MindSync: Translating Brain Dynamics to Text with Discrete EEG Signal Encoding

CS-672 Advanced Deep Learning & Applications
Project Report

to be submitted by

Yash Sharma, B20241

Anurag Maurya, B20183

Avni Mittal, B20088

Under the guidance

of

Dr. Gaurav Jaswal
Human Computer Interaction (HCI), IIT Mandi

SCHOOL OF COMPUTER AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MANDI

KAMAND-175075, INDIA

April, 2024

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Decoding brain states into comprehensible representations has been a central focus of research in neuroscience and cognitive science. Electroencephalogram (EEG) signals offer a particularly attractive avenue for investigation due to their non-invasive nature and ease of recording. Traditionally, EEG decoding techniques have primarily concentrated on categorizing brain states into specific domains such as Motor Imagery (MI), Emotion, Robotic Control, and Gaming. However, these classifications, tethered to particular tasks, are limited in their capacity to facilitate broad-based brain-computer communication.

In recent years, there has been a surge of interest in harnessing EEG data for diverse applications, including natural language processing (NLP). EEG data, with its capacity to capture brain activity, provides a unique lens into human cognitive processes during linguistic tasks such as reading. A particularly intriguing area of exploration is EEG-based translation systems, which aim to directly translate brain signals into text. However, this endeavor presents numerous challenges, including the open-ended nature of language, inter-individual variability in EEG signals, and the imperative of real-time translation.

As the trajectory of research moves towards leveraging large language models (LLM) with increasingly generalized intelligence capabilities, there arises a pressing need to bridge the divide between brain signals and natural language representation. Despite recent advancements, this intersection remains relatively under-explored.

1.1 Motivation

Understanding human cognitive processes during linguistic tasks, notably reading, has long been a fundamental pursuit in cognitive science and neuroscience. EEG offers a unique opportunity to probe these processes in real-time by capturing the brain's electrical activity. Recent strides in EEG technology have facilitated the collection of high-quality EEG data during natural reading tasks, furnishing valuable insights into the neural mechanisms underpinning language comprehension.

One compelling application of EEG data lies in EEG-based translation systems, wherein EEG signals are directly translated into text. Such systems hold transformative potential across various domains, including assistive communication, brain-computer interfaces, and cognitive neuroscience research. However, constructing accurate and efficient EEG-based translation systems poses a formidable challenge owing to the intricate nature of both EEG signals and natural language.

Motivated by these challenges, we introduce MindSync, a novel EEG-based translation model that capitalizes on recent advancements in deep learning and EEG signal processing to accurately translate EEG signals into text. MindSync achieves state-of-the-art performance in EEG-to-Text translation tasks by employing a discrete codex-based approach.

1.2 Related Works

Recent advancements in decoding brain signals into text have seen a variety of methodologies attempting to tackle the inherent complexities of this task. Early approaches involved segmenting brain signals into fragmentary features, often leveraging external event markers such as handwriting or eye-tracking fixations. These methodologies operated at the word level, relying on a limited, closed vocabulary set, which posed constraints due to the misalignment between event markers and language output.

While these initial efforts primarily aimed at achieving high accuracy in decoding brain signals into text, they faced challenges when encountering semantically related words not present in the training set. In response, recent endeavors have shifted

towards transitioning from closed to open vocabulary approaches for EEG-to-Text sequence-to-sequence decoding. Notably, the emergence of large language models has enabled more flexible and adaptable decoding mechanisms.

Furthermore, recent breakthroughs in image generation using EEG data, exemplified by DreamDiffusion [2], have showcased the potential of leveraging advanced techniques such as the CLIP model. DreamDiffusion represents a significant advancement, employing CLIP to align EEG, text, and image spaces, thus enhancing the accuracy and quality of brain signal decoding. On the other hand, DeWave [1] introduces a novel pipeline leveraging VQ-VAE [3] to directly convert raw EEG signals into text, bypassing intermediate steps.

In addition to DreamDiffusion and DeWave, the OpenVocab [4] paper presents an alternative approach focusing on word-level EEG or EEG data supplemented with eye fixations. While OpenVocab has shown promise in certain contexts, it also faces limitations in handling raw EEG signals directly.

