EEG decoding with Transformer for Multi-Modal Natural Language Processing

Guoqing Luo
Department of Computer Science
University of Alberta
gluo@ualberta.ca

Henry Vu
Department of Computer Science
University of Alberta
ddvu@ualberta.ca

1 Introduction

Electroencephalogram (EEG) data, which has rich and multi-dimensional information about brain activities, provides insights into cognitive processes and potential brain-computer interfaces [19]. In recent years, there have been increasing efforts to apply the Transformer architecture [23] to EEG decoding. In particular, the Transformer models have shown good results in decoding EEG data due to their capabilities of capturing complex temporal information within EEG sequences [22, 28, 11, 6].

The Transformer model is originally used in multiple natural language processing (NLP) tasks and gains attention from people all over the world. However, the Transformer models gradually saturate the performance on many NLP tasks, such as machine translation [21] and text summarization [20]. Bisk et al. [2] suggests that relying solely on text for training may reach the point of decreasing returns, and the next step for developing NLP involves utilizing multi-modal information, such as images and audio. In addition, Linzen [12] suggests grounding NLP models in multi-modal settings to compare the generalization abilities of models to human language learning. Following this trend, Hollenstein et al. [8] leverage EEG signals in the semantic language understanding task, showing the potential of multi-modal learning for NLP tasks. However, [8] only use LSTM or CNN models to perform EEG decoding, resulting in the extraction of limited information in EEG data.

To this end, we propose to use the Transformer model for EEG decoding in a multi-modal setting for NLP tasks, including sentiment classification and relation classification. We intend to leverage the ability of domain generalization [26] in Transformer models to not only perform feature extraction but also generate desired answers for a given task. [1]

2 Related Work

Transformer. The Transformer [23] is a prevailing neural architecture that is initially designed as a sequence-to-sequence model with an encoder and a decoder. In particular, the encoder encodes the input sequence into a list of continuous embeddings that incorporate information about the entire input; the decoder then decodes these embeddings for the next-token prediction.

Leveraging the Transformer architecture, researchers have developed three categories pretrained language models (PLMs): 1) encoder-only models such as BERT [3] and RoBERTa [13]; 2) decoder-only models such as GPT-2 [15]; 3) encoder–decoder models such as T5 [16]. Each category of PLMs have distinct architectures and specific application scenarios, highlighting the adaptability of the Transformer architecture.

EEG decoding. Traditionally, researchers have used Convolutional Neural Networks (CNN) for EEG decoding tasks [18, 1, 10, 25]. However, CNNs with large kernels might miss fine-grained information of EEG signals, while CNNs with small kernels have limited receptive fields and are restricted in capturing temporal information inherent within EEG sequences [22].

[1] Our code is in https://github.com/frankdarkluo/eeg-decoding

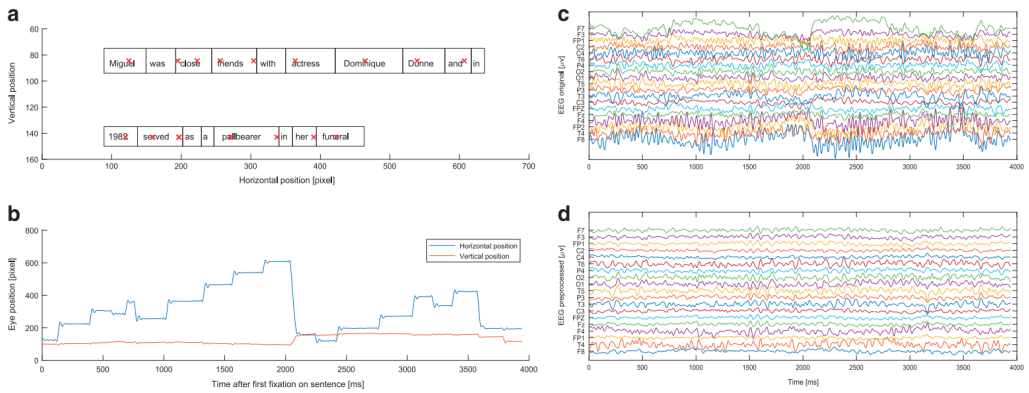


Figure 1: **Visualization of single-trial EEG and eye-tracking data**. **(a)** Single sentence fixation data for a representative subject. Red crosses indicate fixations. Boxes around the words indicate the area in which fixations are allocated to the specific word. **(b)** Raw gaze data of the fixation data plotted above. **(c)** Subset of the raw EEG data during the sentence. **(d)** Same data as in **(c)** after preprocessing.

