Open Vocabulary Electroencephalography-to-Text Decoding and Zero-Shot Sentiment Classification

Zhenhailong Wang, Heng Ji

University of Illinois at Urbana-Champaign

{wangz3, hengji}@illinois.edu

Abstract

State-of-the-art brain-to-text systems have achieved great success in decoding language directly from brain signals using neural networks. However, current approaches are limited to small closed vocabularies which are far from enough for natural communication. Additionally, most of the high-performing approaches require data from invasive devices (e.g., ECoG). In this paper, we extend the problem to open vocabulary Electroencephalography(EEG)-To-Text Sequence-To-Sequence decoding and zero-shot sentence sentiment classification on natural reading tasks. We hypothesize that the human brain functions as a special text encoder and propose a novel framework leveraging pre-trained language models (e.g., BART). Our model achieves a 40.1% BLEU-1 score on EEG-To-Text decoding and a 55.6% F1 score on zero-shot EEG-based ternary sentiment classification, which significantly outperforms supervised baselines. Furthermore, we show that our proposed model can handle data from various subjects and sources, showing great potential for a high-performance open vocabulary brain-to-text system once sufficient data is available. The code is made publicly available for research purpose at https://github.com/MikeWangWZHL/EEG-To-Text.

Introduction

Understanding and decoding the human brain has always been a fascinating topic for centuries. In recent years, Brain-Computer-Interface (BCI) based on motor imagery has gained great success in helping paralytic people restoring motor functionalities such as reaching and grasping (Hochberg et al. 2012; Aflalo et al. 2015; Bouton et al. 2016). Decoding natural language from brain signals, on the other hand, remains a major challenge. We point out that previous approaches on brain-to-text and brain-to-speech decoding (Herff et al. 2015; Anumanchipalli, Chartier, and Chang 2019; Makin, Moses, and Chang 2020; Sun et al. 2019; Panachakel and Ramakrishnan 2021; Nieto et al. 2021; Moses et al. 2021) still have limitations in terms of vocabulary size, device, and articulation dependency, etc. Previous work mainly focuses on achieving high accuracy, thus decodes sentences and words in small closed vocabularies. Moreover, current systems lack the ability to decode semantically close words that do not exist in the training set. In this paper, we extend the problem from closed vocabulary to open vocabulary Electroencephalography(EEG)-To-Text Sequence-To-Sequence decoding as well as zero-shot sentiment classification on natural reading tasks. We enlarge the vocabulary size for more than 100 times from several hundred (Makin, Moses, and Chang 2020; Sun et al. 2019) to 50,265[1]. We utilize data (Hollenstein et al. 2018, 2020) from various subjects and sources recorded by non-invasive devices.

Previous work on brain-to-speech (Herff et al. 2015; Anumanchipalli, Chartier, and Chang 2019; Makin, Moses, and Chang 2020; Moses et al. 2021) decoding has successfully captured low-level auditory features from the movement of our vocal tract to reconstruct words. Instead of only capturing articulation features, another line of work demonstrated that the human brain encodes language into higher dimensional semantic representations (Gauthier and Ivanova 2018; Correia et al. 2014). Interestingly, we have seen similar behavior in large-scale pretrained language models, such as BERT (Devlin et al. 2019), BART (Lewis et al. 2020), GPT2 (Radford et al. 2019) and GPT3 (Brown et al. 2020), which encode words into contextualized semantic embeddings (Ethayarajh 2019). Recent findings on multimodal neurons (Goh et al. 2021) in CLIP (Radford et al. 2021) revealed another level of resemblance between artificial neurons and human brain neurons in the sense that they all respond to highly abstract concepts. The major contributions of the aforementioned large-scale pretrained language models are their transfer learning abilities. By fine-tuning them on specific downstream tasks, we observe a substantial improvement in various NLP tasks, including sequence classification, text generation, etc. Although covariate shift has been generally observed in brain signal data due to intra- and inter-subject variability (Lund et al. 2005; Saha and Baumert 2020), previous work demonstrated promising transfer learning ability in brain signal decoding using deep learning models (Roy et al. 2020; Zhang et al. 2020; Makin, Moses, and Chang 2020). Furthermore, various studies (Muttenthaler, Hollenstein, and Barrett 2020; Hollenstein

[1]We use pretrained BART vocabulary: https://huggingface.co/transformers/model_doc/bart.html?highlight=bart\#transformers.BartConfig

et al. 2021, 2019; Hale et al. 2018; Schwartz and Mitchell 2019) have experimented with connecting brain signal decoding to NLP models, by either using brain signals as an additional modality for improving performance on NLP tasks or using NLP models to understand how the human brain encodes language.

In this paper, we extend previous work to a new level by using pretrained language models for open vocabulary EEG-to-text decoding. The motivation of using pretrained language models is that contextualized representation from pretrained language models carries important linguistic information, including syntactic features, semantic features, and long-distant dependencies (Tenney et al. 2019; Jawahar, Sagot, and Seddah 2019). This existing knowledge obtained from observing large text corpora is particularly useful for our open vocabulary decoding task with scarce data. Based on the assumption that the human brain functions as a special text encoder, we leverage pretrained language models via jointly fine-tuning with additional projection layers. We then further evaluate its power on a novel zero-shot sentence-level sentiment classification task. Detailed definition of the new tasks can be found in task definition section.

