Multimodal Models with Narrow Domain

Team 37

August 30, 2024

Contents

Chapter 1

Introduction

This project focuses on fine-tuning multimodal models to enhance their performance in specialized, narrow-domain applications. The main objective is to ensure that these models can accurately interpret and generate content within specific contexts, such as medical research and social media meme analysis.

Chapter 2

Datasets

2.1 Medical Domain Data

For the medical domain, data from the Human Connectome Project (HCP) was utilized. This dataset provides comprehensive neural imaging data (fMRI), which is essential for advancing research in brain connectivity and related studies.

- **Human Connectome Project (Young Adult Data)**: Human Connectome Project - Young Adult Data Releases

2.2 Meme Data

To study meme-related content, we used datasets that include diverse collections of memes and related metadata. These datasets are crucial for research in social media, cultural analysis, and AI-based meme generation.

- **Meme Captioning Dataset**: Meme-Cap Dataset on GitHub
- **MemeCraft Dataset**: MemeCraft Dataset on GitHub

Chapter 3

Methodologies for Medical Domain Images

3.1 Generative Modeling

3.1.1 Bridge Matching

Bridge Matching involves training on paired fMRI data, specifically resting-state and motor-task functional connectivity matrices. The key idea is to use conditional diffusion models to align functional connectivity patterns by minimizing the distance between the model output and the target connectivity matrices.

3.1.2 Bridge Matching Theory

In conditional diffusion models, we learn to generate samples conditioned on additional information, such as class labels or textual prompts. The training process involves minimizing the loss between the model's prediction and the target using the conditional score functions.

Training Algorithm:

1. Generate a pair $(x_0, x_1) \sim q_{01}(x_0, x_1)$.
2. Sample $t \sim \mathcal{U}[0, 1]$.
3. Sample the interpolation x_t from $p_t(x_t|x_0, x_1)$.
4. Feed the pair (x_t, t) into the neural network to compute $f_\theta(x_t, t)$.
5. Compute the conditional vector field $\beta_t \cdot (x_1 - x_t)/\overline{\sigma}_t^2$.

6. Perform a gradient descent step: $\nabla_\theta \left\| f_\theta(x_t, t) - \beta_t \frac{x_1 - x_t}{\overline{\sigma}_t^2} \right\|^2$.

During generation, start with an initial sample $X_0 \sim q_0$ and evolve it through an SDE using methods such as Euler's method or other solvers. This approach ensures that the generated samples approximate the target distribution.

Loss Function

The loss function for conditional diffusion models represents the weighted deviation between the conditional score function, $s_\theta(X_t, t|Y)$, and the true conditional score function, $\nabla \log p_{X_t|X_0,Y}(X_t|X_0, Y)$. Minimizing this function allows the model to learn to denoise and recover data considering the condition Y.

Main Steps:

1. Conditional Score Function:

$$\nabla \log p_{X_t|Y}(x_t|y) = E\left[\nabla \log p_{X_t|X_0,Y}(X_t|X_0, Y) \mid X_t = x_t, Y = y\right]$$

   This function shows that the score estimate for the current state X_t given Y is equal to the conditional expectation of the noise score added to the initial state X_0 given Y.

2. Loss Function:

$$E\left\| s_\theta(X_t, t|Y) - \nabla \log p_{X_t|X_0,Y}(X_t|X_0, Y) \right\|^2 \to \min_\theta$$

   This loss function represents the mean squared deviation between the model's conditional score function $s_\theta(X_t, t|Y)$ and the true conditional score function $\nabla \log p_{X_t|X_0,Y}(X_t|X_0, Y)$.

   To account for all time steps, the integral (or sum for discrete time) over all time steps is added:

$$\int_0^1 E\left\| s_\theta(X_t, t|Y) - \nabla \log p_{t|0}(X_t|X_0) \right\|^2 \mathrm{d}t \to \min_\theta .$$

   For discrete time, the integral is replaced by a sum:

$$\sum_{t=1}^{T} E\left\| s_\theta(X_t, t|Y) - \nabla \log p_{t|0}(X_t|X_0) \right\|^2 \to \min_\theta .$$

Thus, the loss function for conditional diffusion models is the sum of the mean squared deviations between the model predictions and true values calculated for each time step considering the condition Y. This function is minimized over the model parameters θ to train the model.

