test

LLaV A-MoLE Sparse Mixture of LoRA Experts for Mitigating Data Conflicts

## Abstract

Instruction finetuning on a variety of image-text instruc- tion data is the key to obtaining a versatile Multimodal Large Language Model (MLLM), and different configura- tions of the instruction data can lead to finetuned models with different capabilities. However, we have discovered that data conflicts are inevitable when mixing instruction data from distinct domains, which can result in performance drops for tasks of a specific domain. To address this issue, we propose to apply an efficient Mixture of Experts (MoE) design, which is a sparse Mixture of LoRA Experts (MoLE) for instruction finetuning MLLMs. Within the Transformer layers, we extend the popular Low-Rank Adaption (LoRA) method by creating a set of LoRA experts specifically for the MLP layer, and route each token to the top-1 expert based on a routing function, allowing adaptive choices for tokens from different domains. Since the LoRA experts are sparsely activated, the training and inference cost are kept roughly constant compared to the original LoRA method. By replac- ing the plain-LoRA of LLaVA-1.5 with our MoE design, our final model is named LLaVA-MoLE. Extensive experiments proved that LLaVA-MoLE effectively mitigates the data con- flict issue when mixing multiple distinct instruction datasets with various configurations, and achieves consistent per- formance gains over the strong plain-LoRA baselines. Most importantly, on the mixed datasets, LLaVA-MoLE can even outperform the plain-LoRA baseline trained with twice the samples.

## Introduction

Large language models (LLMs) [1, 3] have demonstrated their remarkable capabilities in following human instruc- tions to complete various tasks, and one of the key to obtain such capability is instruction finetuning (or supervised fine- tuning, SFT) [41]. Similarly, efforts have been made to cre- ate instruction-finetuned multimodal large language models (MLLMs), which connect pre-trained vision encoders with LLMs, resulting in models that are capable of answering questions given visual and textual inputs. 306.322.2949.85 279.762.3660.99 251.213.9118.45 295.868.49117.42 307.372.4122.53 General Benchmark scor e Document Benchmark scor eBioMedicine Benchmark scor eLLaVA-1.5 LLaVA-Doc LLaVA-Med LLaVA-Mix LLaVA-MoLEModelsFigure 1. Model performances on three benchmarks when trained with different data configurations. LLaV A-1.5, LLaV A-Doc, and LLaV A-Med are trained on general multi-task, document, and biomedicine datasets, respectively. While LLaV A-Mix and LLaV A-MoLE are both trained on the mixture of all three datasets. The performance of LLaV A-Mix the document benchmark bene- fits from mixing all datasets, however, the performance all other benchmarks drops after mixing. Our proposed LLaV A-MoLE suc- cessfully resolves data conflicts and maintains high performances on all benchmarks. Although a pre-trained LLM (7B/13B parameters) [9, 39] is usually included in a MLLM, the multimodal in- struction training data still dominates the capability of the trained MLLMs. Thus a large portion of the MLLM- training effort is assigned to constructing high-quality and diverse multimodal instruction data. For example, LLaV A- 1.5 [23] carefully selected a wide range of academic task- oriented data and controlled the data size of each task. The resulting LLaV A-1.5 model demonstrates strong perfor- mances on benchmarks of various common vision-language tasks. Other successful multimodal instruction finetun- ing datasets [6, 22] are also constructed with a carefully designed data configuration. In addition to data, a pop- ular and effective parameter-efficient finetuning method named LoRA (Low-Rank Adaptation) [15] is also the key to LLaV A-1.5’s success. LoRA reduces the number of train- able parameters of Transformers by freezing the pre-trained model weights training only an injected pair of low-rankarXiv:2401.16160v2 [cs.CV] 30 Jan 2024 decomposed weight matrices for each linear layer, which makes it faster to finetune pre-trained large models and is widely adopted in MLLM finetuning [4, 23, 45, 48, 50]. However, when data configuration is critical to MLLMs, we find in our preliminary studies that current MLLMs trained with plain LoRA are sensitive to the training data configuration. As shown in Figure 1, we adopt three in- struction tuning datasets from different domains: 1) a gen- eral multi-task dataset that contains a mixture of various vision-language instructions data, 2) a document-oriented dataset built for chart, table, and document understanding, and 3) a biomedicine dataset consists of question-answer pairs on pathology images. Three models are finetuned on each dataset, the resulting models are named LLaV A- 1.5, LLaV A-Doc, and LLaV A-Med, respectively. To test the finetuned model’s capability on each domain, three in- dividual benchmarks are employed1. When the MLLM is finetuned on each individual dataset, it achieves reasonable performance on the corresponding benchmark. But when mixing the document and biomedicine dataset with the gen- eral dataset, the trained LLaV A-Mix’s performance on the general benchmark drops considerably from 306.3 to 295.8, which means there is a conflict incurred by adding data that are distinctly different from general multi-task instruc- tions. This greatly hinders extending a MLLM’s abilities by adding training data from novel domains. To address the above mentioned issue, we propose to ap- ply a sparse mixture of LoRA experts to LLaV A-1.5 for in- struction finetuning, resulting in our proposed model named LLaV A-MoLE. We extend the common LoRA finetuning paradigm used by LLaV A-1.5 and many other MLLMs. Concretely, we redesign how LoRA is applied to the MLPs in the Transformer layers of the LLM. Instead of adding only one pair of low-rank decomposed matrices to the orig- inal linear layer, we introduce a set of experts with the same structure as the original LoRA but different weights. Then for each token, these experts are sparsely routed by a router function conditioned on the token embedding, i.e., only one LoRA expert is activated and its output of the to- ken is added to the original MLP’s. Since the image and text tokens from different domains can exhibit distinct fea- tures, they are routed to different experts and the MLLM’s ability to handle multiple domains is expanded. In our ex- tensive experiments on various data configurations, we dis- cover that LLaV A-MoLE can effectively mitigate the con- flicts between different instruction datasets, while maintain- ing roughly the same computational cost as the plain-LoRA model. We will further show in Sec.4.3 that under data conflicts, even if the plain-LoRA model is trained on twice the samples (by repeating each dataset in the mixture), its scores on the general benchmark can continue to increase but still fall behind LLaV A-MoLE. In this case, LLaV A- 1The details of these datasets and benchmarks are in Sec 4.2MoLE can achieve better performance with half the training iterations, which is a significant cost reduction. The contributions of this paper are summarized as fol- lows: 1. Based on an advanced MLLM model and large scale datasets, we identify the data conflict issue when instruc- tion finetuning a MLLM on a mixture of distinctly dif- ferent instruction datasets.