In terms of the choice of device, although invasive devices like Electrocorticography (ECoG) generally result in better performance (Burle et al. 2015), we focus on using non-invasive EEG data for a few reasons. Compared to other non-invasive devices like functional magnetic resonance imaging (fMRI), EEG has relatively high temporal resolution with an affordable cost (Zanzotto and Croce 2010; Hecht and Stout 2015; Yi et al. 2013). Analogous to training a language model, the sheer amount of data is essential for learning representation of brain signals (Brown et al. 2020). Compared to data from invasive-devices, EEG data is easier to acquire and more publicly available. However, unlike the abundance of text and image data, we have far from enough brain-text paired data to train a high-performance open vocabulary brain-to-text sequence-to-sequence model. Nevertheless, we have observed a trend in growing availability for open-source EEG-integrated devices[2], which can be a huge potential data source. And we suggest that, shortly, non-invasive BCI will have a larger and wider impact on everyone's life as a new type of human-machine interaction tool in various areas, including Virtual Reality and Augmented Reality (Putze et al. 2020).

To summarize, the main contributions of this paper are as follows.

- We introduce two new tasks: Open vocabulary EEG-To-Text decoding and EEG-based sentence sentiment classification. To the best of our knowledge, this is the first work using open vocabulary setting on brain-to-text decoding.
- We are the first to use pretrained language models for EEG-To-Text decoding. We further propose a novel zero-shot pipeline for EEG-based sentiment classification.
- We show that our proposed framework can leverage data from various subjects and sources, which demonstrates great potential for high-performance open vocabulary EEG-To-Text system.

[2] https://openbci.com/

Symbol	Meaning
$v \in \mathcal{V}$	English token in an open vocabulary
$\mathbf{e} \in \mathcal{E}$	EEG feature vector in an EEG sequence
$c \in \mathcal{C}$	Sentiment label in ternary sentiment classes
$\langle \mathcal{E}, \mathcal{S} \rangle$	Word-level EEG feature sequence and text sentence pair
$\langle \mathcal{E}, c \rangle$	Word-level EEG feature sequence and sentiment label pair
$\langle \hat{\mathbf{e}}, c \rangle$	Aggregated sentence-level EEG and sentiment label pair
$\langle \mathcal{S}, c \rangle$	Text sentence and sentiment label pair

Table 1: List of symbols

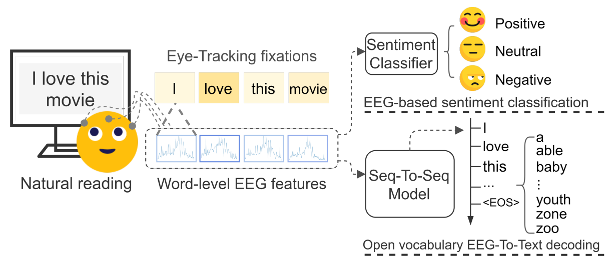

Figure 1: An overview of proposed tasks. Subjects are asked to read text on a screen at their own speed. Simultaneous eye-tracking and EEG data are recorded. The darker the background color, the more fixations are on the word. Word-level EEG features can be extracted by synchronizing with eye-tracking fixations. EEG feature sequences then serve as inputs for sequence-to-sequence decoding or sentiment classification. In this paper, we use ZuCo datasets for experiments, please refer to experiment section for more details.

Task Definition

Open vocabulary EEG-To-Text decoding Given a sequence of word-level EEG features $\mathcal{E}$, the task is to decode the corresponding text tokens from an open vocabulary $\mathcal{V}$ in a Sequence-To-Sequence framework. In this paper, we use EEG-Text pairs $\langle \mathcal{E}, \mathcal{S} \rangle$ recorded in natural reading tasks, e.g., ZuCo dataset (Hollenstein et al. 2018, 2020). During the training phase, such EEG-Text pairs can come from various subjects and various categories of reading materials. During the test phase, the text sentences $\mathcal{S}$ in $\langle \mathcal{E}, \mathcal{S} \rangle$ are totally unseen.

EEG-based sentence sentiment classification Given aforementioned word-level EEG feature sequences $\mathcal{E}$, the task is to predict the sentiment label $c \in \mathcal{C}$ of the corresponding text sentence $\mathcal{S}$. Instead of including text input (Hollenstein et al. 2021; Kumar, Yadava, and Roy 2019) along with brain signal features, we use EEG features as our only input. We further introduce **zero-shot EEG-based sentence sentiment classification**, in which we do not require any EEG-Sentiment pairs $\langle \mathcal{E}, c \rangle$. Instead, we use only EEG-Text pairs

$\langle \mathcal{E}, \mathcal{S} \rangle$ and external Text-Sentiment pairs $\langle \mathcal{S}_{ext}, c \rangle$. Thus, we can leverage EEG-Text pairs without sentiment labels from various sources, as well as existing text-sentiment datasets, such as Yelp[3] and Stanford Sentiment Treebank(Socher et al. 2013). An overview of the proposed tasks can be found in Figure 1.

Method

EEG-To-Text Decoding

Similar to (Makin, Moses, and Chang 2020), we formulate the EEG-To-Text Decoding task in a neural machine translation setting (Sutskever, Vinyals, and Le 2014; Bahdanau, Cho, and Bengio 2015). We try to maximize the probability of the decoded sentence:

$$p(\mathcal{S}|\mathcal{E}) = \prod_{t=1}^{T} p(s_t \in \mathcal{V}|\mathcal{E}, s_{<t}) \tag{1}$$

where T is the length of the target text sequence. The main challenge in our setting is that our vocabulary size $|\mathcal{V}|$ (~ 50000) is significantly larger than previous sequence-to-sequence studies (~ 250) (Makin, Moses, and Chang 2020). In addition, we use more noisy non-invasive EEG data. To address these challenges, we propose a novel framework leveraging pretrained language models. Inspired by the machine translation application using pretrained BART as described in (Lewis et al. 2020), we treat each EEG feature sequence as an encoded sentence by the human brain. We then train an additional encoder to map the embedding from the human brain to the embedding from the pretrained BART. The high-level idea is that we assume the human brain to be a special kind of encoder, as mentioned in introduction, which functions similar to a language model that encodes a sequence of English tokens into contextualized embeddings. One major difference from a traditional machine translation task is that the tokens in the input sequence here are not drawn from a fixed vocabulary, e.g. French words, but rather a continuous feature space. Inspired by Speech Recognition (Hinton et al. 2012), where the acoustic input is also represented as continuous vectors, we use these EEG feature vectors directly as initial word embeddings to feed into the model.

