Research on Pneumonia Patient Condition Classification Using Diffusion Models and CLIP

Group2: Xiaomeng Xu; Wenfei Mao; Yingzhen Wang; Shuoyuan Gao

2024-12-17

Abstract

Pneumonia remains a leading cause of childhood mortality worldwide. Our study explores large language model techniques to classify pediatric chest X-ray images into normal, bacterial pneumonia, and viral pneumonia categories. A dataset of 5,856 radiographs was used, with synthetic images generated via LoRA fine-tuning of a Stable Diffusion model to address data imbalance. A fine-tuned CLIP model achieved modest improvements in training accuracy (from 48.94% to 50.51%), while the test accuracy remained at 37.50%.

1 Introduction

1.1 Background

Pneumonia remains the leading cause of death worldwide among children under five years of age. Despite the implementation of safe, effective, and affordable interventions that have significantly reduced pneumonia mortality from 4 million in 1981 to just over one million in 2013, pneumonia still accounts for nearly one-fifth of all childhood deaths globally (Organization 2014). Chest X-ray imaging has emerged as a promising modality for radiologic diagnosis of pneumonia. However, its role in clinical management and its impact on patient outcomes require further optimization (Kermany et al. 2018).

1.2 Dataset

The dataset utilized for this study comprised 5,856 pediatric chest X-ray images categorized into three classes: Normal, Virus and Bacteria. The numbers of images in these three categories are 1,349, 1,345, and 2,538, respectively. The images were collected from 5,856 patients aged one to five years, all receiving clinical care at the Guangzhou Women and Children's Medical Center. To ensure data quality and reliability, all radiographs underwent preprocessing to eliminate low-quality scans. Furthermore, the images were independently reviewed and classified by two specialist physicians and validated by a third-party expert to minimize the risk of misclassification (Dincer n.d.).

1.3 Imaging Features

The dataset includes radiographic images demonstrating distinct features associated with different lung conditions (Kermany et al. 2018):

- Normal Lungs: These images show no pathological changes.

- Bacterial Pneumonia: Characterized by localized consolidation with well-defined margins, often accompanied by pleural effusion.
- Viral Pneumonia: Identified by bilateral ground-glass opacities, reticular patterns, or patchy infiltrates with poorly defined borders, reflecting its diffuse and interstitial nature.

These imaging features provide critical diagnostic insights and support the development of effective treatment strategies for pneumonia.

1.4 Queation of Interest

Our study investigates the effectiveness of fine-tuning the Stable Diffusion model with LoRA for synthetic image generation and leveraging a CLIP model for robust classification of three chest X-ray categories, as well as exploring their potential limitations and future improvements in pediatric radiologic diagnosis.

2 Method

2.1 Diffusion Model

Diffusion models(Ho, Jain, and Abbeel 2020)draw inspiration from non-equilibrium thermodynamics. They employ a Markov chain of diffusion steps to gradually introduce random noise into the data and then learn to reverse this process, reconstructing desired data samples from the noise. These models are trained using a fixed procedure, and the latent variables operate in a high-dimensional space. Diffusion models consist of two processes: the forward diffusion process and the reverse diffusion process. In the forward process, given a data point sampled from the real data distribution, $x_0 \sim q(\mathbf{x})$, we define a forward diffusion process that progressively adds small amounts of Gaussian noise to the sample over T steps. This process generates a sequence of noisy samples, $x_1, \ldots, x_T$. The step sizes are determined by a variance schedule $\beta_t \in (0,1)_{t=1}^{T}$. The data sample would generally lose its distinguishable features as the step becomes larger. Eventually, x_T would equivalent to an isotropic Gaussian distribution. We will be able to recreate the true sample from a Gaussian noise input, $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and this is reverse diffusion process.

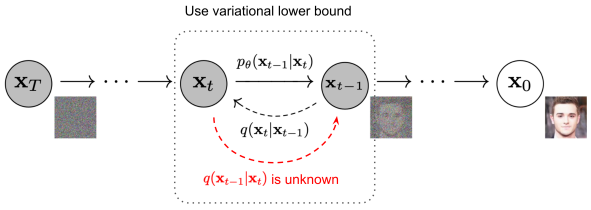

Figure 1: The Markov chain of forward and reverse diffusion process.

Firstly, in our project, we deployed the stable Diffusion Version 2. The model use 865M U-Net as image generator and use OpenCLIP ViT-H/14 as image-text encoder and it could generate 768×768px outputs. Then we utilized LoRA to fined-tuned the model to make the model generate customized images. Specifically, we fine-tuned attention layer and projection layer of U-Net. LoRA keeps the pre-trained model weights fixed while incorporating trainable low-rank decomposition matrices into each layer of the Transformer architecture. Specifically, LoRA introduces two low-rank matrices, LoRA_A and LoRA_B, and only these two low-rank matrices are involved in the fine-tuning process. It could significantly reduce the number of trainable parameters required for the downstream tasks. Typically, the percentage of trainable parameters is about 1%. It could also reduce the GPU requirement by 3 times.(Hu et al. 2021) Finally, after fine-tuning the model, we used it to generate 1,000 synthetic images for normal lung images and viral pneumonia images separately to address the problem of data imbalance. Finally, we combined the original dataset and the synthetic dataset for training the CLIP model.