## Conclusions

In this paper, we first identified the data conflict issue when instruction finetuning multimodal large language models on a mixture of datasets from multiple distinct domains. To address this issue, we propose LLaV A-MoLE, which uses a sparse mixture of LoRA experts to improve the plain-LoRA architecture. It uses a set of LoRA experts for the MLP layers and routes each token to the top- 1 expert. Since only the selected expert is activated to execute computation, the actual computational cost for the entrie model is kept roughly the same as a normal LoRA model. In the meantime, our LLaV A-MoLE ef- fectively mitigates the data conflict and achieves a consis- tent performance improvement over the plain-LoRA base- lines on a variety of data configurations. We further ver- ified that LLaV A-MoLE performs similarly with a dense MoE model while requiring significantly less computa- tional resources, which is particularly advantageous for samples with long context length. For our future work, it would be interesting to apply our method to the multi- task pre-training stage of the MLLMs, where a much larger number of training examples from multiple domains are mixed.

==================================================

MoELoRA Contrastive Learning Guided Mixture of Experts on

## Abstract

Fine-tuning is often necessary to enhance the adaptability of Large Language Models (LLM) to downstream tasks. Nonetheless, the process of updating billions of parameters demands significant computational resources and training time, which poses a substantial obstacle to the widespread application of large-scale models in various scenarios. To address this issue, Parameter-Efficient Fine-Tuning (PEFT) has emerged as a prominent paradigm in recent research. However, current PEFT approaches that employ a limited set of global parameters (such as LoRA, which adds low-rank approximation matrices to all weights) face challenges in flexibly combining different computational modules in downstream tasks. In this work, we introduce a novel PEFT method: MoELoRA. We consider LoRA as Mixture of Experts (MoE), and to mitigate the random routing phenomenon observed in MoE, we propose the utilization of contrastive learning to encourage experts to learn distinct features. We conducted experiments on 11 tasks in math reasoning and common-sense reasoning benchmarks. With the same number of parameters, our approach outperforms LoRA significantly. In math reasoning, MoELoRA achieved an average performance that was 4.2% higher than LoRA, and demonstrated competitive performance compared to the 175B GPT -3.5 on several benchmarks. Keywords: Large Language Models, Mixture of Experts, Parameter Efficient Fine-tuning, Contrastive Learning