Despite these notable advancements, there remains a relative lack of recent developments in the domain of processing raw EEG signals directly into text. This underscores the necessity for continued exploration and innovation in this area to fully harness the potential of EEG-based translation systems and their applications in various domains such as assistive communication, brain-computer interfaces, and cognitive neuroscience research.

1.3 Objective

In this work, we aim to address the aforementioned challenges by proposing MindSync, an innovative EEG-based translation model. MindSync leverages recent advancements in deep learning and EEG signal processing to accurately translate EEG signals into text, with a focus on enhancing performance and robustness in real-world applications.

Chapter 2

Problem Definition and Algorithm

2.1 Problem Definition

Our proposed model, MindSync, is designed to undertake two primary tasks: Firstly, it deciphers open-vocabulary text tokens W from a given sequence of word-level EEG features E, collected during natural reading (Zuco Dataset). Secondly, it addresses two task settings:

1. **Word-level EEG-to-Text** Translation: Here, EEG feature sequences E are fragmented and re-ordered based on eye-fixation F, aligned with each word token w in sequence W.

2. **Raw EEG Waves to Text** Translation: In this scenario, EEG feature sequences E are directly converted into embedding sequences for translation without any event markers. This setting poses a more demanding challenge but is applicable in real-time scenarios.

2.2 Dicrete Codex

Discrete representation is first proposed in VQ-VAE [3]. DeWave is the first work to introduce discrete encoding into EEG signal representation. The discrete representation could benefit both the word-level EEG features and the raw EEG wave translation. Introducing discrete encoding into brain waves could bring two aspects of advantages.

1. It is widely accepted that EEG features have a strong data distribution variance across different human subjects. Meanwhile, the datasets can only have samples from a few human subjects due to the expense of data collection. This severely weakened the generalized ability of EEG-based deep learning models By introducing discrete encoding, we could alleviate the input variance to a large degree as the encoding is based on checking the nearest neighbor in the codex book.

2. The codex contains fewer time-wise properties which could alleviate the order mismatch between event markers (eye fixations) and language outputs.

Inference: Given the EEG waves E, it is first vectorized into embedding as introduced in Section 3.3 with ($X = \Theta(E, F)$) or without ($X = \Theta(E)$) eye fixations F, where X is the embedding sequence. A codex book $\{c_i\} \in \mathbb{R}^{k \times m}$ is initialized with number k of latent embedding with size m. The vectorized feature X is encoded into $z_c(X)$ through a transformer encoder. The discrete representation is acquired by calculating the nearest embedding in the codex of input embedding $x \in X$ as shown in Equation 2.1.

$$z_q(X) = \{z_q(x)\}_i, \quad z_q(x) = c_k, \quad k = \text{argmin}_j \|z_c(x) - c_j\|^2 \tag{2.1}$$

Different from the original VQ-VAE which decodes the original input, MindSync directly decodes the translation output given the representation $z_q(X)$. Given a pretrained language model, the decoder predicts text output with $P(W|z_q(X))$.

Learn: The codex is like a bridge connecting the vectorized EEG feature and the language model. Compared to learning direct EEG-to-text relation, MindSync learns a better discrete codex for the language model. It is easier to learn since we learn the discrete codex by the combination of the loss functions in three parts,

$$L = -\log(p(W|z_q(X))) + \|sg[z_c(x)] - z_q(x)\|_2^2 + \beta\|z_c(x) - sg[z_q(x)]\|_2^2 \tag{2.2}$$

where the loss maximizes the log-likelihood of language outputs $\log(P(W|z_q(X)))$

and minimizes the distance between latent variable z and the codex value z. Here, the sg denotes the stop gradients. The learning is robust for β from 0.1 to 2.0, where we set it as 0.2 throughout the training process.

2.3 Algorithm Definition

In this section we will discuss out proposed pipeline in brief which is also mention in 2.1.

