NAZARBAYEV UNIVERSITY

Master Thesis:

Few-shot Medical Image Classification using Vision Transformers

Author:	Maxat Nurgazin	Master Student at NU
Supervisor:	Nguyen Anh Tu	Assistant Professor, School of Engineering and Digital Sciences at NU
Co-Supervisor:	Min-Ho Lee	Assistant Professor, School of Engineering and Digital Sciences at NU

Outline:

1. Introduction
   - Background
   - Motivation
   - Objectives
2. Related works
   - Proposed idea
3. Methodology
   - Problem definition
   - Few-shot learning
   - System Pipeline
   - Reptile
   - Prototypical Networks
   - Custom ViT
   - Advanced Augmentation techniques
4. Experiments and results
   - Datasets
   - Implementation details
   - Results analysis
5. Conclusions

NAZARBAYEV UNIVERSITY

Introduction

Background

- Medical Image Analysis (MIA) is critical for diagnosing diseases and conditions from medical imaging.
- Machine learning, particularly deep learning, has shown promising results in MIA tasks, like medical image classification (MIC).
- Convolutional Neural Networks (CNNs) have been state-of-the-art in computer vision, including medical imaging.
- Vision Transformers (ViTs) have emerged as an alternative to CNNs, showing impressive performance on various tasks.

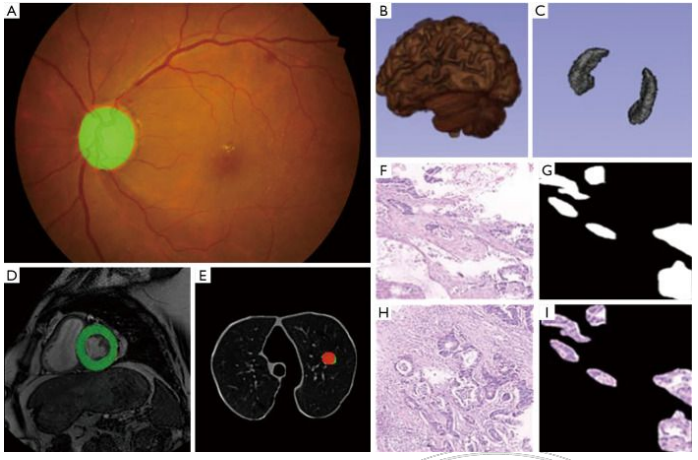


Fig. 1. Deep learning application in medical image analysis. (A) Fundus detection; (B,C) hippocampus segmentation; (D) left ventricular segmentation; (E) pulmonary nodule classification; (F,G,H,I) gastric cancer pathology segmentation. Acquired from [27]

[27] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," Annals of translational medicine, vol. 8, no. 11, 2020.

Motivation

- CNNs struggle with learning long-range pixel relationships due to locality, which ViTs can handle more effectively.
- ViTs lack inductive bias, they can learn better.
- 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.
- Investigating the use of ViTs in few-shot learning for MIC is the main motivation of this thesis.

Research Objectives

- Investigate the performance of ViTs in few-shot learning scenarios for MIC and compare it with traditional CNNs.
- Design a custom ViT architecture and evaluate its performance
- Use different few-shot learning algorithms and assess the performance of ViTs.
- Investigate the effects of advanced data augmentation techniques (Cutout, Mixup, and Cutmix) on ViT performance for FSL.

General Objective

- 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.

Terminology

- MIA: Medical Image Analysis
- MIC: Medical Image Classification
- CNN: Convolutional Neural Network
- ViT: Vision Transformer
- FSL: Few-shot Learning
- ProtoNet: Prototypical Network

NAZARBAYEV UNIVERSITY

Related Works

Few-shot Learning with ViT

Limited number of papers on few-shot learning with ViT

- Hu et al. investigated a simple FSL pipeline with ViT and ResNet50 backbones. Their pipeline with a ViT outperformed the one with a CNN. [1]
- Chen et al. used a vanilla ViT with masking operation to improve few-shot learning performance. It resulted in improved results. [2]

However, this works do not consider medical datasets.

[1] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9068–9077, 2022.
[2] Yuzhong Chen, Zhenxiang Xiao, Lin Zhao, Lu Zhang, Haixing Dai, David Weizhong Liu, Zihao Wu, Changhe Li, Tuo Zhang, Changying Li, et al.
Mask-guided vision transformer (mg-vit) for few-shot learning. arXiv preprint arXiv:2205.09995, 2022.

Medical Image Classification using Vision Transformers

- Krishnan and Krishnan fine-tuned off-the-shelf CNN and ViT models for medical image classification. ViT achieved the highest accuracy. [3]
- Perera et al. proposed a lightweight transformer architecture called POCFormer for COVID-19 detection on portable devices and reported comparable scores with bigger models. [4]
- Duong et al. combined CNN and ViT for Tuberculosis detection in Chest X-ray images and reported high scores. [5]
- Behrendt et al. systematically compared ViTs and CNNs for multi-label medical image classification. DeiT outperformed other models across all dataset sizes. [6]

These works and others show that ViTs can be successfully used for MIC. However, they do not consider FSL.

[3] Koushik Sivarama Krishnan and Karthik Sivarama Krishnan. Vision transformer based covid-19 detection using chest x-rays. In 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), pages 644–648. IEEE, 2021.
[4] Shehan Perera, Srikar Adhikari, and Alper Yilmaz. Pocformer: A lightweight transformer architecture for detection of covid-19 using point of care ultrasound. In 2021 IEEE International Conference on Image Processing (ICIP), pages 195– 199. IEEE, 2021.
[5] Linh T Duong, Nhi H Le, Toan B Tran, Vuong M Ngo, and Phuong T Nguyen. Detection of tuberculosis from chest x-ray images: boosting the performance with vision transformer and transfer learning. Expert Systems with Applications, 184:115519, 2021.
[6] Finn Behrendt, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Data-efficient vision transformers for multi-label disease classification on chest radiographs. Current Directions in Biomedical Engineering, 8(1):34–37, 2022.

