Instructions for ACL 2023 Proceedings

Anonymous ACL submission

Abstract

1 Introduction

Vision–language models (VLMs) combine text and image understanding and have become central to multimodal AI research. Probing studies help to uncover how such models encode semantic information at different layers, complementing evaluations on end tasks. A recent study (Tao et al., 2024) examined decoder-only multimodal models such as Kosmos and LaVIT, and found that global semantics are most strongly represented in intermediate layers, while upper layers shift toward local, token-level information, which can reduce the ability to encode global meaning. In this work, we extend this line of analysis to recently released compact models such as Qwen2-VL-2B and FastVLM-0.5B. These models represent a new generation of efficient VLMs with parameter counts in the low billions, designed for practical deployment under limited resources. Whether their internal representational patterns follow the same global–local layering observed in earlier models, or show different dynamics, remains an open question. To address this, we developed a general probing framework that enables layerwise representation extraction, flexible pooling of token embeddings, and efficient training of lightweight classifiers. Using this framework, we construct two probing tasks on MS COCO (Lin et al., 2014): a caption–image entailment experiment probing global semantics, and an object-category experiment probing local semantics. Following established methodology (Alain and Bengio, 2018), we freeze the VLM parameters, pool hidden states into fixed-size embeddings, and train classifiers on top. We report accuracy, precision, recall, and F1 scores per layer to quantify representational quality. Our experiments provide insight into how current compact VLMs encode global and local information across layers, and allow comparison with trends previously reported for older models (Tao et al., 2024).

2 Methods

2.1 Models

We developed a probing framework for three compact vision–language models: **Qwen2-VL-2B** and **FastVLM-0.5**. All models were used in frozen form, i.e., without updating their parameters, so that only lightweight probing classifiers were trained on top of their internal representations. This allows us to isolate the representational capacity of the models at different layers without confounding effects from fine-tuning.

2.2 Probing Tasks

To investigate the distinction between global and local semantic representations, we designed two probing tasks. The *caption experiment* targets global features by testing whether the model can align an image with a candidate caption. For this task, inputs are constructed using the prompt

```
This image contains: {caption}.
Is this right?
```

and the probe performs binary classification of whether the caption correctly describes the image. Due to limited computational resources, we just use one positve and one negative caption per image in contrast to (Tao et al., 2024). We consider a set of images M with associated ground-truth captions C and a set of randomly sampled negative captions C'. For each image $m \in M$, we select one positive caption $c^+(m) \in C$ and one negative caption $c^-(m) \in C'$. This yields the dataset

$$D = \{(m, c^+(m)), (m, c^-(m)) \mid m \in M\}.$$

The probing classifier is then trained to solve the binary decision problem

$$f : D \rightarrow \{0, 1\},$$

with $f(m, c^+(m)) = 1$ and $f(m, c^-(m)) = 0$. The *category experiment* focuses on local features by probing whether the model can identify the presence of specific objects. For this task, prompts take the form

```
This image contains the following
type of object: {category}
```

and the probe must predict the correctness of the statement. $\mathcal{K}$ denotes the set of object categories, and for each image $m \in M$, the subset $K(m) \subseteq \mathcal{K}$ specifies the categories present in m. The probing task amounts to learning a classifier

$$f : M \to \mathcal{K},$$

such that for an image m, the prediction $f(m)$ corresponds to one of the true categories in $K(m)$. Both tasks are based on the MS COCO dataset (Lin et al., 2014), which provides annotations for images, including captions for each image. We created training and evaluation splits. Due to limited computational resources, we restrict our analysis to a random subset of 20.000 images for the training set and 2.000 images for the evaluation set in both experiments.

2.3 Representation Extraction

Our framework computes hidden representations for every transformer layer of a model given an image–prompt pair. From the token-level hidden states, we derive pooled embeddings that serve as inputs to the probing classifiers. We implemented a general-purpose `pool_tokens` function that supports multiple pooling strategies, including CLS token extraction, mean pooling across valid tokens, max pooling, token-index selection, and a default strategy that retrieves the last non-padding token. Unless otherwise noted, we employ mean pooling, which aggregates information across the entire input sequence while respecting attention masks. This yields a single fixed-size vector per input and per layer, enabling layerwise comparison of representational quality. Moreover, our framework stores all extracted representations in a binary format, facilitating efficient reuse and analysis without redundant computation.

2.4 Probing Classifiers

On top of the pooled embeddings, we train lightweight classifiers that map the representations to task labels. For both experiments, we use a simple linear projection with dropout regularization, optimized with Adam and cross-entropy loss. The caption experiment is framed as binary classification, while the category experiment is treated as multi-label prediction with label and mask vectors indicating valid categories for each image. All probes are trained independently per layer, which allows us to quantify how well each layer captures global or local semantic information. Our trainer additionally computes detailed evaluation metrics, including accuracy for caption entailment and macro-averaged F1, precision, recall, and confusion matrix statistics for category recognition.

2.5 Experimental Workflow

In both experiments, the workflow follows the same high-level structure. Datasets are preprocessed and split into training and evaluation sets, with balanced numbers of positive and negative instances. The target model is then loaded, and representations are computed for all inputs. Probing classifiers are subsequently trained and evaluated for each layer. After finishing one model, GPU memory is released and the next model is processed in the same way. The probing results across layers and models are then aggregated for analysis. This setup ensures that our experiments are efficient, reproducible, and directly comparable across models of different architectures and scales.