3.1.3 Stable Diffusion 1.5

Stable Diffusion 1.5 was utilized to enhance the quality of generated samples by conditioning the diffusion process on target images. In our approach, we added motor-task fMRI connectivity matrices as the target, corresponding to the resting-state matrices of the same patient.

Optimization Approach:

- We integrated target images into the diffusion process, where the model aimed to minimize the difference between the generated images and the target images using diffusion techniques.
- The goal was to align the input images to the target images, effectively refining the diffusion process to produce outputs that closely match the desired target.

This method leverages the strengths of Stable Diffusion to achieve high fidelity and accuracy in generating images that correspond to complex connectivity patterns, crucial for analyzing fMRI data in a medical context.

3.2 Multimodal Models

3.2.1 GILL Model

The GILL model facilitates the integration and analysis of multimodal data, leveraging embeddings and functional connectivity matrices to deliver insightful results.

- **Repository**: GILL on GitHub
- **Overview**: By preparing this multimodal paired dataset and using the GILL model, we aim to build model which can be used as support system.

3.3 Results

As a result of applying the Bridge Matching methodology on paired fMRI data, a model has been developed that aligns functional connectivity patterns between resting-state and motor-task conditions more effectively.

To validate the performance of the model, a series of experiments were conducted involving medical professionals who evaluated the generated connectivity matrices. Participants were provided with a set of generated matrices from both the baseline model and the Bridge Matching model, without being informed of which model produced which matrices. According to the results of this evaluation, the Bridge Matching model demonstrated superior performance in aligning the functional connectivity patterns with the target matrices. However, the overall capability of the models to generate clinically meaningful connectivity patterns was still limited, indicating that further refinement is needed.

For improved results, it is essential to train the Bridge Matching model for a larger number of steps. Additionally, the UNet architecture used in this work, which was originally designed for tasks like MNIST to colored-MNIST conversion, should be enhanced. In our case, we used a different UNet that had previously been employed as a generator for GAN-based models like WcGAN-QC and Pix2Pix. To optimize this model for Bridge Matching, it is crucial to integrate Time Embeddings and Attention layers to better capture low-level information and improve the model's ability to align connectivity patterns accurately.

Regarding the Stable Diffusion 1.5 model, further training is necessary to adapt it to our specific fMRI connectivity matrices. We have already prepared a dataset that includes both the matrices and corresponding descriptions. Our plan is to finetune the Stable Diffusion model on this dataset to enhance its ability to understand and generate meaningful medical images based on these connectivity patterns.

In light of these findings, it is recommended to enhance the dataset by incorporating more diverse and accurately labeled fMRI data, as the current dataset may have constrained the potential of the Bridge Matching model. Additionally, refining the architecture of the UNet and further training the Stable Diffusion model on our specialized dataset will likely lead to significant improvements in model performance.

The **final loss for the Bridge Matching** model after 100 epochs was 129.23.

