A Comparative Study of Vision Transformer Encoders and Few-shot Learning for Medical Image Classification

Maxat Nurgazin and Nguyen Anh Tu
Nazarbayev University, Republic of Kazakhstan

Abstract

Computer vision has been significantly impacted by Vision Transformer (ViT) networks. However, most existing deep-learning-based methods primarily rely on a lot of labeled data to train reliable classifiers for accurate prediction. This requirement might be impractical in the medical field.

This study explores the application of ViT in few-shot learning scenarios for medical image analysis, addressing the challenges posed by limited data availability. We evaluate various ViT models alongside few-shot learning algorithms, perform cross-domain experiments, and analyze the impact of data augmentation techniques.

Our findings indicate that when combined with ProtoNets, ViT architectures outperform CNN-based counterparts and achieve competitive performance against SOTA approaches on benchmark datasets.

Motivation

- Vision Transformers (ViTs) have emerged as an alternative to CNNs, showing impressive performance on various tasks.
- CNNs struggle with learning long-range pixel relationships due to locality, which ViTs can handle more effectively.
- Medical imaging often has limited labeled data, making it difficult to train deep learning models.
- Few-shot learning (FSL) is a promising approach for handling limited labeled data.

Goal

To our knowledge, ViT architectures have not been used in the field of medical image classification in few-shot learning scenarios. Therefore, given their success in other areas of computer vision, it is important to assess their performance in this area under various conditions.

Contribution

- Investigate the efficacy of various ViT models for few-shot medical image classification.
- Study how different few-shot learning algorithms impact the performance of ViT models.
- Analyze the impact of advanced data augmentation techniques on ViT models.
- Explore the effect of a cross-domain scenario on the performance of few-shot learners.
- Framework through experiments when running on the Spark clusters.
- Our methods achieve state-of-the-art performance on challenging medical datasets of few-shot medical image classification.

Methodology

Problem definition:

Let \( D = D_1, D_2, \ldots, D_n \) be a collection of \( n \) medical datasets, with each dataset \( D_i \) consisting of pairs \( (x, y) \) representing an image and its label. Datasets are divided into meta-test set \( D_{\text{meta-test}} \) and meta-train set \( D_{\text{meta-train}} \).

Utilize abundant data in \( D_{\text{meta-train}} \) to learn better initial weights (Reptile) or develop effective embedding space (ProtoNet & MatchingNet).

Goal: Improve performance on problems \( D_{\text{meta-test}} \) with limited data (novel class data)

Overview of the system pipeline

Results

- Datasets: BreakHis (9109 microscopic images of breast tumor tissues from 82 patients with 8 classes), ISIC 2018 (10,015 dermoscopic images of skin lesions across 7 classes), and Pap Smear (917 microscopic images of cervical smears with 7 classes).
- Experimental Settings: Pre-trained models obtained from the timm library.
- ProtoNet: 20 epochs, 500 episodes per epoch, SGD optimizer, learning rate of 10-5 or 10-6, cosine annealing learning rate schedule.
- Reptile: SGD optimizer, learning rate of 10-3 for inner optimization, learning rate of 10-1 for outer meta-update. 1000 meta-iterations, batch size of 10 tasks, 5 and 50 adaptation steps for each task.
- Evaluation metric: Accuracy (%) as evaluation metric. 400 episodes randomly selected from novel categories in the test set. Average accuracy rate for image classification.

References