Medical Image Classification using Few-Shot Learning

- Singh et al. proposed MetaMed, a meta-learning-based approach for few-shot medical image classification using Reptile and a simple CNN. [7]
- Dai et al. introduced PFEMed, a novel few-shot classification method for medical images. This approach surpassed MetaMed on the Pap smear dataset by over 2.63%. [8]
- Cherti and Jitsev investigated the effect of pre-training scale on intra- and inter-domain transfer settings. Demonstrated transfer learning benefited from larger pre-training scales. [9]

These works use CNNs for FSL in MIC. However ViTs are outperforming CNNs.

[7] Rishav Singh, Vandana Bharti, Vishal Purohit, Abhinav Kumar, Amit Kumar Singh, and Sanjay Kumar Singh. Metamed: Few-shot medical image classification using gradient-based meta-learning. Pattern Recognition, 120:108111, 2021.
[8] Zhiyong Dai, Jianjun Yi, Lei Yan, Qingwen Xu, Liang Hu, Qi Zhang, Jiahui Li, and Guoqiang Wang. Pfemed: Few-shot medical image classification using prior guided feature enhancement. Pattern Recognition, 134:109108, 2023.
[9] Mehdi Cherti and Jenia Jitsev. Effect of pre-training scale on intra-and inter-domain full and few-shot transfer learning for natural and medical x-ray chest images. arXiv preprint arXiv:2106.00116, 2021.

Proposed idea

In this work, we aimed to bridge the gap in knowledge by:

- Employing various ViT architectures in few-shot learning for medical image classification
- Evaluating their performance by comparing them with similar CNNs
- Examining the impact of advanced data augmentation techniques

We utilized two FSL algorithms - Prototypical Networks and Reptile

NAZARBAYEV UNIVERSITY

Methodology

Problem Definition

- Let $D = D_1, D_2, ..., D_n$ be a collection of n medical datasets, with each dataset D_k consisting of pairs $(x, y)_j$ representing an image and its label.
- Datasets are divided into meta-test set ($D_{meta\text{-}test}$) and meta-train set ($D_{meta\text{-}train}$)
- Utilize abundant data in $D_{meta\text{-}train}$ to learn better initial weights (Reptile) or develop effective embedding space (ProtoNet)
- Goal: Improve performance on problems $D_{meta\text{-}test}$ with limited data (novel class data)

Few-shot Learning

- Goal: Develop models that generalize effectively to new tasks with limited labeled examples
- Task difficulty: N-way-K-shot (N = number of classes, K = samples per class). Ex: 3-way-3-shot
- Support set for training, Query set for testing

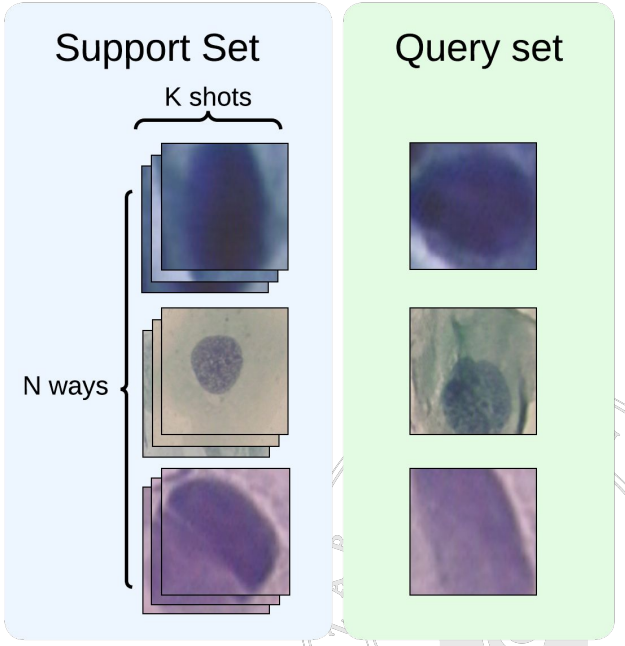
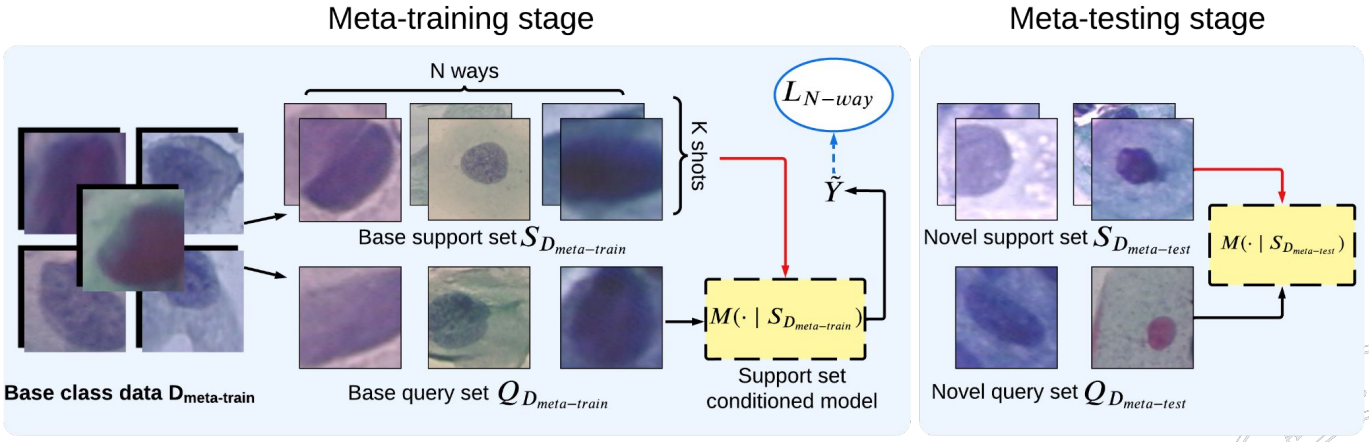


Fig 2. N-way-K-shot task example (3-way-3-shot)

Few-shot Learning Approaches

- Meta-learning - Learn to solve new tasks by drawing experience from previous tasks. **Sharing a learning method**.
  - Models that adapt quickly to new task from few examples
  - Divided into meta-training and meta-testing phases
  - Data presented episodically (one episode is a n-way-k-shot task)
- Transfer learning - Pre-train on a large dataset, then fine-tune on limited noval data. **Sharing learned knowledge.**
  - Less effective when there's a large domain gap between source and target datasets
- Data augmentation - Generate new samples by augmenting limited support set
  - Enhances the limited support set by generating new samples through various techniques

