Adapting Multilingual VQA for Code-Mixed VQA

Sayali Kandarkar, Vedant Bhasin, Mihir Bansal

Carnegie Mellon University, Language Technology Institute Department

ABSTRACT

In this study, we enhance the mBLIP model to address the challenges of code-mixed content in visual question answering (VQA). By implementing Low-Rank Adaptation (LoRA) adaptors, we fine-tune the language mode of mBLIP using the HinGE dataset to improve its language generation in code-mixed scenarios. Crucially, we keep the Q-Former and image encoder frozen. The effectiveness of this adaptation is evaluated using the mcvqa dataset, with the goal of significantly boosting mBLIP's performance in multilingual environments and demonstrating its advanced capability in interpreting and responding to code-mixed VQA tasks.

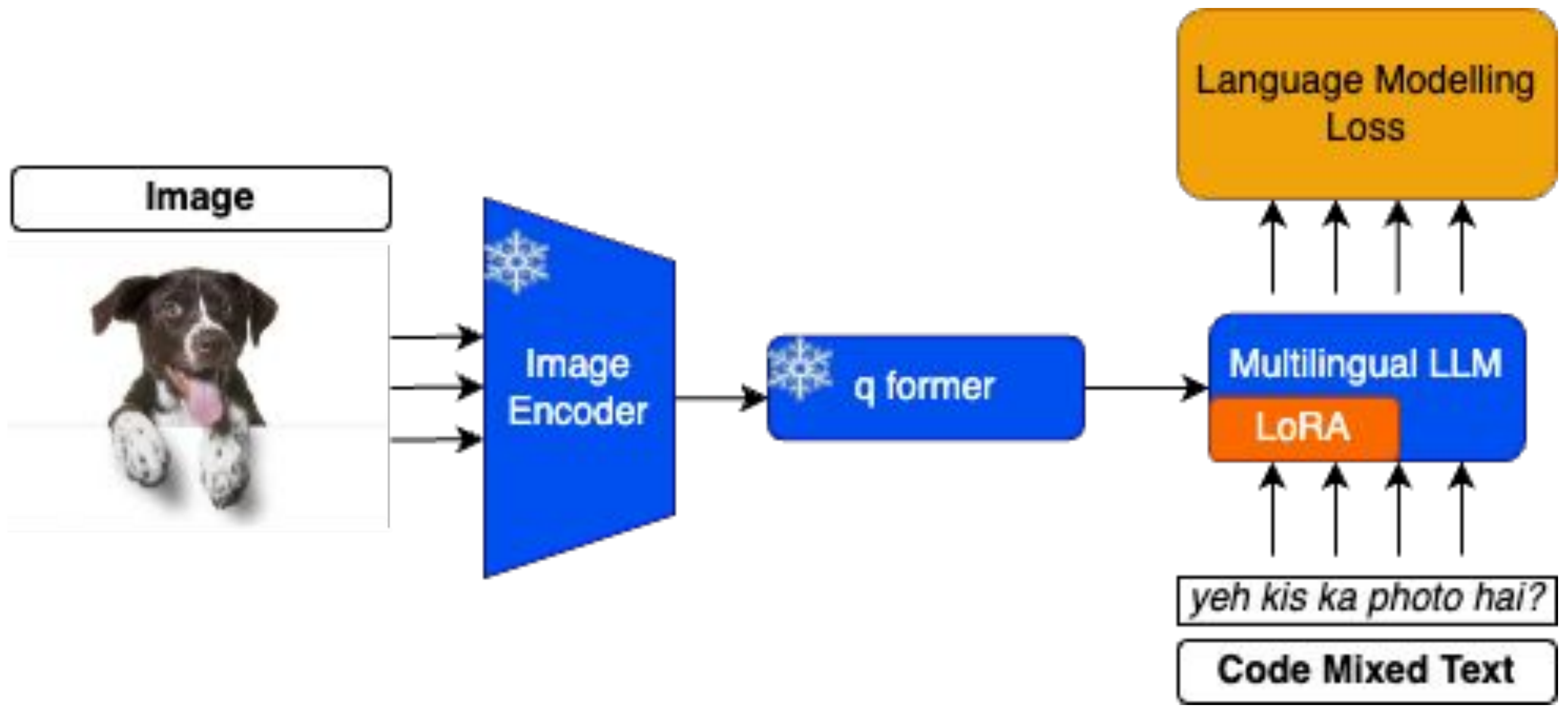
MOTIVATION

The realm of Visual Question Answering (VQA) has witnessed substantial progress in accommodating diverse languages through Multilingual models. However, an unexplored frontier lies in the adaptation of these models to Code-Mixed scenarios, particularly in the linguistic amalgamation of Hindi and English, commonly referred to as Hinglish.

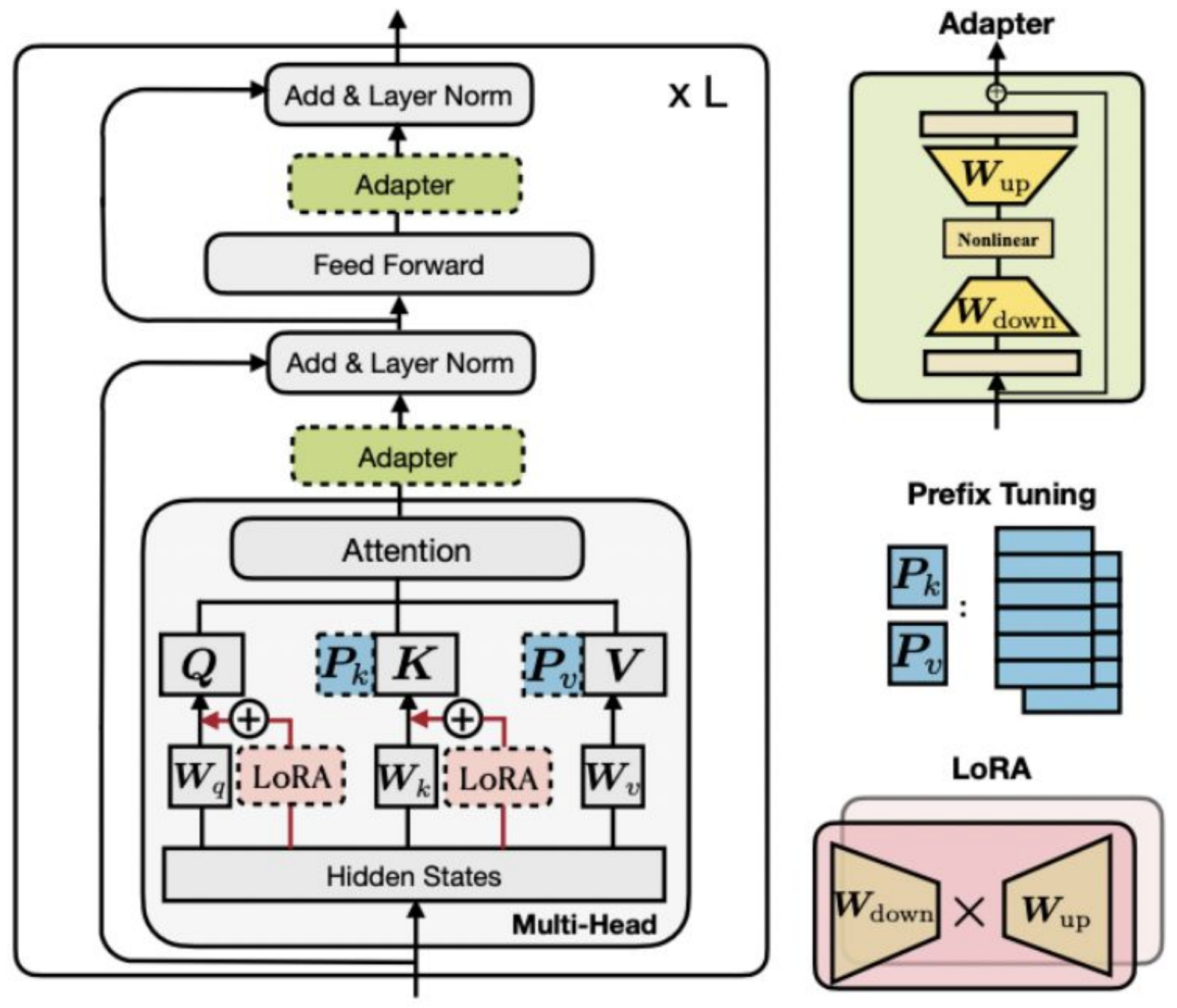
Hinglish, as a vibrant linguistic phenomenon, encapsulates the intricacies of code-mixing, where linguistic elements from Hindi and English coalesce seamlessly. This linguistic fusion is deeply embedded in everyday communication, posing unique challenges for existing VQA models in comprehending the nuances of Code-Mixed inquiries about visual content.