As shown in Figure 2.a, the model contains two major components, a randomly initialized additional encoder, and a pretrained encoder-decoder BART. Given a sequence of EEG features $\boldsymbol{h}_e$, we first feed them into a Multi-layer Transformer Encoder (MTE) and then a single layer feed-forward network to get the mapped embedding $\boldsymbol{h}_m$ of the sequence. Then the mapped embedding is fed into pretrained BART encoder and decoder. Finally, the last hidden states from the BART decoder are fed into a fully connected layer to generate English tokens s_t from pretrained BART vocabulary $\mathcal{V}$.

$$\boldsymbol{h}_m = \text{ReLU}\left((\text{MTE}(\boldsymbol{h}_e))^T \boldsymbol{W}_e\right) \tag{2}$$

$$p(s_t \in \mathcal{V}) = \text{Softmax}\left(\text{BART}(\boldsymbol{h}_m)^T \boldsymbol{W}_d\right) \tag{3}$$

$$\mathcal{L}_{\text{rec}} = -\sum_{t}^{T} \log p(s_t \in \mathcal{V}) \tag{4}$$

where the MTE has 6 layers and 8 attention heads, $\boldsymbol{W}_e$ represents the weights from the fully connected projection layer, $\boldsymbol{W}_d$ represents the weights from the language modeling head, which is a single fully connected layer. All pooling methods use the last hidden states in both MTE and BART decoders. During training, the objective is to minimize the text reconstruction cross-entropy loss as shown in equation 4.

Zero-shot Sentiment Classification Pipeline

We further propose a novel pipeline leveraging the aforementioned Sequence-To-Sequence EEG-To-Text model to do zero-shot EEG-based sentence-level sentiment analysis. As shown in Figure 2.b, the proposed pipeline consists of two modules, a decoder, and a classifier. The idea is to first use an EEG-To-Text model to convert EEG features into noisy text, and then use a text-based classifier to predict the sentiment label. That makes this pipeline zero-shot in the sense that both the decoder and the classifier are trained individually, on EEG-Text pairs $\langle \mathcal{E}, \mathcal{S} \rangle$ and external Text-Sentiment pairs $\langle \mathcal{S}, c \rangle$ respectively, and thus no EEG-Sentiment pairs $\langle \mathcal{E}, c \rangle$ are required. Moreover, the modularized architecture makes it easy to upgrade with improved decoders and classifiers. In this paper, we use a BART-based (Lewis et al. 2020) decoder as mentioned in the previous section. We also experiment with different choices of classifiers based on pretrained BERT (Devlin et al. 2019), BART (Lewis et al. 2020), and RoBERTa (Liu et al. 2019).

The idea of first decoding EEG features into text comes from observations on some preliminary experiments. We find that baselines trained explicitly on EEG-Sentiment pairs $\langle \mathcal{E}, c \rangle$ do not work well on decoding sentiment directly from EEG features. We hypothesize that, compared to previous emotion analysis studies (Koelstra et al. 2011; Liu et al. 2020) where EEG signals are recorded from image/video stimulus, it is more difficult to directly decompose sentiment related features from text semantic features in our natural reading EEG data. However, in our zero-shot pipeline, decoding EEG sequence into text serves as an information filter, which enables us to further leverage the power of finetuned text-based sentiment classifiers.

Experiment

Dataset

We use ZuCo (Hollenstein et al. 2018, 2020) datasets, which contain simultaneous EEG and Eye-tracking data recorded from natural reading tasks. The reading tasks include Normal Reading (NR) and Task-Specific Reading (TSR). Reading material is collected from movie reviews (Socher et al. 2013) and Wikipedia articles. Normal reading with movie

