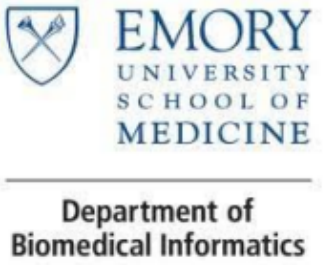

GSoC Proposal for Department of Biomedical Informatics, Emory University

Project Title: Development of an Open-Source EEG Foundation Model

Shreyas S

VIT University, India

Github: https://github.com/Zhreyu

Email: Zhreyas1@gmail.com

Mentors: Mahmoud Zeydabadinezhad, Babak Mahmoudi

Table of Contents:

1. Personal Introduction

1.1 Introduction | Skills

My name is Shreyas, and I'm a sophomore at VIT University AP pursuing an integrated master's degree in computer science. Deep learning has captivated me since the beginning of my studies. I find the concept of teaching machines to tackle complex problems incredibly fascinating. My particular interest lies in the application of deep learning and AI within the MedTech field. As a research-oriented student with experience in deep learning methodologies, I'm constantly seeking opportunities to expand my knowledge and contribute meaningfully to MedTech research.

- **Technologies:** Python,R programming, PyTorch, scikit-learn, Seaborn, Pandas, Numpy, MNE.
- **Skills:** Deep Learning, Data Pipeline Development, Prompt Engineering, Data Analysis, Large Language Models

1.2 Previous Experience

1. **Dubsync.ai** (AIR VIT-AP)

Deep learning and Data Pipeline Intern (Mar 2023 - June 2023)

• Successfully upgraded CodeFormer with ESRGAN 4x architecture, resulting in a 40% improvement in image reconstruction accuracy.

• Implemented a data downloading channel in Python that efficiently retrieved and processed over 3000+ videos, reducing data acquisition time by 50%.

2. **DigitalFortress Private Limited**

AI/ML Researcher (Sep 2023 - Mar 2024)

• Implemented a natural language query agent to efficiently retrieve data from a MongoDB collection utilising Natural Language Processing and prompt engineering LLMs resulting in improved user interaction and data retrieval.

• Developed and implemented pipelines for data preparation and testing for a novel anti-spoof detection model by combining 3 datasets totaling over 55GB, utilising various filters to enhance data quality.
• Designed a custom loss function to optimise model performance, resulting in a 15% increase in accuracy during testing on cross-datasets.

1.3 Why Emory BMI?

My passion for deep learning, particularly in MedTech field, aligns perfectly with Emory BMI's commitment to open-source development in bioinformatics. With my computer science background, I can contribute to their ongoing research while exploring my interest in this field. Emory BMI's impressive track record in GSoC, coupled with their focus on fostering long-term collaboration, makes them an ideal environment for a research-oriented student like myself.

2. Abstract

This project aims to develop an open-source foundational model dedicated to the analysis of EEG (Electroencephalography) data. Such a foundational model is characterised by its extensive pre-training on a broad spectrum of data, enabling it to acquire a comprehensive understanding of that data, regardless of its specific labelling.
This becomes particularly valuable in scenarios marked by a lack of detailed EEG data for specific tasks. The project's focus will be on creating algorithms for the effective processing of EEG signals,feature extraction, and the application of deep learning techniques for pre-training the model using publicly available EEG datasets. The ultimate goal is to develop a foundational EEG model and make it open source.

2.1 State of the art

Currently, there are a few innovative models in large scale EEG data analysis, These models leverage extensive datasets and self-supervised learning techniques to decode the complexities of brain signals. However, they are not without their limitations. One notable challenge is their difficulty in

handling the diverse nature of EEG data and presenting their findings in an easily understandable manner. This issue is particularly pronounced in the analysis of smaller, more specific datasets, where applying the generalised knowledge acquired from broader data proves challenging. Consequently, there remains a degree of uncertainty regarding the applicability of the features learned by these models to varied EEG datasets or tasks that were not part of the initial testing scope.

3.Project

3.1 Objective

In this project, my objective is to develop a foundational EEG model, drawing inspiration from advanced models like Neuro-GPT and EEGFORMER. The focus will be on enhancing the model's ability to effectively interpret and adapt to diverse EEG data. This includes refining the model's proficiency in handling both large-scale and more specific, smaller datasets. The aspiration is to blend the sophistication of current EEG analysis techniques with increased adaptability and user accessibility.