The Transformer model [23], which utilizes the self-attention mechanism, has shown remarkable capabilities in handling sequential data, offering potential enhancements for EEG decoding [28, 9, 22]. These models have shown exceptional performance in various applications, such as natural language processing and computer vision, and have started to gain traction in the field of bio-signal processing. Yang and Modesitt [28] used CNNs to extract features of EEG signals, then added positional embedding to provide the input for further fine-tuning of a Vision Transformer model Dosovitskiy et al. [4]. This resulted in a notable decrease in the Root Mean Square Error (RMSE) of an EEG regression task. Song et al. [22] presented an architecture for EEG decoding that first increases the spatial difference of EEG signals, then exploits the attention mechanism to compute a weighted representation of the spatial features of EEG signals. However, these studies mainly focus on uni-modal applications of Transformer models, while delving into multi-modal integrations in EEG decoding may offer more robust and comprehensive outcomes.

Recently, researchers have worked on integrating the deep learning model with EEG decoding in a multi-modal setting [8, 24]. [7] combined brain recording data and BiLSTM or CNN models trained on language tasks and found that concatenating EEG features with word embeddings can improve classification accuracy. Wang et al. [24] fuse multi-modal data with cross-attention mechanism and train a multi-modal Transformer for human state recognition. Following the multi-modal setting, we use the Transformer model for EEG decoding, which perform EEG feature extraction and word representation generation, and concatenate the representation together for multiple classification tasks.

3 Dataset

In response to the instructor's feedback and to address the limitations we identified in our proposal, we have decided to use an alternative dataset. This new dataset ZuCo [5] consists of EEG and eye-tracking data collected from subjects performing a reading task, as opposed to the visual task originally proposed. There are 2 versions of ZuCo datasets, ZuCo 1.0[2] and ZuCo 2.0 [3].

3.1 Data Acquisition and Preprocessing

To collect the dataset, 12 healthy adult native English speakers (7 male and 5 female right-handed participants aged between 22 and 54) were asked to take part in a series of reading tasks while having electrodes attached to their scalp to record EEG data. Data about eye position and pupil size were recorded with an infrared video-based eye tracker during all the EEG paradigms.

[2]https://osf.io/q3zws/
[3]https://osf.io/2urht/

The EEG data in ZuCo was preprocessed using Automagic [14]. One hundred and five EEG channels were used for scalp recordings (Hence the dimension of a sample of word-level EEG features is 105), and nine EOG channels were used for artifact removal. As for eye-tracking data, default system parameters of the EyeLink1000 tracker programmed by SR-research (https://www.sr-research.com/) were used to identify saccades, fixations and blinks. For the analysis, only fixations that were longer than 100 ms and stayed within the boundaries of each displayed word have been extracted. EEG signals and eye-tracking data were then synchronized using the EYE EEG extension [27].

A visualization of single-trial EEG and eye-tracking data is shown in Figure 1.

3.2 Experimental Paradigms

Three different experimental paradigms were used to compile the ZuCo dataset: Normal Reading-Sentiment (SR), Normal Reading-Wikipedia (NR) and Task-Specific Reading-Wikipedia (TSR)

- **Normal Reading-Sentiment:** Participants were asked to read 400 positive, negative and neutral sentences (from the *Stanford Sentiment Treebank* movie review data) to analyze emotional responses during reading. The control condition for this task requires the participants to rate the quality of 47 of all described movies. An example sentence for task 1 is: "It's the best film of the year so far, the benchmark against which all other Best Pictures contenders should be measured.".
- **Normal Reading-Wikipedia:** Participants were presented with sentences that contained semantic relations. As a control condition, participants had to answer 68 multiple-choice questions about the content of the previous sentence. An example sentence is "As a child, his hero was Batman, and as a teenager his interests shifted towards music.". For this sentence, a question could be "Who was his childhood hero?".
- **Task-Specific Reading-Wikipedia:** Participants were presented with the same sentences in Task 2 but while reading, they focused on a specific type of relation (e.g. education, employer). As a control condition, participants had to report whether the sentence contained a specific relation. An example sentence for Task 3 is "He won a Nobel Prize in Chemistry in 1928" and the participant would be asked whether the "award" relation was mentioned.

