Fine-tuning the Foundational Model for Image Classification using Few Samples

Computer Vision (CS6350)

TPA-17

1. Introduction

Foundation models (FMs) such as CLIP, DINOv2, and SAM have transformed the landscape of computer vision by learning universal representations from massive web-scale datasets. These models generalize well across domains, but often struggle with high accuracy in specialized domains (e.g., medical imaging, remote sensing, or manufacturing inspection) due to domain shift.

This project investigates how to efficiently fine-tune such foundation models to perform high-accuracy classification in low-data regimes, balancing performance with computational and parameter efficiency.

2. Problem Statement

To evaluate and compare fine-tuning strategies on foundation models (CLIP, DINOv2, ViT) for a downstream image classification task in a domain with limited labeled data. The project will explore fine-tuning and adapter-based methods (e.g., LoRA) for optimal performance and resource trade-offs.

3. Scope of Work

Choose 1–2 pretrained foundational models:

- CLIP (OpenAI or OpenCLIP)
- DINOv2 (Meta)
- ViT (Google/MAE)

- Select a domain-specific dataset:
  - E.g., chest X-ray classification, skin cancer (ISIC), crop disease (PlantVillage), aerial imagery

- Implement multiple adaptation strategies:
    - Zero-shot evaluation
    - Linear probing
    - Full fine-tuning
    - Adapter tuning (LoRA, BitFit, prefix-tuning)
- Compare across:
    - Classification accuracy (Top-1, F1-score)
    - Number of trainable parameters
    - Training time/resource usage
    - Generalization (on held-out or shifted domain samples)

4. Expected Input and Output

- A reproducible training and evaluation pipeline
- Visualization of learned representations (e.g., t-SNE, Grad-CAM)
- Comparative tables and plots showing trade-offs

5. Dataset

1. **FC100 (Few-shot CIFAR-100)**
2. **Meta-Dataset (**https://github.com/google-research/meta-dataset**)**
3. **Flowers102-FewShot (https://www.robots.ox.ac.uk/~vgg/data/flowers/102/index.html)**
4. **ISIC 2018 Few-Shot (https://challenge.isic-archive.com/)**

Explore the dataset and choose any 2-3 to compare the methods and demonstrate the results.

6. References

1. https://medium.com/@nischaydiwan1026/exploring-parameter-efficient-fine-tuning-for-foundation-models-in-image-segmentation-49a7701a012a
2. https://www.databricks.com/blog/efficient-fine-tuning-lora-guide-llms
3. https://medium.com/@caterine/fine-tuning-a-vision-model-with-pytorch-part-1-56123db13c85

4. Y. Wei et al., "Improving CLIP Fine-tuning Performance," 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 5416-5426, doi: 10.1109/ICCV51070.2023.00501.
5. Q. Wu, J. Qi, D. Zhang, H. Zhang and J. Tang, "Fine-Tuning for Few-Shot Image Classification by Multimodal Prototype Regularization," in IEEE Transactions on Multimedia, vol. 26, pp. 8543-8556, 2024, doi: 10.1109/TMM.2024.3379896.
6. Davila, Ana, Jacinto Colan, and Yasuhisa Hasegawa. "Comparison of fine-tuning strategies for transfer learning in medical image classification." *Image and Vision Computing* 146 (2024): 105012.
7. Tanveer, Muhammad Suhaib, Muhammad Umar Karim Khan, and Chong-Min Kyung. "Fine-tuning darts for image classification." *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021.
8. Dong, Xiaoyi, et al. "Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet." *arXiv preprint arXiv:2212.06138* (2022).