EMORY
UNIVERSITY
SCHOOL OF
MEDICINE

GSOC 2024 Project Proposal

Proposal Title

Development of an Open-Source EEG Foundation Model

Abstract / Project Summary:

The project aims to add libcamera support to OpenCV as opposed to the current V4L for better functionality.
This will be a great chance for me to learn and master how libcamera and OpenCV both function on a lower-level. Moreover it will give me great exposure to how linux development is done.

Introduction

Contributor Name	Aayush Jignesh Shah
Contributor Email	2001aayushshah@gmail.com
GitHub	https://github.com/aayushjshah
Time Zone	IST (UTC+5:30)
PortFolio	https://aayushjshah.github.io/devPortFolio/

Potential Mentor:Mahmoud Zeydabadinezhad (mzeydab -at- emory.edu) and Babak Mahmoudi, PhD

Personal Background (Brief CV)

Current - CitiBank - Software Development Engineer
Education : *Veermata Jijabai Technological Institute*
2023 Undergraduate , B.Tech in Information Technology
CGPA- 8.23/10

Relevant Experience:

International Gemological Institute(IGI) , *Computer Vision(AI) Intern*
Feb'22 - June'22
Hands on experience through ML Lifecycle from EDA to model deployment .Designed optimal hardware for capturing light-sensitive images, developed support software for hardware operation, pre-processing techniques for images in hand right from image

processing to annotation of the dataset,researched multiple models and approaches , and developed a Deep Learning based multi-class light sensitive classifer model which was sucessfully deployed in use .

Control Dynamics and Research Lab - VJTI Research Assistant
March'23-May'23
Leveraging Graph Neural Networks to optimize the performance of Power Systems through advanced computational modeling.
Citi, *SDE (Spring,Angular)* | (Pune, India)
May 2023 - Aug 2023
• Enhanced Citi's Tradelink Mutual Funds platform for APAC region with 30 enhancements across **Spring Microservices** backend, **Angular-based micro** frontends, and SQL Databases.Experienced in entire SDLC pipeline.
• **ProjectPlus(GenAI)** - Developed two variants of Large Language Models based Document Classifier using **Langchain, vector DB,python and prompt chaining**; participated in Citi (PBWM) wide GenAI hackathon as the **only fresher team**.

Barclays, *Software Developer Intern* | Pune (India) Onsite
06th June'22 - 31st July'22
• Enhanced website performance by Engineereing a **Python-based machine learning** utility to parse user logs, predict error probabilities, and optimize website performance.
• Designed Kibana based dashboards over , Enhanced Code Quality using Sonar for Java. **Tech : Java ,Python**

Relevant Projects

ProfitSense | Deep Learning ,ML, Python - Final Year project
• Designed a Deep Learning-based trading setup that recommends trades for the Indian Stock Market, delivers associated risk factors, and delivers an average profit of 4% per trade.Implemented SVM,Random-Forest etc Machine Learning based models
• Co-authored and published literature survey as well as a *research paper* for the project.

Grounding-Dino and Segment Anything Model based Zero shot Image Segmentation and object detection. March 2024

Smart India Hackathon(India’s Largest Hackthon) - National Finalists
Spearheaded the development of NLP(Natural Language Processing),web-based solution that generates summaries from online and local documents.

Power prognosis – *Deep Learning* Nov 2022 - Dec 2022
• Streamlined Deep LSTM based forecasting of remaining Useful Battery Life . Dataset : NASA Lithium-Ion Battery Dataset

Graphify - *NLP,BERT*   Oct 2022 - Dec 2022
• Leveraged advanced NLP(Natural Language Processing techniques to generate accurate and detailed knowledge graphs from text inputs using tools such as Bert, Graph convolutions, and WikiData

Project Goals / Major Contributions

Goal :