[3] https://www.yelp.com/dataset

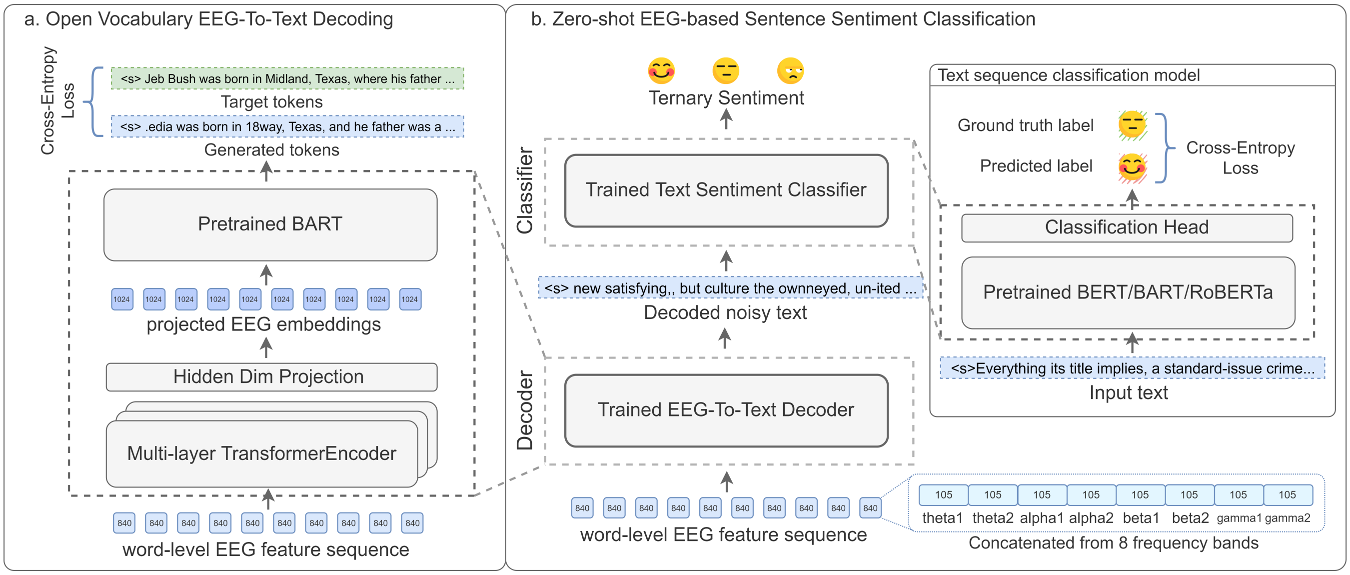

Figure 2: (a) Our proposed framework for EEG-To-Text decoding. (b) Our proposed pipeline for zero-shot EEG-based sentence sentiment classification. EEG-To-Text model trained on EEG-Sentence pairs in (a) is plugged in as Decoder in (b). Sequence classification model trained on external Text-Sentiment pairs is plugged in as Classifier in (b).

Reading Task	#Unique Sentences	#Training Samples	#Testing Samples
SR v1.0	400	3391	418
NR v1.0	300	2406	321
NR v2.0	349	4456	601
TSR v1.0	407	3372	350

Table 2: Dataset Statistics. SR: Normal Reading (sentiment), NR: Normal Reading (wikipedia), TSR: Task Specific Reading (wikipedia). We used data from 12 subjects in v1.0 and 18 subjects in v2.0.

reviews has ground-truth ternary sentiment labeling: positive, neutral, or negative. In this paper, we use concatenated word-level EEG feature sequences aggregated by gaze duration (GD). For more details on EEG input data please refer to the Appendix. We further clean up the dataset by omitting sentences that contain NaN values. And then we split each reading task's data into *train*, *development*, *test* (80%,10%,10%) by unique sentences, that is, the sentences in *test* set are totally unseen. The final statistics of the dataset can be found in Table 2.

EEG-To-Text Decoding Evaluation

We train our EEG-To-Text model on EEG-Text sequence pairs. Analogous to typical NLP fine-tuning tasks, we hypothesize that the model should be able to benefit from expansion of the training corpora. So we gradually increase the training dataset size by adding data from various reading tasks. We also include data from different subjects to test the model's robustness against inter-subject variability. We report the BLEU scores and ROUGE-1 scores in Table 3.

Results The results show that, by increasing the scale of the training set, the model achieves significantly better performance. More importantly, we find that the model can handle data from various subjects and materials. By comparing models trained on "*SR v1.0 (half$_{1st}$)*" with "*SR v1.0 (half$_{1st}$)+NR v1.0 (half$_{1st}$)*" and "*SR v1.0*", we find that the model achieves more improvements when adding data from diverse source (BLEU-1: from 33.5% to 37.3%) than from uniform source (BLEU-1: from 33.5% to 34.5%). And we observe similar behavior even when adding data from different subjects as shown by comparing "*SR v1.0 (half$_{1st}$)*" with "*SR v1.0 (half$_{1st}$)+NR v1.0 (half$_{2nd}$)*" (BLEU-1: from 33.5% to 37.4%). Furthermore, our best-performed model is trained and tested on "*SR v1.0+NR v1.0+NR v2.0*" data, which is from 30 different subjects reading movie reviews and Wikipedia articles.

We also observe that the choice of the reading task is important. Comparing models trained on "*SR v1.0+NR v1.0*" and on "*SR v1.0+NR v1.0+TSR v1.0*", we can see that adding Task-Specific Reading (TSR) data does not result in visible improvements. The reason is that in the TSR setting, subjects are asked to read text prompt with a preliminary question such as "Does this sentence contain the *Nationality* relation?". We reason that such a setting can make subjects focus on specific parts of a sentence, which results in highly different distributions in EEG features compared with data from Normal Reading tasks (SR, NR). In addition, we showcase the power of using a pretrained language model by reporting scores from the same model trained from scratch (*w/o pretrained weights*).

Discussion By taking a closer look at the decoding results (Table 4), we find that the model can sometimes correctly capture named entities, such as "*George W. Bush*" in

Reading Task	**#Training Sample**	**BLEU-N(%)** $N=1$	$N=2$	$N=3$	$N=4$	**ROUGE-1(%)** P	R	F
SR v1.0 (half$_{1st}$)	1771	33.5	16.6	7.2	3.4	22.8	21.6	22.0
SR v1.0 (half$_{1st}$) + NR v1.0 (half$_{1st}$)	3129	37.3	20.2	10.2	5.0	27.8	25.4	26.5
SR v1.0 (half$_{1st}$) + NR v1.0 (half$_{2nd}$)	3058	37.4	18.8	8.9	4.0	26.3	25.2	25.5
SR v1.0	3391	34.5	17.2	7.2	3.0	23.9	22.0	22.8
SR v1.0 + NR v1.0	5797	37.1	20.0	10.4	5.3	27.7	26.0	26.8
SR v1.0 + NR v1.0 + TSR v1.0	9169	37.4	20.0	10.5	5.7	27.4	24.6	25.9
SR v1.0 + NR v1.0 + NR v2.0	10710	**40.1**	**23.1**	**12.5**	**6.8**	**31.7**	**28.8**	**30.1**
w/o pretrained weights	10710	24.7	7.3	2.4	1.0	19.4	20.2	18.9

Table 3: EEG-To-Text sequence-to-sequence model evaluation: (half$_{1st}$) and (half$_{2nd}$) indicate using data from the first half of or the second half of the subjects respectively; no parenthesis means using data from all subjects.

