Enhancing VQA in Multimodal Large Language Models through PEFT and RLHF

Geerisha Jain, Kritanjali Jain, Shaurya Singh

Language Technologies Institute, School of Computer Science, Carnegie Mellon University

Motivation

The ability to generate contextually accurate and human-like responses to visual questions is a cornerstone of multimodal artificial intelligence applications. However, current Multimodal Large Language Models (MLLMs) often struggle with hallucination and misalignment between visual inputs and textual outputs, which undermines their reliability and trustworthiness.

In this project, we aim to enhance the qualitative performance of MLLMs in Visual Question Answering (VQA) tasks by integrating Parameter-Efficient Fine-Tuning (PEFT) techniques, specifically LoRA, with Reinforcement Learning from Human Feedback (RLHF) via Direct Preference Optimization (DPO). Additionally, prompt-templating strategies will be employed to structure inputs effectively.

By leveraging human-preference-aligned datasets, our work aspires to develop models that not only produce accurate and visually grounded answers but also produce more human-like responses.

Literature Review

Efforts to align Multimodal Large Language Models (MLLMs) with human preferences and factual correctness have advanced significantly. Sun et al. (2023) [1] introduced Factually Augmented RLHF, incorporating image captions to reduce hallucinations and enhance vision-language alignment using Proximal Policy Optimization (PPO) . Kosmos-1 [2] unified multimodal tasks under a single architecture using self-supervised pre-training, focusing on parameter-efficient alignment.

Direct Preference Optimization (DPO) has emerged as a robust alternative to PPO for human preference alignment, addressing challenges like modality conflicts and catastrophic forgetting in MLLMs. Studies comparing DPO and PPO [3] show their complementary strengths, with PPO excelling in real-world applications and DPO achieving better generalization. CLIP-DPO [4] further refined preference optimization by leveraging CLIP models to rank outputs, reducing hallucinations without external data.

Parameter-efficient methods like PE-RLHF, employing LoRA, have drastically lowered computational costs while maintaining competitive performance, particularly in VQA tasks. Frameworks like Muffin [5] and datasets like UniMM-Chat have pushed the boundaries of multimodal instruction-following by integrating vision-language models as universal assistants. Together, these approaches inform our strategy of combining PEFT with RLHF to enhance MLLMs for robust, human-aligned VQA.

