Exploring multimodal vision-language models with CLIP: Implementation, Challenges, and Methodological Insights

1. Implementation Overview

Our implementation follows the evaluation protocols detailed in the original CLIP paper [1] and leverages the official OpenAI CLIP repository. Two primary evaluation modes were adopted:

- **Zero-shot classification:** Where class labels are expressed as natural language prompts and image-text similarity is used for prediction.
- **Linear probing:** Where image features from CLIP's visual encoder are frozen, and a linear classifier is trained on top using CIFAR-10 training data.

We used the `"ViT-B/32"` model variant and reported the following performances:

Method	Accuracy (reported)	Accuracy (observed)
Zero-shot	**89.83%***	**89.59%**
Linear probe	**95.1%**	**95.02%**

While attempts were made to get the same results as reported, a small reduction was observed.
*This is based on statistics shared by OpenClip, a third part library

2. Experimental Studies

To deepen our understanding of CLIP's zero-shot performance, we conducted several experiments exploring variations in:

- **Text Descriptions**: We evaluated the effect of altering the **prompt structure** (e.g., "This is a photo of a dog" vs. "An image showing a dog"). The best performance (89.59%) was observed with the variant "An image showing a {label}."
- **Class Names**: Replacing the original CIFAR-10 class names with **synonyms or closely related terms** (e.g., "airplane" → "aircraft", "dog" → "canine") produced a small improvement, highlighting CLIP's sensitivity to prompt formulation.
- **Combined Prompts**: We combined the best-performing text structure with the best-performing class variants, expecting additive gains. However, the resulting accuracy (89.31%) was slightly lower than the standalone best cases.

3. Augmentation-based Attempt

We experimented with using **two image versions**, i) the original and ii) an augmented one (horizontal flip + rotation) and averaged their cosine similarity scores with the text embeddings.

This ensemble-like approach **slightly reduced** accuracy (from 89.59% to 89.42%), suggesting the limitations of naive augmentation strategies when working with pre trained contrastive models like CLIP.

4. Implementation Challenges

- **Prompt Sensitivity**: CLIP's performance has minor dependence on the structure and semantics of the textual prompt. Designing "optimal prompts" becomes a trial-and-error process without prompt engineering tools.
- **Requirement for large dataset**: This model was trained on WebImageText (WIT) dataset which is a closed source dataset curated by OpenAI and contains 400M images [1]. Researchers have reported improved performance but relied on larger datasets [2].
- **Limited Control in Zero-Shot**: Without fine-tuning, there's no way to directly optimize performance for a benchmark, only prompt engineering is possible. MetaCLIP [2] did achieve improvements by curating a new dataset with a better data distribution than WIT.

5. Methodological Improvements

While our experiments focused on dataset-specific variations and augmentations, we propose more **concrete architectural and methodological modifications** for future improvements. (Some strategies regrettably have already been performed by other groups):

1. **Prompt ensembling with learnable weights**: Rather than averaging prompts heuristically, implement a small learnable model that weights each prompt variant based on its agreement with the image representation. This retains zero-shot generalisability while adapting slightly to the dataset [3].
2. **Attention-guided prompt tuning**: Integrate attention mechanisms that allow the model to soft-select parts of the prompt that are most aligned with the image embedding. This could be implemented as a lightweight adapter module over the text encoder [4].
3. **Joint feature alignment via contrastive re-calibration**: Fine-tune only a shallow projection layer after both image and text embeddings using a small unlabeled or weakly labeled dataset. This aligns the embeddings better for downstream classification without full model training. Currently, we freeze all pretrained parameters during a linear probe.
4. **Curriculum prompting**: Design a gradual prompting pipeline where the model is successively exposed to increasingly descriptive or context-rich prompts, aiming to improve generalisation on ambiguous or out-of-distribution classes.

6. Conclusion

Our exploration confirms CLIP's robustness and replicability in zero-shot and linear probing setups. Although dataset-level and augmentation tricks yielded marginal gains, deeper architectural or prompt-encoding improvements offer more promising avenues. Future work will focus on integrating light-weight prompt learning methods that can adapt CLIP to new domains with minimal compute and annotation cost.

REFERENCES

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.

[2] Xu, H., Xie, S., Tan, X. E., Huang, P. Y., Howes, R., Sharma, V., ... & Feichtenhofer, C. (2023). Demystifying clip data. arXiv preprint arXiv:2309.16671.

[3] Agnolucci, L., Baldrati, A., Todino, F., Becattini, F., Bertini, M., & Del Bimbo, A. (2023). Eco: Ensembling context optimization for vision-language models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2811-2815).

[4] Brouwer, E., van Woerden, J. E., Burghouts, G., Valdenegro-Toro, M., & Zullich, M. (2024). Adaptive Prompt Tuning: Vision Guided Prompt Tuning with Cross-Attention for Fine-Grained Few-Shot Learning. arXiv preprint arXiv:2412.14640.