(1)	Ground Truth: He is a prominent **member of** the *Bush family*, the younger brother of **President George W. Bush**...
	Model Output: was a former **member of** the *American family*, and son brother of **President George W. Bush**...
(2)	Ground Truth: Raymond Arrieta (born March 26, 1965 in San Juan, **Puerto Rico**) is considered by many to be one of **Puerto Rico's greatest** comedians.
	Model Output: mond wasaga,19 in 17, 18) New Francisco, **Puerto Rico**) is a one many to be the of the **Rico's greatest** poets.
(3)	Ground Truth: He was first *appointed* to fill the Senate **seat** of Ernest Lundeen who had **died** in office.
	Model Output: was a *elected* to the the position **seat** in the Hemy in **died** died in 18 in
(4)	Ground Truth: Adolf Otto Reinhold Windaus (December 25, 1876 - June 9, 1959) was a significant *German chemist*.
	Model Output: rian Hitler,hardt,eren18 18, 1885 – January 3, 18) was a *German figure*- and
(5)	Ground Truth: It's *not a particularly* **good** film, but neither is it a *monsterous* one.
	Model Output: was a a *bad* **good** story, but it is it *bad bad*. one.

Table 4: EEG-To-Text decoding examples on unseen test sentences. (1-4) are biographical sentences from Wikipedia in NR v1.0, v2.0. (5) is movie review in SR v1.0. **Bold** words indicate exact match, *Italic* words indicate semantic resemblance, and Underline words indicate error match.

(1) and "*Puerto Rico*" in (2), which does not even exist in the training set. In the cases that the model fails to decode the correct entity mentions, e.g., "*San Juan[LOCATION]*" vs "*New Francisco[LOCATION]*" in (2), "*Adolf Otto Reinhold Windaus[PERSON]*" vs "*Hitler[PERSON]*" in (5), the entity types are correctly captured.

To systematically evaluate this interesting behavior, we perform named entity recognition[4] on both ground truth and model output. We calculate a *longest common subsequence (LCS)* based matching score and a *multiset cosine similarity (Multiset)* score on the extracted sequences of entity types, as shown in equation 5 and 6.

$$Sim_{LCS} = \frac{\text{LCS}(X, Y)}{\max(|X|, |Y|)} \tag{5}$$

$$Sim_{Multiset} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}, i \in \mathcal{I} \tag{6}$$

where X, Y are sequences of named entity types, e.g. Geopolitical entities (GPE), PERSON, from ground truth and model output respectively. $LCS(X, Y)$ returns the length of the longest common subsequence of X, Y. x_i, y_i are the number of instances on type i. $\mathcal{I}$ is a union of the sets of extracted entity types from ground truth and model output. The results can be found in Table 5.

Reading Task	**LCS**(%)	**Multiset**(%)
SR v1.0	14.9	17.8
SR v1.0 + NR v1.0	29.2	50.2
SR v1.0 + NR v1.0 + NR v2.0	**35.6**	**55.7**

Table 5: Named entity type matching results. *LCS* refers to Longest Common Subsequence based matching score, *Multiset* refers to Multiset Cosine Similarity.

Apart from that, we find that the model can generate semantically close words or synonyms even though they do not exactly match the ground truth, as shown in *Italic* words in Table 4. For example, in (5), although "*monsterous*" does not appear in the training set and is actually a typo, the model is able to generate "*bad*" which semantically resembles "*monstrous*". Because the available training data is far from enough for fine-tuning a large language model like

[4]We use off-the-shelf named entity recognizer from Spacy: https://spacy.io/usage/linguistic-features\#named-entities

BART, the model still struggles to correctly reconstruct most of the fine-grained entities, as shown in the underlined words in Table 4. And the decoded sentences still suffer from grammatical errors.

EEG-based Sentiment Classification Evaluation

We first implement a few baselines that are explicitly trained on EEG-Sentiment pairs $\langle \mathcal{E}, c \rangle$. And then we evaluate the baselines and our proposed zero-shot pipeline on the same test set from reading task SR v1.0.

Baselines

- **MLP** A Multi-Layer Perceptron baseline with three fully connected layers, ReLU activation, and one dropout layer. MLP is trained on aggregated sentence-level EEG-Sentiment pairs $\langle \hat{\mathbf{e}}, c \rangle$, where $\hat{\mathbf{e}}$ is an average over all word-level EEG features in a sequence.
- **Bi-LSTM** A Bi-directional-LSTM baseline with four stacked LSTM layers and a single-layer classification head.
- **BERT** BERT-based baselines with additional multi-layer transformer encoders (MTE). We also experiment with training from scratch using randomly initialized weights and fine-tuning from pretrained weights. Both LSTM and BERT baselines are trained on EEG-Sentiment sequence pairs $\langle \mathcal{E}, c \rangle$ from reading task SR v1.0.
- **Text Baselines** For more comprehensive comparison, we also report scores from text-based sentiment classification models based on pretrained BERT/BART/RoBERTa. We train and test these baselines on Text-Sentiment pairs $\langle \mathcal{S}, c \rangle$ from reading task SR v1.0.

