Integrating Vision Transformer-Based Self-Supervised Pre-training with Few-Shot Learning on Meta-Album

Abdulaziz Alyahya
abalyahya@gwu.edu
The George Washington University
Washington, District of Columbia, USA

Qichang Dong
qdong12@gwu.edu
The George Washington University
Washington, District of Columbia, USA

1 INTRODUCTION

The rapid advancement in deep learning has significantly impacted various domains, with machine learning models requiring extensive labeled data for training. Few-shot Learning (FSL) addresses this by training models on limited labeled data, although challenges like overfitting persist due to data scarcity. Self-supervised Learning (SSL) has emerged as a powerful method for automatic feature extraction from unlabeled data, showing promise in computer vision and natural language processing.

While many researchers [2, 4, 6, 10–12] have advanced FSL and explored SSL, there is a noticeable gap in studies systematically investigating their combined benefits, especially with Vision Transformers (ViT). This research aims to bridge this gap, exploring the integration of FSL and SSL with ViT pre-training on the Meta-Album[14] dataset.

1.1 Research Question

Will the model pre-trained with SSL using Vision Transformers (ViT) on Meta-Album[14] outperform the model using only standard FSL techniques?

2 RELATED WORK

Li et al. (2022)[7] introduced the Ranking Distance Calibration (RDC) method, a **state-of-the-art** approach in Cross-Domain Few-Shot Learning, enhancing discriminative representation through innovative calibration techniques. Vinyals et al. (2016)[16] emphasized the integration of non-parametric structures to bolster adaptability, leading to the inception of Matching Networks.Furthermore, Bertinetto et al. (2018)[1] presented R2-D2, a differentiable ridge regression base learner, which showcased its prowess in high-dimensional, few-shot scenarios, outperforming several contemporary methods. Moreover, Zhai et al. (2019)[17] underscored the potential of SSL, which, when amalgamated with limited labels, rivaled the performance achieved using a substantial portion of ImageNet labels. Moreover, Triantafillou et al. (2019)[13] unveiled the META-DATASET environment, shedding light on the variability in model performance across diverse data sources and their sensitivity to test data volume. Furthermore, Lu et al. (2022)[9] brought forth a paradigm shift by introducing an FSL approach devoid of base dataset labels. By centering on maximizing mutual information, they achieved results that rivaled top-tier supervised methods, prompting a reevaluation of the significance of base dataset label information in FSL. Moreover, Boudiaf et al. (2020)[3] and Li et al. (2019) [8] made noteworthy contributions by setting new benchmarks in FSL without intricate meta-training and by innovating with recursive learning strategies, respectively, both of which have paved the way for future research trajectories in this domain. Moreover, Dosovitskiy et al. (2021)[5] introduced the Vision Transformer, excelling in image classification without image-specific biases, except for initial patch extraction, as illustrated in Figure 1.

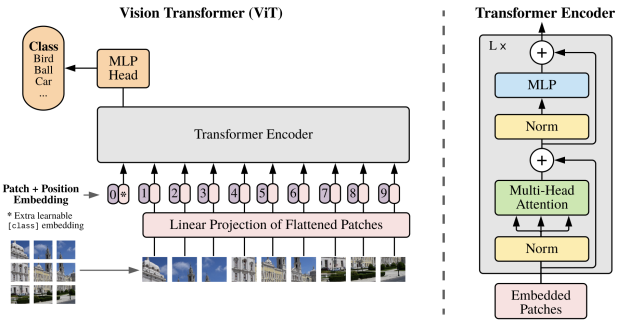

Figure 1: Vision Transformer (ViT) by Dosovitskiy et al. (2021)[5] inspired by Vaswani et al. (2017)[15]

3 METHOD

3.1 Model Architecture and Preparation

As can be seen from Figure 2., two models are employed: Model_A, a Convolutional Neural Network (CNN) optimized for FSL, and Model_B, SSL pre-trained based on the Vision Transformer (ViT) architecture. The core of the ViT architecture is the self-attention mechanism, it is defined by formula (1):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. This mechanism enables the model to focus on different parts of the input image, considering the global context and relationships between patches

3.1.1 Model_A (CNN for Few-Shot-Learning): Model_A follows standard FSL tasks, serving as the baseline model for this study.

3.1.2 Model_B (ViT with Self-Supervised Pre-training).

- Phase 1 (Self-Supervised Pre-training):

Model_B undergoes SSL pre-training using the ViT architecture, focusing on contrastive learning to enhance feature representations by distinguishing between augmented views of images.

Mutual Information (MI), a crucial concept in SSL for learning representations, quantifies variable dependency, as shown in formula (2).

$$MI(X;Y) = \sum_{x \in X, y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{2}$$

MI enhances feature representation learning in SSL by maximizing shared information across data views, crucial for tasks like contrastive learning.

- Phase 2: Few-Shot Learning:

Post SSL pre-training, Model_B undergoes fine-tuning via FSL tasks and Matching Networks, leveraging learned representations to calculate similarities between test and support set examples, as depicted in formula (3).

$$\hat{y} = \sum_{i=1}^{K} a(\hat{x}, x_i) y_i \tag{3}$$

Where $\hat{y}$ is the prediction for a new data point $\hat{x}$, K represents the number of examples in the support set, x_i are the data points in the support set, y_i are the corresponding labels, and $a(\hat{x}, x_i)$ is the attention mechanism that computes the similarity between $\hat{x}$ and x_i.