System Pipeline



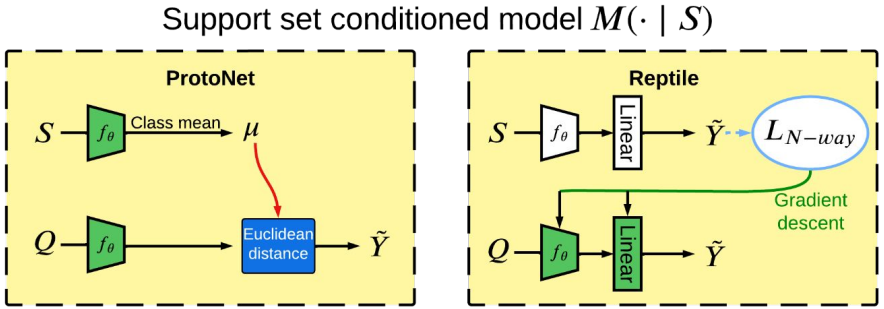
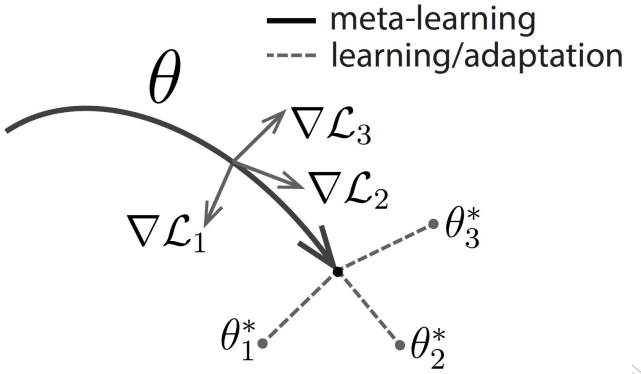


Fig 3. Overall Pipeline

Reptile [10]

- Meta-learning algorithm designed for few-shot learning
- Find init. params that can be quickly adapted to new tasks
- Two-stage process: inner loop updates and outer loop updates
- Quick training and simple implementation compared to MAML
- (1) Cross entropy loss used during meta-training and meta-testing phases



$$\mathcal{L}_{T_i}(f_\phi) = -\sum_{x_i, y_i \sim T_i} y_i log(f_\phi(x_i) + (1 - y_i) log(1 - f_\phi(x_i)) \quad (1)$$

Algorithm 1 Reptile

1: Initialize model weights $\boldsymbol{\theta}$
2: **for** iteration $= 1, 2, \ldots, N$ **do**
3: Sample task T_i from task distribution $p(T)$
4: Perform K steps of SGD on T_i to obtain updated weights $\boldsymbol{\theta}'$
5: Update $\boldsymbol{\theta}$ using $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \epsilon(\boldsymbol{\theta}' - \boldsymbol{\theta})$

[10] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999, 2018.

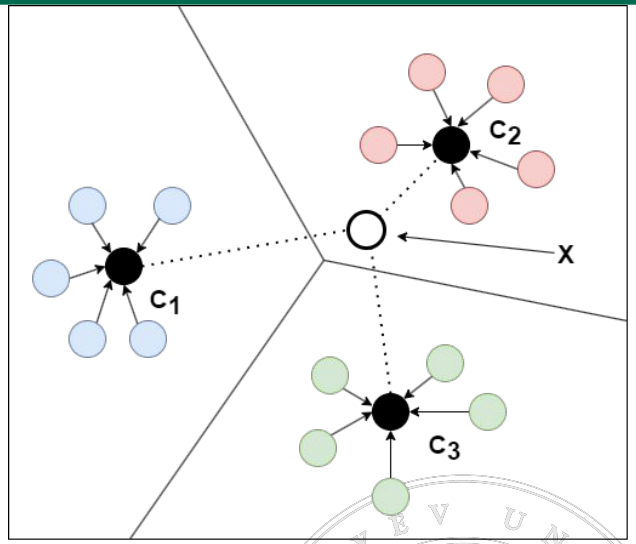
Prototypical Networks [11]

- Learn a prototype for each class in the embedding space

Algorithm 2 Prototypical Networks

1: **for** each class k in D_{train} **do**
2: Extract N_s support samples for class k
3: Calculate the class prototype c_k as the mean of N_s samples using (1)
4: **for** each sample q in D_{query} **do**
5: Predict class for q using (2)

- The figure on the right demonstrates the classification procedure



$$c_k = \frac{1}{N_k} \sum_{(x_i, y_i) \in S, y_i = k} \phi(x_i) \quad (1)$$

$$\hat{y} = \arg\min_k \|\phi(x) - c_k\|^2 \quad (2)$$

[11] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. Advances in neural information processing systems, 30, 2017.

Custom ViT

- Add Squeeze & Excitation (SE) block to ViT_small [18] and ConViT_small [28]
- Adapted from [29], where it showed an improvement over vanilla ViT.
- We also use SE block with ConViT.
- ConViT is similar to original ViT
  - But uses gated positional self-attention in some layers for convolutional inductive bias of locality.

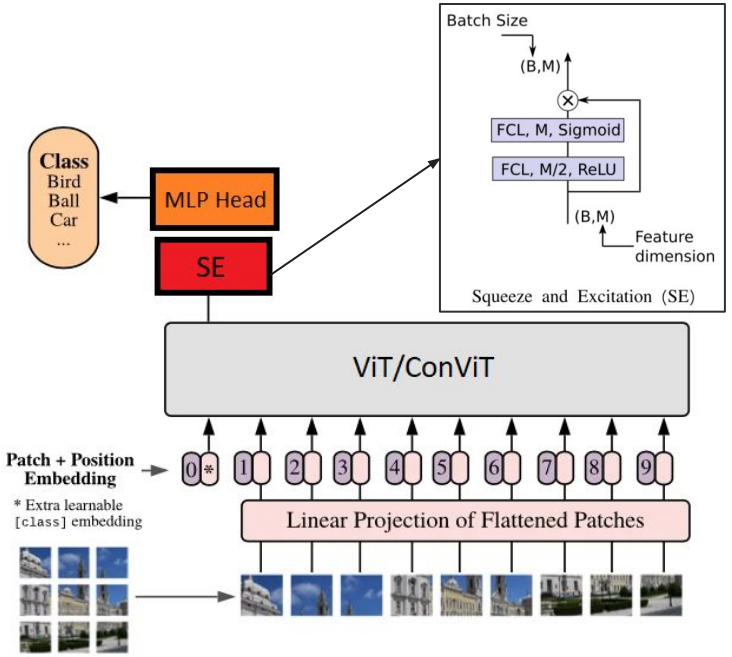


Fig. 4. Architecture of Custom ViTs. Figure is adapted from [18], [28], and [29].