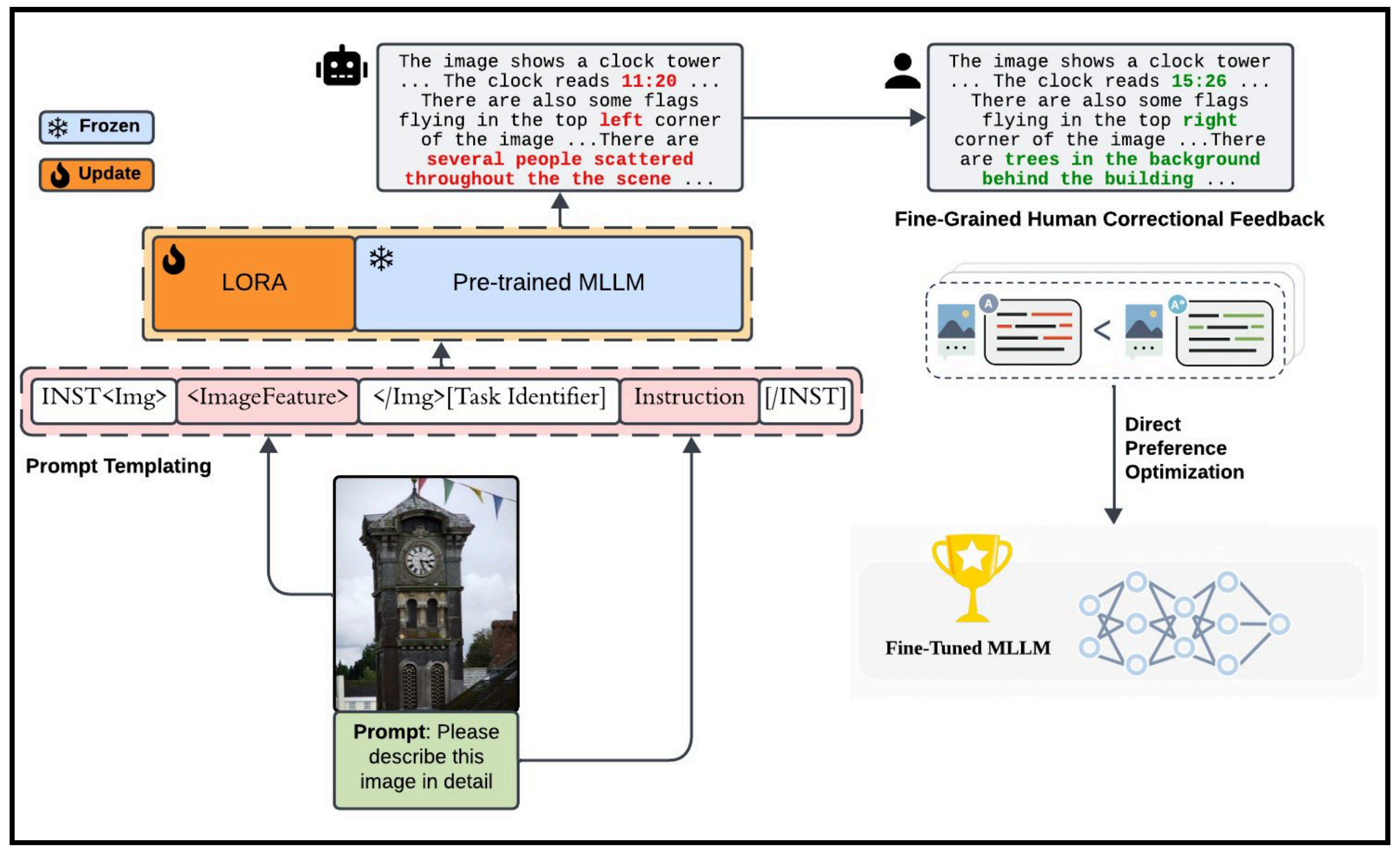
Training and Evaluation Datasets

The RLHF-V-Dataset is a specialized resource for fine-tuning Visual Question Answering models using Reinforcement Learning from Human Feedback (RLHF). It contains 4,733 preference pairs for training, capturing human corrections across diverse instructions to align model outputs with human expectations. For evaluation, a subset of 1,000 preference pairs is used to provide detailed feedback, enabling thorough performance assessment. The dataset is designed to reduce hallucinations by 34.8% while preserving informativeness, making it highly effective for robust training and evaluation. By incorporating challenging real-world examples, the RLHF-V-Dataset enhances model alignment with human-like reasoning and mitigates biases in multimodal tasks.

Methodology

Reinforcement Learning from Human Feedback (RLHF) with Direct Preference Optimization (DPO)

We fine-tune the LLaVa model using human preference annotations from the RLHF-V-Dataset. DPO directly aligns the model's outputs with human preferences, bypassing the need for reward models and policy gradient methods like PPO. This approach simplifies the fine-tuning process while improving response alignment with human expectations, particularly in Visual Question Answering (VQA) tasks.


Frozen
Update
The image shows a clock tower ... The clock reads 11:20 ... There are also some flags flying in the top left corner of the image ...There are several people scattered throughout the the scene ...
The image shows a clock tower ... The clock reads 15:26 ... There are also some flags flying in the top right corner of the image ...There are trees in the background behind the building ...
Fine-Grained Human Correctional Feedback
LORA
Pre-trained MLLM
INST<Img>
<ImageFeature>
</Img>[Task Identifier]
Instruction
[/INST]
Prompt Templating
<
Direct Preference Optimization
Fine-Tuned MLLM
Prompt: Please describe this image in detail

System Architecture: Overview of System

LoRA for Parameter-Efficient Fine-Tuning (PEFT)

Low-Rank Adaptation (LoRA) is employed as a PEFT technique to fine-tune LLaVa efficiently. By introducing trainable low-rank matrices to adapt a subset of the model's parameters, LoRA significantly reduces computational overhead while maintaining performance. This method is ideal for optimizing large multimodal models with limited resources.

