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 neural activities, provides insights into cognitive processes and potential brain-computer interfaces [13]. Recent advancements have seen the successful application of the Transformer architecture [18] 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 [16, 23, 9, 5].

The Transformer model, with the use of the self-attention mechanism to capture global, long-range dependencies, has shown remarkable progress in the Natural Language Processing (NLP) domain. However, the model seems to have saturated on many NLP tasks, such as machine translation [15] and text summarization [14]. Bisk et al. [2] suggests that relying solely on texts may reach the point of decreasing returns, and the next step for developing NLP involves utilizing multi-modal information, such as images and audio. Following this multi-modal setting, Hollenstein et al. [6] leverages EEG signals in semantic language understanding tasks and demonstrates the potential of multi-modal learning for NLP tasks. However, [6] utilize only LSTM and CNN architectures in their work to perform EEG decoding, limiting the amount of information extracted from the 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 [21] in Transformer models to not only perform feature extraction but also generate desired answers for a given task. [1]

1.1 Related Work

Traditionally, researchers have used Convolutional Neural Networks (CNN) for EEG decoding tasks [12, 1, 8, 20]. 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 [16].

The Transformer model [18], which is a prevailing sequence-to-sequence architecture with the utilization of the self-attention mechanism, has shown remarkable capabilities in handling sequential data and offering potential enhancements for EEG decoding [23, 7, 16]. These models have shown exceptional performance in various applications, such as natural language processing and computer vision, and have started to gain attraction in the field of bio-signal processing. Yang and Modesitt [23] 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 [3]. This resulted in a notable performance gain of an EEG regression task. [16] 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

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

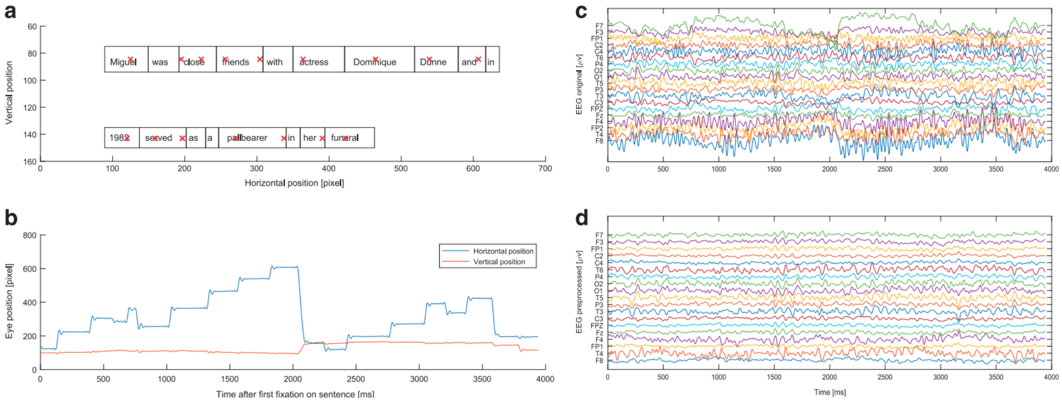

Figure 1: **Visualization of single-trial EEG and eye-tracking data**. **(a)** Single sentence fixation data for a representative subject. Red crosses indicate fixations. Boxes around the words indicate the area in which fixations are allocated to the specific word. **(b)** Raw gaze data of the fixation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

on uni-modal applications of Transformer models, while delving into multi-modal integration in EEG decoding may offer more robust and comprehensive outcomes.

Recently, researchers have worked on integrating deep learning models with EEG decoding in a multi-modal setting [6, 19]. [6] combine 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. In addition, Wang et al. [19] 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 performs EEG feature extraction and word representation generation, and concatenate the representation together for multiple classification tasks.

2 Dataset

We follow [6] and use the dataset ZuCo [4]. This dataset 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].

2.1 Data Acquisition and Preprocessing

To collect the dataset for ZuCo 1.0, 12 healthy adult native English speakers (18 for ZuCo 2.0) 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.