## Introduction

With the rapid advancement of Large Language Models (LLMs) such as GPT3 (Brown et al., 2020), BLOOM (Scao et al., 2022) and LLaMA (Touvron et al., 2023), the successful application of self- supervised pretraining on unlabeled text data has presented unprecedented opportunities for enhanc- ing downstream tasks. However, to fully harness the potential of these LLMs in practical applications, it is also necessary to continuously fine-tuning (Wei et al., 2021; Chung et al., 2022) the LLMs based on the training data of specific tasks to meet the per- formance requirements of downstream tasks. The substantial number of parameters, often exceeding one billion, makes fine-tuning these LLMs a costly endeavor, demanding a significant investment in computational resources (Figure 1a). [1]Therefore, in recent years, Parameter-Efficient Fine-Tuning (PEFT) (Mangrulkar et al., 2022; Zhang et al., 2023) techniques have emerged with the aim of reducing the cost of fine-tuning by freezing cer- tain model weights or introducing smaller trainable modules. In the continual exploration within this field, a series of methods such as LoRA (Hu et al., 2021), AdaLoRA (Zhang et al., 2023), [2]Adamix (Wang et al., 2022), QLoRA (Dettmers et al., 2023) and LoRAHub (Huang et al., 2023) have emerged, each offering unique perspectives on efficiently fine-tuning Large Language Models for better ap- \* Equal Contributions.plicability in downstream tasks. LoRA (Figure 1b) introduces the concept of LoRA rank to reduce the number of trainable parameters. AdaLoRA builds upon LoRA’s foundation, [3]achieving a search-free approach that greatly simplifies the fine-tuning pro- cess. Adamix combines the MoE with Adapters to surpass the performance of LoRA. LoRAHub employs a gradient-free method (Liu et al., 2020) to perform weighted combinations of multiple LoRA weights, thereby better adapting to new down- stream tasks. However, current PEFT approaches that employ a limited set of global parameters face challenges in flexibly combining different computational mod- ules in downstream tasks. Inspired by methods such as Mixture of Experts (MoE), Adamix, and LoRAHub, we propose a novel PEFT approach named MoELoRA. This method considers LoRA as a Mixture of Experts, leveraging the modeling capabilities of multiple experts for complex data domains, as well as utilizing LoRA’s parameter- efficient characteristics. As well as Figure 1c, dur- ing both training and inference, only the LoRA se- lected by the gating network will be activated and only these "experts" relevant to specific tasks will participate in gradient updates or forward inference. However, applying MoE to LoRA presents chal- lenges. Firstly, under the MoE architecture, gating network doesn’t exhibit a preference for a particular expert, leading to a certain level of routing random- ness (Zuo et al., 2021). Secondly, guiding experts to learn distinct features poses a challenging task.arXiv:2402.12851v1 [cs.CL] 20 Feb 2024 Pretrained Weights A1B1 AnBn A2B2 Gating Network Input HiddenGate Select+ Output HiddenLoad Balance LossContrastive Loss(C) MoELoRA Pretrained Weights AB Input Hidden+ Output Hidden(b) LoRA Pretrained WeightsΔW Input Hidden+ Output Hidden(a) Fine -TuningFigure 1: The Different Architectures for (a)Fine-Tuning, (b)LoRA and (c)proposed method MoELoRA. ∆Wdenotes the gradient increment for the downstream tasks. LoRA decomposes ∆Winto two matrices AandBand our proposed MoELoRA can select AiandBicorresponding to a specific task for better adaptation. In order to differentiate the capabilities of different experts, we employed contrastive learning on the outputs of the experts. To address these issues, we introduce con- trastive learning among experts. Through this contrastive learning approach, we treat the out- puts of the same expert as positive samples and the outputs of different experts as negative sam- ples, encouraging experts to learn distinct features. In the end, we achieve performance surpassing LoRA under the same number of parameters. In math reasoning, MoELoRA averaged 4.2% higher performance than LoRA, and in common-sense reasoning, it averaged 1.0% higher than LoRA. Furthermore, MoELoRA exhibits competitive per- formance compared to the 175B GPT -3.5 on a few benchmarks. In summary, our work makes the following con- tributions: (1) We consider LoRA as Mixture of Experts and propose a novel PEFT method named MoELoRA, which leverages the MoE architecture to achieve dynamic combinations of multiple LoRA modules, better catering to the requirements of downstream tasks. (2) In response to the random routing issue in using Mixture of Experts (MoE) for LoRA fusion, we propose employing contrastive learning to en- courage experts to learn distinct features. (3) We conduct experiments on 11 datasets for math reasoning and common-sense reason- ing tasks, demonstrating that our approach outper- forms LoRA in all tasks. The results of ablation ex- periments also show improvement in downstream tasks with contrastive learning. Furthermore, we perform tracking analysis of MoE routing to under- stand the impact of our method on the model’s decision-making process.2. Related Work

==================================================

# References

==================================================

[1] Nima Aghli and Eraldo Ribeiro. Combining weight pruning and knowledge distillation for cnn compression. In Proceedings of the IEEE/CVF conference on computer vision and patternrecognition , pages 3191–3198, 2021.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Has-  
son, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring,  
Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne  
Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand  
Sharifzadeh, Mikoł aj Bi ´nkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and  
Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo,  
S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neu-  
ral Information Processing Systems , volume 35, pages 23716–23736. Curran Associates,  
Inc., 2022. URL https://proceedings.neurips.cc/paper\_files/paper/2022/file/  
960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf .

[3] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? Proceedings of machine learning and systems , 2:129–146, 2020.