Precision Fine-Tuning: Leveraging LoRA for Text-Only Adaptation in Multi-Modal Medical Models

Wenru Lu

School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, United Kingdom
wl23796@essex.ac.uk

Abstract. Large Multimodal Model (LMM), which has the ability to process visual and textual information, has great potential in medical and other professional fields. However, adapting these complex models to specific sub domains or tasks faces many challenges. Due to the high demand for computer resources and the risk of destroying the pre-training model, it is often difficult to achieve full fine-tuning. This paper proposes a new "precision fine tuning" method, which uses Low-Rank Adaptation (LoRA) technology to achieve efficient and directional model adaptation. This technology only applies LoRA to the text decoder layer inside the medgemma multimodal model, which can avoid and do not change the visual encoder. This project is based on a small and carefully selected data set of 98 Medical Abstracts. The experimental results show that the model training process converges stably, the training loss is significantly reduced, and the text generation component can be successfully adapted. At the same time, the qualitative evaluation also shows that the coherence, relevance and the use of language in the professional field of the generated text have been significantly improved. This method only updates a small part of the total model parameters, and significantly improves the parameter efficiency compared with the total fine-tuning. This study confirmed that directional LoRA is a potential technology, which can improve the ability of multimodal medical model text generation in a directional and efficient way.

Keywords: Large Multi-Modal Model, Low-Rank Adaptation, Parameter-Efficient Fine-Tuning, Medical Domain Adaptation, Precision Fine-Tuning.

1 Introduction

The large-scale multimodal model integrating visual and text information shows excellent performance in various tasks. In medical and other professional fields, MedGemma and other models have been applied to tasks such as medical image interpretation and report generation. Such large-scale models usually need fine-tuning to adapt to specific sub domains or tasks, but full-scale fine-tuning is extremely expensive in terms of computational cost and resource consumption. Therefore, the development of efficient adaptive technologies is essential to give full play to the potential of these powerful models. Low rank adaptation (LoRA) provides a method

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_91

with high parameter efficiency. Adaptation can be achieved by updating only a few model parameters through the low rank matrix [1]. However, when applying LoRA to complex multimodal architectures, it is necessary to select adaptive components strategically. It is an accurate and efficient strategy to adapt only to the text generation component (i.e. text decoder) while retaining the visual encoder, which can focus on the aspects that need to be improved most in the text task. This paper discusses the application of LoRA in the implementation of this "pure text" adaptation strategy in multimodal medical model.

This method aims to optimize the text generation ability of medical domain models, which may achieve faster adaptive adjustment and reduce the risk of damaging visual understanding. This strategy is implemented by updating the text decoding layer inside the model using the medical domain specific text corpus. This method has been applied in MedGemma. This paper shows the feasibility and effectiveness of this precise tuning method. Recent studies have also explored efficient fine-tuning methods for visual language models in the medical field. Peft technology, such as LoRA, has become a common strategy for large-scale Visual Language Model (VLM) to adapt to specific tasks or fields [2]. In the medical field, Zhang et al. [3] systematically compared the performance of various peft methods, including adapter and prefix tuning, on general medical VLM, and analyzed their trade-offs between efficiency and effect. Li et al. [4] further applied LoRA to MedGemma model and proved that fine tuning the whole model (including visual and language components) can effectively improve its performance in specific clinical downstream tasks (such as classification and retrieval). However, the existing research mainly focuses on using LoRA to improve the understanding ability of the model or its adaptability to multimodal joint representation. In contrast, there is still a lack of fine-tuning research on the pure language generation module in VLM, which aims to improve the quality of text generation in specific fields such as biomedicine. The method proposed in this study aims to fill this gap by applying LoRA fine-tuning only to MedGemma's internal language decoder, and to explore a more focused and efficient method to enhance the domain specific text generation ability of the model.

2 Methodology

2.1 Background and Motivation

Multimodal models such as MedGemma integrate independent components for different input types. Usually, visual encoders for images are combined with language decoders for text generation. This inherent modularity shows that the adaptation strategy can be implemented for specific parts of the model according to the main requirements of downstream tasks. In many key applications in the medical field - such as interpreting patient narratives, generating clinical summaries, or answering diagnostic questions based on text symptoms - the core requirement is to generate high-quality, domain specific text. In addition, in many real clinical interaction scenarios, patient information is mainly presented in the form of oral narration or clinical records,

and is not always accompanied by medical image data (such as CT scan and X-ray film). Therefore, the text generation ability of the optimization model becomes crucial to expand its practical value in text centric application scenarios.