The EEG data in ZuCo was preprocessed using Automagic [10]. 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. EEG signals and eye-tracking data were then synchronized using the EYE EEG extension [22].

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

2.2 EEG Frequency Bands

Following Hollenstein et al. [6], we want to investigate the effect of using different EEG frequency bands by splitting the EEG data into 4 frequency bands: *theta* (4-8 Hz), *alpha* (8.5-13 Hz), *beta*

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

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

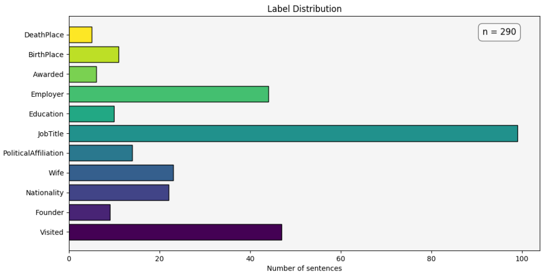

Figure 2: Distribution of labels in ZuCo 2.0

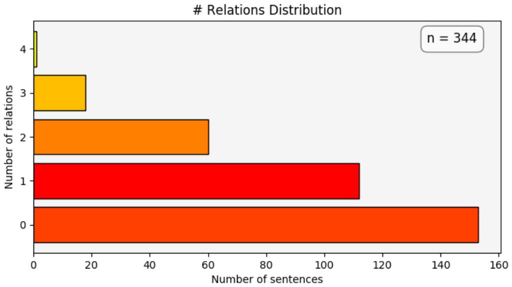

Figure 3: Distribution of # labels of each sentence.

	ZuCo 1.0	**ZuCo 2.0**
	Task 1: Normal Reading - Sentiment	Task 1: Normal Reading
Participants	12	18
Sentences	400	344

Table 1: Statistics for each ZuCo task.

(13.5-30 Hz) and *gamma* (30.5-49.5 Hz). Different egg frequency band has been shown to correspond to different cognitive aspects of language processing. The *theta* frequency band is responsible for memory encoding and retrieval, which are crucial in language comprehension. The *alpha* band is associated with attentional processes, aiding in tasks such as semantic retrieval and syntactic processing. The *beta* band is implicated in the motor aspects of speech and may also play a role in the predictive coding of language. The *gamma* band is linked to higher cognitive functions, including the integration of sensory information, which is vital in processing complex language structures.

2.3 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).

- **ZuCo 1.0 Task 1: Normal Reading - Sentiment**
  Participants were asked to read very positive, very negative and neutral movie reviews from the *Stanford Sentiment Treebank* to analyze emotional responses during reading. The control condition for this task requires the participants to rate the quality of some 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.".
- **ZuCo 1.0 Task 2 and ZuCo 2.0 Task 1: Normal Reading - Wikipedia**
  Participants were presented with sentences that contained semantic relations. As a control condition, participants had to answer 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?".

3 Methodology

3.1 Framework

Our framework is shown in Figure 4. Specifically, our EEG decoder consists of a convolutional module and a Transformer self-attention module for extracting both spatial and temporal information of EEG data, which is different from the convolutional and recurrent architecture.[4]

In addition, we use four algorithms to produce the word embedding, namely random initialization, GloVe, BERT and RoBERTa. We then use an EEG decoding component to decode the corresponding

[4] Please refer to Hollenstein et al. [6] for architecture comparison.