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
[28] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in International Conference on Machine Learning, 2021, pp. 2286–2296.
[29] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, and R. Seguier, "Learning vision transformer with squeeze and excitation for facial expression recognition," arXiv preprint arXiv:2107.03107, 2021.

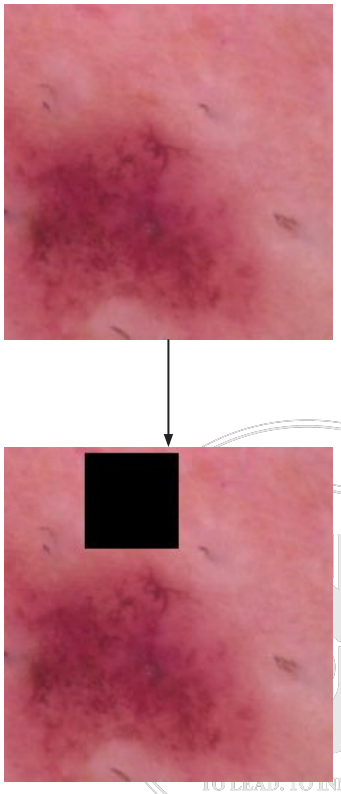
Advanced Augmentation Techniques

- Encourage model to learn more generalized representations by providing more diverse and robust training data
- Techniques used: Cutout, Mixup, and Cutmix

- Note: Only Cutout is compatible with ProtoNet algorithm

Cutout [12]

- Cutout is a data augmentation technique that randomly removes rectangular regions from input images during training.



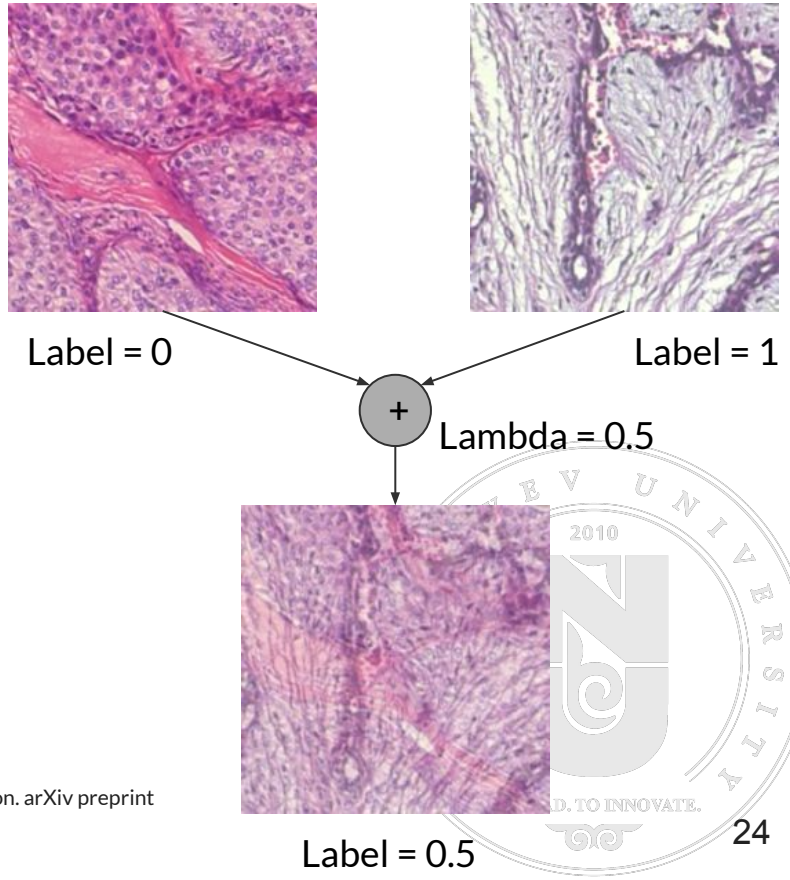
[12] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017.

Mixup [13]

- Mixup is a method that involves generating new training examples by taking a linear combination of two randomly chosen input images and their corresponding labels.

$$x_{\text{mixup}} = \lambda x_1 + (1 - \lambda) x_2$$

$$y_{\text{mixup}} = \lambda y_1 + (1 - \lambda) y_2$$



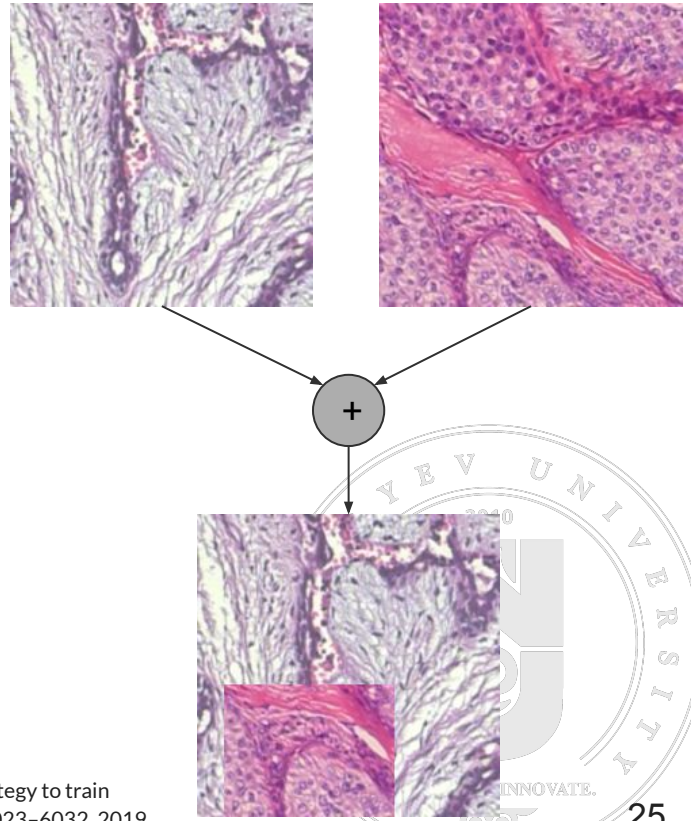
[13] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017.

