Thought2Text - Semantic Language Generation from non-Invasive Neural Activity Readings

Rajath Rao, Pranav Chitale, Ashutosh Tiwari

1 Introduction

Speech impairments affect a significant portion of the global population, with varying prevalence rates across regions and age groups. The World Health Organization (WHO) estimates that approximately 1% of the world's population has moderate to severe speech impairments, while up to 10% experience some form of communication disorder during their lifetime [1]. We propose a framework to assist individuals with paralysis-induced speech impediments, driven by the urgent need to enable effective self-expression for these individuals.

Our framework, Thought2Text, addresses the sequence-to-sequence task of extracting nuanced semantic representations from textual inputs. Humans tend to process words in chunks than treating words unlike traditional language models that rely on tokenizers. And so, we generate candidate sequences of tokens by considering historical context tokens and the user's brain state (fMRI response).

We delve into semantics and transformers, experimenting with fine-tuning *distilgpt2* to auto-regressively extract stimulus features from text for encoding fMRI responses. Our exploration involves classifying fMRI responses using neural networks, considering the potential of non-linear neural patterns between textual stimuli and fMRI responses. We explore the use of a multi-layer perceptron (MLP) instead of regularized L2 ridge regression for this purpose.

2 Background

In mid-2023, UT Austin researchers developed a framework to predict words a subject heard or imagined from fMRI data. The system interprets the data to produce textual representations summarizing what the subject heard [2]. Meta AI resumed similar work [3, 4] where magnetoencephalogram (MEG) waves were used for the encoding-decoding task. This method is currently state-of-the-art due to the rich temporal resolution that MEGs provide.

Our focus is on generating semantic language from non-invasive neural activities, particularly leveraging functional magnetic resonance (fMRI) responses for their rich spatial information about the brain [2]. An fMRI is a voxel-based data type that contains spatial and temporal information about brain activity, acquired through electromagnetic medical imaging techniques that measure changes in Blood Oxygen Level Dependent (BOLD) within the brain [5]. fMRI responses have limited temporal resolution (0.5 Hz or 10 seconds), while humans speak at around 3 words per second [2, 4]. To address this, we use Language Models conditioned on temporal brain responses, akin to prior user states in Human Language Modeling.

3 Data

The data for this project came from two primary sources. The first, was a Deep fMRI Dataset was curated at UT Austin by the authors of [2, 5] and is available on OpenNeuro UT Austin. The second dataset was the CNNDailyMail [6, 7] dataset from HuggingFace.

3.1 Deep fMRI Dataset

The Deep fMRI Dataset [8] involved an extensive study with 8 independent subjects listening to stories from an audiobook while their fMRI responses were scanned. The audiobooks contained 153,908 words, totaling 165,678 tokens. Each subject spent approximately 6.4 hours listening, encompassing 219,511 audible phonemes. Additionally, the study recorded fMRI responses while subjects imagined speech, not just during perceived speech.

3.2 CNNDailyMail

The CNNDailyMail dataset consists of story-like documents composed of news articles from CNN and the Daily Mail [6, 7]. Each element includes an **id**, a **full text article**, and **highlights** summarizing the article. For our project, we use individual articles as documents for fine-tuning our model. With an average length of 786 tokens per article, we work with the first 10,000 articles from the training dataset, providing over 7M tokens. We chose this dataset for its story-like nature, relevant to the auditory stimuli in our experiments.

Table 1: Dataset Information

Dataset	**Statistics**
Deep fMRI Responses	81126 fMRI voxels, (2.66×2.66×2.6) mm voxel size
Deep fMRI Stories	6.4 Hours, 153908 words, 219511 phonemes
CNNDailyMail	287,113 articles, 781 mean tokens

4 Methodology

4.1 Overview

Throughout this project, we primarily follow the methodology of [2] and update individual components based on experimental results. Our model card for our approach can be found in Figure 1 The paper presents an encoder-decoder framework where encoder weights condition the decoder for auto-regressive text generation. The decoder incorporates modules such as the Encoder Model (EM), Language Model (LM), and Beam Search algorithm. Given an fMRI response:

1. We start with empty candidate sequences.
2. The LM proposes a set of next tokens for each sequence. New candidate sequences are formed by appending the tokens to respective sequences.
3. The EM transforms the word embeddings of these potential candidate sequences to BOLD feature space.
4. We shortlist candidate sequences for the next iteration using Beam Search based on their similarity with the actual fMRI BOLD response.

4.2 Encoder Model (EM)

The Encoder Model as seen in Figure 2 plays a key role in the Thought2Text process by learning a transformation from semantic features to predicted BOLD responses. We use two different approaches for EM.