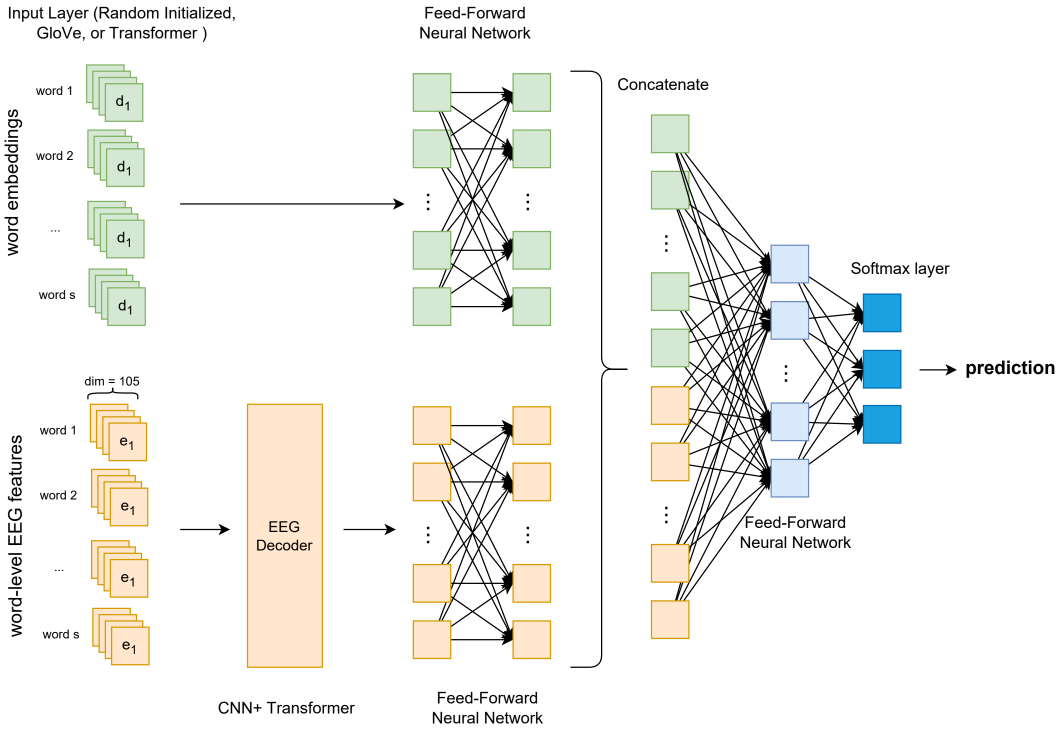

Figure 4: Our proposed framework. We propose to use the CNN+Transformer model to decode EEG signals into an embedding.

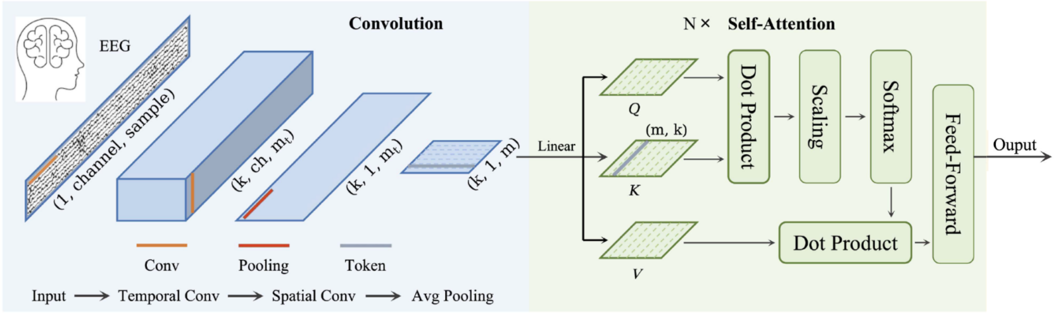

Figure 5: The Conformer architecture [16] that our proposed architecture is mainly based on.

EEG signals into embeddings and concatenate the two embeddings before passing the unified embedding into a fully connected layer and a softmax layer to produce the classification result.

3.2 Natural Language Processing Task

In this segment, we outline the natural language processing tasks utilized to evaluate the performance of our proposed framework. The tasks selected for this project are two commonly used NLP tasks, namely sentiment classification and relation detection. The two tasks receive similar input data which is tokenized sentences augmented with features from the EEG recordings.

Task 1: Sentiment Classification. The goal of sentiment classification is to identify affective states and subjective information from a given piece of text. By leveraging the sentences recorded in Task 1: Normal Reading - Sentiment of ZuCo 1.0, we run experiments on both binary (positive/negative) and ternary (positive/negative/neutral) sentiment classification.

