensor value at time t. We concatenate multiple meshes along the time axis to create a tensor $T \in \mathbb{R}^{N \times L \times D}$, where D is the number of meshes M_t. Due to the multi-input nature of our model, we use multiple tensors P, resulting in an input $I \in \mathbb{R}^{P \times N \times L \times D}$. AA-CascadeNet is a multi-input model, where every input goes through three convolutional and two LSTM layers before being merged in a cascade fashion, as described in [12]. While the convolutions extract spatial features, the LSTM's extract temporal features. Lastly, a softmax classifier classifies the extracted features. We augment the first convolutional layer in each input with multi-head self-attention and add global attention between the two LSTM layers in each input. Contrary to [12], we use 1, 2 and 4 filters for the three convolutional layers, respectively, as well as two attention heads and 125 fully connected units. Further, the optimal number of meshes in the input (D) found is 100. This was empirically found to be the optimal configuration.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

4 Data Description

MEG data comes from the database of the Human Connectome Project (HCP)[9]. Specifically, we use the *1200 Subjects Release (S1200)* dataset. It counts with 248 magnetometer channels, recording at a sampling rate of 2034.51 Hz. We picked the subjects that have MEG data, and filtered them to 18, discarding those with a significantly small amount of data. The subjects are in 4 different states during the recording, requiring activity from separated brain regions: (i) Resting state, (ii) Story vs. Math state, (iii) Working Memory state and (iv) Motor state. For more details on the data, please refer to the official documentation.

5 Experiments

5.1 Data Preprocessing

Since the order of magnitude of MEG data is considerably small, we apply normalization for A-CascadeNet and scaling of $\times 10^5$ for the AA-EEGNet. For AA-EEGNet, every recording is segmented into smaller segments and then classified individually. We use segments of 0.7s with 33% overlapping between them. For AA-CascadeNet, every mesh described in section 3.2 represents one time-step. Also, a 50% overlapping between the number of inputs, i.e. P (see section 3.2), across data samples is used.

5.2 Experimental setup

In an intra-subject setup, we considered different MEG recordings of individual subjects for training and test, achieving perfect intra-subject classification. In the inter-subject setup, we selected a group of 12 and 6 subjects for training and test, respectively. In the latter case, the same number of recordings per subject were used. Specifically, the duration of every recording per subject is as follows: • **Resting state**: 3 runs of 6 minutes each. • **Story vs. Math state**: 2 runs of 7 minutes each. • **Working Memory state**: 2 runs of 10 minutes each. • **Motor state**: 2 runs of 14 minutes each. Our objective is to perform a multi-class classification of these 4 tasks. We use categorical cross-entropy as a loss function, Adam to optimize it and a learning rate $1e^{-4}$. When training AA-CascadeNet, we use a batch size of 64, and when training AA-EEGNet, 16.

6 Results and discussion

The obtained accuracy for the models on the test set are tabulated in Table 1.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Table 1: Cross subject classification accuracy of all the models.

Model	No Attention	Self Att.	Self & Global Att.
AA-EEGNet	0.83 ±0.15	0.90 ± 0.08	0.90 ± 0.08
AA-CascadeNet	0.91 ± 0.07	0.92 ± 0.08	0.93 ± 0.06

From the presented results, we can notice that the incorporation of attention enhances the performance of both models. When using the base models (no attention), AA-CascadeNet is significantly superior to AA-EEGNet. Due to its LSTM operations, AA-CascadeNet can capture long-range dependencies more effectively. However, when attention is included, their performances are comparable. This is thanks to the attention mechanisms that are especially beneficial to address the locality issue in CNN's. The convolutional filters in the AA-EEGNet and feature maps from the convolutional layers in the AA-CascadeNet are shown in Fig. 2 (a) and (b) respectively. While the visualizations at the top of the figure correspond with the base networks (no attention), the ones at the bottom correspond with the attention augmented networks. In Fig. 2 (a), one can

