Fine-grained Image-Text Feature Alignment for large Vision-Language Models

Abstract

Large vision–language models often struggle with fine-grained image–text alignment in low-resource settings, leading to mode collapse and reduced output diversity. We address this by applying LoRA-based fine-tuning to BLIP-2 on Flickr8K, finetuning the model less than 5 % parameters. On the 2,000 images evaluation set, the lightweight approach consistently improves BLEU, ROUGE, and CIDEr scores while mitigating mode collapse, outperforming prompt engineering and sampling baselines. Results demonstrate LoRA's effectiveness for robust, fine-grained alignment under data constraints.

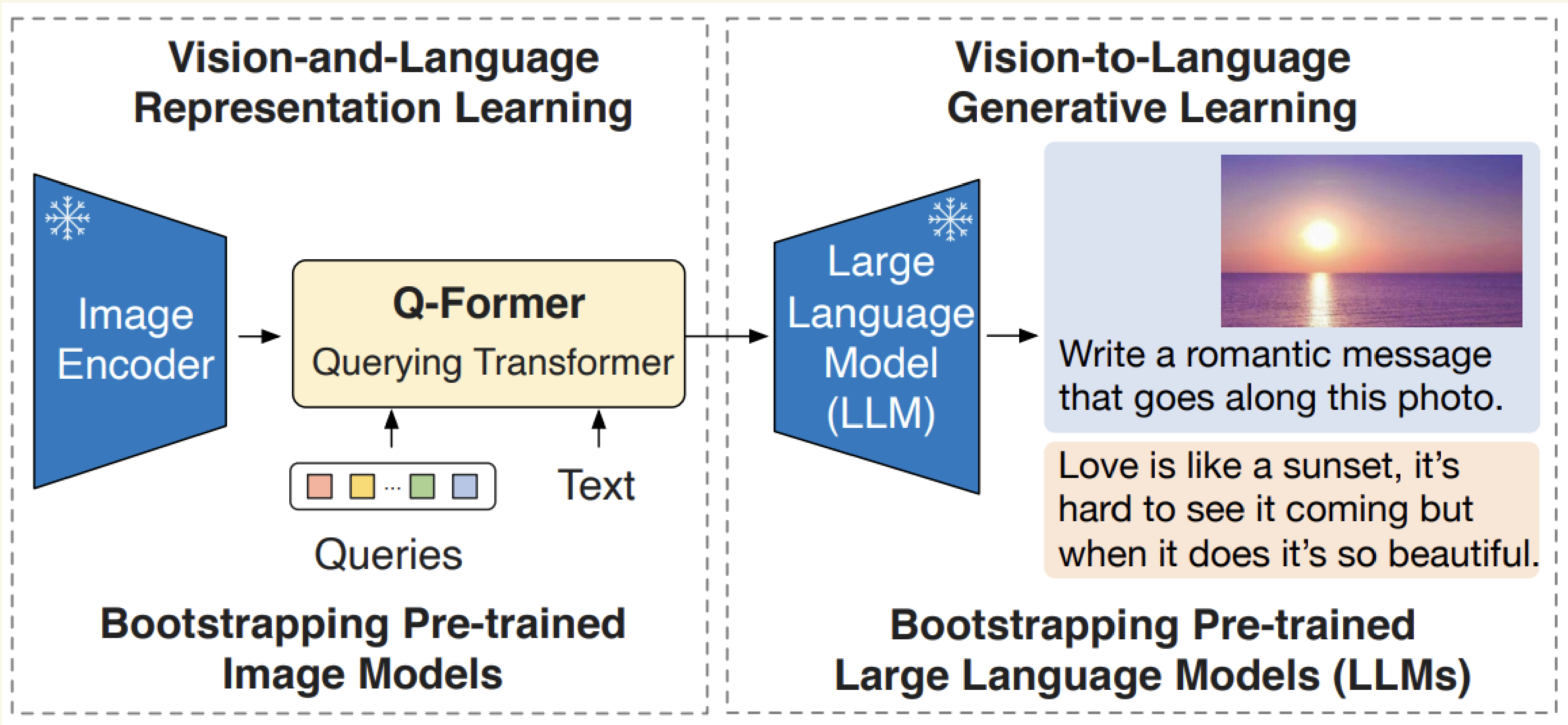
Introduction

We improve BLIP-2's performance on the small Flickr8K dataset using LoRA fine-tuning, achieving better captioning metrics and reducing repetitive outputs. Alternative methods like prompting and sampling proved less reliable, underscoring LoRA's efficiency in low-data scenarios.

Model Architecute & Method

BLIP-2 (Bootstrapping Language-Image Pre-training) is an efficient vision-language pre-training framework that bridges a frozen vision encoder and a frozen large language model through a lightweight *Querying Transformer* (Q-Former):

- **Vision Encoder** – A pre-trained visual backbone (e.g., ViT-G/14) extracts high-dimensional image features.
- **Q-Former** – Equipped with learnable query vectors, it transforms visual embeddings into language-compatible representations and enables cross-modal interaction.
- **Language Model** – A frozen pre-trained language model (e.g., Flan-T5, OPT) generates natural language descriptions, leveraging rich linguistic knowledge.


Vision-and-Language Representation Learning
Vision-to-Language Generative Learning
Image Encoder
Q-Former
Querying Transformer
Queries
Text
Large Language Model (LLM)
Bootstrapping Pre-trained Image Models
Bootstrapping Pre-trained Large Language Models (LLMs)

Write a romantic message that goes along this photo.

Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.

Figure: Overview of BLIP-2's framework [1].

To enhance task-specific performance, we adopt **LoRA** (Low-Rank Adaptation) [2] for efficient fine-tuning:

- **Method** – Freeze most pre-trained parameters and insert low-rank trainable adapters into selected projection layers of the Q-Former and language model. Fine-tuning is performed on 6,000 task-specific images.

- **Results**
  - Significant improvements across multiple evaluation metrics (**BLEU**, **ROUGE**, **METEOR**, **CIDEr**).
  - Stable training with a smooth loss curve.
  - Reduced mode collapse issues.
- **Trade-off** – Slight decrease in vocabulary diversity (lower **TTR**), but notable gains in accuracy and task relevance.

Results & Analysis

- Evaluation is conducted by comparing the zero-shot baseline with the fine-tuned model using several metrics:

Metric	Baseline	Finetuned (LoRA)
BLEU-1	0.5077	**0.6867**
BLEU-2	0.4182	**0.5820**
BLEU-3	0.3548	**0.4972**
BLEU-4	0.3099	**0.4379**
ROUGE-1 (F1)	0.3486	**0.4519**
ROUGE-2 (F1)	0.1259	**0.2022**
ROUGE-L (F1)	0.3157	**0.4272**
METEOR	0.2871	**0.4126**
CIDEr	0.2315	**0.2468**
TTR	0.0586	**0.0487**
Repetition Rate ↓	0.1220	**0.1326**

- **Qualitative Results**
  - **w/ fine-tune:** "a person sitting on the ground with a tv in front of them"
  - **w/o fine-tune:** "A man is sitting in front of a television with a plate of food in front of him."
- **Key Improvements**
  - Enhanced grammatical structure (e.g., tense, subject-verb agreement).
  - Greater lexical and syntactic variety.
  - Better alignment between visual content and text.

Reference

[1] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," in *Proc. International Conference on Machine Learning (ICML)*, pp. 19730–19742, 2023.

[2] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2022). Lora: Low-rank adaptation of large language models. ICLR, 1(2), 3.

Wenqi Su, Binyu Li, Ruiying Ma
Supervisor: Wei Wang
SURF-2025-0239

SURF
Summer Undergraduate Research Fellowship