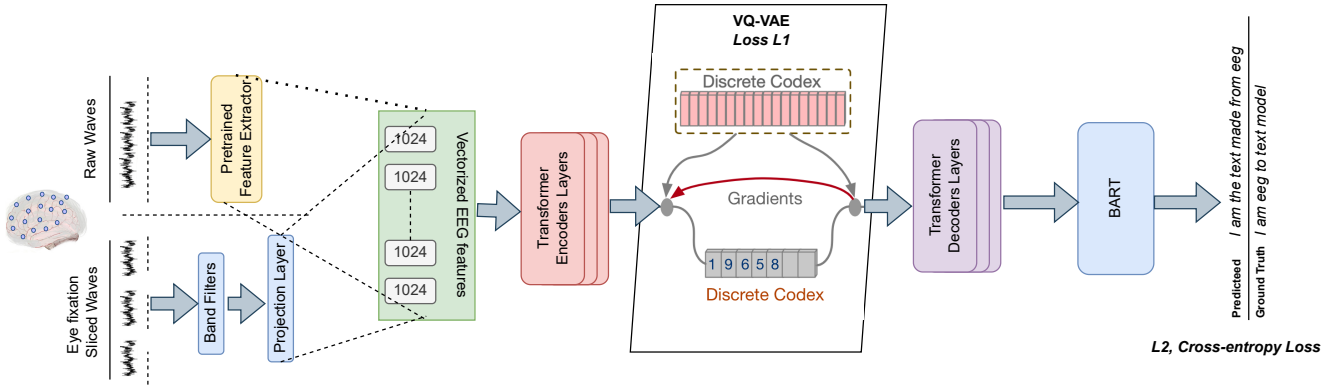

Fig. 2.1: Proposed Pipeline

2.3.1 Feature Extraction

Feature extraction is a crucial step in distilling relevant information from the raw and often noisy EEG signals, facilitating further processing. While word-level feature extraction has already been accomplished, the processing of raw EEG data necessitates the integration of a dedicated feature extractor into our pipeline. To address this, we employed a pretrained feature extractor, meticulously trained on a vast corpus of EEG data. Our decision to adopt this approach was influenced by the work of Bai et al. [2].

The feature extractor operates by ingesting the raw EEG wave, typically presented in a (128, #time) dimensional format, and outputting a sequence of extracted features, each comprising 1024 dimensions. This extraction process serves the dual purpose of noise reduction within the EEG signals and facilitating the comprehension of the provided EEG data. Notably, this step is unnecessary for word-level features, as they

have already undergone processing and band-wise filtration.

2.3.2 Vector Quantized Variational Autoencoder (VQ-VAE)

Moving forward, both the extracted features, whether from raw EEG or word-level, are directly inputted into a Transformer Encoder to enhance their representation. Subsequently, these representations are passed through the VQ-VAE block, as described in Section 2.2, aiming to improve the quantized representation of the input. Consequently, discrete EEG representations are learned. We define the loss incurred during this step as L_1, which is minimized during the training process.

$$L_1 = -\log(p(W|z_q(X))) + \|\text{sg}[z_c(x)] - z_q(x)\|_2^2 + \beta\|z_c(x) - \text{sg}[z_q(x)]\|_2^2 \tag{2.3}$$

The decoder endeavors to reconstruct the input from the obtained quantized output. The codebook is learnable, which is updated during the training process.

2.3.3 Bidirectional and Auto-Regressive Transformers (BART)

We employed the pre-trained BART model [**?**] on a large-scale text corpus as the generative language model for translation output. Given the limited EEG-to-text translation data, leveraging the BART model could incorporate prior knowledge of text relations. In such a scenario, the translation system only needs to learn a codex representation for the language model, simplifying the learning process. The codex representations are fed into the pre-trained BART model, and the output hidden states are obtained. A fully connected layer is applied to the hidden states to generate English tokens from the pre-trained BART vocabulary V.

Moving forward, the quantized EEG features are utilized to fine-tune the large BART model (*facebook/bart-large*). The learning rate in this case is kept small to avoid significant alterations to the pre-trained weights. The cross-entropy loss, denoted as L_2, is calculated between the predicted and ground truth texts.

2.3.4 Loss Combination

The two aforementioned losses become integral components of the total loss, denoted as L, which is backpropagated during the training of the complete pipeline: $L = L_1 + \alpha \cdot L_2$, where $0 < \alpha < 1$.