Cutmix [14]

- Cutmix is a method that combines the strengths of both Cutout and Mixup. The idea behind Cutmix is to replace a portion of an input image x_1 with another image x_2, while also adjusting the corresponding labels accordingly.

$$x_{\text{cutmix}} = \lambda x_1 + (1 - \lambda) x_2$$

$$y_{\text{cutmix}} = \lambda y_1 + (1 - \lambda) y_2$$



[14] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6023–6032, 2019.

NAZARBAYEV UNIVERSITY

Experiments and Results

Datasets: ISIC 2018

- Three publicly available medical imaging datasets: BreakHis, ISIC 2018, and Pap Smear
- Images downsampled to 224x224 for compatibility with pre-trained models
- Each dataset contains at least six classes for 2- and 3-way n-shot learning

ISIC 2018 [15]:

- 10,015 dermoscopic images of skin lesions across 7 classes
- 4 meta-train and 3 meta-test classes

[15] Jinyi Zou, Xiao Ma, Cheng Zhong, and Yao Zhang. Dermoscopic image analysis for isic challenge 2018. arXiv preprint arXiv:1807.08948, 2018.

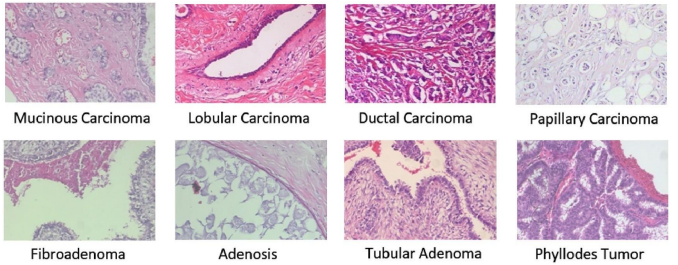
BreakHis and Pap smear

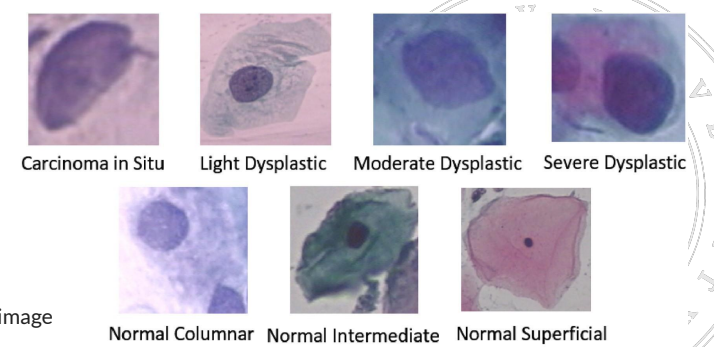
BreakHis [16]:

- 9109 microscopic images of breast tumor tissues from 82 patients
- 8 classes, with 5 meta-train and 3 meta-test classes

Pap Smear [17]:

- 917 microscopic images of cervical smears
- 7 distinct classes, with 4 meta-train and 3 meta-test classes





[16] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. Ieee transactions on biomedical engineering, 63(7):1455–1462, 2015.
[17] Jan Jantzen, Jonas Norup, Georgios Dounias, and Beth Bjerregaard. Pap-smear benchmark data for pattern classification. Nature inspired Smart Information Systems (NiSIS 2005), pages 1–9, 2005.

Models

- ViT family [18]: ViT_tiny, ViT_small, and ViT_base
- Other ViT architectures: Mobile_ViT (MViT_v2_0.5) [19], DeiT_base [20], and Swin_base [21]
- CNN models: ResNet50 [22] and VGG16 [23]
- All models pre-trained on the ImageNet1k dataset

Model	Dim	Parameters
ViT_tiny	192	5.5m
MViT_v2_0.5	384	1.4m
ViT_small	384	22m
ViT_base	768	85m
DeIT_base	768	85m
Swin_base	1024	86m
ResNet50	2048	23.5m
VGG16	4096	134m

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
[19] Sachin Mehta and Mohammad Rastegari. Separable self-attention for mobile vision transformers, 2022.
[20] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, pages 10347–10357. PMLR, 2021.
[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10012–10022, 2021.
[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

Implementation details and Settings

- Python programming language
- PyTorch framework [24]
- Pre-trained models obtained from the timm library [25]
- ProtoNet experiments conducted using the easyfsl library [26]

- Hardware specifications:
  - PC: NVIDIA RTX 3060 Ti, Intel i5-10400 CPU, 16GB RAM
  - Google Colab Pro Platform: NVIDIA Tesla T4 or A100

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019.
[25] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.
[26] Etienne Bennequin. easyfsl. https://github.com/sicara/easy-few-shot-learning, 2021.

Training

- Utilized pre-trained model checkpoints
- Employed data augmentation techniques
- 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

- Accuracy (%) as evaluation metric
- 400 episodes randomly selected from novel categories in the test set
- Average accuracy rate for image classification
- Tested 2- and 3-way 2-, 5-, and 10-shot few-shot learning scenarios

Pretrained ViTs without Meta-training

- Investigates results of pre-trained models in few-shot classification tasks without meta-training
- Focus on ISIC 2018 and BreakHis x100 datasets
- Pure transfer learning, no fine-tuning to meta-datasets
- Models directly used as a backbone of a ProtoNet
- Serves as a baseline for further sections

Pretrained ViT without Meta-training: Observations

- Models with more parameters generally show better performance
- Mobile ViT (MViT_v2_0.5) has the lowest score, followed by ViT_tiny
- ViT and CNN models show comparable results
- Results only serve as an initial baseline and should not be used to judge overall performance of models in few-shot learning

