COSE474-2024F: Final Project Proposal
"Research on Improving Prompt Learning for Vision-Language Models"

Nayoung Kim [1]

1. Introduction

Recently, pre-trained vision-language models have demonstrated zero-shot generalization capabilities across various computer vision tasks. Models like CLIP (Contrastive Language-Image Pre-training) have shown the ability to understand both images and text simultaneously, which improved the performance a lot. However, fine-tuning these models for specific tasks can lead to overfitting problems when trained on small datasets. This issue can be addressed with prompt learning techniques, and therefore this project aims to implement and improve prompt learning methods based on the CLIP model. Specifically, I will focus on understanding and implementing basic prompt learning approaches such as CoOp (Context Optimization).

2. Problem definition & challenges

The goal is to fully understand the prompt learning methods based on pre-trained vision-language models (CLIP) and implement and improve it that can be effectively applied to various downstream tasks. In particular, the project will be focused on exploring methods that can learn task-specific knowledge while maintaining generalization performance, even with small datasets. After impelementing CLIP and CoOp, CoCoOp and ProMetaR will also be considered.

3. Related Works

1. Jinyoung Park, Juyeon Ko, Hyunwoo J. Kim, "Prompt Learning via Meta-Regularization" (2024) This is the main previous study to which I would refer.

2. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision" (2021) This is the previous study about CLIP (Contrastive Language-Image Pre-training).

3. Zhou et al., "Learning to Prompt for Vision-Language Models" (2021) This is the previous study about CoOp (Context Optimization).

4. Datasets

Below are public datasets that could be used for experiments. In particular they will test the model's ability to distinguish between visually similar classes, a common challenge in real-world applications. Therefore, these datasets will evaluate the performance of vision-language models and prompt learning methods.

1. ImageNet Large-scale image classification dataset with 1000 classes

2. CIFAR-100 Small-scale image classification dataset with 100 classes

3. Caltech101 Object recognition dataset with 101 categories

4. Oxford Flowers Flower image dataset with 102 species

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

To evaluate the effectiveness of the proposed methods, I will compare the methods against baselines and state-of-the-art approaches. This project's primary baseline will be CLIP, evaluated in a zero-shot setting, pre-trained without any fine-tuning. CoOp based on the offical code will also bwe implemented and then be compared.

6. Schedule

Weeks 1-2: Understanding and implementing the CLIP model
Weeks 3-4: Implementing and experimenting with the CoOp method
Weeks 5-6: Dataset preparation and baseline experiments
Weeks 7-8: Result analysis and report writing

[*]Equal contribution [1]Department of Media & Communication, Double Major: Department of Computer Science, Korea University, Seoul, Korea. Correspondence to: Nayoung Kim <tkfkd9008a@korea.ac.kr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.