3.2 Approach

In my approach, I will begin by reviewing existing EEG analysis techniques, exploring, and experimenting with various algorithms/architectures suitable for EEG data on a small scale to test their effectiveness in adapting to the complexities of EEG data. Once the most suitable algorithm or architecture is identified, I will proceed with acquiring EEG data, followed by conducting an in-depth Exploratory Data Analysis (EDA). This step is very important for gaining a good understanding of the data's unique patterns and characteristics. With these insights, I will move on to data pre-processing and model construction. The final stage will involve model fine-tuning, ensuring its ability to interpret and analyse EEG data across various use cases.

It's important to acknowledge that this approach may necessitate adjustments based on the data and the project's evolving nature. I remain committed to a flexible and adaptable research methodology, continuously

evaluating and refining my approach as new information arises or challenges are encountered. This iterative process allows for the optimization of the model's development and ultimately leads us towards achieving the desired outcomes. I will regularly evaluate my progress and stay prepared to refine or even change our chosen methods. Engagement with mentors and the community will be key to navigating these decisions.

3.3 Anticipated Challenges:

Data Heterogeneity and Quality: EEG data can vary greatly in quality and format, depending on the source and collection methods. This heterogeneity poses a challenge in creating a universally applicable model.
Mitigation: To address this, we will establish robust preprocessing pipelines that standardise data formats and apply noise reduction techniques. Additionally, we will use a diverse dataset, such as the TUEG Corpus, to ensure our model can adapt to different types of EEG data.

Model Overfitting: Given the complexity and variability of EEG signals, there's a risk of overfitting our model to specific dataset characteristics.
Mitigation: We will employ regularization techniques and Dimensionality Reduction strategies to prevent overfitting. Later Cross-validation will be used to ensure the model's generalizability across different datasets.

4.Timeline:

From May 10th to May 20th, I will be in the midst of my end-semester exams, which means I will be able to commit only around 10-15 hours per week to the project during this period. However, I plan to compensate for this reduced commitment in the upcoming summer break starting May 26th. For the two months of summer, I intend to dedicate full-time hours, approximately 40-45 hours per week, to ensure steady progress and make up for the time during the exams.

4.1 Community Bonding Period

Between May 1st and May 26th, I will be laying the groundwork for my EEG project. Here's how I plan to unfold this phase:

- I will set up my development environment, ensuring all necessary software and tools are in place for efficient workflow.
- During this time, I will apply for access to relevant EEG datasets and handle any other prerequisites required for the project.
- I will strategize and plan the execution of tasks, laying out a clear roadmap to guide the project's progression.
- I plan to start a blog to document my journey, capturing the challenges and milestones, which will serve as a valuable reference for the community and peers.
- Additionally, I will explore initial algorithms, engaging with my mentors to finalise the best approach to take forward for the EEG foundational model.
- I will also discuss with my mentors and the community to outline evaluation criteria for selecting the most suitable deep learning architecture. This will ensure our project remains adaptable and on the right track

4.2 Development Phase

Week 1 (May 27 - June 2):

- Refine Initial Model Choice: Briefly revisit and solidify the chosen model architecture based on its suitability for handling diverse EEG data.
- Begin EDA: Explore EEG datasets to understand their characteristics, identify potential noise patterns, and explore frequency bands relevant to your use case.

Week 2 (June 3 - June 9):

- Continue EDA & Feature Engineering: Leverage EDA insights to identify relevant features from the EEG data and potentially implement feature engineering techniques if necessary.

Week 3 (June 10 - June 16):

- Model Utility Development: Develop utility functions and scripts specifically designed to assist with the training process of your EEG model.

Week 4 (June 17 - June 23):

- Data Pre-processing Development: Focus on developing data loaders for efficient data loading and pre-processing pipelines based on insights from EDA and feature engineering.

Week 5 (June 24 - June 30):

- Model Implementation: Begin building the foundational model framework based on the chosen architecture and incorporating findings from EDA and feature engineering.