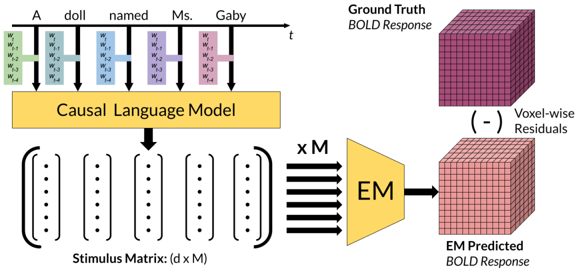

Figure 2: Encoder Model Diagram

We employ a Ridge Regression model following [2]. Inputs are semantic features from the Language Model (LM), each 768 in size, corresponding to feature vectors from 2s, 4s, 6s, and 8s earlier. Concatenated into a 3072-sized vector, these serve as input to the regression model. Each input maps to one of 81126 BOLD response features, resulting in weights of size (3072, 81126). This facilitates assessing the similarity between actual and predicted BOLD responses in the decoder.

Secondly, we propose employing a Multi-Layer Perceptron (MLP), which maps the same inputs to outputs, incorporating non-linear activations in the network. We argue that thoughts are complex constructs, and non-linear models can better capture these feature spaces.

Thought2Text - Model Card

- **Model Details**
  A DistilGPT2 [9] GPT Transformer with the Ridge Regression Layer replaced with MLP
- **Intended Use**
  Intended to create BOLD mappings from textual input, providing a contextual representation of the text.
- **Factors**
  fMRI readings are unique to an individual.
- **Metrics**
  Evaluation metrics comprise Mean Squared Error and Mean Absolute Error, along with the utilization of WER, BLEU, METEOR, and BERT scores.
- **Training Data**
  - CNNDailyNews [6, 7]
  - TextGrids [2]
- **Evaluation Data** Perceived Speech BOLD voxels [8]
- **Ethical Considerations** Unconsented analysis of BOLD voxels has the potential to compromise mental privacy
- **Caveats and Recommendations** Outputs are personalized to an individual's fMRI scans, and obtaining fMRI data can be challenging due to its susceptibility to noise.
- **Qualitative Analysis**
  - Better performance with GPT2 + MLP compared to baseline GPT1 + Ridge for **MSE** and **MAE** evaluation
  - Very poor performance improvements compared to baseline for the scoring methods.

Figure 1: Model Card

4.3 Language Model (LM)

We aim to investigate the use of fine-tuned *DistilGPT2* [9] to extract stimulus features from perceived speech. In the decoder, this model will be employed for Causal Language Modeling (CLM), generating short bursts of sequence tokens instead of next-token prediction. *DistilGPT2* is a distilled version of GPT-2 [10] with 88.2M parameters, compared to GPT-2's 137M.

We fine-tune *DistilGPT2* on 10,000 articles from the CNNDailyMail dataset. In contrast, previous work [2] used GPT-1 [11], fine-tuning it on an unreleased dataset of 200 million words from Reddit posts and 240 autobiographical stories.

As a feature extractor, the LM computes rich semantic representations from stimulus words. Specifically, for DistilGPT2, we use the second last transformer block's hidden activations for each stimulus word. These layers offer superior semantic representations compared to the last layer, which is more task-oriented. DistilGPT2 is also employed for CLM in the Decoder to generate candidate sequences one token at a time.

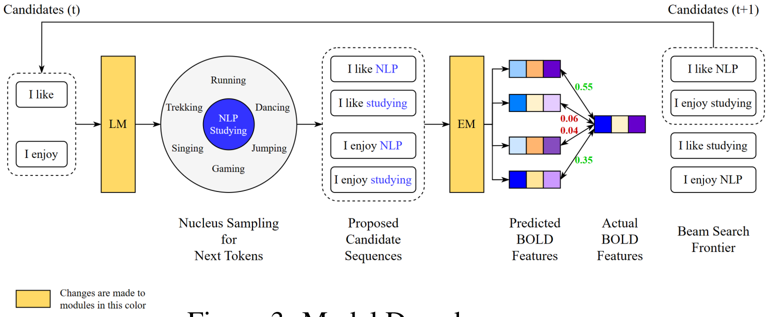

Figure 3: Model Decoder

4.4 Beam Search

Beam search selects from proposed candidate sequences. Following [2], we utilize Nucleus

Table 2: Ablation Study Encoder Results

Model	Metric	
	MSE	MAE
Baseline:	...	...
- GPT1 + Ridge Regression	0.939	0.757
Improvement:	...	...
- GPT1 + MLP	0.774	0.695
- GPT2 + Ridge Regression	0.978	0.773
- GPT2 + MLP	0.693	0.657

Table 3: Ablation Study Decoder Results

Model	Metric			
	WER	BLEU	METEOR	BERT
Baseline:	...	...	...	...
- GPT1 + Ridge	12.6	5.75	6.45	12.8
Improvement:	...	...	...	...
- GPT1 + MLP	1.92	-0.55	1.24	2.75
- GPT2 + Ridge	-13.1	-22.3	-16.0	-33.9
- GPT2 + MLP	...	...	...	...

sampling [12] for selecting the next words in our decoder as shown in Figure 3. Nucleus sampling, also called top-p sampling, ensures diverse yet coherent output by selecting the most probable words below a dynamic threshold ”p”. This improves upon traditional sampling methods. Next tokens are extracted from the nucleus and appended to current candidate sequences. The likelihood of a candidate sequence ($P(R|S)$) as the actual thought is determined by the EM converting sentences to BOLD features. The sentence with the least residual BOLD differences is chosen. Based on beam size, the next set of candidate sequences is selected from those with the least residual BOLD differences.