BreakHis X100 Pretaining Only						
Model	2-way			3-way		
	3-shot	5-shot	10-shot	3-shot	5-shot	10-shot
ViT_tiny	71.25	77.35	78.50	57.27	61.53	67.87
ViT_small	74.71	79.42	83.24	63.22	69.25	73.91
ViT_base	74.50	80.70	84.90	63.90	69.17	75.50
Swin_base	77.95	83.20	65.37	72.77	80.30	82.3
ResNet50	79.62	83.31	85.72	68.75	73.09	77.61
VGG16	70.40	79.15	81.75	60.70	65.40	71.67

ISIC 2018 Pretaining Only						
Model	2-way			3-way		
	3-shot	5-shot	10-shot	3-shot	5-shot	10-shot
ViT_tiny	70.25	74.60	76.15	54.83	59.51	65.41
MViT_v2_0.5	59.30	63.10	67.80	46.13	48.27	49.90
ViT_small	77.40	81.89	85.95	63.67	69.84	75.28
ViT_base	74.75	77.70	82.45	60.73	65.73	69.97
DeIT_base	71.75	79.40	81.75	58.33	61.87	69.47
Swin_base	75.10	80.15	82.00	62.27	67.67	71.50
ResNet50	72.66	76.17	79.15	56.69	62.31	65.81
VGG16	72.45	78.60	81.30	60.00	65.87	68.20

Meta-Training Results

ISIC 2018

Algorithm	Model	2-way			3-way		
		3-shot	5-shot	10-shot	3-shot	5-shot	10-shot
Protonet	MViT_v2_0.5	74.64	76.94	81.50	60.60	64.23	69.23
	ViT_tiny	81.03	83.61	86.52	67.84	71.82	77.68
	ViT_small	84.35	86.70	89.72	72.10	76.18	81.45
	ViT_base	83.94	86.02	90.26	72.75	77.69	81.99
	Swin_base	82.49	84.17	89.12	70.75	74.67	79.92
	ResNet50	66.62	68.65	72.81	51.43	53.83	58.34
	VGG16	72.32	76.04	80.69	57.81	61.86	66.92
PN w/o Pretraning	ViT_small	56.19	57.55	60.17	39.87	41.08	41.88
Reptile 5 steps	ViT_small	71.23	76.65	81.38	66.20	72.23	78.10
	ResNet50	59.50	62.80	65.78	42.62	43.22	44.13
Reptile 50 steps	ViT_small	76.05	80.3	85.55	67.5	73.15	77.37
	ResNet50	66.68	72.13	77.03	53.63	57.03	60.18

BreakHis X100

Algorithm	Model	2-way			3-way		
		3-shot	5-shot	10-shot	3-shot	5-shot	10-shot
Protonet	MViT_v2_0.5	76.89	79.60	84.65	64.51	71.43	77.05
	ViT_tiny	75.34	79.44	83.53	62.64	69.88	75.18
	ViT_small	80.64	83.80	87.62	69.39	75.91	81.47
	ViT_base	79.33	81.65	84.62	68.52	73.27	76.38
	Swin_base	79.46	82.86	86.26	68.34	74.28	80.51
	ResNet50	68.62	72.12	73.31	55.80	60.28	61.88
	VGG16	67.06	69.70	74.74	52.89	57.94	61.15
Reptile 5 steps	ViT_small	66.90	74.20	81.80	47.37	57.17	68.47
	ResNet50	64.90	67.60	73.25	34.70	36.33	38.23
Reptile 50 steps	ViT_small	73.45	77.9	86.18	55.05	63.38	75.92
	ResNet50	72.15	76.63	80.33	60.33	63.45	68.47

Pap Smear

Algorithm	Model	2-way			3-way		
		3-shot	5-shot	10-shot	3-shot	5-shot	10-shot
Protonet	MViT_v2_0.5	80.84	84.36	86.88	68.04	73.24	78.37
	ViT_tiny	84.65	86.96	88.86	74.33	77.92	81.17
	ViT_small	92.40	94.05	94.90	86.38	89.09	90.62
	ViT_base	92.05	93.26	93.94	85.21	88.48	89.47
	Swin_base	85.42	87.56	89.78	75.73	79.88	82.46
	ResNet50	70.49	71.75	69.61	57.74	58.48	59.60
	VGG16	87.95	90.11	91.45	78.21	81.81	84.32
Reptile 5 steps	ViT_small	83.35	87.05	91.96	72.52	81.13	87.94
	ResNet50	71.44	74.59	78.39	48.00	49.86	50.44
Reptile 50 steps	ViT_small	85.85	88.33	92.55	76.75	82.58	86.92
	ResNet50	86.60	90.38	90.85	65.73	67.75	73.83

Meta-Training Results

- Analyzing test results of few-shot classification models using ProtoNet and Reptile meta-learning algorithms
- ViTs paired with ProtoNet showed noticeable performance gains across all datasets and FSL tasks
- Mobile ViT and ViT, being the smallest models, showed lower results
- ViT_small demonstrated the highest results in most cases, often outperforming bigger models
- CNNs perform worse after meta-training than before and are generally non-competitive

Meta-Training Results: ProtoNet

- Importance of pretraining highlighted
- ProtoNet with ViT_small backbone pretrained on ImageNet1k has accuracy scores up to 30% higher when meta-trained
- Indicates learning a more discriminative feature representation space
- ViT_small outperforms ResNet50 in most tasks across datasets

