COSE474-2024F Final: MIP(Meta-Information Based Prompt tuning)

Joo-Young Park

Abstract

Nowadays, the volume and cost of model training have increased significantly to enhance the performance of transformer models, leading investigators to focus on transfer learning[1] which makes pre-trained model available to build dataset-optimized models. Extending from that perspective, prompt tuning is an effective method for using pre-trained foundation models such as CLIP[2] at computer vision domain that show high cost efficiency with enhanced accuracy.
Conventional prompt tuning uses some fixed prefix script such as "a photo of aclass". BY introducing learnable prompts through CoOp[3], unlike existing prompts, prompt tuning infers the context vectors while not modifying the layers of pre-trained models. Tuning context vector achieves better performance on the given datasets revising its context prompt. By this methodology, prompts are now also perceived as parameterizable factors in prompt transfer learning.
Motivated by this flow of prior studies, our approach to achieve further improved training results from a given CLIP model and dataset is to regularize the prompt and initialize the context vector with meta-data from the prompt learner itself.
Our research introduced mathematical functions that has parameters from the pre-trained model and investigated how the accuracy of classification changes belong to the changes of the trainer. By observing the internal shifts , we propose how posterior researches address the problem of formulating the prompts.
Code is available at github.com/hihello122/MIP. You can verify the entire structure of our model here and how the project was conducted on the Google Colab environment by jupyter notebook.

1. Introduction

As learnable prompt, another tunable factor of the machine learning is introduced to research domain, requirement of regularization or formalize to guarantee the effeciency of training also has emerged. Our project address the methodology to enable the weight to be translated properly and the context vector regularized in properly trainable region.
In this paper, we propose an approach to overcome presented problem situation in two perspectives. Our novel approaches start on the initialization of context vector. The other idea is about handling the degree of how the derivatives of prompts transfers by regularization with meta-data of the prompt adequately.
Our research would use CLIP model as a pre-trained model and compares the result of few-shot training on different datasets with basements of our model, CoOp an CoCoOp. Since our objective is about the efficiency of training we are going to compare accuracy on the base classes. The basement model will be Vit(vision transformer)[4] on our experiment being also capable to be extended to other models such as RESNET.
Our research has its meaning on proposing investigation potential of regularizing the prompt and handling of the prompt data. After the context vector outcome, prompt is interpreted as not only text-information but also the weighted vector for the optimization. So, we expect that expanded proposal or research can be held in the future about prompt learner regularization and initialization.

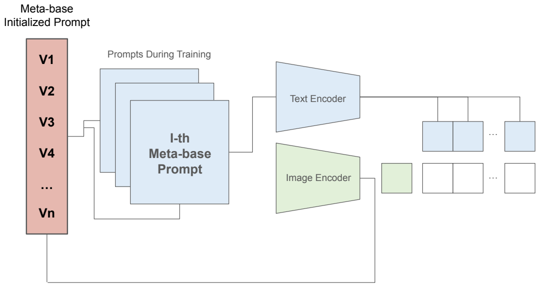
2. Related Works

Vision Transformer Since the emergence of the transformer model, the base of state of the art models in various fields of deep learning has changed to the transformer model from existing structures. Therefore computer vision fields have also looked forward to adapting the model structure to object detection and classification. ViT(Vision Transformer) suggested the method to build an image classifier based on the transformer model. Furthermore some modern models use text-based features and image-extracted features simultaneously such as the CLIP(Contrastive Language-Image Pre-training) model.

Prompt tuning The domainant tendency of the preference to the transformer models led the cost and requirement of model training extremely high, so bypassing methodology for performance improvement are now quite interesting areas of study. Since investigating transfer learning that uses some part of a pre-trained large model to apply to different datasets or training set, researches are being conducted with a intense focus on fine-tuning and prompt tuning. First, fine-tuning updates the entire parameters of the pre-trained model to appropriately fit to the given dataset[5] by training slightly on the dataset. Simultaneously causes extensive train cost for general purpose to update the whole parameters of pre-trained model.