4 Methodology

4.1 Framework

Our framework is shown in Figure 2. To decode EEG signals into embeddings, we use the transformer model as an EEG Decoder. We also use a Transformer model to encode the corresponding word stimuli into word embeddings. Then, we concatenate the two embeddings and feed it into a transformer decoder, which predicts a desired textual answer for the multi-label classification task.

We plan to fully leverage the capabilities of Transformers to understand the complex pattern of EEG signals and then generate the embedding with useful information. Detailedly, we plan to use several language models, such as RoBERTa-base and GPT-2, to encode the EEG signals. Then, we concatenate the two embeddings and feed the conncatenated embedding into a fully connected layer and finally a softmax layer for classification tasks.

To evaluate the performance of our models and make comparisons between them, since all tasks that we perform are classification-based (sentiment analysis and relation classification), we follow Hollenstein et al. [6] and report the F_1 score.

5 Experiments

5.1 Implementation details

To assess the impact of the EEG signals, the hyperparameters are tuned individually for all baseline models as well as for EEG augmented models. The hyper-parameters are presented in Table 1. All results are reported as means over three independent runs with different random seeds. Due to the limited amount of data, we perform three-fold cross validation on a 80% training and 20% test split. The best hyperparameters were selected according to the accuracy of the model on

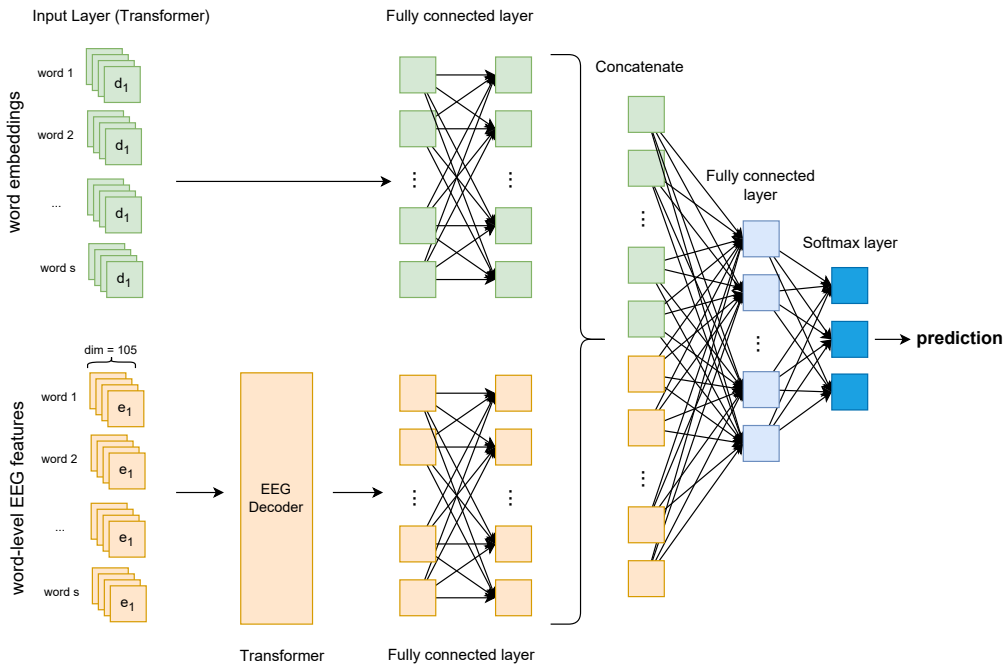


Figure 2: Our proposed framework. We use the Transformer model to perform feature extraction on the EEG signals into an embedding and use the same Transformer model to encode the sentences into word embeddings.