3.2 Dataset

Table 1: Meta-Album [14] Datasets Across Domains

Domain	Dataset 1	Dataset 2	Dataset 3
Large Animals	Birds	Dogs	Animals with Attributes
Small Animals	Plankton	Insects 2	Insects
Plants	Flowers	PlantNet	Fungi
Plant Diseases	Plant Village	Medicinal Leaf	PlantDoc
Microscopy	Bacteria	PanNuke	Subcel. Human Protein
Remote Sensing	RESISC	RSICB	RSD
Vehicles	Cars	Airplanes	Boats
Manufacturing	Textures	Textures DTD	Textures ALOT
Human Actions	73 Sports	Stanford 40 Actions	MPII Human Pose
OCR	Omniprint-MD-mix	Omniprint-MD-5-bis	Omniprint-MD-6

The present study will use the Meta-Album [14] dataset, detailed in table 1. This dataset contains 30 subsets across 10 domains, each with 3 unique datasets [14]. The experiment will go through main steps as shown in Figure 2.

3.3 Training and Evaluation

The models are trained using 70% of the data as the support set, with 15% allocated for validation and 15% for testing (query set). Performance is evaluated based on classification accuracy and F1-score.

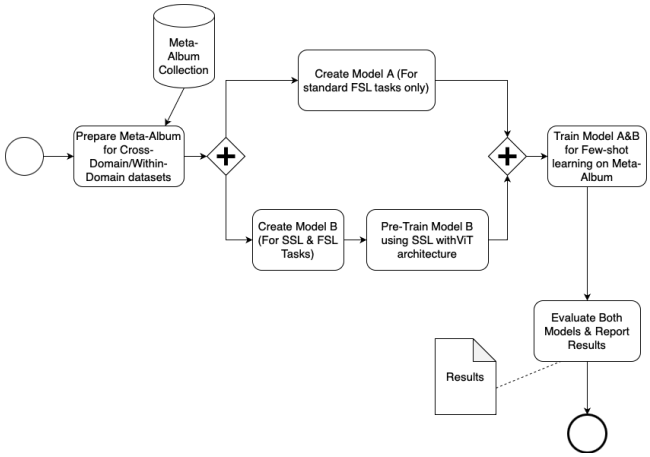

Figure 2: The Main Steps of The Experiment

Table 2: Comparison of "within domain" few-shot learning methods from Meta-Album by Ullah et al. (2022)[14]

Method	1-shot	5-shot	10-shot	20-shot
TrainFromScratch	32%	39%	41%	41%
Finetuning	45%	57%	59%	62%
MatchingNet	43%	55%	59%	63%
ProtoNet	51%	61%	64%	66%
FO-MAML	41%	50%	54%	57%
Vit-Based SSL & FSL(TBD) (Ours)	TBD	TBD	TBD	TBD

4 RESULTS

Upon concluding this experiment, we anticipate revealing the potential advantages of integrating self-supervised pre-training which employ ViT as the backbone architecture with few-shot classification on the Meta-Album dataset [14] as showen in Table 2.

5 DISCUSSION AND CONCLUSION

In this study, using the Meta-Album dataset [14], we aim to explore the benefits of integrating self-supervised pre-training which employ ViT as the backbone architecture with few-shot classification. By comparing Models A and B, we hope to provide insights and set a foundation for future research in this domain

REFERENCES

[1] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. 2018. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136* (2018).

[2] Ondrej Bohdal, Yinbing Tian, Yongshuo Zong, Ruchika Chavhan, Da Li, Henry Gouk, Li Guo, and Timothy Hospedales. 2023. Meta Omnium: A Benchmark for General-Purpose Learning-to-Learn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7693–7703.

[3] M Boudiaf, IM Ziko, J Rony, J Dolz, P Piantanida, and IB Ayed. [n. d.]. Transductive information maximization for few-shot learning. arXiv 2020. *arXiv preprint arXiv:2008.11297* ([n. d.]).

[4] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. 2019. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729* (2019).

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg

Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[6] Vincent Dumoulin, Neil Houlsby, Utku Evci, Xiaohua Zhai, Ross Goroshin, Sylvain Gelly, and Hugo Larochelle. 2021. Comparing transfer and meta learning approaches on a unified few-shot classification benchmark. *arXiv preprint arXiv:2104.02638* (2021).

[7] Pan Li, Shaogang Gong, Chengjie Wang, and Yanwei Fu. 2022. Ranking distance calibration for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9099–9108.

[8] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. 2019. Learning to self-train for semi-supervised few-shot classification. *Advances in neural information processing systems* 32 (2019).

[9] Yuning Lu, Liangjian Wen, Jianzhuang Liu, Yajing Liu, and Xinmei Tian. 2022. Self-supervision can be a good few-shot learner. In *European Conference on Computer Vision*. Springer, 740–758.

[10] Carlos Medina, Arnout Devos, and Matthias Grossglauser. 2020. Self-supervised prototypical transfer learning for few-shot classification. *arXiv preprint arXiv:2006.11325* (2020).

[11] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. 2018. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems* 31 (2018).

[12] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems* 30 (2017).

[13] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. 2019. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096* (2019).

[14] Ihsan Ullah, Dustin Carrión-Ojeda, Sergio Escalera, Isabelle Guyon, Mike Huisman, Felix Mohr, Jan N van Rijn, Haozhe Sun, Joaquin Vanschoren, and Phan Anh Vu. 2022. Meta-album: Multi-domain meta-dataset for few-shot image classification. *Advances in Neural Information Processing Systems* 35 (2022), 3232–3247.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[16] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems* 29 (2016).

[17] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. 2019. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867* (2019).