This research endeavor addresses a gap in existing methodologies, recognizing the need for VQA systems to navigate the complexities of linguistic diversity within the Code-Mixed landscape. Our goal is to establish a robust academic foundation for adapting Multilingual VQA models to the unique challenges presented by Code-Mixed (Hindi+English) VQA scenarios.

PROPOSED METHODS



We chose the mBLIP model as our baseline due to its strong performance in vision-language tasks and its ability to handle multiple languages. However, mBLIP showed limited effectiveness in tasks involving code-mixing of English and Hindi. To improve this, we are implementing a focused fine-tuning approach to specifically enhance mBLIP's performance in code-mixed Visual Question Answering (VQA) tasks. We hypothesize the main issue lies in the model's language component, and that the image encoder and qformer are robust. Our proposed method targets the language models ability to generate and integrate code-mixed language with visual data.



To finetune the BLOOM-7B language model We utilize LoRA adaptors targeting the attention mechanism of this language model. The training involves using the HinGE dataset, which offers high-quality, human-generated code-mixed sentences, alongside their separate English and Hindi counterparts. The goal of this method is to significantly improve mBLIP's handling of code-mixed languages, leading to better and more contextually accurate responses in VQA tasks. This should make the model more versatile and effective in multilingual contexts.

RESULTS

Using Gregor/mblip-bloomz-7b model, and MaXM dataset, we got 48.25% accuracy.
When we switched to the xGQA dataset, we got 35.91% accuracy, both for the English language.

Model	Dataset	Accuracy (%)
Gregor/mblip-bloomz-7b	MaXM	48.25
Gregor/mblip-bloomz-7b	xGQA	35.91

Using Gregor/mblip-bloomz-7b model, on Hindi devnagari language of MaXM, we got 50% accuracy, but when we used indic-transliteration library, and converted devnagari to roman scripture, we got really bad results, just 3% accuracy.

Model	Script	Accuracy
Gregor/mblip-bloomz-7b	Hindi (Roman)	0.50
	Hindi (Devanagari)	0.03
Gregor/mblip-mt0-xl	Hindi (Roman)	0.43
	Hindi (Devanagari)	0.29

So, there is a lot of room for improvement as far as code-mixing for even high-resource languages such as Hindi is concerned.

NEXT STEPS

The next phase of our research involves finetuning the mBLIP model using HinGE dataset.

We also plan on employing GPT-3.5's few-shot learning capability by providing examples of code-mixed captions in prompts to generate diverse and context-aware responses, facilitating evaluation of linguistic adaptability.

We anticipate that the incorporation of code-mixed captions will enhance the model's adaptability to diverse linguistic contexts, ultimately leading to improved performance in multilingual image captioning tasks

.