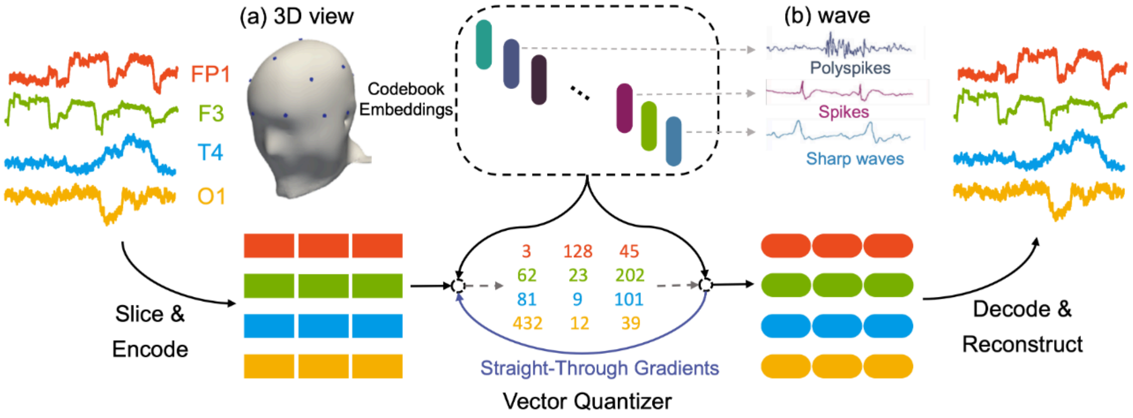
The goal of the project is to create an open-source foundation model for analyzing EEG (Electroencephalography) data. This involves developing algorithms for signal processing, automatic feature extraction, and implementing deep learning-based techniques to pre-train a model on publicly available EEG datasets. The ultimate aim is to enhance the robustness and versatility of EEG data analysis, particularly in scenarios where labeled data for specific tasks is limited. Through this project, the aim is to contribute to advancements in EEG data analysis and promote open collaboration in the field of medical data analysis.

Plan :

To get started I have surveyed a couple of approaches to bring achieve the desired outcomes of our project.

1) EEGFormer: Towards Transferable and Interpretable Large-Scale EEG Foundation Model

EEGF TRANSFORMER, pretrained on large-scale compound EEG data. The pretrained model cannot only learn universal representations on EEG signals with adaptable performance on various downstream tasks but also provide interpretable outcomes of the useful patterns within the data



Overview of EEGF

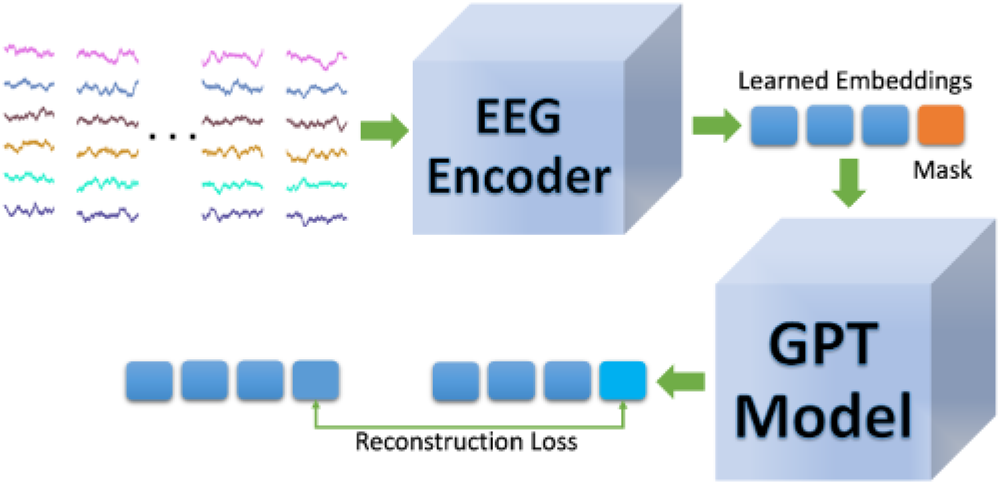
Transformer,Initially, multi-variate EEG signals are segmented into patches, which are then passed through a Transformer encoder. Subsequently, a vector-quantized model is employed to generate discrete indices. These indices are then fed into a shallow Transformer decoder.

2) Neuro GPT Model

Neuro-GPT, a foundation model consisting of an

EEG encoder and a GPT model.

The foundation model is pre-trained on a large-scale data set using a self-supervised task that learns how to reconstruct masked EEG segments.They then fine-tune the model on a motor imagery classification task to validate its performance in a low-data regime (9 subjects). Experiments demonstrate that applying a foun-dation model can significantly improve classification per-formance compared to a model trained from scratch, which provides evidence for the generalizability of the foundation model and its ability to address challenges of data scarcity and heterogeneity in EEG.



Neuro-GPT Pipeline: the EEG encoder takes chunks of EEG data as input and generates embeddings as tokens for the GPT model. The last embedded chunk in the sequence is masked. The GPT model then predicts the masked chunk and a reconstruction loss is computed between the prediction and the original embedding token.

Project TASKS AND SCHEDULE

Community Bonding Period : May 1 - May 26

During the community bonding period, efforts will focus on building relationships within the project community, understanding project goals, and making design decisionsof model's and architecture and workflow. Finalising of research material for literature survey.This time will involve getting to know mentors and contributors, refining the project plan, and setting up communication channels for effective collaboration.