ISIC 2018							
Algorithm	Model	2-way			3-way		
		3-shot	5-shot	10-shot	3-shot	5-shot	10-shot
PNw\oPT	ViT_small	56.19	57.55	60.17	39.87	41.08	41.88
Protonet	MViT_v2_0.5	74.64	76.94	81.50	60.60	64.23	69.23
	ViT_tiny	81.03	83.61	86.52	67.84	71.82	77.68
	ViT_small	84.35	86.70	89.72	72.10	76.18	81.45
	ViT_base	83.94	86.02	90.26	72.75	77.69	81.99
	Swin_base	82.49	84.17	89.12	70.75	74.67	79.92
	ResNet50	66.62	68.65	72.81	51.43	53.83	58.34
	VGG16	72.32	76.04	80.69	57.81	61.86	66.92
BreakHis X100							
Protonet	MViT_v2_0.5	76.89	79.60	84.65	64.51	71.43	77.05
	ViT_tiny	75.34	79.44	83.53	62.64	69.88	75.18
	ViT_small	80.64	83.80	87.62	69.39	75.91	81.47
	ViT_base	79.33	81.65	84.62	68.52	73.27	76.38
	Swin_base	79.46	82.86	86.26	68.34	74.28	80.51
	ResNet50	68.62	72.12	73.31	55.80	60.28	61.88
	VGG16	67.06	69.70	74.74	52.89	57.94	61.15
Pap Smear							
Protonet	MViT_v2_0.5	80.84	84.36	86.88	68.04	73.24	78.37
	ViT_tiny	84.65	86.96	88.86	74.33	77.92	81.17
	ViT_small	92.40	94.05	94.90	86.38	89.09	90.62
	ViT_base	92.05	93.26	93.94	85.21	88.48	89.47
	Swin_base	85.42	87.56	89.78	75.73	79.88	82.46
	ResNet50	70.49	71.75	69.61	57.74	58.48	59.60
	VGG16	87.95	90.11	91.45	78.21	81.81	84.32

Meta-Training Results: Reptile

- Performance highly dependent on proper hyperparameter selection
- Noticeable performance increase when task-adapted for more steps (5 to 50)
- ViTs adapt faster (with fewer steps)
- ViTs with Reptile still lower in performance compared to ProtoNets
- ResNet50 showed much better results with Reptile
- ProtoNet with a ViT backbone is a better option than a CNN paired with Reptile due to ease of use, training, better performance, and lower complexity

ISIC 2018							
		2-way			3-way		
Algorithm	Model	3-shot	5-shot	10-shot	3-shot	5-shot	10-shot
Protonet w/o Pretraning	ViT_small	56.19	57.55	60.17	39.87	41.08	41.88
Reptile 5 steps	ViT_small	71.23	76.65	81.38	66.20	72.23	78.10
	ResNet50	59.50	62.80	65.78	42.62	43.22	44.13
Reptile 50 steps	ViT_small	76.05	80.3	85.55	67.5	73.15	77.37
	ResNet50	66.68	72.13	77.03	53.63	57.03	60.18
BreakHis X100							
Reptile 5 steps	ViT_small	66.90	74.20	81.80	47.37	57.17	68.47
	ResNet50	64.90	67.60	73.25	34.70	36.33	38.23
Reptile 50 steps	ViT_small	73.45	77.9	86.18	55.05	63.38	75.92
	ResNet50	72.15	76.63	80.33	60.33	63.45	68.47
Pap Smear							
Reptile 5 steps	ViT_small	83.35	87.05	91.96	72.52	81.13	87.94
	ResNet50	71.44	74.59	78.39	48.00	49.86	50.44
Reptile 50 steps	ViT_small	85.85	88.33	92.55	76.75	82.58	86.92
	ResNet50	86.60	90.38	90.85	65.73	67.75	73.83

Data Augmentation Techniques - ISIC 2018 Dataset

- For ProtoNet:
  - Cutout resulted in lower scores for most tasks (ViT_small and ResNet50)
- For Reptile:
  - Cutout led to lower performance in most cases, except for 3-way k-shot tasks of ResNet50
  - CutMix generally resulted in lower scores for the majority of tasks
  - Mixup showed an uplift in accuracy scores in 4 tasks (ResNet50) and 3 tasks (ViT_small) out of 6
  - Mixup performs better than other techniques and is recommended as a good data augmentation technique

ISIC 2018										
			2-way				3-way			
Algorithm	Model	FSL	Standart	CutOut	MixUp	CutMix	Standart	CutOut	MixUp	CutMix
ProtoNet	ViT_small	3 shot	84.35	81.73	-	-	72.10	70.55	-	-
		5 shot	86.70	85.89	-	-	76.18	76.23	-	-
		10 shot	89.72	89.22	-	-	81.45	81.13	-	-
	ResNet50	3 shot	66.62	65.52	-	-	51.43	49.32	-	-
		5 shot	68.65	68.75	-	-	53.83	53.81	-	-
		10 shot	72.81	72.18	-	-	58.34	57.74	-	-
Reptile	ViT_small	3 shot	76.05	75.30	77.50	74.85	67.50	64.87	66.20	67.40
		5 shot	80.30	80.35	79.40	77.75	73.15	69.97	71.33	72.57
		10 shot	85.55	83.95	85.75	85.65	77.37	76.53	77.87	79.63
	ResNet50	3 shot	70.28	68.73	70.75	70.10	54.47	55.70	55.00	53.70
		5 shot	75.78	73.60	74.15	74.60	58.22	59.90	60.65	58.92
		10 shot	78.83	76.58	78.03	77.95	61.58	64.67	64.95	63.62

Comparison with MetaMed [7]

- Focused on ViT_small and ResNet50 models with both ProtoNet and Reptile
- Meta-training without advanced augmentation techniques
- Note: MetaMed used a simple CNN model with only 3840 parameters, making the comparison not entirely fair

- Key Observations:
  - ViT_small outperforms other models in all cases with ProtoNet, and in most cases with Reptile
  - ResNet50 lags behind the performance of other models, including those presented in the MetaMed paper

ISIC 2018

Algorithm	Model	2-way 3-shot	2-way 5-shot	2-way 10-shot	3-way 3-shot	3-way 5-shot	3-way 10-shot
ProtoNet	ViT_small	84.35	86.70	89.72	72.10	76.18	81.45
	ResNet50	66.62	68.65	72.81	51.43	53.83	58.34
Reptile	ViT_small	76.05	80.30	85.55	67.50	73.15	77.37
	ResNet50	70.28	75.78	78.83	54.47	58.22	61.58
	MetaMed	72.75	75.62	81.37	54.83	59.33	69.75

BreakHis X100

Algorithm	Model	2-way 3-shot	2-way 5-shot	2-way 10-shot	3-way 3-shot	3-way 5-shot	3-way 10-shot
ProtoNet	ViT_small	80.64	83.80	87.62	69.39	75.91	81.47
	ResNet50	68.62	72.12	73.31	55.80	60.28	61.88
Reptile	ViT_small	73.45	77.90	86.18	55.05	63.38	75.92
	ResNet50	72.15	76.63	80.33	60.33	63.45	68.47
	MetaMed	78.75	81.38	83.88	63.08	66.42	74.08