Prompt Templating

Prompt templating is used to structure input-output formats, enhancing the model's ability to integrate visual and textual information. These templates guide the model in producing consistent and contextually relevant responses, ensuring alignment with the desired answer structure in VQA tasks. This method supports improved coherence and contextual grounding in the generated outputs.

Evaluation Metrics

The proposed MLLM for Visual Question Answering (VQA) will be evaluated using a combination of quantitative and qualitative metrics. Quantitative metrics include **ROUGE** and **BLEU**, which measure n-gram overlap, precision, and brevity to assess how closely the generated responses match reference answers.

CIDEr evaluates semantic alignment by rewarding content that is consistent with the reference answers and penalizing hallucinated or irrelevant content, making it particularly suited for assessing context-aware VQA.

METEOR provides a nuanced evaluation of linguistic quality by accounting for synonyms, paraphrasing, stemming, and word order, offering deeper insights into the semantic equivalence of model outputs.

Results & Analysis

Metric	Base LLaVa	LLaVa with PPO	LLaVa with DPO and LoRA	LLaVa with DPO and LoRA and Prompt Templating
ROUGE-1 Score	0.4415	0.5610	0.4713	0.3487
ROUGE-2 Score	0.1936	0.2738	0.2131	0.1369
ROUGE-L Score	0.3224	0.4924	0.3220	0.2434
BLEU Score	0.1791	0.2092	0.1894	0.2243
CIDEr	0.4752	0.5102	0.5690	0.2182
METEOR	0.3101	0.3590	0.3082	0.2182

- LLaVa with PPO: Best overall performance; highest ROUGE and METEOR scores, indicating strong improvements.
- LLaVa with DPO and LoRA: Mixed results; notable CIDEr increase (0.5690) but lower ROUGE scores, suggesting trade-offs.
- LLaVa with DPO, LoRA, and Prompt Templating: BLEU score improves (0.2243), but other metrics decline, indicating potential syntactic gains at the expense of semantic quality.

Future Work

Advanced prompt engineering techniques can be explored. Beyond the current use of prompt templating, gradient-based search methods could be used to optimize discrete prompt representations, making desired outputs more likely. Additionally, prompt tuning, which involves directly optimizing embeddings fed into the LLM, could bypass the need for discrete prompts altogether, resulting in more adaptive and efficient input representations.

The model can be extended to handle a wider range of vision-language tasks, aiming for improved response quality across diverse scenarios. This focus on generalizability would enable the model to perform effectively in varied multimodal challenges, broadening its applicability.

Future evaluations could prioritize real-world performance by shifting from traditional metrics to comprehensive benchmarks that simulate practical tasks. This would provide a more accurate reflection of how an agent based on the MLLM performs in real-world scenarios, ensuring its reliability and utility in practical applications.

References

[1] Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., et al. (2023). Aligning large multimodal models with factually augmented RLHF. arXiv preprint arXiv:2309.14525.
[2] Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., et al. (2023). Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:2302.14045.
[3] Xu, S., Fu, W., Gao, J., Ye, W., Liu, W., et al. (2024). Is DPO superior to PPO for LLM alignment? A comprehensive study. arXiv preprint arXiv:2404.10719.
[4] Ouali, Y., Bulat, A., Martinez, B., & Tzimiropoulos, G. (2024). CLIP-DPO: Vision-language models as a source of preference for fixing hallucinations in LVLMs. arXiv:2408.10433.
[5] Yu, T., Hu, J., Yao, Y., Zhang, H., Zhao, Y., et al. (2023). Reformulating vision-language foundation models datasets towards universal multimodal assistants. arXiv:2310.00653.