Pipeline In our zero-shot pipeline, we experiment with various combinations of decoders and classifiers. For the decoder, we use the BART-based EEG-To-Text decoding model trained on SR v1.0 EEG-Text pairs $\langle \mathcal{E}, \mathcal{S} \rangle$. We demonstrate the importance of the additional Multi-layer Transformer Encoder (MTE) component by removing it from the model. We also experiment with the decoder using a randomly initialized BART model. For the classifier, we experiment with sequence classification models based on BERT, BART and RoBERTa, which have identical architectures as the text baselines mentioned in the baselines section. Here, instead of being trained on SR v1.0 data, the classifiers are trained on a subset of Text-Sentiment pairs $\langle \mathcal{S}_{ext}, c \rangle$ from Stanford Sentiment Treebank (Socher et al. 2013). We include sentences with a sentiment score in the range of very negative ($[0, 0.2]$), neutral ($(0.4, 0.6]$) or very negative ($(0.8, 1.0]$), and assign a ternary sentiment label $c \in \{0, 1, 2\}$ to them respectively. And then we exclude those sentences that are already in the ZuCo SR v1.0 dataset to make sure that the sentences $\mathcal{S}_{ext}$ in $\langle \mathcal{S}_{ext}, c \rangle$ do not overlap with the sentences $\mathcal{S}$ in $\langle \mathcal{E}, \mathcal{S} \rangle$, which are used to train the Decoder. This setting guarantees that we do not require any EEG-Sentiment pairs $\langle \mathcal{E}, c \rangle$. The results are presented in Table 6.

Results One major observation from EEG-based baselines is that traditional sequence classification models struggle to converge when decoding sentiment directly from EEG features. As shown in scores from *Bi-LSTM* and *MLP*, they hardly outperform random guess. And we can see that the *Bi-LSTM* model can not take advantage of the additional sequential information from EEG feature sequences, since we do not observe improvements compared with simply doing classification on a single averaged feature vector using *MLP*. By comparing scores from $BERT_{rand}$ and $BERT_{fine}$, we find that directly applying pre-trained language model to EEG features does not help much. We hypothesize that since the amount of labeled data is very limited ($\sim$ 400 unique sentences) compared with a typical sentiment dataset such as Stanford Sentiment Treebank (Socher et al. 2013)($\sim$ 10k unique sentences), it is simply too noisy for the models to extract useful features out of the EEG sequences. We qualitatively verify this hypothesis by observing high bias in the predicted label distribution, as shown in Table 7.

On the other hand, we find that our best-performed zero-shot pipeline (DEC_{BART}+CLS_{BART}) significantly outperforms all fully supervised baselines by a large margin of over 20%. We also observe that our pipeline is more robust to noisy data as shown in Table 7. As discussed in method section, we think the reason why our zero-shot pipeline works much better than baselines is that, by first decoding EEG sequence into text, the decoder filters out noisy information from the EEG features and enables the model to effectively leverage the power of text-based classifiers. Furthermore, we show that our zero-shot pipeline is highly modularized in the sense that improving a single component can result in improvements on the whole model. For example, the BART-based classifier (CLS_{BART}), which achieves top performance on text input, also constitutes our best-performed pipeline (DEC_{BART}+CLS_{BART}). The results from the crippled versions of decoder (DEC_{BART}) show that the pretrained language model and the additional Multi-layer TransformerEncoder (MTE) are essential to the pipeline, as we observe substantial performance drop on the ones *w/o MTE*, and *w/o pretrained weights*.

Related Work

Related work on brain-to-speech and brain-to-text decoding can be categorized into three major approaches by the features they are capturing: motor imagery based, overt speech based, and inner speech based. Various kinds of brain signals have been explored including EEG, ECoG, and fMRI. We point out that previous approaches still have limitations in terms of vocabulary size, articulation dependency, speed and device. Additional details can be found in Appendix.

Motor imagery based systems, e.g., point-and-click (Pandarinath et al. 2017; Jarosiewicz et al. 2015) and imaginary handwriting (Willett et al. 2021) has high accuracy but relatively low typing rate. Overt speech based approaches for decoding or synthesizing speech achieve faster communication rate. However, these approaches either require subjects to physically move their vocal tract (Herff et al. 2015; Anumanchipalli, Chartier, and Chang 2019; Makin, Moses, and Chang 2020), or require the subjects to imagine the

Model	Test Input	Is Zero-Shot	Precision(%)	Recall(%)	F1(%)	Accuracy(%)
MLP (Hollenstein et al. 2019)	$\hat{\mathbf{e}}$	No	32.8	33.6	27.5	31.8
Bi-LSTM (Hollenstein et al. 2021)	$\mathcal{E}$	No	33.9	34.1	17.4	30.9
BERT$_{rand}$	$\mathcal{E}$	No	38.2	33.5	30.0	35.5
BERT$_{fine}$	$\mathcal{E}$	No	23.7	34.5	27.2	36.6
DEC$_{BART}$ + CLS$_{BERT}$	$\mathcal{E}$	Yes	61.0	50.4	50.1	49.1
DEC$_{BART}$ + CLS$_{RoBERTa}$	$\mathcal{E}$	Yes	58.2	51.2	51.9	50.9
DEC$_{BART}$ + CLS$_{BART}$	$\mathcal{E}$	Yes	**62.4**	**56.5**	**55.6**	**55.3**
w/o pretrained + CLS$_{BART}$	$\mathcal{E}$	Yes	10.0	33.3	15.4	30.0
w/o MTE + CLS$_{BART}$	$\mathcal{E}$	Yes	41.3	40.9	39.3	40.4
CLS$_{BERT}$	$\mathcal{S}$	No	76.0	74.5	74.1	75.4
CLS$_{RoBERTa}$	$\mathcal{S}$	No	72.5	71.3	70.6	72.7
CLS$_{BART}$	$\mathcal{S}$	No	**79.3**	**78.3**	**77.4**	**79.7**

Table 6: Ternary sentiment classification results on SR v1.0 testset. In *Test Input* column, $\hat{\mathbf{e}}$ means aggregated sentence-level EEG features, $\mathcal{E}$ is word-level EEG feature sequence, $\mathcal{S}$ is text sentences. In the Model column, the first section contains baselines explicitly trained on EEG-Sentiment pairs, where subscript $_{rand}$ indicates it's randomly initialized, and $_{fine}$ indicates it's fine-tuned from pretrained checkpoint. The second section contains our proposed Zero-shot pipelines with different choices of decoder (DEC) and classifier (CLS). The third section contains text-based baselines.

