COSE474-2024F: Final Project Proposal
" Prompt learning optimized for General Vision Language Model "

1. Introduction

Previous state of the art vision models, including Convolutional neural network models such as ResNet(He et al., 2015) and Vit, vision transformer(Dosovitskiy et al., 2021) that introduce transformer architecture into vision models, solely trained on image and usually later fine-tuned on text labeled training set. More Recently, using contrastive language-image pre-training, Vision Language model CLIP(Radford et al., 2021) combined text label and counterpart image to enable multi-modal and more extended model usage with attention based evaluation. CLIP also show potential of proper prompt, by adding a phrase rather than just give text encoder the class name that we want to evaluate.
In this research, I will try to find an architecture that can lead to more better performance, with zero or few shot model for downstream task.

2. Problem definition & challenges

In case of CLIP like model, the main problem or challenge is that how and what to choose from prompt that will feed into text encoder and image encoder. The structure of CLIP implies that we can apply same model into many other related tasks. Since such process could greatly improve the performance of CLIP like models, many papers after CLIP tries to find better architecture that works on trained dataset, and generalizes well within the constraint of zero or few shot learning. As previous researchers did, I will try to find proper architecture that is on par with current papers.

3. Related Works

With the idea started from CLIP, CoOp(Zhou et al., 2022b) presented adding learnable vector before or after class keyword to automate prompt engineering, and Co-CoOp(Zhou et al., 2022a) added meta network that is added onto those vectors so that they can benefit from image representation. MaPLe(Khattak et al., 2023) extended CLIP's work by adding learnable part not only text encoder but also image encoder and relating two part by putting linear net and image encoder gets linear net transformed information from prompt. RPO(Lee et al., 2023) allowed only prompt to see original feature by masking attention.

4. Datasets

Currently planning to test model on various dataset including CIFAR, ImageNet dataset and many more, at least 11 datasets, that includes Caltech101, Oxford Pets, Stanford Cars, Flowers102, Food101, FGVC, Aircraft, SUN397, DTD, EuroSAT, UCF101, that were commonly done in CLIP like models. If time is enough, I'm planning to do more test on datasets that were done in other previous similar research.

5. State-of-the-art methods and baselines

CLIP baseline model accuracy on 11 datasets in order of base and novel was 69.34,74.22. CoOp showed 82.69, 63.22 percentage correctness and CO-CoOp was 80.47, 71.69 each. RPO's harmonic mean score was 77.78, MaPle's score was 78.55(Khattak et al., 2023).
For now, SoTA harmonic mean score is 83.73 from PromptKD: Unsupervised Prompt Distillation for Vision Language Models(Li et al., 2024) that utilized huge teacher model to firstly learn from large data and save text encoder vectors, and later use that model with vectors to align student model on specifically on image encoder to follow teacher model.

6. Schedule & Roles (if you have a teammate)

After the Mid-term exam period, I will try previous researches that provides codes as well as study how previous research really worked. Before the 11/15, I will work on architecture or models that can be applied on CLIP model. Then by testing some models, I'm planning to seek improvement on top of those previous works, at least not as bad as before, and will try to get some result before the beginning of December. At last weeks, I will write final project paper based on the work that will be done.

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer,

M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. Maple: Multi-modal prompt learning, 2023.

Lee, D., Song, S., Suh, J., Choi, J., Lee, S., and Kim, H. J. Read-only prompt optimization for vision-language few-shot learning, 2023. URL https://arxiv.org/abs/2308.14960.

Li, Z., Li, X., Fu, X., Zhang, X., Wang, W., Chen, S., and Yang, J. Promptkd: Unsupervised prompt distillation for vision-language models, 2024. URL https://arxiv.org/abs/2403.02781.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. 2021.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16816–16825, June 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, July 2022b. ISSN 1573-1405.