Hyperparameter	Range
LSTM layer dimension	256
Number of LSTM layers	1
CNN filters	14
CNN kernel sizes	[1,4,7]
CNN pool sizes	3, 5, 7
Dense layer dimension	128
Dropout	0.3
Batch size	60
Learning rate	10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}
Random seeds	13, 22, 42
Threshold	0.5

Table 1: Hyperparameters for the model. *Threshold only applies to relation classification.*

the validation set (10% of the training set) across all three-folds. We also applied an early stopping mechanism with a patience of 30 epochs and a minimum difference in validation accuracy of 10^{-7}. The validation set is used for both hyperparameter tuning and early stopping.

5.2 Preliminary Results

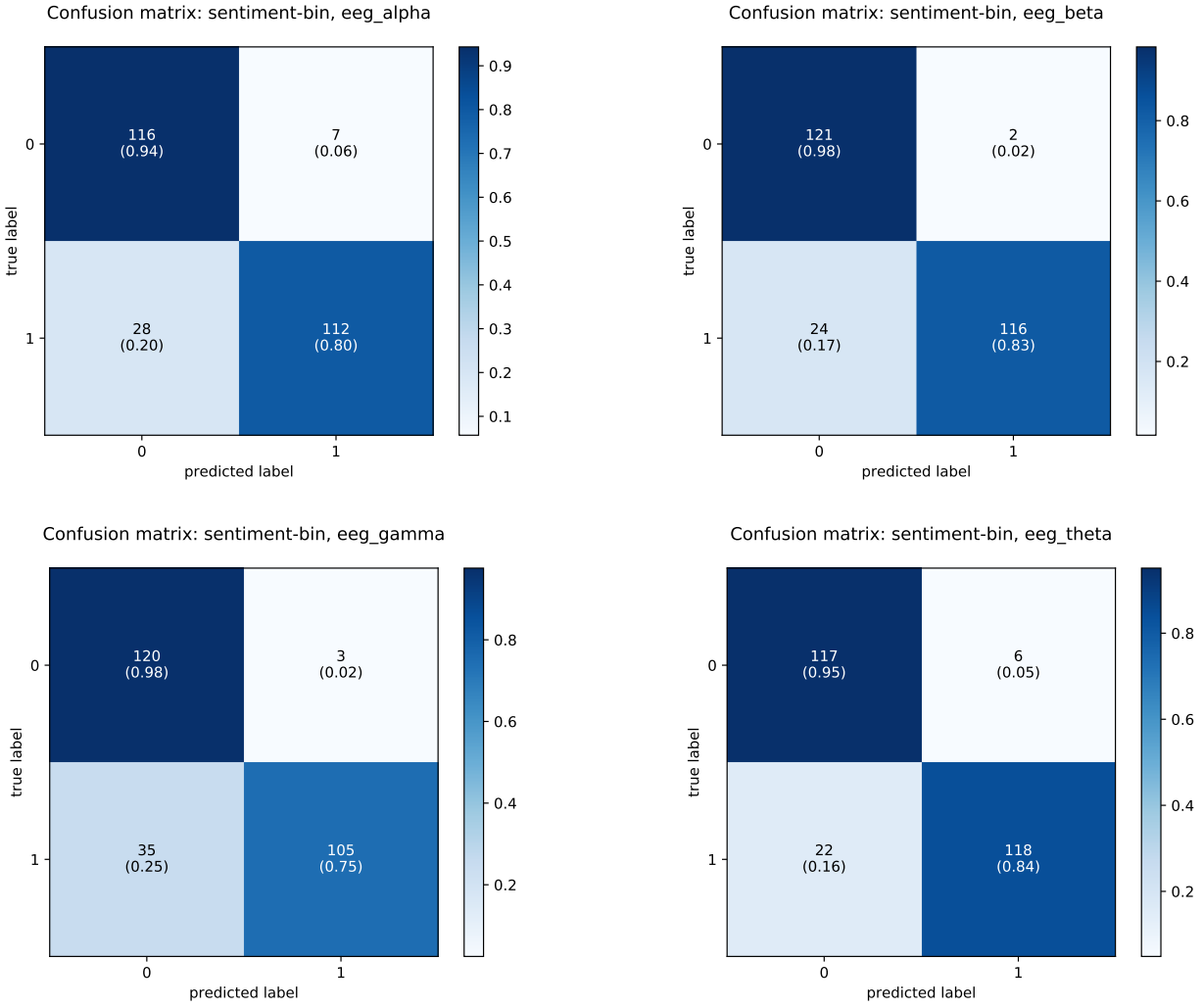
We used the codebase from Hollenstein et al. [6] and replicated the experiments by preprocessing the EEG and text data on the binary sentiment classification task. Additionally, we use a stronger language model RoBERTa to encode the words. The F1-score results are reported in Table 2. We observe that different variants of EEG data are all contributive to the sentiment classification task, showing the potential of decoding EEG for NLP. We also observe that the addition of EEG β contributes to most of the word embeddings, including randomly initialized embeddings, GloVe,

