Adapting Small Vision–Language Models for Japanese Image Captioning

Allie Okumura
1009846419, allie.okumura@mail.utoronto.ca
Adapting Small Vision-Language Models for Japanese Image Captioning

Introduction

Vision–language models (VLMs) have achieved impressive performance in English image captioning, yet their capabilities in other languages remain underexplored. Japanese presents particular challenges due to its writing system and tokenization complexity, which are often poorly supported in multilingual models. At the same time, there is strong motivation to adapt small VLMs, as they are more efficient, accessible, and deployable on limited hardware such as a single T4 GPU. The goal of this project is to adapt the small VLM Qwen2-VL-2B-Instruct to Japanese for the task of image captioning. By applying low-rank adaptation (QLoRA), we hypothesize that compact models can be specialized effectively for Japanese, outperforming their zero-shot multilingual baseline performance while retaining computational feasibility. Deep learning provides the right framework for this task because it enables transfer learning, reusing pretrained multimodal representations and specializing them to a new linguistic domain.

Illustration

The overall model architecture is depicted in Figure 1. An input image is encoded by a frozen vision backbone, followed by a trainable projector that maps image embeddings into the language model space. The language model is adapted using QLoRA on the top layers to produce fluent Japanese captions. The figure highlights the distinction between frozen and trainable components.

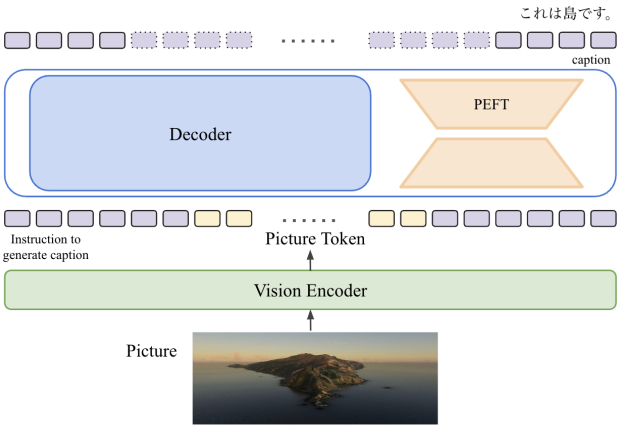

Figure 1: Overview of the proposed architecture. Frozen components are shown in grey, trainable modules in blue.

Background & Related Work

Image captioning has evolved from CNN–RNN architectures to transformer-based models. Early work by Karpathy and Fei-Fei demonstrated deep visual–semantic alignments for English captions,

while Xu et al. introduced attention mechanisms for better visual grounding. Yoshikawa et al. released the STAIR Captions dataset, enabling large-scale Japanese captioning and showing the limits of English-to-Japanese machine translation pipelines. More recent frameworks such as BLIP-2 highlight the efficiency of combining frozen vision encoders with trainable language projections, while the Qwen2-VL family demonstrates that small-scale multimodal models can handle diverse tasks with multilingual capabilities. Our project builds on this body of work by applying efficient adaptation techniques to Japanese, a language where direct captioning resources exist but are underutilized in modern VLM adaptation.

Data Processing

We will primarily use the STAIR Captions dataset, which contains 820,000 Japanese captions for 164,000 images derived from MS COCO. The YJ Captions dataset, which comprises about 26,000 Japanese captions, will serve as an out-of-domain test collection. Data preprocessing will include normalization of punctuation and Japanese full/half-width characters, which is necessary to avoid tokenization inconsistencies. We will remove duplicate captions and noisy artifacts such as URLs or emojis, and truncate overly long captions to maintain consistent input lengths. Each caption will be tokenized with the Qwen2-VL tokenizer, which supports Japanese subwords, ensuring compatibility with the base model. STAIR will be divided into training, validation, and test splits using a Karpathy-style partition (113k/5k/5k images), while the entire YJ dataset will be held out for zero-shot generalization testing. All preprocessing scripts and splits will be documented in the project repository to ensure reproducibility.

Architecture

The proposed model is Qwen2-VL-2B-Instruct with QLoRA adaptation. The vision encoder, based on a ViT (Vision Transformer) backbone, remains frozen. A projector layer mapping vision features to the language space is fully trainable. Within the language model, the top six transformer layers will be adapted using LoRA with rank 32, scaling factor 32, and dropout 0.05. Training will use cross-entropy loss with label smoothing, and optimization will be performed with AdamW under a cosine schedule. During evaluation, captions will be generated with beam search of size three. This configuration balances efficiency with sufficient flexibility to achieve meaningful improvements under Colab T4 constraints.

Baseline Model

Our baselines will consist of stand-alone Japanese VLMs evaluated in zero-shot mode, including the Qwen2-VL-2B-Instruct model without fine-tuning, LLaVA-JP-1.3B, Japanese Stable VLM, Asagi-2B, and PaliGemma-3B. These models are capable of producing captions in Japanese directly from images. We will compare their outputs against our fine-tuned Qwen2-VL-2B to highlight the benefits of language-specific adaptation in small VLMs.

Ethical Considerations

The datasets are derived from MS COCO, which may encode cultural biases and skew towards Western imagery. As a result, captions may underrepresent Japanese-specific content. The adapted model may also produce hallucinated objects or literal but unnatural phrasing. Since this work is conducted for research and educational purposes, we will not deploy the system in real-world applications without further safeguards. We will report common failure modes and provide human evaluations to highlight potential risks. Proper attribution of the datasets and models will be respected.

References

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2016). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Yoshikawa, Y., Shigeto, Y., & Takeuchi, A. (2017). STAIR Captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of ACL*.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *Proceedings of ICML*.

Bai, J., Dai, W., Guo, D., et al. (2024). Qwen-VL: A versatile vision–language model for real-world applications. *arXiv preprint arXiv:2407.10224*.