Expose: Probing Small Vision-Language Models for Global and Local Semantic Representations

Tsun Wai Wong Jonathan Schwab

This project investigates how small vision-language models (1–4B parameters) encode global semantics versus local semantics at the object level. Building on [1], which showed that large models often store richer global information in intermediate layers while upper layers focus on token-level prediction, we examine whether such trends hold for recent compact models like Qwen2-VL, MiniCPM-V, Phi-4-mini, and Gemma-3. These models, attractive for computationally constrained applications, remain underexplored in terms of representational behavior. We construct two probing tasks on MS COCO, a dataset consisting of images and captions [2]: an image-text entailment setup for global semantics, where positive captions are original and negatives are semantically similar captions from other images, and an object recognition task for local semantics, where the model receives an image and a shuffled list of present objects and must decide whether that objects appears. For both, VLM parameters are frozen and layer-wise hidden states extracted. A lightweight classifier (logistic regression or small MLP) is trained to predict labels, revealing the distribution of task-relevant information across layers [4]. Preprocessing aligns with model requirements: images are resized to match encoders, captions cleaned and transformed appropriately, and prompts adapted to each task. Accuracy is used for global evaluation, macro-F1 [3] for local to account for category imbalance. We expect that intermediate layers in small VLMs will, like their larger counterparts, capture the richest global representations, while upper layers emphasize local or token-prediction cues, though new architectures may yield different patterns. The findings could guide architectural or training adjustments to preserve global meaning in resource-friendly multimodal systems.

References

[1] M. Tao *et al.*, “Probing Multimodal Large Language Models for Global and Local Semantic Representations,” 2024. [Online]. Available: `https://arxiv.org/abs/2402.17304`

[2] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common Objects in Context,” *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: `http://arxiv.org/abs/1405.0312`

[3] J. Opitz and S. Burst. Macro F1 and Macro F1. *arXiv preprint arXiv:1911.03347*, 2021. URL: `https://arxiv.org/abs/1911.03347`.

[4] G. Alain and Y. Bengio, “Understanding intermediate layers using linear classifier probes,” *arXiv preprint arXiv:1610.01644*, 2018.