Hyperparameter	Range
Batch size	60
Learning rate	10^{-5}
Random seeds	13, 22, 42, 66, 78
Threshold	0.5

Table 2: Hyperparameters used for training the model. Threshold only applies to relation detection.

Task 2: Relation Detection. For relation detection, the objective is to correctly identify the semantic relations within a sentence. We use the data recorded in Task 1: Normal Reading - Wikipedia of ZuCo 2.0, where each sentence could include any combination of 11 relation types shown in Figure 2. Therefore, relation detection is treated as a multi-class and multi-label classification task. The distribution of each label and the number of labels for each sentence is shown in Figure 2 and Figure 3.

3.3 Implementation Details

The hyperparameters are presented in Table 2. All results are averaged over five runs with different random seeds. Due to the limited amount of data, we perform five-fold cross-validation on an 80% training and 20% test split. The best hyperparameters were selected according to the accuracy of the model on the validation set (10% of the training set) across all five folds. We also applied an early stopping mechanism with 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.

3.4 Evaluation

Consistent with the conventional approach and evaluation metric in NLP research, we report the F1-score as our performance metric. The F1-score is a harmonic mean of precision and recall, balancing between the model's ability to identify positive instances correctly (precision) and its capability to find all positive instances (recall): $F_1 \text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$.

4 Results

We present the results of all augmented models compared to the baseline results in detail, which are reported separately in Table 3 and Table 4 for binary and ternary sentiment classification, as well as in Table 5 for relation detection. The first row of each table represents the text-only baseline. This is followed by the multi-modal models augmented with three types of EEG decoders on seven components (uniform random noise and six types of EEG signals). As for the embedding types, we implement a stronger language model, RoBERTa, to not only encode sentences into contextual embeddings but also explore whether EEG data can still contribute to the task or not.

Sentiment Classification. In both binary and ternary sentiment classification tasks, our decoder achieves the best performance on three of the word embedding types, compared to the RNN and CNN decoder, showing the effectiveness of our approach. We first observe that different variants of EEG data all make modest contributions in different decoders, showing the potential of EEG decoding for NLP tasks. Second, the combination of EEG data and CNN decoder yields higher overall results than that with RNN decoder. These two observations are consistent with the results in [6].

Relation Detection. In the relation detection task (Table 5), our decoder still achieves the best performance compared to the text-only baseline and the other two decoders. In addition, We observe that the addition of different types of EEG data is not helpful for randomly initialized embeddings, but is beneficial with the combination of GloVe, BERT, and RoBERTa embeddings.

However, we observe the combination of EEG data and CNN decoder yields lower overall results than those using RNN decoder; and EEG $\theta + \alpha + \beta + \gamma$ yields inconsistent results with all three decoders, which requires further investigation.

Setting	Model	F1-score			
		Random Initialized	GloVe	BERT	RoBERTa
Baseline	Text-only	0.574	0.748	0.88	0.901
w/ RNN decoder [6]	+ Random noise	0.504	0.628	0.894	0.907
	+ EEG full	0.558	0.7	0.896	0.914
	+ EEG θ	0.538	0.75	0.904	0.911
	+ EEG α	**0.568**	0.746	0.904	0.91
	+ EEG β	0.558	0.76	0.9	0.915
	+ EEG γ	0.56	0.764	0.888	0.92
	+$\theta + \alpha + \beta + \gamma$	0.544	0.746	0.898	0.901
w/ CNN decoder [6]	+ Random noise	0.496	0.692	0.814	0.822
	+ EEG full	0.536	0.716	0.896	0.915
	+ EEG θ	0.538	0.756	0.896	0.913
	+ EEG α	0.55	0.762	**0.912**	0.918
	+ EEG β	0.55	0.764	0.892	0.911
	+ EEG γ	0.546	0.74	0.906	0.921
	+ $\theta + \alpha + \beta + \gamma$	0.522	0.762	0.904	0.905
w/ Ours	+ Random noise	0.498	0.702	0.83	0.852
	+ EEG full	0.54	0.74	0.908	0.918
	+ EEG θ	0.538	0.76	0.902	0.919
	+ EEG α	0.55	0.76	0.91	0.921
	+ EEG β	0.552	0.762	0.902	0.92
	+ EEG γ	0.564	**0.768**	**0.912**	**0.924**
	+ $\theta + \alpha + \beta + \gamma$	0.548	0.768	0.907	0.917

Table 3: For the binary sentiment classification task, we report the F_1-score averaged on five runs. The best results per column are marked in bold.