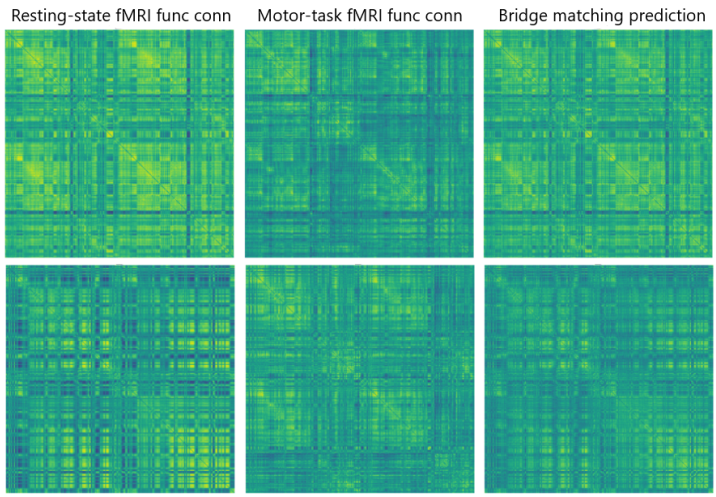

Figure 3.1: Performance of Bridge Matching Model

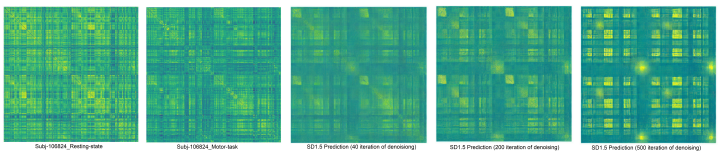

Figure 3.2: Stable Diffusion 1.5 with image condition performance

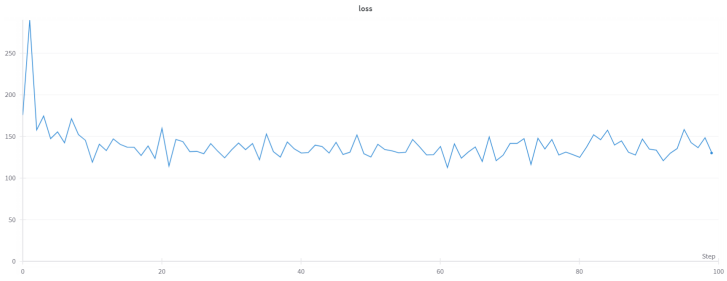

Figure 3.3: Loss Curves for Bridge Matching Model

Chapter 4

Methodologies for Memes Domain Finetuning

The preparation of training examples included the search and selection of available meme datasets, with a description of what is happening on them. The data was prepared in a format that is compatible with additional finetuning, after which LoRA was utilized. To test the performance of the finetuned model, a competition was organized between the baseline model and the finetuned one.

4.1 Multimodal Models

4.1.1 TinyLLaVA-Phi-2-SigLIP-3.1B

Repository: TinyLLaVA-Phi-2-SigLIP-3.1B on GitHub

Overview: TinyLLaVA Factory is an open-source modular codebase for small-scale large multimodal models (LMMs), implemented in PyTorch and HuggingFace, with a focus on simplicity of code implementations, extensibility of new features, and reproducibility of training results.

By preparing a multimodal paired dataset and using the TinyLLaVA-Phi-2-SigLIP-3.1B model, we aim to enhance the accuracy and interpretability of memes, enabling a deeper understanding of the context and hidden meaning of the meme.

4.2 Results

As a result of finetuning the model (Figure 4.1) on the collected meme dataset, a new model has been developed that presumably better under-

stands the context of images.

The validation of the results was conducted through a competition of models involving respondents. Each participant was shown a set of memes along with descriptions of those memes from both the baseline model and the finetuned model, without specifying which model was finetuned and which was not. According to the results of the described test (Figure 4.2), the finetuned model performs better in describing memes and understanding the meaning of images. However, the overall assessment of the tested models' ability to describe memes turned out to be quite low. Therefore, the task of improving multimodal models for understanding memes remains relevant and requires additional solutions.

To continue working on this task, it is proposed to create a comprehensive dataset of memes with clear descriptions— it is assumed that the lack of such a dataset has been a significant limitation for the finetuning of the model in this work.