Model	Pos(%)	Neu(%)	Neg(%)
BERT$_{rand}$	68.6	26.5	4.9
BERT$_{fine}$	72.4	27.0	0.6
DEC$_{BART}$+CLS$_{BART}$	19.3	54.4	26.3
Ground Truth	37.5	30.0	32.5

Table 7: Predicted label distribution: *Pos*, *Neu*, *Neg* means *Positive*, *Neutral*, *Negative* respectively. Model names are consistent with Table 6.

physical articulation of the sentence (Moses et al. 2021). This makes the decoding system language-dependent, since the same concept may have totally different pronunciations in different languages. Another line of work tries to address articulation dependency by decoding language from imagined speech or read text (Panachakel and Ramakrishnan 2021; Sun et al. 2019; Nieto et al. 2021). A major limitation for most of the aforementioned approaches is that current experiments are often constrained with a small closed vocabulary, with ten to a few hundred unique words. Moreover, most state-of-the-art high-performance brain-computer-interface (Willett et al. 2021; Makin, Moses, and Chang 2020; Pandarinath et al. 2017) for language communication use invasive devices such as Electrocorticography (ECoG). Although compared to non-invasive devices like Electroencephalogram (EEG), ECoG has a higher temporal and spatial resolution as well as a higher signal-to-noise ratio, it is hard to collect large-scale dataset, and is currently impractical to extend the benefit to healthy people.

Our work is distinguished from previous work by the fact that we decode text from EEG features in a Sequence-To-Sequence manner over the whole English vocabulary. And our model is able to decode unseen sentences and handle EEG data from various sources and subjects.

Related work on sentiment/emotion analysis from brain signals traditionally focuses on video or image stimulus (Koelstra et al. 2011; Liu et al. 2020). Previous attempts using text-elicited brain signals often treat brain signals only as an additional input along with other traditional modalities like text and image (Hollenstein et al. 2021; Kumar, Yadava, and Roy 2019; Gauba et al. 2017).

Our work on zero-shot sentiment discovery extends previous work by using only EEG features as input and by the fact that we do not require any EEG-Sentiment labeled pairs.

Conclusions and Future Work

In this paper, we introduce two challenging tasks, namely, open vocabulary EEG-To-Text Sequence-To-Sequence decoding and EEG-based sentence sentiment classification. We propose novel frameworks leveraging pretrained language models which show great scalability and zero-shot ability. Future work is needed on collecting larger-scale EEG-Text datasets as well as extending the current framework to multilingual settings. Furthermore, previous work (Segaert et al. 2012) demonstrated that certain frontal regions, e.g., Broca's area, show activation during both language comprehension and language production processes. One future direction is to apply our current model on decoding inner speech in an open vocabulary setting. A dedicated sentence-level inner speech dataset with a larger vocabulary size is needed, since current datasets (Nieto et al. 2021) on inner speech decoding have very limited word coverage.

Acknowledgements

This research is based upon work supported in part by U.S. DARPA GAILA Program No. HR00111990058. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Aflalo, T.; Kellis, S.; Klaes, C.; Lee, B.; Shi, Y.; Pejsa, K.; Shanfield, K.; Hayes-Jackson, S.; Aisen, M.; Heck, C.; et al. 2015. Decoding motor imagery from the posterior parietal cortex of a tetraplegic human. *Science*, 348(6237): 906–910.

Anumanchipalli, G. K.; Chartier, J.; and Chang, E. F. 2019. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753): 493–498.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations*.

Bouton, C. E.; Shaikhouni, A.; Annetta, N. V.; Bockbrader, M. A.; Friedenberg, D. A.; Nielson, D. M.; Sharma, G.; Sederberg, P. B.; Glenn, B. C.; Mysiw, W. J.; et al. 2016. Restoring cortical control of functional movement in a human with quadriplegia. *Nature*, 533(7602): 247–250.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*.

Burle, B.; Spieser, L.; Roger, C.; Casini, L.; Hasbroucq, T.; and Vidal, F. 2015. Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view. *International Journal of Psychophysiology*, 97(3): 210–220.

Correia, J.; Formisano, E.; Valente, G.; Hausfeld, L.; Jansma, B.; and Bonte, M. 2014. Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *Journal of Neuroscience*, 34(1): 332–338.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*, 4171–4186.

Ethayarajh, K. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 55–65.

Gauba, H.; Kumar, P.; Roy, P. P.; Singh, P.; Dogra, D. P.; and Raman, B. 2017. Prediction of advertisement preference by fusing EEG response and sentiment analysis. *Neural Networks*, 92: 77–88.

Gauthier, J.; and Ivanova, A. 2018. Does the brain represent words? An evaluation of brain decoding studies of language understanding. *Conference on Cognitive Computational Neuroscience*.

Goh, G.; Cammarata, N.; Voss, C.; Carter, S.; Petrov, M.; Schubert, L.; Radford, A.; and Olah, C. 2021. Multimodal neurons in artificial neural networks. *Distill*, 6(3): e30.

Hale, J.; Dyer, C.; Kuncoro, A.; and Brennan, J. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2727–2736.

Hecht, E.; and Stout, D. 2015. Techniques for studying brain structure and function. In *Human Paleoneurology*, 209–224. Springer.

Herff, C.; Heger, D.; De Pesters, A.; Telaar, D.; Brunner, P.; Schalk, G.; and Schultz, T. 2015. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, 9: 217.

Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A.-r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T. N.; et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6): 82–97.