2.2 Contrastive Language-Image Pre-training(CLIP)

After combined the original and synthetic dataset, we used these data to fine-tuned the projection layer and full connected layer of CLIP model with LoRA. CLIP is a neural network trained on a variety of (image,text) pairs. It can be instructed in natural language to predict the most relevant text snippet, given an image, similarly to the capabilities of GPT-2 and GPT-3.(Radford et al. 2021) In our project, we deployed CLIP-ViT-large-patch14. The model use ViT-L/14 Transformer as image encoder and use masked self-attention Transformer as text encoder. CLIP model was pre-trained on a larger-scale dataset. Therefore, it could achieve very impressive results on many computer vision tasks. However, it would achieve much better results after fine-tuning, because our dataset is a medical dataset. After fine-tuning the CLIP model, we used fine-tuned model to classify test dataset into three categories. The basic principle of image classification in the CLIP model is that it calculates the cosine similarities between images and texts. The image and text with the highest similarity are matched, thus completing the classification. Therefore, we need to create some prompts to use for classification. We tried different prompts and found that 'An image of [Type] chest X-ray,' where the type could be one of three: normal, bacteria, or virus, worked best. Besides, We should also pay attention to the case sensitivity in "Type," which should be consistent with the case of the training dataset folder. Finally, we successfully used the fine-tuned model to classify the test dataset.

2.3 Experiment Setup

Our experiments were conducted on a server equipped with an NVIDIA RTX 3090 GPU with 24GB memory, running CUDA 12.2 Toolkit. The model was implemented using PyTorch 2.5.1 and Python 3.9.19. We fine-tuned the CLIP model for 3 epochs using the AdamW optimizer, with a batch size of 4 and a learning rate of 5e-5. For parameter-efficient fine-tuning, we adopted LoRA with a configuration of LoRA alpha equals to 32 and LoRA dropout equals to 0.1. More detailed parameter settings can be found on Github.