However, there are many challenges in adapting MedGemma and other large models through standard full-scale fine-tuning, that is, updating all parameters. This not only requires a lot of computing resources and memory, but also is cumbersome in iterative development or deployment in resource constrained environments. More importantly, updating all weights indiscriminately may damage the stable functions of other components, especially when the pre-training ability of the visual encoder is very strong and the target task does not rely heavily on the new visual data mode. This highlights the urgent need for more refined and targeted adaptation technology.

The parameter efficient fine tuning (PET) method can solve these limitations [5]. Only a small part of the parameters in the model needs to be updated. Among many efficient parameter tuning methods, LoRA is particularly prominent: its core principle is to inject the trainable low rank matrix into the specific weight layer of the pre-training model (usually located in the attention mechanism and multi-layer perceptron module). This mechanism greatly reduces the number of parameters that can be trained and the computational burden. Its performance is often equivalent to that of full fine-tuning, and even better than that of full fine-tuning in specific tasks. Since the MedGemma architecture clearly distinguishes between visual and language modules, and text generation dominates in the above key application scenarios, LoRA can choose to adjust only the language decoder (language model). This precise component specific optimization method is called "target fine-tuning", which aims to improve the quality of text output of the model in a specific medical field without interfering with the visual processing path of pre training. By focusing the adaptive work on the text generation layer, this method has many advantages, such as faster convergence speed, lower computational cost, and can minimize the risk of damaging the core visual understanding ability of the model.

2.2 Data Preparation

In order to apply medgemma model to the field of medical text generation, the paper compiled a medical literature corpus from the open access subset [6] of PubMed Central (PMC). The source documents were obtained in the standard NLM JATS XML format.

A total of 100 XML files were selected in this pilot study to be used as experimental and analytical data sets. The experiment uses a special preprocessing script (preprocess_pmc_xml.py) to extract and clean up the core text content required for language model adaptation, especially the extraction and use of article titles and abstracts.

The core of the script is the (extract_text_from_xml("file_path")) function. This function uses the xml.etree.elementtree library to parse a single xml file. Its operation process is shown in Figure 1:

After the extraction and download process is completed, the preprocess_pmc_xml.py script will traverse the first 100 XML files specified by NUM_FILES_TO_USE. For each file, the extract_text_from_xml function is called.

The text strings extracted and cleaned up successfully will be summarized into a list (all_texts). Of the 100 files processed, 98 generated valid text output, and the other two had errors. Finally, the list of 98 strings will be serialized through torch.save() and saved to the file preprocessed_pmc_dataset_100.pt.

It is worth noting that the script saves the list of original text strings, not the tensor after word segmentation or the pre-built dataset object. This design choice provides greater flexibility for the subsequent training process, enabling the dynamic application of the word splitter associated with medgemma model in the data loading process. This flexible method can effectively transform the unstructured XML source file into a structured list of plain text fragments, which is suitable for fine-tuning the target language model.

