VICT: Visual In-Context Tuning

Young Kyun Jang[1], Dat Huynh[1], Zihang Meng[1], and Ser-Nam Lim[1,2]

Meta[1], University of Central Florida[2]

kyun0914@gmail.com, badat0202@gmail.com, zihangm@cs.wisc.edu, sernam@gmail.com

Abstract

Bridging the gap between vision and language understanding, we present a novel approach to integrate Large Language Models (LLMs) with visual understanding capabilities, establishing Vision-LLMs. Drawing inspiration from instruction-following tuning methodologies, we design an innovative framework that aligns a new LLM with an existing vision encoder and bridging module, using only a limited amount of training data and a small number of additional trainable parameters. A key component of our work is the introduction of Visual In-Context Tuning (VICT), a mechanism that exploits the inherent context of an image during the fine-tuning process to enhance the model's ability to deliver accurate, context-grounded responses. Demonstrating improved performance in multimodal evaluation protocols such as image captioning and visual question answering, our approach underscores the potential of Vision-LLMs in advancing the field of vision-language understanding.

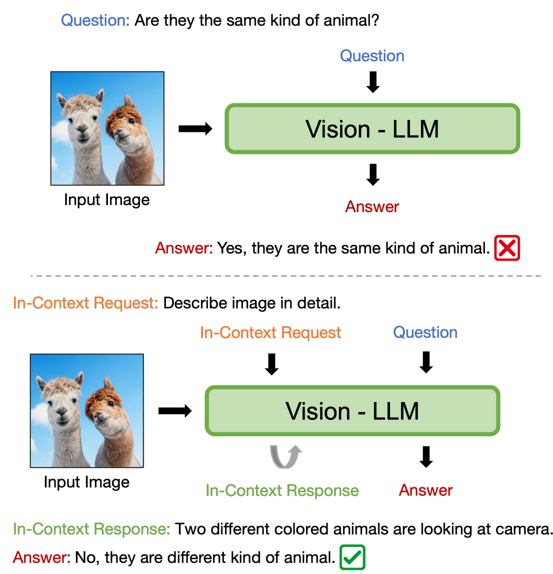

Figure 1: In-context reasoning on visual data induce better understanding for LLM.

1. Introduction

Large Language Models (LLMs), such as billion-scale auto-regressive transformers, have exhibited extraordinary proficiency in comprehending and generating natural language [3, 17, 4, 15]. By utilizing in-context few-shot examples, LLMs can successfully carry out a diverse array of Natural Language Processing (NLP) tasks by reasoning with inherent intelligence. Furthermore, the method of instruction-following tuning on LLMs has demonstrated significant success. This is evident in models like InstructGPT [12] - the technology underpinning ChatGPT [11] - and Stanford's Alpaca [14] and Vicuna [1]. These techniques focus on teaching the models to harness the vast knowledge embedded in LLMs through specific instructions, allowing them to generate appropriate answers, improving their adaptability and utility across a wide range of NLP cases.

Expanding on this, integrating visual understanding into LLMs to create *Vision-LLMs* has emerged as an interesting research direction. Since images cannot be directly interpreted as inputs for LLMs, strategies have been developed to bridge this domain gap. By keeping the pre-trained vision encoder and LLM fixed, and learning the alignment between them using a large volume of web-collected image-text paired data, methods such as Flamingo [2] and BLIP-2 [9] have demonstrated decent performance. Their capabilities in zero-shot and few-shot generalization have shown particular promise in open-ended vision and language tasks.

Drawing inspiration from the impressive success of instruction tuning in NLP, our study aims to apply similar principles to fine-tune a LLM that achieves visual understanding, even with a restricted amount of training data. Specifically, our objective is to incorporate a newly developed LLM, known as LLaMA [15], along with the fixed vision encoder and the bridging module Q-Former (QF) of BLIP-2 [9]. However, since QF has been pre-trained to connect the vision encoder with different LLMs such as OPT [17] and Flan-T5 [4], it could pose a challenge. To this end, we provide extra room for LLaMA to be fine-tuned and enhance the understanding of QF outputs. Particularly, we fix the original LLaMA parameters and incorporate small, trainable parameters using Parameter-Efficient Fine-Tuning (PEFT) methods [6, 16] This approach not only provides extra training capacity but also mitigates the risk of catastrophic forgetting of LLM, caused by misalignment between language and visual data.