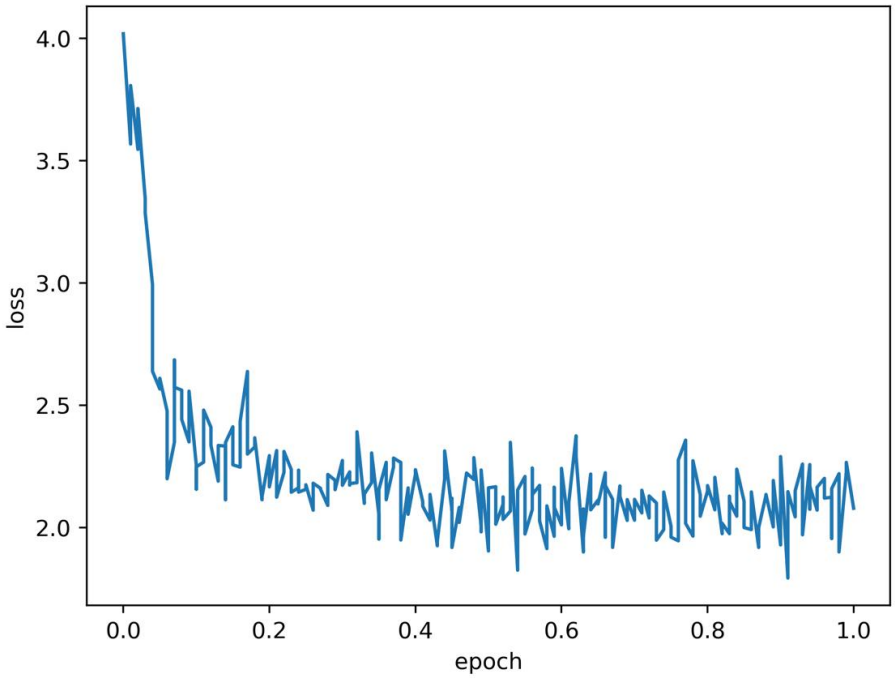

Figure 4.1: Loss values during model fine-tuning

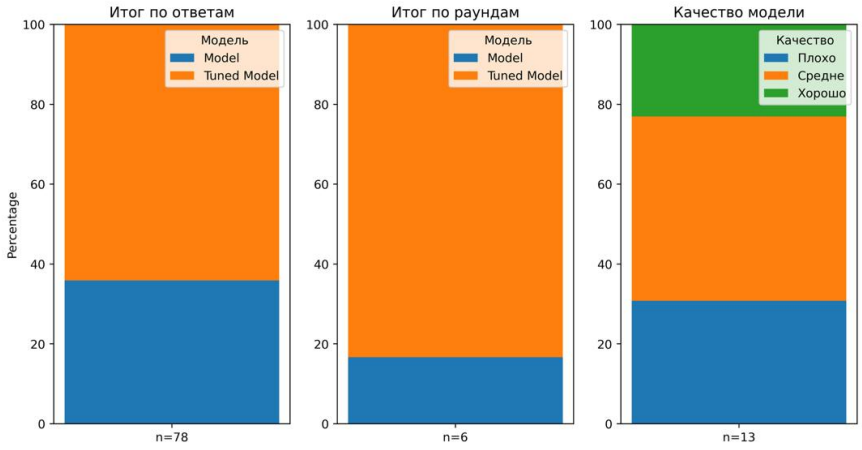

Figure 4.2: Comparison of model responses based on the results of human evaluation. The first column shows the share of votes for each model. The second column shows the share of rounds won for each model. The third column shows the qualitative assessment of the models by respondents for meme understanding tasks.

4.3 Example

The baseline and finetuned models were given a meme (Figure 4.3) and the prompt "Why is this picture funny?". The responses of both models are provided below.

4.3.1 Baseline Model:

The picture is funny because it features a police officer in a black uniform, who is trying to give a ticket to a person in a car. The officer is holding a ticket in his hand, and the person in the car is looking at him with a surprised or confused expression. The situation is humorous because it is not common for a police officer to give a ticket to a person in a car, especially if the person appears to be innocent or unaware of the violation. The unexpected interaction between the officer and the person in the car creates a lighthearted and amusing scene.

4.3.2 Finetuned Model:

The picture is funny because it features a police officer in a black uniform, who is trying to give a ticket to a car. However, the car is parked on the wrong side of the road, which is a common mistake that can lead to traffic violations. The humorous aspect of the image comes from the fact that the police officer is trying to enforce the rules, but the car's position on the wrong side of the road makes it difficult for him to do so. The situation is comical because it is not the typical scenario where a police officer would issue a ticket, and it highlights the importance of paying attention to traffic rules and signs to avoid such situations.

Figure 4.3: Example of a meme used to compare the performance of the baseline and finetuned models

Chapter 5

Conclusion

This project aimed to improve multimodal models for narrow-domain applications by focusing on medical imaging and meme analysis. The use of specialized datasets and advanced methodologies has led to notable advancements in both fields, although challenges remain.

In the medical domain, the application of the Bridge Matching methodology has shown promise in aligning functional connectivity patterns between resting-state and motor-task fMRI data. Despite the positive results, there is a need for further refinement. Specifically, increasing the number of training steps, enhancing the UNet architecture, and integrating Time Embeddings and Attention layers are crucial steps for improving the model's performance. Additionally, further training of the Stable Diffusion 1.5 model with a dataset of fMRI matrices and corresponding descriptions is expected to enhance its ability to generate accurate and meaningful medical images.

For the meme analysis, finetuning the multimodal models has improved their ability to understand and generate contextually relevant descriptions. The competition between the baseline and finetuned models highlighted the improvements made but also underscored the limitations in the current datasets and models. The need for a comprehensive dataset with well-described memes is evident, and addressing this gap will be essential for further advancements.

Overall, this project demonstrates the potential of multimodal models in specialized applications but also points to areas needing further development. Future work should focus on refining model architectures, expanding and improving datasets, and conducting more extensive validation to fully realize these models' capabilities in their respective domains.