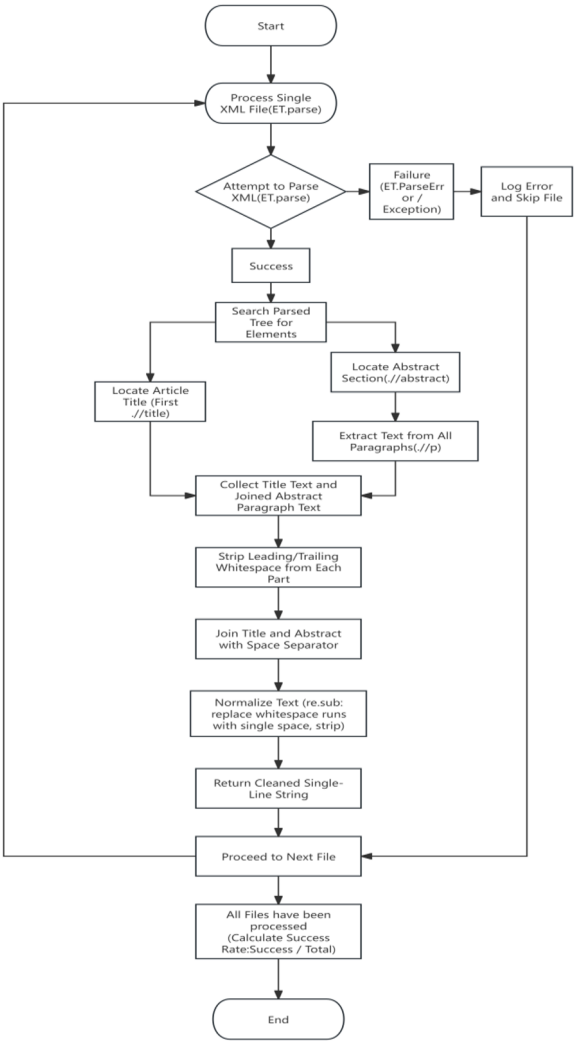

Fig. 1. Preprocessing Script Flowchart (Picture credit: Original)

2.3 Targeted LoRA Application

This section describes in detail the method of accurately identifying the target layer in the model, the process of fully configuring the LoRA adapter according to the actual implementation, and the software and hardware training settings related to the model.

First, on the basis of the motivation to carry out adaptive adjustment focusing on language generation components, the paper implemented a targeted LoRA strategy for the internal language decoder of medgemma model.

Medgemma model architecture integrates visual transformer (VIT) encoder and language decoder based on Gemma 3 architecture. In order to achieve this precise fine-tuning, it is necessary to accurately identify the corresponding language decoder module in the whole model framework. To this end, the paper used a special script (inspect_language_model_layers.py) to deeply analyze the model, so as to systematically check each named module in the model. The analysis results confirm that the text generation sub module is located under the hierarchical path of model.language_model. After further in-depth analysis, it is found that the internal structure of the sub module is composed of a series of gemma3decoderlayer and other objects, and the layers required for adapter injection are located in the model.language_model.layers.x file.

After determining the target layer, inject the LoRA adapter into all relevant linear transformation layers in the selected gemma3decoderlayer component. Specifically, according to the configuration defined in the tuning script (lora_finetune_medgemma.py), as shown in table 1, the adapter is added to the following sub modules of the decoder layers 0, 1, 2 and 33 [7]:

Table 1. Adapter configuration ($X \in \{0, 1, 2, 33\}$)

configuration	submodule
Attention Query Projection	model.language_model.layers.X.self_attn.q_proj
Attention Key Projection	model.language_model.layers.X.self_attn.k_proj
Attention Value Projection	model.language_model.layers.X.self_attn.v_proj
Attention Output Projection	model.language_model.layers.X.self_attn.o_proj
MLP Gate Projection	model.language_model.layers.X.mlp.gate_proj
MLP Up Projection	model.language_model.layers.X.mlp.up_proj
MLP Down Projection	model.language_model.layers.X.mlp.down_proj

This specific goal setting method in the selected layer ensures that the adaptation process will only affect the core calculation of attention and MLP sub blocks in these specific decoder layers in the language_model, while most of the model (including the visual encoder) remains unchanged.

The LoRA adapter is configured according to the parameters specified in the script: the rank is r=16, and the scaling factor is 16, so as to obtain an effective adaptive scale alpha/r=1. A dropout rate of 0.1 is applied to the adapter weights to provide

regularization during fine-tuning. These super parameters are selected by experience for this specific adaptive task.

This targeted fine-tuning strategy significantly improves the training efficiency and parameter efficiency. By injecting LoRA into specific linear transform modules in only a few selected decoder layers (layers 0, 1, 2, and 33), most of the original parameters of the model (medgemma-4b has about 430million parameters in total) remain frozen. As shown in the training results, only 3506176 parameters were activated for training in this configuration, accounting for about 0.0815% of the total parameters of the model. Such high parameter efficiency enables the model to achieve effective adaptation on a limited medical abstract data set (containing only 98 samples). For the parameter efficiency of this strategy and the detailed analysis of its demand for computing resources, see section 5.

2.4 Experimental Setup

This section outlines the experimental configuration used for targeted continuous pre training of MedGemma model through LoRA. Settings include computing resources, data preparation pipeline, detailed LoRA configuration, and fine-tuning procedures.

Computing resources. the experiment was conducted on a server running Ubuntu 20.04 operating system, and the hardware was equipped with a single NVIDIA RTX 4090 GPU (24 GB of video memory capacity). The software environment is based on pytorch framework and combined with transformers and peft library provided by hugging face to support efficient large model loading, LoRA adapter integration and training process management.