To validate the feasibility of our approach, we performed a sanity check, examining the distribution of distances between the image token embeddings generated by QF and the word embeddings in the LLM's vocabulary. We found that even when a QF embedding is most closely aligned with a word in the LLM's vocabulary, the distance between them remains substantial. It implies that QF captures broad and rich nuanced range of semantics from images, not confined to specific vocabulary words. This capability of encoding images into robust vision-language space suggests the potential of QF to be applicable across various language models.

Finally, we introduce **V**isual **I**n-**C**ontext **T**uning (VICT). Unlike conventional in-context few-shot learning methods in the NLP field, which typically exclude training, VICT is specifically designed to incorporate in-context examples based on the image while fine-tuning Vision-LLM. In this case, *"in-context"* refers to the inherent context of the image itself, as illustrated in Figure 1. During inference, for example, when the model is provided with an image and a corresponding request, such as a question, a detailed description of the image in natural language is initially generated. Then, this description serves as the in-context basis for further reasoning, ultimately aiding the model in delivering more accurate and focused responses. This process extends the model's reasoning capabilities and supports to produce targeted responses that are deeply grounded in the visual context.

2. Visual Instruction Tuning on LLM

Overview. Our goal is to empower Large Language Model (LLM) with the ability to perform visual understanding. To achieve this, we align output image embedding achieved from pre-trained vision encoder and Q-Former (QF) [9] to serve as inputs for LLM. Additionally, we introduce LLM's PEFT for fine-tuning, enabling LLM to leverage its capabilities for multi-modal reasoning. Ultimately, we establish Vision-LLM configured with LLaMA [15].

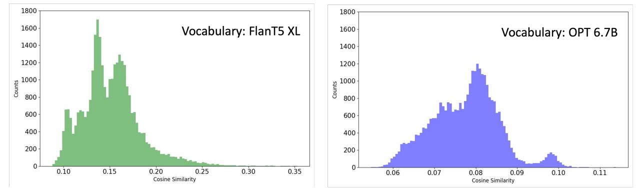

Figure 2: Histogram of cosine similarity values between Q-Former's output tokens and the nearest word embedding.

2.1. Investigation

Before applying the fixed vision encoder and QF to the new LLM, we first investigate the distribution of the image token embeddings generated by QF that go into the matching LLM (OPT [17] and Flan-T5 [4]). As shown in Figure 2, we measure the cosine similarity between the QF tokens and the word embeddings in the vocabulary, and plot the highest similarity scores as a histogram. Upon analyzing the results, we observe in both cases that QF tokens do not correspond to specific words, the maximum similarity is 1 but on average it is near 0. This outcome suggests that QF is capable of encapsulating a wide and detailed spectrum of image semantics, not limited to explicit vocabulary words. It also implies that even if a new LLM has a different vocabulary, it could potentially reason from QF tokens as long as it's able to transform the space into one that it can interpret through linear mapping.

2.2. Framework

Inspired by the above observation, we configure our Vision-LLM fine-tuning framework, utilizing QF to bridge between vision encoder and LLM, as depicted in Figure 3. Depending on the specific vision encoder used in the training of QF, we utilize either ViT-L [13] or ViT-G [5]. The parameters of vision encoder and QF are kept fixed during training to preserve QF's vision-language alignment capabilities. In addition, to preserve QF's representation and simultaneously enhance the interpretability toward new LLM, we fix the original query tokens that are used as prompts for QF's and add extra trainable prompt tokens. Given that the vocabularies of the LLMs used in training QF (OPT, Flan-T5) differ from that of the target LLM (LLaMA), we incorporate a trainable single linear projection layer to ensure appropriate mapping between them. In order to provide extra room for LLaMA to understand QF output tokens, we add a small number of trainable parameters (delta), through the Parameter Efficient Fine-Tuning (PEFT) technique, LoRA [6] and LLaMA-Adapter [16]. By integrating the entire components, we establish single stage end-to-end training framework for Vision-LLM tuning.

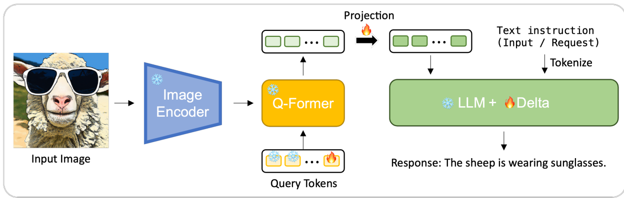

Figure 3: Overall Vision-LLM tuning framework. We denote trainable components as (flame) and non-trainable ones as (snowflake). Delta in LLM represents parameters from PEFT.