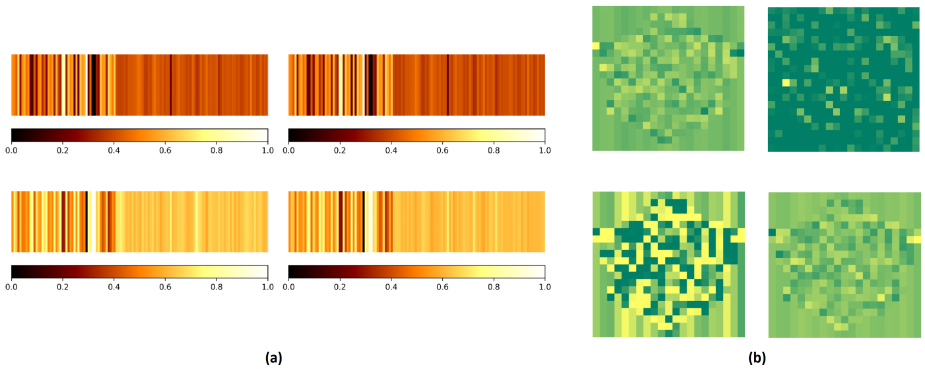

Fig. 2: The four images at the top are generated with the base networks and the four at the bottom are generated with the models incorporating attention. (a) Conv. Filters AA-EEGNet. (b) Feature Maps AA-CascadeNet: 1^{st} Conv. layer at the left and 3^{th} Conv. layer at the right.

observe that the model without attention weights more intensely a few specific time-steps (lighter areas). In contrast, the model that incorporates attention weights the input time-steps in a completely different manner. In the same way, we can notice a difference in the AA-CascadeNet's feature maps. In the first column of Fig.2 (b) (feature maps of 1^{st} Conv layer), some input channels are intensely weighted when attention is included. Thus the network identifies the most relevant channels for the task. Moreover, in the second column of Fig.2 (b) (feature maps of 3^{th} Conv layer), we can see how the mesh representation is lost as the data goes through several convolutional layers. Nevertheless, the mesh representation is kept when attention is included in the model.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

7 Conclusion

In this paper, different architectures for analyzing EEG brain signals are adapted and enhanced with various attention mechanisms to perform MEG-based brain decoding. Here, we found that integrating attention in the models leads to enhanced performance. In particular, models that solely rely on convolutional operations can benefit greatly from using this mechanism. Moreover, that attention mechanisms have shown to be a convincing technique to achieve a more efficient extraction of features without the need of having domain knowledge. Finally, we also show a potentially good model transferability between EEG decoding and MEG decoding, even though the nature of the data is different.

References

[1] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le. Attention augmented convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3286–3295, 2019.

[2] C. Davatzikos, D. Shen, R. C. Gur, X. Wu, D. Liu, Y. Fan, P. Hughett, B. Turetsky, and R. Gur. Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. *Archives of general psychiatry*, 62(11):1218–1227, 2005.

[3] R. N. Henson, D. G. Wakeman, V. Litvak, and K. J. Friston. A parametric empirical bayesian framework for the eeg/meg inverse problem: generative models for multi-subject and multi-modal integration. *Frontiers in human neuroscience*, 5:76, 2011.

[4] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack Jr, J. Ashburner, and R. S. Frackowiak. Automatic classification of mr scans in alzheimer's disease. *Brain*, 131(3):681–689, 2008.

[5] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance. Eegnet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[6] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[7] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 2019.

[8] R. Schirrmeister, J. Springenberg, L. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball. Deep learning with convolutional neural networks for brain mapping and decoding of movement-related information from the human eeg. arxiv, 2017. *arXiv preprint arXiv:1703.05051*, 2017.

[9] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

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

[11] S. Waldert, H. Preissl, E. Demandt, C. Braun, N. Birbaumer, A. Aertsen, and C. Mehring. Hand movement direction decoded from MEG and EEG. *Journal of neuroscience*, 28(4):1000–1008, 2008.

[12] D. Zhang, L. Yao, X. Zhang, S. Wang, W. Chen, R. Boots, and B. Benatallah. Cascade and parallel convolutional recurrent neural networks on EEG-based intention recognition for brain computer interface. In *Thirty-Second AAAI Conference on AI*, 2018.