Literature Review and Research (1 week): May 27 - June 3

- Research existing open-source foundation models for medical data, specifically EEG data.
- Identify current limitations and challenges in the field.
- Review self-supervised methods for multivariate time series data.
- Study various encoding mechanisms such as Transformers and EEG-specific encoders.

Designing Model Architecture (3 weeks): June 4 - June 11 , June 12 - June19,June 20 - June 27

- Design the architecture of the EEG foundation model.
- Implement one of the surveyed approaches using PyTorch.
- Consider incorporating self-supervised learning techniques and encoding mechanisms identified during the literature review.

Development of Data Pipeline (1 weeks): June 28-July 4th

- Survey EEG datasets and select suitable ones for model training.
- Standardise data format and preprocess using techniques like Fast Fourier Transforms for conversion to the frequency domain.
- Implement a data pipeline to handle data ingestion, preprocessing, and augmentation.

MID TERM Evaluation : July 12th 2024

Model Training and Testing (2 weeks): July 5th - July 20th (+1 Day MidTerm Evaluation)

- Set up the model architecture using PyTorch.
- Train the model using the selected EEG datasets.
- Test the developed architecture on various downstream tasks such as seizure detection, abnormal detection, and emotion recognition.

Evaluation and Optimization

- Evaluate the performance of the model on the selected tasks.
- Optimise the model architecture and hyperparameters based on performance metrics.
- Incorporate feedback and iterate on the model design if necessary.

Hosting and Documentation (2 weeks): Till August 14th

- Evaluate hosting platforms such as cloud services and Hugging Face for making the model publicly available.
- Document the model setup, installation process, and usage instructions.
- Create examples and tutorials demonstrating the usage of the model.

Project Management and Reporting (Throughout):

- Use project management tools to track progress and manage tasks.
- Maintain regular communication with mentors and stakeholders.
- Write weekly progress reports and keep documentation up to date.

August 14 - August 21
Buffer period for completing additional feature integrations or for completing any remaining backlogs.
Documenting every aspect of the coding period.

Finalise the project deliverables, including the model, documentation, and reports.
Prepare and submit the final project report and code to the GSOC platform.

Aug 26 - September 2
Mentors submit final GSoC contributor evaluations. (standard coding period)
Finalization and Submission (Last week):

Planned GSoC work hours

As mentioned in the idea lists Development of an Open-Source EEG Foundation Model is an entirely new project and expected work hours are 350 hours (Hard category)
Current Personal Commitments : I work as a Software Developer @ Citibank and have 9am IST to 5 pm IST Time commitment for Monday to Friday towards the same

Proposed Work Hours for GSOC : 7:00 pm IST to 10:00/ 10:30 pm IST i.e. 3hrs -4hrs) over Monday to Friday (accounting for 15-20 hours) and (15 -20) hours over the weekend.

With an ongoing job together , emergencies may arise. For the same I intend to dedicate hours over the additional 2 months prior to completion of mainstream GSOC 2024 say the Project got extension over those 2 months.

SkillSet :

Project Goals	**Relevant Major Contributions**
● **Strong programming skills, preferably in Python.**	**VJTI - BTech IT (8.23 C.G.P.A**
● **Experience with deep learning frameworks, preferably PyTorch.**	**International Gemological Institute -AI based Diamond Grading , Grounding Dino-Sam based Image Segmentation**
● **Knowledge of self-supervised learning and large language models**	**Citi - Project Plus, SIH Hackthon summarizer, LLM Udemy Course**
● **Knowledge in signal processing, neuroscience, or related fields.**	**Enrolled (CET)Common Entrance Test for Engineering in 12th Grade for Subject Physics , Chemistry , Mathematics, Biology. PCB - 96.99%tile**

Programming languages: Python, Java, C++,HTML,CSS,JavaScript,Nodejs , Java Swing API ,
Machine Learning: Regression, Classification, Clustering, Auto Encoders, Neural Networks,Attention Networks,CNN , LSTM,GNN, ML Life Cycle
Computer Vision : edge detection, image segmentation, object detection, feature extraction,
image enhancement, color space transformations, Frequency Domain based applications, Fourier Transformation Applications
Frameworks and Libraries: TensorFlow, Keras, PyTorch, Scikit-learn,OpenCv , pyTorch geometric,Pillow,numpy,pandas,matplotlib,Jupyter ,Darknet,Cuda
Software Development: Git, Debugging,Testing,Code Maintenance,OOPS
Database Management: SQL, MongoDB,Vector Databases
Analytical and Problem-Solving Skills: Root cause analysis, Troubleshooting
Communication and Collaboration: Verbal and written communication, Teamwork , MS Office
, Tableau
Continuous Learning: Online Courses, Tech Blogs
Abstract : Leadership, Creative, Can Do Attitude, Ingenuity, Optimistic