Model	F1-score			
	Random Initialized	GloVe	BERT	RoBERTa
Baseline	0.574	0.68	0.89	0.907
+EEG full	0.574	0.71	**0.917**	0.924
+EEG θ	0.575	0.77	0.900	0.921
+EEG α	0.588	0.76	0.913	0.920
+EEG β	**0.602**	**0.77**	0.897	**0.935**
+EEG γ	0.590	0.75	0.917	0.930
+$\theta + \alpha + \beta + \gamma$	0.580	0.71	0.893	0.903

Table 2: For the binary sentiment classification task, we report the F_1-score averaged on three runs. The best results per column are marked in bold.

and RoBERTa embeddings. The good performance of the EEG β features may be explained by the emotional connotation of words on the β response [17].

In addition, we also present visualization of confusion matrices of EEG α, β, γ, and θ variants respectively. We observe that these four variants achieve all has a false positive problem where around 20% of the sentences with positive labels are predicted to be negative.



6 Originally Proposed Work Plan

6.1 Timeline

- *September 22 - September 29*: Explore the codebase for similar projects that use EEGEyeNet and further preprocess the EEGEyeNet dataset to fit our application (Completed - Henry).
- *September 30 - October 5*: Implement our framework with the Transformer model to do feature extractions for the eye-tracking and EEG data (Completed - Guoqing).

7 Current Work Plan

7.1 Timeline - Completed

- *October 7 (After feedback) - October 10*: Explore a different dataset where EEG data is for NLP tasks.
- *October 10 - October 15*: Read related research work done using the ZuCo dataset ([8] in our case) and look into the codebase.
- *October 15 - October 25*: Write code for data preprocessing and perform feature extraction for EEG signals; Refactor the codebase for running experiments to replicate some baseline results for the binary sentiment classification task.
- *October 26 - October 27*: Write code for using RoBERTa for the binary sentiment classification task.

7.2 Current Contributions

Guoqing Luo:

- Literature review and use a different dataset after feedback.
- Model architecture implementation.
- Conduct experiments on the classification tasks with the RoBERTa model as the new results.

Henry Vu:

- Build up EEG feature preprocessing pipeline
- Conduct experiments on the classification tasks to get the baseline results
- Write code for getting confusion matrices of classification tasks.

7.3 Timeline - Todo

- *October 27 - October 31*: Obtain baseline results in terms of F_1 scores for all model variants (including our proposed ones) for the ternary sentiment classification and relation classification tasks.
- *November 1 - November 8*: Fully implement the complete proposed architecture in PyTorch. We plan to use three categories of Transformer models for EEG feature extraction: RoBERTa, GPT-2 and T5.
- *November 8 - November 19*: Conduct experiments to see if our proposed framework achieves better results. It is also possible that we need to incorporate a CNN component to perform a more effective EEG feature extraction.
- *November 20 - November 25*: Interpret the effects of different hyperparameters (e.g. depth of layer, # of training examples), and different categories of language models on the performance.
- *November 26 - November 30*: Complete the final report.

References

[1] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning representations from EEG with deep recurrent-convolutional neural networks. In *International Conference on Learning Representations*, 2016. URL `http://arxiv.org/abs/1511.06448`.

[2] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In *Proceedings of the Conference*

on Empirical Methods in Natural Language Processing, pages 8718–8735, 2020. URL https://aclanthology.org/2020.emnlp-main.703.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019. URL https://aclanthology.org/N19-1423.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=YicbFdNTTy.

[5] Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13, 2018. URL https://www.nature.com/articles/sdata2018291.