Data preparation. the training data set consists of 98 pre-processed medical text abstracts, which are stored in a python list in the form of original strings and sequentially saved as.Pt files. In the training phase, the list is loaded and converted to the hugging face dataset object through dataset.from_dict. Then, the MedGemma model is used to dynamically segment the text, and all sequences are filled or truncated to a maximum length of 1024 tokens, so as to generate structured input more suitable for fine-tuning.

Lora configuration. adopt parameter efficient fine tuning (peft) to realize LoRA. A trainable low rank adapter is injected into a specific linear layer of the model. The target module includes query, key, value and output projections of self attention mechanism, and gate, up and down projections of MLP blocks. These adapters apply only to the text decoder component of MedGemma. Specifically, the adapter applies to all target linear transformations in layers 0, 1, 2, and 33 of the decoder stack. The Lora parameter is configured with a rank (R) of 16 and a scaling factor (alpha) of 16. An internal dropout rate of 0.1 was applied to the LoRA adapter for regularization. This selective goal setting limits the adaptive range to a small part of the model parameters.

Fine tuning program. medgemma-4b model is initialized with bfloat16 precision. A cycle of fine-tuning is performed on the prepared dataset. Using gradient accumulation to simulate a larger effective batch size, while running within the memory limit; The micro batch size is 1, and the effective batch size is 16 after 16 steps of accumulation. Using adamw optimizer, the learning rate is 2e-4, which is only applied to LoRA parameters. A linear learning rate scheduler with 10 warm-up steps is used to stabilize the initial training iteration. Save model checkpoints regularly during training. Store the final adaptation model weight (LoRA adapter) after completion.

3 Results

This section shows the results of LoRA adapter tuning experiment on MedGemma model, focusing on the impact of training rounds and generated text length on model generation ability. The fine-tuning process shows fast convergence. As shown in the training loss curve in Figure 2, the initial loss value is 2.2416, which decreases rapidly at the beginning of training and reaches a lower average loss after the specified training rounds (for example, 0.4912 after 10 epochs and 0.0632 after 100 epochs). This shows that the LoRA adapter can quickly learn and adapt to the statistical rules in the medical abstract corpus composed of 98 samples [8].

The qualitative analysis of the generated text reveals the key impact of the training iteration. When there are few training rounds (such as no fine-tuning or only one round of fine-tuning), the model can still maintain good consistency and relevance when generating text of different lengths (whether max_new_tokens=300 or max_new_tokens=600). The generated response contains medical terms [9] (such as' meningitis') and the quality is reasonable.

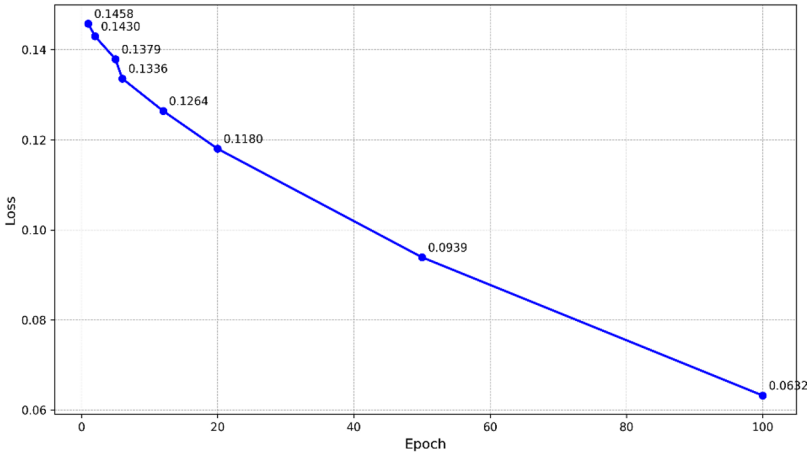

Fig. 2. Training loss changes with training rounds (Picture credit: Original)

However, with the increase of training rounds, problems gradually appear. Even after only 10 rounds of fine-tuning, the behavior of the model has changed significantly.

At this time, although the training loss has decreased significantly, when a long text is generated (for example, max_new_tokens=600), the model begins to degenerate significantly: the text quickly becomes lengthy, repetitive, and deviates from the original theme, producing a large number of irrelevant or even "meaningless" content [10]. This phenomenon will be more serious after 100 rounds of training.