The another approach, prompt-tuning[6] adds some prompts to the pre-trained model to enhance the model performance and data suitability being widely used regardless of the field. CoOp and CoCoOp are consistent with such a perspective of prompt tuning that is more economical than fine-tuning. Prior method leverage context information of the text to get better classification, while the other uses the additional meta-network to improve outcomes. CoCoOp shows more powerful outcomes than the CoOp in the unseen class classification but CoOp has better results in some cases making some trade-off.

3. Experiments



Before explaining the methodology of our research, we briefly give an overview of the CLIP, CoOp and CoCoOp which is the basement of our research.

3.1 CLIP Model which is contrastive language-image pre-training uses text encoder and image encoder simultaneously to use both-sided features connecting the labels and the images. Although a diverse image classification model can be utilized as the image encoder, our subsequent research uses the vision transformer of the backbone model of images.

3.2 CoOp and CoCoOp are consistent with a perspective of prompt tuning. CoOp suggests some methodology of using prompt as vector which has dimension of word embeddings rather than lexical prompts. This context vector is differentiable which makes it able to propagate to update the context-vector while preserving the layer of pre-trained model, CLIP frozen. CoCoOp provides better generalization performance by introducing Meta-Net, which is a dynamic token of prompt parameterized by context vector updated during the entire training process. Our implementaion is based on CoCoOp in this project. Our proposal has two perspectives to enhance the classification performance at the given datasets as follows. 1)Initializing the prompts with the meta-data of pre-trained model. We expect properly set up state by initializing the vector with information of clip model. 2)Prompt learner get the meta-data of itself so prompt simultaneously updated independent with pre-trained model regularized by the regularization function which has its meta-data as parameter.

3.3 Datasets On this project we used 3 different datasets to evaluate the performance of our methodology, oxfordpets, oxfordflowers and eurosat. To implement the model training and testing uploaded source code on Google Colab and trained on Colab Pro environment. So the specifications of CPU, GPU, OS and python environment are as follows.

Table 1. **Detailed specification of the training environment.**

CPU	Intel(R) Xeon(R) CPU @ 2.2GHz
GPU	A-100 GPU
OS	Ubuntu 22.04.3 LTS (x86_64)
ENV	Python == 3.10.12
	torch == 2.5.1+cu121
	torchvision == 0.20.1+cu121
	torchaudio==2.5.1+cu121
	torchsummary==1.5.1
	flake8==3.7.9
	numpy==1.26.4
	optree==0.13.1

To get reliable data to initialize the context vector, we extracted parameters of CLIP from output layer and normalized it. The equation to derive input vector of prompter learner is as shown below. It mean that extracted dimension of the output layer of the clip model is imported as divisor of initialization and weight become the initial weights.

$$f_{\text{initialization}} = \frac{y}{\|y\|_1}, \quad y = \sum_{x=1}^{i} \sigma x \tag{1}$$

With the meta-initialized input vector, trainer forwards the prompt during the iterations. In this progress, we divide the context and bias by the square root of vector size to regularize the transformation of tokenized prompt. Below equation explains this prompt scaling process and this proposal is to control the size

$$f_{\text{regularization}} = \frac{T(p)}{\sqrt{\|y\|_l}} \tag{2}$$

With the meta-initialized input vector, trainer forwards the prompt during the iterations. In this progress, we divide the context and bias by the square root of vector size to regularize the transformation of tokenized prompt. Above equation explains this prompt scaling process and this proposal is regularize the variance of prompts by the flow. .

Our benchmark includes OxfordPets, Oxfordflowers for classification on specific class and EuroSAT for classification on datasets from satellite. Due to the training circumstances, we can not use much larger dataset such as ImageNet. We seperated given datasets to three class train, validation and test for few-shot training with the same condition of prior experiments. The number of shot was fixed to 16 to use table from research conducted earlier, extracting average value from three seeds as it has done.