5 Discussion

5.1 Effectiveness of Adding EEG Features

We conducted a comparative analysis to assess the effect of concatenating EEG features with word embeddings on classification tasks, which involves evaluating the models with EEG-augmented input against two baselines: a text-only input and a random-noise-augmented input. We find that concatenating EEG features results in a performance increase across all tested embeddings except for randomly initialized embedding. Notably, the best-performing EEG frequency band always outperforms both the text-only and random noise baselines. This demonstrates the potential of using EEG data, and more broadly a multi-modal approach in natural language processing.

As expected, concatenating random noise to word embeddings provides no improvement over the text-only baseline; however, similar to Hollenstein et al. [6], we observe that BERT (and RoBERTa) can handle added noise effectively as concatenating noise sometimes increases the performance. We hypothesize that random noise serves as a regularization term when concatenating with contextualized word embeddings.

5.2 EEG Frequency Band Analysis

We investigate whether there is a change in model performance using a theoretically motivated approach of dividing the broadband EEG signal into 4 individual bands, as each frequency band corresponds to different cognitive functions during language processing.

When considering the specific contribution of an individual EEG frequency band, we observe that the addition of EEG γ and EEG β contributes the most in binary (Table 3) and ternary classification (Table 4) tasks respectively. These results are consistent with Hollenstein et al. [6]. The strong performance of the EEG β and EEG γ features can be explained by the emotional connotation of words in the β response, and the higher cognitive function in γ response [11]. For relation detection, the best-performing EEG frequency band is EEG θ. This could be due to the θ frequency band's role in cognitive processes such as memory, navigation, and attention, which are crucial for understanding the relations between various concepts.

Setting	Model	F1-score			
		Random Initialized	GloVe	BERT	RoBERTa
Baseline	Text-only	0.356	0.496	0.656	0.667
w/ RNN decoder [6]	+ Random noise	0.332	0.432	0.658	0.671
	+ EEG full	0.344	0.472	0.648	0.667
	+ EEG θ	0.34	0.498	0.644	0.662
	+ EEG α	0.328	0.508	**0.668**	0.67
	+ EEG β	0.338	0.498	0.638	0.668
	+ EEG γ	0.334	0.522	0.662	0.67
	+ $\theta + \alpha + \beta + \gamma$	0.318	0.496	0.662	0.688
w/ CNN decoder [6]	+ Random noise	0.324	0.5	0.588	0.607
	+ EEG full	0.342	0.486	0.648	0.672
	+ EEG θ	0.324	0.514	0.654	0.678
	+ EEG α	0.33	0.512	0.636	0.679
	+ EEG β	0.348	0.52	0.656	0.692
	+ EEG γ	0.312	0.51	0.658	0.685
	+ $\theta + \alpha + \beta + \gamma$	0.32	0.498	0.638	0.687
w/ Ours	+ Random noise	0.336	0.504	0.592	0.61
	+ EEG full	0.346	0.488	0.652	0.675
	+ EEG θ	0.326	0.518	0.658	0.682
	+ EEG α	0.332	0.516	0.65	0.681
	+ EEG β	**0.35**	**0.524**	0.66	**0.696**
	+ EEG γ	0.314	0.512	0.662	0.688
	+ $\theta + \alpha + \beta + \gamma$	0.322	0.502	0.642	0.689

Table 4: For the ternary sentiment classification task, we report the F_1-score averaged on five runs. The best results per column are marked in bold.

Our results show that using individual EEG frequency bands is beneficial, however, it is still unclear whether this approach can generalize well to other NLP tasks, or if there exists a frequency band that always outperforms the others.