[6] Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:1912.00903*, 2019. URL https://arxiv.org/abs/1912.00903.

[7] Nora Hollenstein, Cedric Renggli, Benjamin Glaus, Maria Barrett, Marius Troendle, Nicolas Langer, and Ce Zhang. Decoding eeg brain activity for multi-modal natural language processing. *Frontiers in Human Neuroscience*, page 378, 2021.

[8] Nora Hollenstein, Cedric Renggli, Benjamin Glaus, Maria Barrett, Marius Troendle, Nicolas Langer, and Ce Zhang. Decoding eeg brain activity for multi-modal natural language processing. *Frontiers in Human Neuroscience*, page 378, 2021. URL https://www.frontiersin.org/articles/10.3389/fnhum.2021.659410/full.

[9] Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15, 2021. ISSN 1662-5161. URL https://www.frontiersin.org/articles/10.3389/fnhum.2021.653659.

[10] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. URL https://dx.doi.org/10.1088/1741-2552/aace8c.

[11] Young-Eun Lee and Seo-Hyun Lee. Eeg-Transformer: Self-attention from transformer architecture for decoding EEG of imagined speech. In *International Winter Conference on Brain-Computer Interface*, pages 1–4, 2022. URL https://ieeexplore.ieee.org/abstract/document/9735124.

[12] Tal Linzen. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, 2020. URL https://aclanthology.org/2020.acl-main.465.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

[14] Andreas Pedroni, Amirreza Bahreini, and Nicolas Langer. Automagic: Standardized preprocessing of big EEG data. *NeuroImage*, 200:460–473, 2019. URL https://www.sciencedirect.com/science/article/abs/pii/S1053811919305439.

[15] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[16] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified

text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[17] Michele Scaltritti, Caterina Suitner, and Francesca Peressotti. Language and motor processing in reading and typing: Insights from beta-frequency band power modulations. *Brain and Language*, 204:104758, 2020. URL https://www.sciencedirect.com/science/article/pii/S0093934X20300171.

[18] R Schirrmeister, L Gemein, K Eggensperger, F Hutter, and T Ball. Deep learning with convolutional neural networks for decoding and visualization of EEG pathology. In *IEEE Signal Processing in Medicine and Biology Symposium*, pages 1–7, 2017. URL http://ieeexplore.ieee.org/document/8257015/.

[19] Donald L. Schomer and Fernando H. Lopes da Silva. *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Oxford University Press, 2017. ISBN 9780190228484. URL https://doi.org/10.1093/med/9780190228484.001.0001.

[20] Raphael Schumann, Lili Mou, Yao Lu, Olga Vechtomova, and Katja Markert. Discrete optimization for unsupervised sentence summarization with word-level extraction. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pages 5032–5042, 2020. URL https://aclanthology.org/2020.acl-main.452.

[21] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, 2016. URL https://aclanthology.org/P16-1162.

[22] Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for EEG decoding. *arXiv preprint arXiv:2106.11170*, 2021. URL https://arxiv.org/abs/2106.11170.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[24] Ruiqi Wang, Wonse Jo, Dezhong Zhao, Weizheng Wang, Baijian Yang, Guohua Chen, and Byung-Cheol Min. Husformer: A multi-modal transformer for multi-modal human state recognition. *arXiv preprint arXiv:2209.15182*, 2022. URL https://arxiv.org/abs/2209.15182.

[25] Nicholas Waytowich, Vernon J Lawhern, Javier O Garcia, Jennifer Cummings, Josef Faller, Paul Sajda, and Jean M Vettel. Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials. *Journal of Neural Engineering*, 15(6):066031, 2018. URL https://dx.doi.org/10.1088/1741-2552/aae5d8.

[26] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=yzkSU5zdwD.

[27] Irene Winkler, Stephanie Brandl, Franziska Horn, Eric Waldburger, Carsten Allefeld, and Michael Tangermann. Robust artifactual independent component classification for bci practitioners. *Journal of neural engineering*, 11(3):035013, 2014.

[28] Ruiqi Yang and Eric Modesitt. ViT2EEG: Leveraging hybrid pretrained vision transformers for EEG data. *arXiv preprint arXiv:2308.00454*, 2023. URL https://arxiv.org/abs/2308.00454.