2.4 Training Paradigm

MindSync undergoes a multi-stage training process, bifurcated into two stages. In the initial stage, the language model is not involved in weight updates. The objective of this stage is to train a proper encoder projection θ_{codex} and a discrete codex representation C for the language model. In the subsequent stage, the gradient of all weights, including the language model θ_{BART}, is opened to fine-tune the entire system. The latter stage employs a considerably lower learning rate compared to the first stage.

Chapter 3

Experimental Evaluation

3.1 Dataset

We utilized the ZuCo datasets [5], which include EEG and Eye-tracking data collected during natural reading tasks, encompassing Normal Reading (NR) and Task-Specific Reading (TSR). The reading materials are sourced from movie reviews and Wikipedia articles. Normal reading tasks with movie reviews are labeled with sentiment (positive, neutral, or negative). The dataset is divided into multiple versions, each containing a varying number of unique sentences and corresponding training and testing samples.

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 3.1: Dataset Statistics. SR: Normal Reading (sentiment), NR: Normal Reading (Wikipedia), TSR: Task Specific Reading (Wikipedia). Data from 12 subjects in v1.0 and 18 subjects in v2.0.

We aggregated word-level EEG feature sequences based on gaze duration (GD). The EEG input data underwent a cleaning process where sentences containing NaN values were omitted. Subsequently, we partitioned the data from each reading task into train, development, and test sets, with proportions of 80%, 10%, and 10%, respectively, ensuring that the sentences in the test set were entirely unseen during training. All

the results are obained using the dataset *SRv1.0, NRv1.0 and NRv2.0.*

3.2 Results

We assess translation effectiveness by employing NLP metrics, specifically BLEU and ROUGE, as depicted in Table 3.2. In our examination of word-level EEG features, we ensure fairness by comparing our outcomes with EEG-to-Text, using a consistent language model.

Source	**Model**	**BLEU-N**				**ROUGE-1**		
		N=1	**N=2**	**N=3**	**N=4**	**R**	**P**	**F**
Word-Level features	EEG-to-Text	27.84	23.18	12.61	6.80	28.84	31.69	30.10
	DeWave**	41.35	24.15	13.92	8.22	28.82	33.71	30.69
	MindSync(Ours)	**29.11**	**15.7**	**8.4**	**4.2**	**34.39**	**34.02**	**33.8**
Raw waves	EEG-to-Text	13.07	5.78	2.55	1.10	15.22	18.08	16.36
	Wave2Vec	18.15	8.94	3.89	2.04	18.96	23.86	20.07
	DeWave	20.51	10.18	5.16	2.52	21.18	29.42	24.27
	MindSync (Ours)	**29.6**	**15.65**	**8.343**	**4.33**	**33.225**	**33.266**	**32.811**

Table 3.2: Evaluation metrics of EEG-to-Text translation under both word-level features input and raw waves input. For a fair comparison, these results keep the same teacher-forcing evaluation setting as EEG-to-Text. DeWave [1] has officially withdrawn its submission due to errors in its evaluation scheme.

Generated Samples

Decoding Results with word-level features

Predicted: Baldwin was young, he began to Unitedidas Utah to where a of a fatherboy foreign exchange program.

Ground truth: When Bush was seventeen, he went to Leon, Mexico, as part of his school's student exchange program.

Predicted: Was in the United States Army from a Republican from Wyoming from was((((((

Ground Truth: He served in the United States Senate as a Republican from Minnesota.

Predicted: was married president executive at and at at at the Louisiana of Wyoming West of Columbia. and also the the of the College End Columbia's Department. program..((((((((((

Ground Truth: He later became an educator, teaching music theory at the University of the District of Columbia; he was also director of the District of Columbia Music Center jazz workshop band.

Decoding results on raw waves

Predicted: Was the member member of the H family, and son brother of President George W. Bush and the grandfather cousin of President President John H. W. Bush. The Bush. isgggggggggg))))))

Ground Truth: He is a prominent member of the Bush family, the younger brother of President George W. Bush and the second son of former President George H. W. Bush and Barbara Bush.