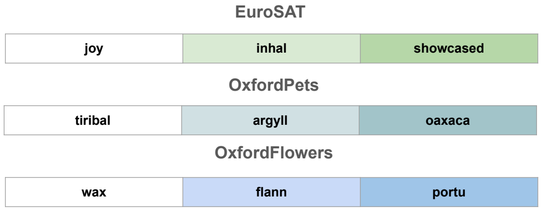
Table 2 shows the accuracy of each models on the given datasets. We can check that in general cases, base class classification is well performed by CoOp while CoCoOp does on unseen classes.

Table 2. **Comparision of CLIP,CoOp, CoCoOp in the base-to-new generalization setting** On the given environment, we trained prompt of MIP from base classed 16 shots and the result of other models is extracted from the base research. H: Harmonic mean (to highlight the generalization trade-off)

		Base	New	H
OxfordPets.	CLIP	91.17	97.26	94.12
	CoOp	93.67	95.29	94.47
	CoCoOp	**95.20**	**97.69**	**96.43**
	MIP	90.17		90.17
		Base	New	H
OxfordFlowers.	CLIP	72.08	**77.80**	74.83
	CoOp	**97.60**	59.67	74.06
	CoCoOp	94.87	71.75	**81.71**
	MIP	92.90		92.90
		Base	New	H
EuroSAT.	CLIP	56.48	**64.05**	60.03
	CoOp	**92.19**	54.74	68.69
	CoCoOp	87.49	60.04	**71.21**
	MIP	71.03		71.03

Unlike our expectation of the proposal of the project, the accuracy of MIP on general domain is lower than existing methods. It seems that regularization of the prompt was not well mathematically designed or not propagated appropriately . Also synchronization with the tokenizer of pre-trained model may be the cause of the issue.

Given below 3 tables images shows top-3 matched words for each datasets in order of EuroSAT, Oxfordpets and OxfordFlowers. Before getting matched words from the trained model, our expectation was that similar structure dataset would map to prompts that has the similar tendency. Therefore OxfordPets and OxfordFlowers seems to be have similar kind of prompts rather than the EuroSAT. In contrast, the actual result was quite different that even in the same trained model, it was hard to understand the connection between matched prompts.



EuroSAT

joy	inhal	showcased

OxfordPets

tiribal	argyll	oaxaca

OxfordFlowers

wax	flann	portu

4. Limitation

The first limitation comes from the tokenizer of CLIP model. In our design of the regularization function, tokenizer act as an oracle that quantifies that prompt while calculation of parameter was proceeded independently. So the regularization would be an approximation of dimensional alignment. However it seems to be that the prompt regularizing with optimizing the model tokenizer in the later research is worth to investigate.

The other one is the insufficiency of variation of datasets. Our model training and testing are done on 3 atasets, two of them is specific domain dataset and the other is images from specialized source, satillites. So the generalized domain dataset is absent as a hardness from building datasets and envrionment in virtual environment. So setting local environment and datasets seems to make follow-up research possible on the more diverse datasets.

The experiment design has its own limitation also. It should be the equivalent environment to train and test the model with the given dataset. However, result of existing models came from the papers not the simulated, that can infer the comparision of the results of the models.

5. Future proposal

To overcome the limitations mentioned above, subsequent researched is recommended to operate whole train and test procedure in the united environment. Otherwise, consider the tokenizer of foundation model structure to regularize the data and derivative distribuition to hyperparameterize the variance and mean of the tokenized prompts.

References

[1]Long, M., Cao, Y., Wang, J., Jordan, M. (2015, June). Learning transferable features with deep adaptation networks. In International conference on machine learning (pp. 97-105). PMLR.

[2]Chen, T., Kornblith, S., Norouzi, M., Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

[3]Zhou, K., Yang, J., Loy, C. C., Liu, Z. (2022). Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16816-16825).n machine learning (pp. 8748-8763). PMLR.

[4]Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., ... Tao, D. (2022). A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence, 45(1), 87-110.

[5]Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[6]Jia, M., Tang, L., Chen, B. C., Cardie, C., Belongie, S., Hariharan, B., Lim, S. N. (2022, October). Visual prompt tuning. In European Conference on Computer Vision (pp. 709-727). Cham: Springer Nature Switzerland.