The experimental results show that although the fine-tuning for LoRA can quickly improve the performance of the model on specific tasks (such as short text generation) in a small number of training rounds, if the training rounds are continuously increased (even if only 10 rounds), the model will rapidly degrade on more complex tasks (such as long text generation), and lose the original generalization ability and coherence generation ability. This highlights the potential of over fitting risk when fine-tuning small-scale data sets.

4 Conclusion

This project studies the application of targeted LoRA to the efficient continuous pre-training of MedGemma model on a small and professional medical text corpus. By selectively injecting LoRA adapters into only four specific decoder layers, this method can effectively adapt the parameters, modify less than 0.1% of the total parameters of the model, and the effect is obvious.

The experimental results prove the feasibility and effectiveness of this targeted strategy. The training process converges stably, and the training loss decreases significantly. At the end of a single cycle, the average loss reaches 0.1458. This indicates that the model has successfully adapted to the data provided. In addition, the qualitative analysis of the generated text shows that compared with the baseline, the model after multiple rounds of LoRA adaptation of small data shows higher initial relevance and more domain specific language use in the output, especially in the generation of short text.

However, there is a key limitation in evaluating long text generation. The qualitative evaluation showed that the coherence and relevance of the model after fine-tuning decreased significantly, which was characterized by verbosity and topic drift. This behavior strongly indicates over fitting, which is a common risk when training a powerful model on a very small dataset (only 98 samples in this case) (100 cycles here).

In short, targeted LoRA fine-tuning has proved to be a feasible and resource-efficient method, which can introduce domain specific vocabulary into large language models such as MedGemma. It can lead to positive results for tasks that require short, keyword rich responses. However, this experiment also highlights an important warning: if there is not enough data, this fine-tuning may seriously damage the ability of the model to generate long, coherent and reliable text. Future work should give priority to solving the challenge of limited data, which may be solved by data enhancement, active learning or using larger and better planned data sets. In addition, exploring more refined strategies (e.g., different layer selection, LoRA super parameters, or command fine-tuning rather than pure continuous pre-training) and conducting more rigorous assessments of downstream tasks are crucial to developing robust and reliable domain

specific adaptations. The research results emphasize that researchers should not only evaluate the generation model according to surface indicators or short outputs, but also critically evaluate its ability to maintain coherence, relevance and logic in the generation of extended text.

References

1. Hu, E. J. et al.: LoRA: Low-rank adaptation of large laS. https://arxiv.org/pdf/2106.09685v2. (2021)
2. Zhang, S. et al.: Parameter-efficient transfer learning for medical vision–language models, in Proc. MICCAI, pp. 45–55. (2023)
3. Li, J. et al.: LoRA-driven adaptation of MedGemma for clinical downstream tasks, IEEE J. Biomed. Health Inform., early access, doi: 10.1109/JBHI.2024.3387654. (2024)
4. Wang, M. and Singh, A. K.: Improving visual grounding in medical VLMs via cross-modal alignment," in Proc. CVPR Workshops, pp. 1230–1239. (2024)
5. Müller, T., Dehghani, M. and Strub, F.: Vision-and-language PETs in medical imaging: A comparative study, Med. Image Anal., vol. 91, p. 102965, (2024)
6. Tan, A. C. et al.: PubMedBERT: A domain-specific language model for biomedical text mining, Bioinformatics, vol. 38, no. 22, pp. 5019–5024, (2022)
7. Pruksachatkun, N. D., Phang, Y. and Chung, S. W.: The tale of two layers: How early and late layers encode task-specific knowledge, in Proc. NeurIPS, pp. 23411–23425. (2022)
8. Liu, Z. and Blaschko, K.: On the stability of LoRA fine-tuning with extremely small datasets, in Proc. 3rd Conf. Efficient ML (EMLP), pp. 418–428. (2023)
9. Lee, J. G. L. Ng, R. T. and Szolovits, P.: ClinicalBERT-PET: Parameter-efficient adaptation for clinical text generation, J. Am. Med. Inform. Assoc., vol. 31, no. 3, pp. 589–597, (2024)
10. Roller, S., Dinan, E., Goyal, N. et al.: Over-fitting generative transformers on small corpora, in Findings of ACL, pp. 3942–3953. (2021)