Predicted: Film of movie that should youeless, a bad grade, is is is is (((((((()))))))) r r

Ground Truth: The sort of movie that gives tastelessness a bad rap.

Predicted: was in the United States Army from a Republican from 1955 from is is is (((((((((

Ground Truth: He served in the United States Senate as a Republican from Minnesota.

Chapter 4

Conclusion and Future Work

In this paper, we introduce MindSync, a pioneering model for EEG-to-text translation, which addresses two challenging tasks: open vocabulary [4] EEG-to-text sequence decoding and EEG-based sentence sentiment classification. Leveraging novel frameworks built upon pretrained language models, MindSync exhibits remarkable scalability, outperforming existing models on raw EEG data with a notable BLEU-1 score of 30, surpassing the current best of 21.

Looking ahead, future endeavors should focus on gathering larger-scale EEG-text datasets and extending the current framework to multilingual settings. Drawing inspiration from previous studies on brain regions involved in language processes, such as Broca's area, we envision applying MindSync to decode inner speech in an open vocabulary setting. This direction necessitates dedicated datasets with expanded vocabulary coverage, considering the limitations of current inner speech decoding datasets.

While our work builds upon the concepts of open-vocabulary translation and discrete codex encoding introduced by DeWave [1], it expands the task to decode raw EEG waves without relying on eye fixation markers. Despite these advancements, the quality of brain decoding results remains a significant challenge, particularly in a teacher forcing setting for fair comparison.

Moving forward, our ongoing research aims to explore more realistic settings that remove teacher forcing for both training and testing. Additionally, we plan to incorporate a "neural-feedback" mechanism into EEG-to-text research to further enhance

its scientific value. By drawing inspiration from both open vocabulary approaches and DeWave, we aim to push the boundaries of EEG-to-text translation, advancing our understanding of brain-computer interfaces and their potential applications.

4.1 Limitations

Despite the progress achieved by MindSync in EEG-to-text translation, leveraging Dream Diffusion instead of Wave2Vec for feature extraction from raw EEG waves, its accuracy still falls short of real-life scenarios compared to traditional language-to-language translations. Additionally, to maintain a fair comparison with EEG-to-text methods, this paper adopts a teacher-forcing setting during evaluation. While this setting simplifies the task by eliminating accumulation errors and transforming sequence decoding into a word-level classification task, it may not fully capture the complexities of real-world applications.

Moreover, the experiments conducted in this paper are restricted to publicly available neural reading data, which may not fully align with the concept of "silent speech" and direct thought translation from human brains. The current ZuCo dataset is gathered by presenting reading stimuli to participants, thus focusing on decoding text from external stimuli rather than extracting thoughts directly from the brain.

This paper primarily focuses on introducing the Dream Diffusion approach for feature extraction from raw EEG waves and proposes the use of discrete codex representations for EEG-to-text translation. However, a key scientific challenge in this domain remains finding better methods for accomplishing the "silent speech" task. Our ongoing research is dedicated to exploring solutions to this problem.

4.2 Code Availability and Experiment Logs

The code implementation of MindSync, along with scripts for running experiments and reproducing results, can be found on GitHub at the following link: `https://github.com/yashcode00/eeg2text`.

Additionally, experiment logs containing detailed information about loss metrics, model performance, and other relevant experiment data are available on Weights & Biases (wandb) platform. The experiment logs can be accessed and visualized using the following link: `https://wandb.ai/b20241/ben10?nw=nwuserb20241`.

We encourage readers interested in further exploration or replication of our work to refer to the provided resources on GitHub and Wandb.

Bibliography

[1] Y. Duan, J. Zhou, Z. Wang, Y.-K. Wang, and C.-T. Lin, "Dewave: Discrete eeg waves encoding for brain dynamics to text translation," 2024.

[2] Y. Bai, X. Wang, Y. pei Cao, Y. Ge, C. Yuan, and Y. Shan, "Dreamdiffusion: Generating high-quality images from brain eeg signals," 2023.

[3] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," 2018.

[4] Z. Wang and H. Ji, "Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification," 2024.

[5] N. Hollenstein, M. Troendle, C. Zhang, and N. Langer, "Zuco 2.0: A dataset of physiological recordings during natural reading and annotation," 2020.