3 Result

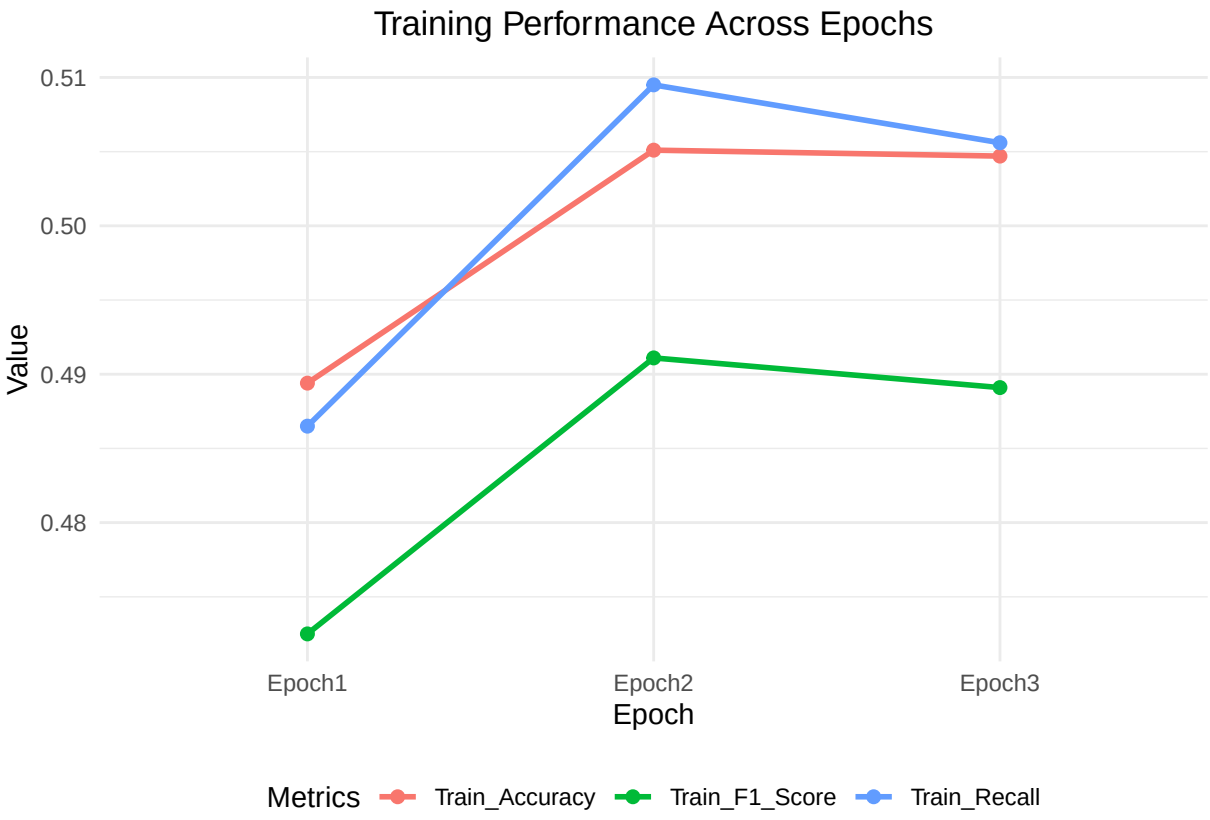

Figure 2: Training performance metrics (Accuracy, F1-Score, Recall) across epochs.

From the result, we can see the image shows that as the number of training epochs increases, the training accuracy, recall, and F1 score gradually improve. The accuracy rises from 0.4894 to 0.5047, recall increases from 0.4865 to 0.5056, and the F1 score grows from 0.4725 to 0.4891. This indicates that the model is progressively learning the features of the training images. The final accuracy on the test dataset is 0.375, and the recall is 0.5, indicating that approximately half of the images in the test set were correctly classified.

4 Conclusion

From the above results, we can see that the final classification outcome is promising. Our project leverages a multimodal large model approach, which, compared to traditional CNN networks, fully utilizes information from different modalities, thereby achieving an ideal image classification performance. Moreover, multimodal LLMs possess more powerful feature representation capabilities, enabling them to learn finer-grained feature representations, which allows for better handling of complex tasks. Compared to CNNs, multimodal LLMs exhibit stronger generalization abilities and better interpretability, making the classification results easier to understand and explain.

While the proposed approach demonstrates promising results, it still has certain limitations, and we propose some future work to address these limitations. Firstly, the stable diffusion model uses DDPM (Denoising Diffusion Probabilistic Models) for training. The process requires approximately 1,000 steps to complete, which is time-consuming and has high GPU requirements. However, the consistency models(Song et al. 2023) require only 5 steps or fewer to complete this process because it follows a certain route. This makes it up to 50 times faster compared to the stable diffusion model. Recently, at CVPR 2024, UC Berkeley (Frans et al. 2024) proposed a one-step diffusion model. Therefore, there are many

other promising approaches to explore in future work. Secondly, after fine-tuning the model, we can apply reinforcement learning methods, such as DPO or PPO, to further optimize the model. Reinforcement learning methods has achieved great success in post-training stage of large language models. For example, OpenAI's recent o1 model has been optimized using reinforcement learning and has significantly outperformed GPT-4 in multiple tasks. Therefore, we could adopt such methods to further improve the classification results. Finally, we adopted LoRA to fine-tune the model due to limited computing resources. However, it may introduce noises to the training dataset because it cannot handle more customized images. Therefore, in future work, full parameters fine-tuning or reasoning with reinforced fine-tuning (Luong et al. 2024) may be better options compared to LoRA, as they can achieve better fine-tuning results and reduce the impact on the training dataset.

5 Contribution

Xiaomeng Xu: Code editing; abstract; introduction

Wenfei Mao: Code editing; Diffusion Model; CLIP; Conclusion

Yingzhen Wang: Code editing; Results; Diffusion model

Shuoyuan Gao: Code editing; Experiment Setup; Conclusion

Github Link: https://github.com/xxm12345666/biostat625-group2-project

References

Dincer, Tolga. n.d. "Labeled Chest x-Ray Images." https://www.kaggle.com/datasets/tolgadincer/labeled-chest-xray-images.

Frans, Kevin, Danijar Hafner, Sergey Levine, and Pieter Abbeel. 2024. "One Step Diffusion via Shortcut Models." https://arxiv.org/abs/2410.12557.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. 2020. "Denoising Diffusion Probabilistic Models." https://arxiv.org/abs/2006.11239.

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. "LoRA: Low-Rank Adaptation of Large Language Models." https://arxiv.org/abs/2106.09685.

Kermany, Daniel S., Michael Goldbaum, Wenjia Cai, Catarina C. S. Valentim, Huiying Liang, Sally L. Baxter, and Kang Zhang. 2018. "Classification of Images of Childhood Pneumonia Using Convolutional Neural Networks." *PLoS One* 13 (2): e0192361. https://doi.org/10.1371/journal.pone.0192361.

Luong, Trung Quoc, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. "ReFT: Reasoning with Reinforced Fine-Tuning." https://arxiv.org/abs/2401.08967.

Organization, World Health. 2014. *Revised WHO Classification and Treatment of Pneumonia in Children at Health Facilities: Evidence Summaries.* Geneva, Switzerland: World Health Organization.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 2021. "Learning Transferable Visual Models from Natural Language Supervision." https://arxiv.org/abs/2103.00020.

Song, Yang, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. "Consistency Models." https://arxiv.org/abs/2303.01469.