Hochberg, L. R.; Bacher, D.; Jarosiewicz, B.; Masse, N. Y.; Simeral, J. D.; Vogel, J.; Haddadin, S.; Liu, J.; Cash, S. S.; Van Der Smagt, P.; et al. 2012. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485(7398): 372–375.

Hollenstein, N.; de la Torre, A.; Langer, N.; and Zhang, C. 2019. CogniVal: A Framework for Cognitive Word Embedding Evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 538–549.

Hollenstein, N.; Renggli, C.; Glaus, B.; Barrett, M.; Troendle, M.; Langer, N.; and Zhang, C. 2021. Decoding EEG Brain Activity for Multi-Modal Natural Language Processing. *Frontiers in Human Neuroscience*.

Hollenstein, N.; Rotsztejn, J.; Troendle, M.; Pedroni, A.; Zhang, C.; and Langer, N. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1): 1–13.

Hollenstein, N.; Troendle, M.; Zhang, C.; and Langer, N. 2020. ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 138–146.

Jarosiewicz, B.; Sarma, A. A.; Bacher, D.; Masse, N. Y.; Simeral, J. D.; Sorice, B.; Oakley, E. M.; Blabe, C.; Pandarinath, C.; Gilja, V.; et al. 2015. Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface. *Science translational medicine*, 7(313): 313ra179–313ra179.

Jawahar, G.; Sagot, B.; and Seddah, D. 2019. What does BERT learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; and Patras, I. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1): 18–31.

Kumar, S.; Yadava, M.; and Roy, P. P. 2019. Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction. *information fusion*, 52: 41–52.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.

Liu, J.; Wu, G.; Luo, Y.; Qiu, S.; Yang, S.; Li, W.; and Bi, Y. 2020. EEG-based emotion classification using a deep neural network and sparse autoencoder. *Frontiers in Systems Neuroscience*, 14: 43.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lund, T. E.; Nørgaard, M. D.; Rostrup, E.; Rowe, J. B.; and Paulson, O. B. 2005. Motion or activity: their role in intra- and inter-subject variation in fMRI. *Neuroimage*, 26(3): 960–964.

Makin, J. G.; Moses, D. A.; and Chang, E. F. 2020. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, 23(4): 575–582.

Moses, D. A.; Metzger, S. L.; Liu, J. R.; Anumanchipalli, G. K.; Makin, J. G.; Sun, P. F.; Chartier, J.; Dougherty, M. E.; Liu, P. M.; Abrams, G. M.; Tu-Chan, A.; Ganguly, K.; and Chang, E. F. 2021. Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria. *New England Journal of Medicine*, 385(3): 217–227.

Muttenthaler, L.; Hollenstein, N.; and Barrett, M. 2020. Human brain activity for machine attention. *arXiv preprint arXiv:2006.05113*.

Nieto, N.; Peterson, V.; Rufiner, H. L.; Kamienkowski, J.; and Spies, R. 2021. " Thinking out loud": an open-access EEG-based BCI dataset for inner speech recognition. *bioRxiv*.

Panachakel, J. T.; and Ramakrishnan, A. G. 2021. Decoding Covert Speech From EEG-A Comprehensive Review. *Frontiers in Neuroscience*, 15: 392.

Pandarinath, C.; Nuyujukian, P.; Blabe, C. H.; Sorice, B. L.; Saab, J.; Willett, F. R.; Hochberg, L. R.; Shenoy, K. V.; and Henderson, J. M. 2017. High performance communication by people with paralysis using an intracortical brain-computer interface. *Elife*, 6: e18554.

Putze, F.; Vourvopoulos, A.; Lécuyer, A.; Krusienski, D.; Bermúdez i Badia, S.; Mullen, T.; and Herff, C. 2020. Brain-computer interfaces and augmented/virtual reality. *Frontiers in human neuroscience*, 14: 144.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

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

Roy, S.; Chowdhury, A.; McCreadie, K.; and Prasad, G. 2020. Deep learning based inter-subject continuous decoding of motor imagery for practical brain-computer interfaces. *Frontiers in Neuroscience*, 14.

Saha, S.; and Baumert, M. 2020. Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review. *Frontiers in computational neuroscience*, 13: 87.

Schwartz, D.; and Mitchell, T. 2019. Understanding language-elicited EEG data by predicting it from a fine-tuned language model. 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)*, 43–57.

Segaert, K.; Menenti, L.; Weber, K.; Petersson, K. M.; and Hagoort, P. 2012. Shared syntax in language production and language comprehension—an fMRI study. *Cerebral Cortex*, 22(7): 1662–1670.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.

Sun, J.; Wang, S.; Zhang, J.; and Zong, C. 2019. Towards sentence-level brain decoding with distributed representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7047–7054.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Tenney, I.; Xia, P.; Chen, B.; Wang, A.; Poliak, A.; McCoy, R. T.; Kim, N.; Durme, B. V.; Bowman, S.; Das, D.; and Pavlick, E. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Willett, F. R.; Avansino, D. T.; Hochberg, L. R.; Henderson, J. M.; and Shenoy, K. V. 2021. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858): 249–254.

Yi, G.; Wang, J.; Bian, H.; Han, C.; Deng, B.; Wei, X.; and Li, H. 2013. Multi-scale order recurrence quantification analysis of EEG signals evoked by manual acupuncture in healthy subjects. *Cognitive neurodynamics*, 7(1): 79–88.

Zanzotto, F. M.; and Croce, D. 2010. Comparing EEG/ERP-like and fMRI-like techniques for reading machine thoughts. In *International Conference on Brain Informatics*, 133–144. Springer.

Zhang, K.; Xu, G.; Zheng, X.; Li, H.; Zhang, S.; Yu, Y.; and Liang, R. 2020. Application of transfer learning in EEG decoding based on brain-computer interfaces: a review. *Sensors*, 20(21): 6321.