Week 6 (July 1 - July 7):

- Initial Model Testing: Test the model with initial datasets to assess its baseline performance and identify areas for improvement.

Week 7 (July 8 - July 14):

- Test Result Analysis & Optimization Planning: analyse test results and pinpoint areas where the model can be optimised based on performance metrics and adaptation to diverse data.

Week 8 (July 15 - July 21):

- Model Optimization & Architecture Enhancements: Implement optimizations and potentially refine the model architecture based on the analysis from Week 7.

Week 9 (July 22 - July 28):

- Fine-Tuning Data Preparation: Acquire and pre-process data for fine-tuning the model.

Week 10 (July 29 - August 4):

- Fine-Tuning: Conduct fine-tuning by adapting the pre-trained model to the new, low-level dataset acquired in Week 9.

Week 11 (August 5 - August 11):

- Testing & Refinement: Continue the iterative process of testing the fine-tuned model analysing results and implementing further optimizations.
- Codebase Cleanup: This section focuses on ensuring the code is well-structured, readable, and includes comments for future reference and maintainability.

Week 12 - 13 (August 12 - August 25):

- Model Consolidation: Finalize the core structure of the foundational EEG model based on the iterative testing and optimization process.
- Final Testing & Reporting: Conduct final rounds of testing with various use cases and metrics to create a robust performance report for the model.

Note: Throughout this project, I will be documenting my process, challenges, and solutions in a way that's easy for others to understand and build upon. By sharing this journey, I hope to contribute to the broader conversation around large scale EEG data analysis and learn from the community as well.

5.Previous Learnings:

5.1 NeuroGPT:

While reading the Neuro-GPT paper, I found the integration of an EEG encoder with a GPT model for pre-training on extensive EEG datasets particularly insightful. This method, designed to address EEG data's scarcity and diversity, employs self-supervised learning to fill in masked EEG segments. What stands out is how it treats EEG signal chunks as tokens for the GPT model, a technique borrowed from natural language processing but modified for EEG's unique challenges.After understanding, I had a thought of specifically focusing on enhancing the encoder, I practically applied these concepts by experimenting with similar architectures.

GitHub: click here Colab Links: NeuroGPT encoder implementation

5.2 fMRI-S4:

While reading the fMRI-S4 paper, I was intrigued by its application of 1D convolutions and State Space Models (SSMs) for fMRI data, focusing on capturing both short- and long-range dependencies in neural signals. This approach appeared highly relevant for EEG data analysis, which similarly requires deciphering complex temporal patterns. Inspired to adapt this methodology for EEG,We can consider how the combination of 1D

convolutions for short-term patterns and SSMs for capturing the broader temporal landscape could enhance EEG analysis. The paper's insights into managing temporal dynamics provided a robust framework that I believed could be tailored for EEG, aiming to improve the accuracy and depth of interpretations derived from EEG data.

Colab Link : S4EEG POC GitHub: Modified fMRI repository for EEG data

5.3 EEGFormer:

The EEGFORMER paper introduced me to the concept of discrete representation learning, which I found particularly compelling for EEG signal processing. A method that applies discrete representation learning through a vector-quantized Transformer model for EEG data analysis. This strategy, utilizing large-scale EEG datasets for pretraining, offered an insightful way to improve model scalability and make the outcomes more interpretable. Inspired by EEGFORMER's methodology of encoding EEG signals into discrete tokens for better interpretation and reuse, Here, we can recognize the potential to extend these concepts towards developing an efficient and broadly accessible model. The emphasis on creating transferable and interpretable representations from extensive datasets highlights a promising area for new ideas, especially in making these models use less computing power and in improving how well they can explain their findings.

6. References

Based on what I've learned, I intend to discuss with mentors to integrate these concepts effectively. These studies have inspired me to further explore and experiment with enhancing the state-of-the-art foundational models. I am deeply grateful for the opportunity to learn and grow in this fascinating field so far.

https://github.com/wenhui0206/NeuroGPT
https://github.com/state-spaces/s4
https://github.com/elgazzarr/fMRI-S4
https://arxiv.org/pdf/2111.00396.pdf
https://arxiv.org/pdf/2311.03764.pdf
https://arxiv.org/pdf/2208.04166.pdf https://arxiv.org/abs/2401.10278