5.3 Comparison between Embedding Types

The experimental results in Table 3, 4, and 5 show that contextualized BERT and RoBERTa embeddings outperform the non-contextual methods across all the classification tasks. We visualize the difference in Figure 6, showing that embedding type has a large impact on the baseline performance.

In addition, we compare the percentage of F1-score increase between the text-only baseline and different EEG frequency bands with BERT or RoBERTa. Table 7 shows our results. We observe that randomly initialized embedding in the addition of EEG data always results in a performance decrease. Then, we find contextualized BERT embedding can always achieve larger performance gain than RoBERTa embedding in the addition of EEG data, showing that BERT is a more effective embedding type than RoBERTa.

5.4 Ethical Implications

Word Embedding Bias: Word embeddings, learned from extensive text corpora, have been shown to exhibit human biases. For instance, gendered occupational terms such as "nurse" or "engineer" strongly correlate with words representing women or men, respectively [17]. This association can influence tokenized words in our relation detection task.

Representation Bias: The ZuCo dataset is recorded from English-speaking adults primarily from Western, developed nations such as Australia, Canada, the UK, and the USA. This demographic focus may lead to concerns about the generalizability of the data across a broader population, which includes individuals of varied native languages and cultural interpretations of words.

As our primary focus is on improving the EEG decoding architecture, there are no immediate ethical concerns arising from the results of this study. However, these findings could potentially encourage further utilization of EEG data as a multi-modal source of information.

Setting	Model	F1-score			
		Random Initialized	GloVe	BERT	RoBERTa
Baseline	Text-only	0.156	0.286	0.422	0.435
w/ RNN decoder [6]	+ Random noise	0.144	0.254	0.596	0.602
	+ EEG full	0.108	0.294	0.612	0.62
	+ EEG θ	0.166	0.352	0.63	0.635
	+ EEG α	0.154	0.35	0.628	0.64
	+ EEG β	0.164	0.34	0.624	0.632
	+ EEG γ	0.168	0.346	0.62	0.625
	+ $\theta + \alpha + \beta + \gamma$	0.252	0.298	0.448	0.455
w/ CNN decoder [6]	+ Random noise	0.142	0.286	0.328	0.33
	+ EEG full	0.148	0.324	0.602	0.61
	+ EEG θ	0.148	0.336	0.628	0.635
	+ EEG α	0.16	0.358	0.622	0.63
	+ EEG β	0.148	0.344	0.618	0.625
	+ EEG γ	0.146	0.316	0.622	0.628
	+ $\theta + \alpha + \beta + \gamma$	**0.258**	0.29	0.38	0.385
w/ Ours	+ Random noise	0.142	0.263	0.332	0.342
	+ EEG full	0.16	0.35	0.63	0.625
	+ EEG θ	0.172	**0.362**	**0.635**	**0.644**
	+ EEG α	0.17	0.36	0.634	0.642
	+ EEG β	0.168	0.358	0.632	0.642
	+ EEG γ	0.166	0.356	0.63	0.64
	+ $\theta + \alpha + \beta + \gamma$	0.27	0.312	0.38	0.392

Table 5: For the relation detection classification task, we report the F_1-score averaged on five runs. The best results per column are marked in bold.

6 Future work

Potential avenues for future research in this area include the following considerations:

- Cut EEG data into more segments for data augmentation.
- Analyze which EEG frequency band has a higher attention score in the transformer module.
- Analyze the classification performance of different layers of transformer architecture: input, middle, and final layer.
- Combine LSTM+self-attention module for experiments.

If time is not a concern, an interesting approach could be considered:

- First, collect a large amount (e.g., 1M samples) of high-quality EEG data of people reading natural sentences.
- Then, pre-train from scratch or fine-tune an off-the-shelf language model given the EEG data (EEG-LM).

