COSE474-2024F: Final Project Proposal
"Training CLIP with Fewer Dataset using Meta-Learning"

2019320006 HeeWoong Ahn

1. Introduction

In recent years, pre-trained foundation models have demonstrated significant capabilities in various domains by leveraging large-scale datasets and advanced architectures. CLIP (Contrastive Language-Image Pretraining), a vision-language model, is one such model that has shown strong performance in zero-shot classification. However, in more focused domains, fine tuned models such as Fine-Tuning DARTS still lead in image classification on a dataset such as FashionMNIST.

While models of focused domains are superior in performance, we cannot ignore the computation power and amount of data it needs to be trained. Thus, we want to introduce Meta-learning on CLIP in order to make foundation models to adapt faster with smaller set of data on focused domains. Meta-learning refers to the study of methods that enable learning systems to improve their learning performance over time by adapting the learning process itself based on experience (Vilalta & Drissi, 2002). We aim to explore whether Few-shot learning of Meta-Learning framework applied to CLIP can achieve performance comparable to SOTA methods, specifically Fine-Tuning DARTS, while having less computational effort and data requirements.

2. Problem definition & challenges

2.1. Problem Definition

The goal of this study is to investigate whether applying a few-shot learning approach within a meta-learning framework to a foundation model like CLIP can produce comparable performance to fine-tuned models such as Fine-Tuning DARTS, but with reduced computational effort and fewer data requirements. By comparing these models on a common benchmark dataset (e.g., FashionMNIST), we aim to evaluate whether the meta-learning approach to fine-tuning CLIP can match the domain-specific SOTA (state-of-the-art) performance of Fine-Tuning DARTS, especially when training resources are limited.

2.2. Challenges

Optimizing Inner Loop and Epoch Settings: The number of inner loops and epochs directly impacts computational cost and model performance. Striking the right balance is crucial to avoid underfitting or excessive resource usage while ensuring meaningful predictions.

Determining Sample Size: In few-shot learning, selecting the optimal number of training samples is critical. Too few samples may result in poor performance, while too many could undermine the purpose of few-shot learning.

Choosing the Number of Training Tasks: The number of tasks used during meta-training affects the model's generalization ability. Finding an optimal number of tasks ensures effective meta-training without overfitting or excessive computation.

Setting an Accuracy Threshold: Defining a sufficient accuracy level is essential for validating the effectiveness of the meta-learning approach, ensuring the model performs well with minimal data and resources compared to state-of-the-art methods like Fine-Tuning DARTS.

3. Related Works

MAML is a meta-learning algorithm designed to enable a model to quickly adapt to new tasks with minimal task-specific data. The core mechanism involves training the model's parameters such that, after only a few gradient descent updates on new task data, the model can effectively perform on that task(Finn et al., 2019)

Inner Loop: In the inner loop, the model updates its parameters using a few steps of gradient descent for each task. This adaptation allows the model to specialize to each specific task with just a small amount of data.

Outer Loop: The outer loop of MAML updates the model's initialization parameters by evaluating how well the model performed after adaptation in the inner loop. The aim is to find initial parameters that enable fast adaptation across a variety of tasks.

4. Datasets

The *FashionMNIST* dataset, introduced by Zalando Research, consists of 70,000 grayscale images (28x28 pixels) divided into 60,000 training and 10,000 test samples. Each image belongs to one of ten fashion item categories, such as T-shirt, dress, and sneaker, serving as a more complex alternative to the MNIST dataset. FashionMNIST retains the same format and train-test split as MNIST, making it a drop-in replacement for benchmarking image classification models. This dataset is commonly used to evaluate the performance of machine learning algorithms, particularly in tasks involving real-world image classification (Xiao et al., 2017).

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

Recent advancements in Neural Architecture Search (NAS) have shown significant improvements in image classification tasks. The Differentiable Architecture Search (DARTS) method, in particular, has been a popular stochastic NAS technique due to its computational efficiency. However, it makes several approximations, leading to performance limitations. To overcome these issues, Fine-Tuning DARTS has been proposed, incorporating fixed operations like attention modules, resulting in improved classification accuracy across multiple datasets, including FashionMNIST.

On FashionMNIST, Fine-Tuning DARTS achieved superior accuracy (96.91) compared to traditional methods, such as ResNet (95.99), VGG (95.47), and DeepCaps (94.46). (Tanveer et al., 2020)

6. Schedule

Read discussion papers related to meta-learning (~11.15)
Make meta-learning model and test it on CoLAB (~12.01)
Write discussion paper and finalize the paper (~12.14)

References

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. Online meta-learning. 2019.

Tanveer, M. S., Khan, M. U. K., and Kyung, C.-M. Fine-tuning darts for image classification. 2020.

Vilalta, R. and Drissi, Y. A perspective view and survey of meta-learning. 18(2):77–95, 2002.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.