5 Results

In this section, we delve into the baselines and enhancements. Our experiments focus on employing different methods for Language Modeling (LM) and Encoding Model (EM) compared to those used in [2]. We concentrate on the task of perceived speech and evaluate the models based on Subject 1’s responses.

We evaluate encoder performance using Mean Squared Error (MSE) and Mean Absolute Error (MAE) with a baseline of GPT-1 LM + Ridge Regression (RR) EM from [2] code[1]. We experiment with GPT-2 LM and MLP-based EM, detailed in Table 2. DistilGPT-2 + MLP shows the most significant improvements, enhancing MSE and MAE by 26.2% and 13.2%, respectively. This enhancement is attributed to DistilGPT2’s superior capability with fewer parameters, resulting in improved semantic representations. An MLP-based EM proves more effective than simple ridge regression, leveraging multiple layers and non-linear outputs for better mapping. These improvements are illustrated in Figure 4.

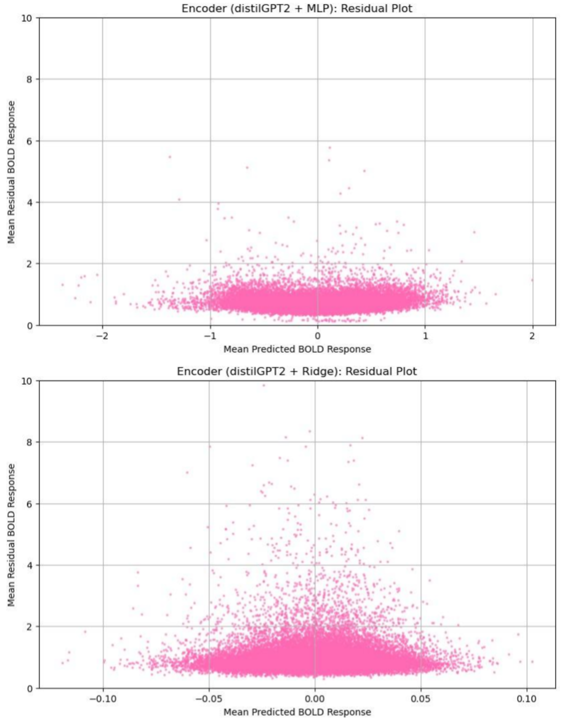

Figure 4: *(Top)* residual plot for distilGPT2 + Ridge Regression. *(Bottom)* residual plot for distilGPT2 + MLP for encoder predictions.

Next, we examine the Decoder, employing Beam Search with LM and EM. We utilize the same baseline models as the Encoder evaluation. The metrics, including WER, BLEU ($n = 1$),

[1] https://github.com/HuthLab/semantic-decoding

METEOR, and BERT, are obtained using the code from [2]. Notably, we observe a substantial 84.7% improvement in Word Error Rate (WER), although other metrics show relatively poor performance. We attribute this to BOLD features lacking temporal information about thoughts, resulting in inaccurate word ordering. This discrepancy leads to higher unigram-overlap BLEU scores but also higher WER, indicating ordering inaccuracies. We speculate that our models encounter a similar issue.

An unresolved issue with DistilGPT2 leads to premature End of Sequence (EOS) tokens during beam search, preventing us from obtaining results. While we could evaluate it for the Encoder module, which doesn't involve text generation, we encountered difficulties with the Decoder module.

6 Conclusion

We've successfully shown that replacing the final Ridge regression layer of both GPT1 and DistilGPT2 models with an MLP layer enhances the encoding of text semantics into fMRI BOLD voxels. Notably, despite having fewer parameters than GPT1, DistilGPT2 excels in extracting deep semantic representations from textual stimuli.

References

[1] World Health Organization et al. *World report on disability 2011*. World Health Organization, 2011.

[2] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

[3] Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023.

[4] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812*, 2023.

[5] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

[6] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[7] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701, 2015.

[8] A LeBel, L Wagner, S Jain, A Adhikari-Desai, B Gupta, A Morgenthal, J Tang, L Xu, and AG Huth. An fmri dataset during a passive natural language listening task. *OpenNeuro. doi*, 10, 2021.

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*, 2019.

[10] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[11] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[12] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020.