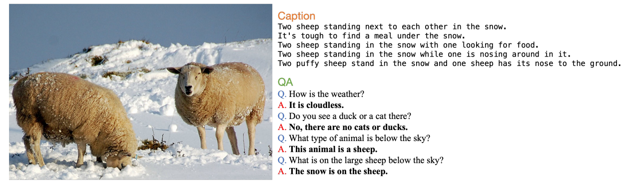

Figure 4: Image-text paired data example for visual instruction tuning.

2.3. Training

Visual Instruction Example. Instead of using noisy yet vast amounts of image-text paired data, we leverage two types of training datasets that are smaller in volume but meticulously annotated. The first is the image captioning dataset COCO[10], which features a few different captions describing a single image. This dataset is well-suited for helping the model gain an understanding of the diverse and general descriptions that can be derived from an image. The second dataset, GQA [7], used for Visual Question Answering (VQA), comprises compositional questions pertaining to a single image, with each question corresponding to two types of answers: detailed full answer and brief short answer. It requires complex but focused understanding, such as spatial understanding and multi-step inference for the scene, and is designed to be more challenging than general VQA tasks [8], which only guesses short answers, by preventing educated guesses using language and world priors. It is important to note that the two datasets share some images while having different annotations as shown in Figure 4, allowing for simultaneous captioning and VQA on a single image, which in turn facilitates in-context visual reasoning.

Visual In-Context Tuning. In the realm of vision-language understanding, key challenges lie in accurate scene understanding, finding alignment between image and text, and generating precise language to describe these visual elements. As LLMs do not directly perceive images, they need to be tuned to perform visual reasoning. Additionally, our objective is to fine-tune the LLM that can generate responses to a wide range of questions or provide a comprehensive description of a single image. In doing so, the model will develop the capacity to understand and infer visual in-context relevant to the specific task.

For a given image $\mathbf{X}$ and its corresponding caption $\mathbf{C}$, question $\mathbf{Q}$, and answer $\mathbf{A}$, we aim to train Vision-LLM to predict $\mathbf{C}$ or $\mathbf{A}$, regarding the input instruction prompt. We configure five prompts: I_{cap}, I_{qa}, $\mathrm{I}_{\text{qa-cap}}$, $\mathrm{I}_{\text{cap+q-a}}$, and $\mathrm{I}_{\text{qa+q-a}}$ where I_{cap} prompts the model to describe an image to produce caption, and I_{qa} prompts the model to provide an sentence-like answer to a given question. $\mathrm{I}_{\text{qa-cap}}$, $\mathrm{I}_{\text{cap+q-a}}$, and $\mathrm{I}_{\text{qa+q-a}}$ takes extra inputs that provides context inherent in the image, such as caption or another question-answer pair. In the case of $\mathrm{I}_{\text{qa-cap}}$, model takes $\mathbf{X}$, $\mathbf{Q}$ and $\mathbf{A}$ as inputs and predicts $\mathbf{C}$, and for the case of $\mathrm{I}_{\text{cap+q-a}}$, model takes $\mathbf{X}$, $\mathbf{C}$ and $\mathbf{Q}$ as inputs and predicts $\mathbf{A}$, and for the case of $\mathrm{I}_{\text{qa+q-a}}$, model takes $\mathbf{X}$, $\mathbf{Q}'$, $\mathbf{A}'$ (different question-answer pair from the same image) and $\mathbf{Q}$ as inputs and predicts $\mathbf{A}$. We refer to this approach—wherein the model generates outputs based on cross-referencing—as "Visual In-Context Tuning" (VICT), given that the model's responses are predicated on in-context speculations that can be drawn from the image. The training objective here is as same with the original auto-regressive cross entropy loss [3, 15], where we only compute gradient on the model outputs similar to instruction tuning [9, 12, 14, 1].

During the inference time, the Vision-LLM trained via VICT can conduct multi-modal reasoning by utilizing in-context examples as inputs. Users can supply either a caption or a question-answer pair, enabling Vision-LLM to maintain context throughout the conversation. Alternatively, VICT can autonomously execute on the Vision-LLM side, self-improved reasoning manner. This method generates predictions using either I_{cap}, or I_{qa} with the given question, and employs these predictions as input for subsequent VICT prompts ($\mathrm{I}_{\text{qa-cap}}$, $\mathrm{I}_{\text{cap+q-a}}$, and $\mathrm{I}_{\text{qa+q-a}}$), in order to produce the correct response.

