Towards A Language Model for Brain Activity Recordings

Jans Solano [1], Josue Ortega [2], David van Dijk[2]

[1]Engineering Department, Pontifical Catholic University of Peru
[2]Wu Tsai Institute, Yale University

Background

In neuroscience, functional Magnetic Resonance Imaging (fMRI) is a powerful tool for studying how cognition and behavior arise from brain activity. It uses magnets and radio waves to detect changes in blood flow, producing brain activity images. Due to its complex spatiotemporal dynamics, extracting clinical insights from fMRI is challenging.

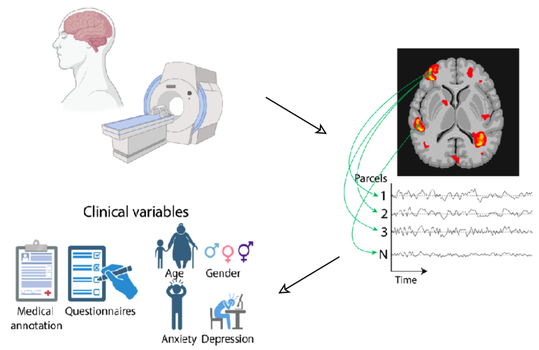

Figure 1. Functional Magnetic Resonance Imaging (fMRI) recordings background.

As shown in Figure 1, fMRI data can be segmented into signals by brain regions. While recent models use this to predict clinical variables, they lack the text comprehension capabilities of a Large Language Model (LLM).

Objectives

- Analyze the performance of Visual Language Models in fMRI clinical variable prediction.
- Train a quantized image modeling framework for fMRI reconstruction.
- Employ adversarial training to align text and fMRIs in a shared vocabulary.
- Fine-tune a LLM with the shared vocabulary as tokenizer for clinical variable prediction.

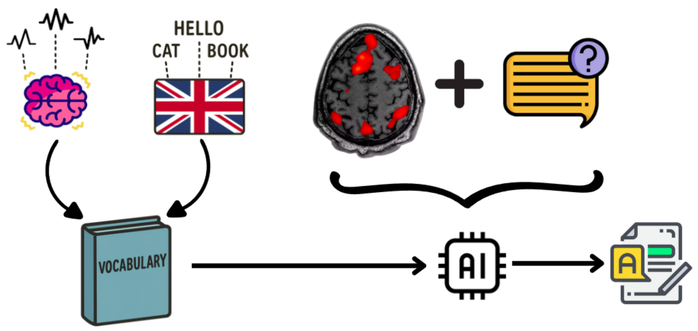

Figure 2. Overview of the method.

Visual Language Models Performance

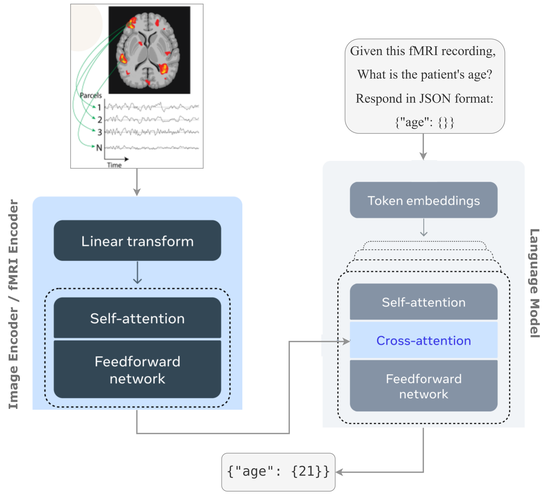

Figure 3. Multimodal architecture for clinical variable finetuning.

For this analysis, two Vision-Language Models (VLMs) were evaluated: **LLaMA 3.2 11B** and **Qwen 7B**. We assessed each model's performance in predicting the patient's age after fine-tuning, either using the pre-trained VLM visual encoder or an fMRI encoder paired with **GPT-2** as the language model. Our results in Figure 4 show that:

1. LLaMA's output distribution is narrow, likely due to tokenizer and limited numerical pretraining data.
2. Qwen outperforms LLaMA despite being smaller and predicts a wider age range.
3. fMRI encoders yield better performance than VLM visual encoders, showing the data's complexity.