7 Work distribution

Guoqing Luo:

- Implementing our CNN+Transformer decoder and conducting experiments on sentiment classification and relation detection.
- Implementing RoBERTa for encoding texts and conducting experiments.
- Analysis of comparison between BERT and RoBERTa.
- Providing visualizations for results

Henry Vu:

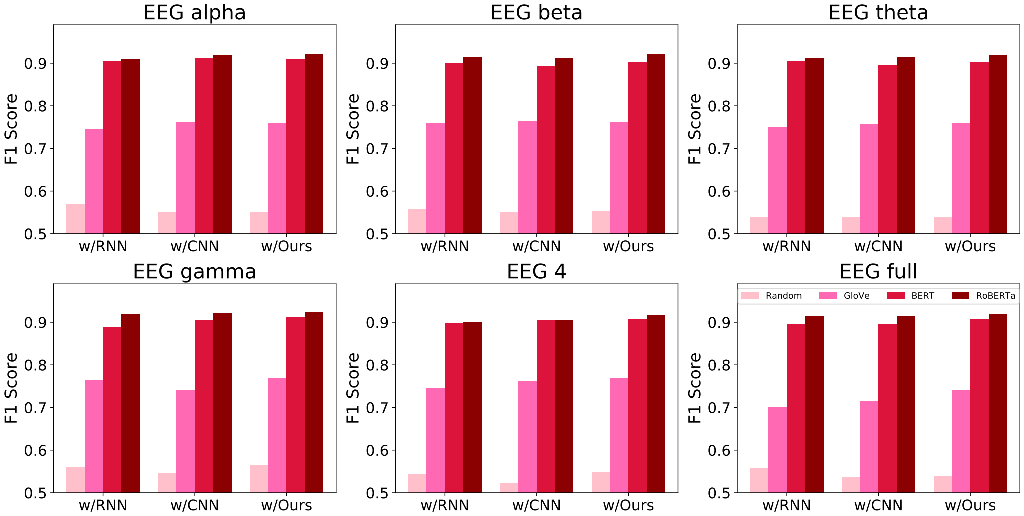

Figure 6: Visualization of different embedding types across various EEG frequency bands on binary sentiment classification task under our decoder. The legend is shown in the bottom right figure. Similar trend is found under RNN and CNN decoder.

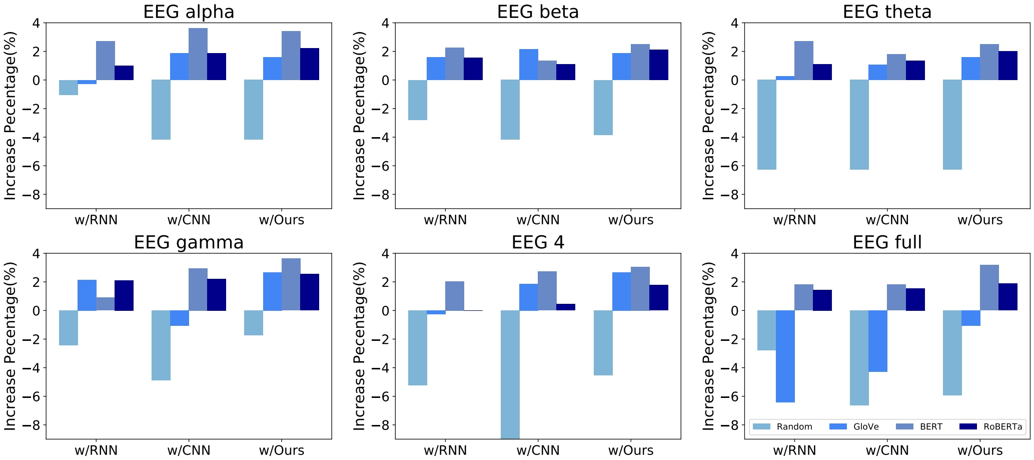

Figure 7: Increase in percentage across various EEG frequency bands over the text-only baseline on binary sentiment classification task. The legend is shown in the lower right figure. Similar percentage is found under RNN and CNN decoder.

- Extracting EEG features for experimental use.
- Conducting experiments on the sentiment classification and relation detection tasks to get the baseline results.
- Analysis of effects of different EEG frequency bands.
- Providing visualizations for data and results.

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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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/.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9. 2021. URL https://dl.acm.org/doi/pdf/10.1145/3465416.3483305.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.