3. Experiments and Conclusion

In this section, we examine the potential effectiveness of our proposed Visual In-Context Tuning (VICT) method, which is designed to align the vision encoder and q-former with a new Language Model, with the aim of producing an effective Vision-LLM efficiently. Our primary objective is to investigate whether the Vision-LLM is capable of generating valid text derived from images. To explore this, we conduct experiments using two benchmark tasks, namely captioning and visual question answering.

As shown in Table 1 and 2, our approach leverages the principles of instruction-following tuning to align a new LLM with a pre-existing vision encoder and bridging mod-

Table 1: Image captioning results. In each group, the best scores have been highlighted in bold.

Method	BLEU-1	BLEU-4	METEOR	ROUGE$_L$	CIDEr	SPICE
The original BLIP-2 paired with different LLMs						
BLIP-2 OPT2.7	0.699	0.363	0.273	0.576	1.270	0.217
BLIP-2 OPT6.7	0.634	0.301	0.247	0.535	1.093	0.196
BLIP-2 Flan-T5XL	**0.850**	**0.420**	**0.312**	**0.614**	**1.495**	**0.248**
BLIP-2 Flan-T5XXL	0.793	0.360	0.289	0.588	1.302	0.228
LLM: LLaMA 7B, PEFT: LLaMA-Adapter, (Q-Former Source)						
VICT (OPT2.7)	0.833	0.419	0.315	**0.617**	**1.497**	**0.250**
VICT (OPT6.7)	0.819	0.405	0.303	0.601	1.409	0.233
VICT (Flan-T5XL)	0.830	0.414	**0.312**	0.612	1.448	0.242
VICT (Flan-T5XXL)	0.804	0.391	0.299	0.593	1.372	0.232
LLM: LLaMA 7B, PEFT: LoRA, (Q-Former Source)						
VICT (OPT2.7)	**0.839**	**0.427**	**0.315**	**0.617**	**1.505**	0.242
VICT (OPT6.7)	0.824	0.407	0.303	0.603	1.412	0.236
VICT (Flan-T5XL)	0.828	0.410	0.308	0.607	1.453	**0.244**
VICT (Flan-T5XXL)	0.810	0.395	0.297	0.595	1.384	0.233

Table 2: Visual question and answering results. Acc. denotes accuracy. In each group, the best scores have been highlighted in bold.

Method	Acc. short	Acc. short	Acc. short
	Without context	+ Given caption	+ Generated caption
The original BLIP-2 paired with different LLMs			
BLIP-2 OPT2.7	0.445	0.475	0.478
BLIP-2 OPT6.7	0.457	0.470	0.481
BLIP-2 Flan-T5XL	0.508	**0.524**	**0.519**
BLIP-2 Flan-T5XXL	**0.509**	0.500	0.497
LLM: LLaMA 7B, PEFT: LLaMA-Adapter, (Q-Former Source)			
VICT (OPT2.7)	0.605	0.605	0.605
VICT (OPT6.7)	0.583	0.569	0.584
VICT (Flan-T5XL)	0.604	**0.608**	0.603
VICT (Flan-T5XXL)	0.609	0.607	**0.607**
LLM: LLaMA 7B, PEFT: LoRA, (Q-Former Source)			
VICT (OPT2.7)	0.603	0.607	0.604
VICT (OPT6.7)	0.586	0.580	0.579
VICT (Flan-T5XL)	0.608	0.592	0.608
VICT (Flan-T5XXL)	**0.614**	**0.612**	**0.615**

ule (Q-Former), and demonstrates the improved performances in captioning and VQA. We will further explore the capability of VICT in various multi-modal tasks.

References

[1] Vicuna an open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org/. 1, 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 1

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 1, 3

[4] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 1, 2

[5] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 2

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 2

[7] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 3

[8] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 2017. 3

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2, 3

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[11] OpenAI. Chatgpt. https://openai.com/blog/chatgpt/. 1

[12] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022. 1, 3

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. 2

[14] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca an instruction-following llama model. https://crfm.stanford.edu/2023/03/13/alpaca.html. 1, 3

[15] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2, 3

[16] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 2

[17] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 1, 2