Pap Smear

Algorithm	Model	2-way 3-shot	2-way 5-shot	2-way 10-shot	3-way 3-shot	3-way 5-shot	3-way 10-shot
ProtoNet	ViT_small	92.40	94.05	94.90	86.38	89.09	90.62
	ResNet50	70.49	71.75	69.61	57.74	58.48	59.60
Reptile	ViT_small	83.35	87.05	91.96	72.52	81.13	87.94
	ResNet50	71.44	74.59	78.39	48.00	49.86	50.40
	MetaMed	85.37	86.50	89.37	70.58	72.42	83.00

[7] Rishav Singh, Vandana Bharti, Vishal Purohit, Abhinav Kumar, Amit Kumar Singh, and Sanjay Kumar Singh. Metamed: Few-shot medical image classification using gradient-based meta-learning. Pattern Recognition, 120:108111, 2021.

Custom ViT Preliminary Results

- Custom ViTs were pre-trained on ImageNet1k, then on CIFAR 100.
- Generally, lower results when compared with unmodified models.
- CIFAR 100 is too small for a proper pre-training
- Later, use bigger datasets.

ISIC 2018							
Algorithm	Model	2-way			3-way		
		3-shot	5-shot	10-shot	3-shot	5-shot	10-shot
Protonet	ViT_small	84.35	86.70	89.72	72.10	76.18	81.45
	ViT_small_SE	77.84	80.66	84.36	64.30	68.24	74.66
	ConViT	76.33	78.89	82.94	63.17	67.07	71.96
	ConViT_SE	75.21	77.18	81.71	60.94	65.11	69.60

NAZARBAYEV UNIVERSITY

Conclusion

Conclusion and Key Contributions

In this work we have shown that:

- ViTs can be effectively used for few-shot medical image classification outperforming comparable CNNs.
  - Especially with the ProtoNet FSL algorithm.
  - Reptile performance depends highly on hyperparameter selection.

Additionally, we demonstrated that:

- ViTs reach high performance with a simpler ProtoNet.
- Tested a custom ViT architecture with SE block for FSL.
  - Have only preliminary results. Lower performance than unmodified.
- Advanced augmentation techniques showed mixed results
  - Mixup improved accuracy scores in most cases.
  - Cutout and Cutmix showed positive results in less than 50% of tasks.

Future Directions

- Designing a ViT architecture fit for FSL.
- Investigating usage synthetic data augmentation/generation techniques (e.g., Variational Autoencoders, Generative Adversarial Networks) in the pipeline.

Reference List

[1] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9068–9077, 2022.
[2] Yuzhong Chen, Zhenxiang Xiao, Lin Zhao, Lu Zhang, Haixing Dai, David Weizhong Liu, Zihao Wu, Changhe Li, Tuo Zhang, Changying Li, et al.
Mask-guided vision transformer (mg-vit) for few-shot learning. arXiv preprint arXiv:2205.09995, 2022.
[3] Koushik Sivarama Krishnan and Karthik Sivarama Krishnan. Vision transformer based covid-19 detection using chest x-rays. In 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), pages 644–648. IEEE, 2021.
[4] Shehan Perera, Srikar Adhikari, and Alper Yilmaz. Pocformer: A lightweight transformer architecture for detection of covid-19 using point of care ultrasound. In 2021 IEEE International Conference on Image Processing (ICIP), pages 195– 199. IEEE, 2021.
[5] Linh T Duong, Nhi H Le, Toan B Tran, Vuong M Ngo, and Phuong T Nguyen. Detection of tuberculosis from chest x-ray images: boosting the performance with vision transformer and transfer learning. Expert Systems with Applications, 184:115519, 2021.
[6] Finn Behrendt, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Data-efficient vision transformers for multi-label disease classification on chest radiographs. Current Directions in Biomedical Engineering, 8(1):34–37, 2022.
[7] Rishav Singh, Vandana Bharti, Vishal Purohit, Abhinav Kumar, Amit Kumar Singh, and Sanjay Kumar Singh. Metamed: Few-shot medical image classification using gradient-based meta-learning. Pattern Recognition, 120:108111, 2021.
[8] Zhiyong Dai, Jianjun Yi, Lei Yan, Qingwen Xu, Liang Hu, Qi Zhang, Jiahui Li, and Guoqiang Wang. Pfemed: Few-shot medical image classification using prior guided feature enhancement. Pattern Recognition, 134:109108, 2023.
[9] Mehdi Cherti and Jenia Jitsev. Effect of pre-training scale on intra-and inter-domain full and few-shot transfer learning for natural and medical x-ray chest images. arXiv preprint arXiv:2106.00116, 2021.
[10] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999, 2018.
[11] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. Advances in neural information processing systems, 30, 2017.
[12] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017.
[13] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017.
[14] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6023–6032, 2019.

Reference List

[15] Jinyi Zou, Xiao Ma, Cheng Zhong, and Yao Zhang. Dermoscopic image analysis for isic challenge 2018. arXiv preprint arXiv:1807.08948, 2018.
[16] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. Ieee transactions on biomedical engineering, 63(7):1455–1462, 2015.
[17] Jan Jantzen, Jonas Norup, Georgios Dounias, and Beth Bjerregaard. Pap-smear benchmark data for pattern classification. Nature inspired Smart Information Systems (NiSIS 2005), pages 1–9, 2005.
[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
[19] Sachin Mehta and Mohammad Rastegari. Separable self-attention for mobile vision transformers, 2022.
[20] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, pages 10347–10357. PMLR, 2021.
[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10012–10022, 2021.
[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.
[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019.
[25] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.
[26] Etienne Bennequin. easyfsl. https://github.com/sicara/easy-few-shot-learning, 2021.
[27] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," Annals of translational medicine, vol. 8, no. 11, 2020.
[28] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in International Conference on Machine Learning, 2021, pp. 2286–2296.
[29] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, and R. Seguier, "Learning vision transformer with squeeze and excitation for facial expression recognition," arXiv preprint arXiv:2107.03107, 2021.

NAZARBAYEV UNIVERSITY

Q&A