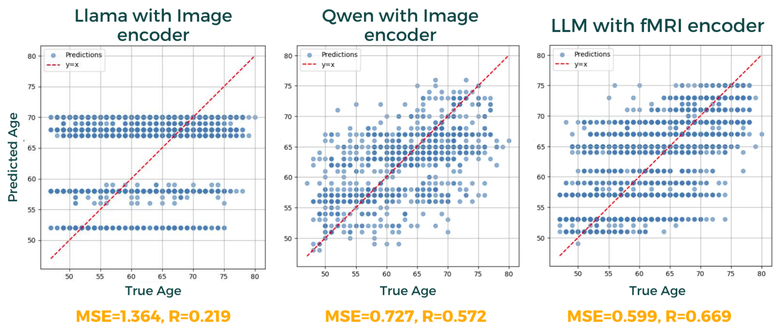

Figure 4. Comparative results of image and fMRI encoder.

From this comparative study, we conclude that a specialized fMRI encoder is necessary for a better understanding of fMRI dynamics. Therefore, in the following results, we adopt an image reconstruction framework.

Adversarial training for Text-fMRI alignment

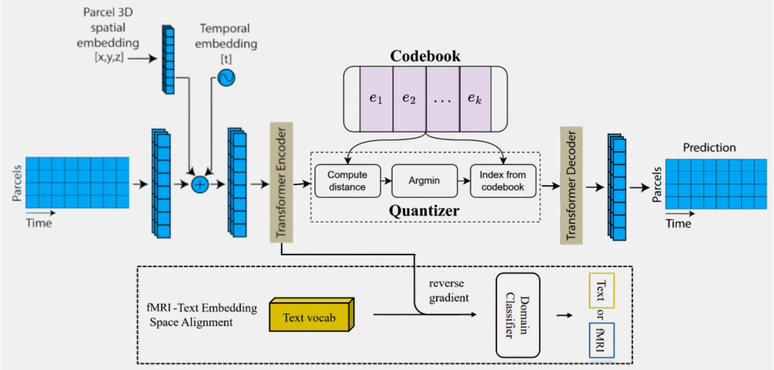

Figure 5. Adversarial Vector Quantized Autoencoder architecture.

The architecture consists of an **Autoencoder** for learning complex representations of fMRI, a **Vector Quantizer** for discretizing them and creating a vocabulary, and a **Domain Classifier** for modality alignment. Our results show that:

1. An R^2 of 0.25 was achieved, which is 0.03 lower than the baseline model without the domain classifier.
2. The discriminator achieved an accuracy of approximately 50%, indicating confusion and thus alignment between the text and fMRI representations.

Conclusions

- A specialized fMRI encoder is necessary for an accurate encoding learning.
- Vector Quantizer enables efficient data compression and fast encoding.
- Adversarial training enables simultaneous feature representation and cross-modal alignment.

Future work

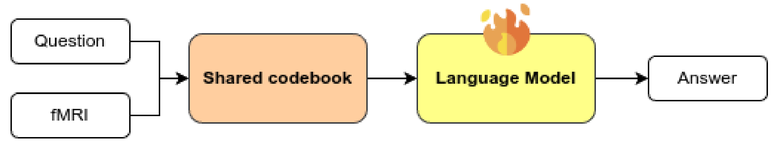

Figure 6. Finetuning procedure with shared codebook.

References

[1] A. van den Oord et al, "Neural discrete representation learning," in *International Conference on Neural Information Processing Systems*, 2018.

[2] K. H. et al, "Masked autoencoders are scalable vision learners," in *Conference on Computer Vision and Pattern Recognition*, 2021.

[3] W. J. et al, "Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals," in *International Conference on Learning Representations*, 2025.

[4] J. O. C. et al, "BrainLM: A foundation model for brain activity recordings," in *International Conference on Learning Representations*, 2024.

luis.vega@yale.edu