3 Results

To evaluate the performance of the probing classifiers across different layers and models, we generated a series of plots that visualize key metrics. For the caption experiment, which focuses on global semantic features, the accuracy per layer across all three models is shown in Figure 1. Thus the positive and negative examples are balanced, a random baseline would achieve 50% accuracy. We can see that both probes show a similar evulution in accurancy over the most layers, with Qwen2-VL-2B performing slightly better than FastVLM-0.5B. The accurancy increases from the first towards the later layers. Interstingly, the accurancy of FastVLM-0.5B drops significantly in the last layers, while it remains relatively stable for Qwen2-VL-2B.

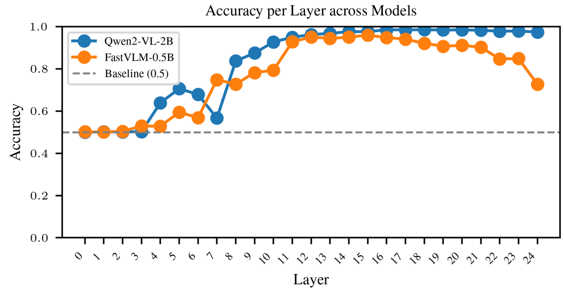

Figure 1: Accuracy per layer across all three models.

In 2, we present a heatmap that summarizes the precision, recall, and F1 scores for the caption experiment across all layers and models. Here we can also observe a similar trend as for the accuracy. The scores are increases from the first towards the later layers and FastVLM-0.5B shows a significant drop in the last layers.

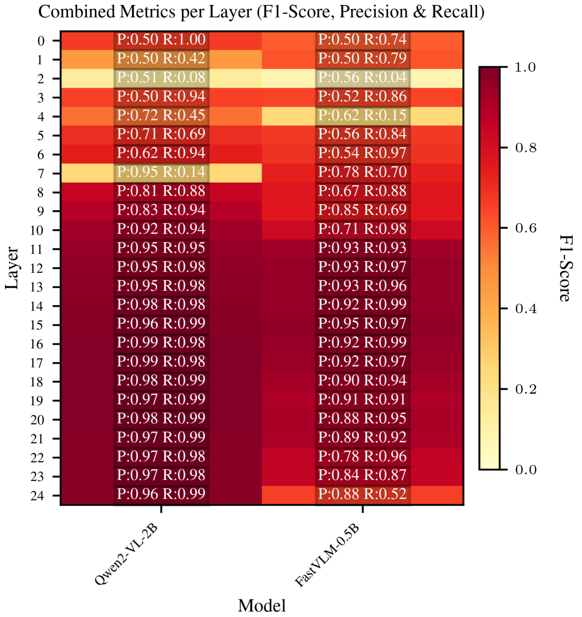

Figure 2: Precision, Recall and F1 per layer across all three models.

4 Discussion

5 Conclusion

6 References

References

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, abs/1405.0312.

J. Opitz and S. Burst. 2021. Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

M. Tao et al. 2024. Probing multimodal large language models for global and local semantic representations. *arXiv preprint*.

A Appendix

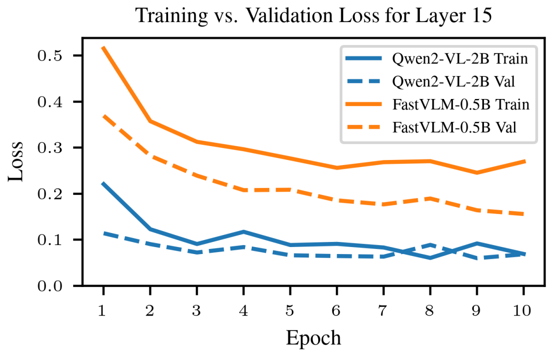

Figure 3: A comparison of training and validation loss for the layer 15 probe (layer with best accuracy scores), analyzing model performance in the global caption experiment. The Qwen2-VL-2B model (blue) demonstrates a significantly lower and more stable loss than the FastVLM-0.5B model (orange). The small gap between the training (solid) and validation (dashed) curves for the Qwen2 model suggests robust training with no signs of overfitting. Moreover we can see that the strongest learning happens in the first 2 epochs.

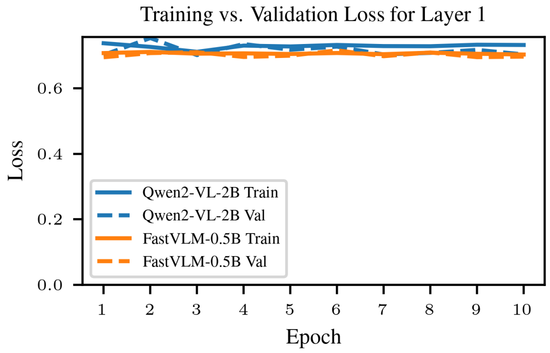

Figure 4: A comparison of training and validation loss for the layer 1 probe (layer with lowest accuracy scores), analyzing model performance in the global caption experiment. Both models show no significant learning, with losses remaining high and stable across epochs. This indicates that the representations at this early layer do not contain sufficient information for the caption entailment task, leading to poor probe performance.