Vision Transformers for Galaxy Morphology Classification: Fine-Tuning Pre-Trained Networks vs. Training From Scratch

Rahul Kumar*, Md Kamruzzaman Sarker*†, Sheikh Rabiul Islam†

Department of Computing Sciences, University of Hartford, CT, USA
{rkumar,sarker,shislam}@hartford.edu

Abstract. In recent years, the Transformer-based deep learning architecture has become extremely popular for downstream tasks, especially within the field of Computer Vision. However, transformer models are very data-hungry, making them challenging to adopt in many applications where data is scarce. Using transfer learning techniques, we explore the classic Vision Transformer (ViT) and its ability to transfer features from the natural image domain to classify images in the galactic image domain. Using the weights of models trained on ImageNet (a popular benchmark dataset for Computer Vision), we compare the results of two distinct ViTs: one base ViT (without pre-training) and another fine-tuned ViT pre-trained on ImageNet. Our experiments on Galaxy10 dataset show that by using the pre-trained ViT model, we can get better accuracy compared to the ViT model built from scratch and do so with a faster training time. Experimental data further shows that the fine-tuned ViT model can achieve similar accuracy to the model built from scratch while using less training data.

Keywords: Deep Learning · Vision Transformer (ViT) · Transfer Learning · Pre-Trained · Image Classification

1 Introduction

The formation of galaxies has always been of particular interest to astronomers and astrophysicists throughout history. Known as galaxy morphology, examining galactic structures can help us understand the processes that go into star creation and indicate potential areas for the formation of life outside the Milky Way [1]. With the launch of the James Webb Space Telescope in December 2021, the observatory will allow us to understand the history of the universe through the use of infrared light. The James Webb Space Telescope can peer billions of years into the past to uncover the first stars and galaxies ever created by magnifying the infrared light with its 18 gold-plated mirrors while simultaneously blocking off the sun's heat rays using an expansive heat shield. Using this advanced technology, images taken of the oldest galaxies in the universe can offer valuable insight into the evolution of galaxies from the beginning of time to the present day. However, with the life expectancy of this mission being around five

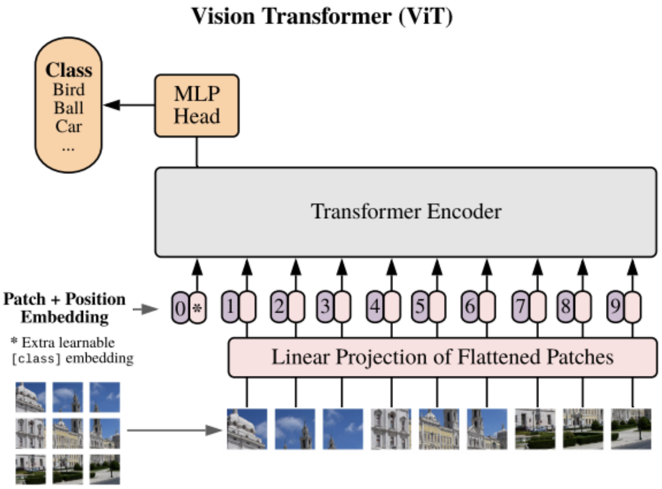

Fig. 1. Vision Transformer Architecture from Dosovitskiy et al.'s *An Image is Worth* 16×16 *Words: Transformers For Recognition At Scale*

to ten years due to fuel capacity, collecting image data on just galaxies is limited due to the endless reservations of astronomers needing the observatory for research purposes [2]. One solution to address this issue of limited data involves leveraging transfer learning. Transfer Learning is a subset of machine learning which uses pre-trained models to transfer knowledge from one domain to a new problem. We aim to determine whether we can improve the classification process of galaxies using transfer learning and the Vision Transformer architecture to save time and money on data collection. For the classifier model, we choose the Vision Transformer (ViT) [3] due to it's improved performance compared to the CNN model in recent years. In order to train our model to classify galaxies, we use the Galaxy10 dataset [7]. To evaluate the transfer capability of the ViT in galaxy classification, we devise two hypothesis: 1: Can we achieve a better accuracy by using a pre-trained model? 2: How much data does the pre-trained model need to obtain a similar performance to the model built from scratch?

Below, we outline the dataset along with how it is pre-processed, the steps that went into pre-training and fine-tuning the models, as well as the results of their performance when fine-tuned on galactic data. In Section 3, we discuss the data being used in this paper. Section 4 discusses the experimental approach and the methods for pre-training, fine-tuning, and calculating accuracy. In Section 5, we discuss the results from each of these experiments. Finally, the paper ends with a discussion and a brief conclusion in Section 6.

2 Related Work

2.1 Vision Transformer

We start the discussion with the introduction of Vision Transformer. Introduced by Dosovitskiy et al., the Vision Transformer (ViT) [3] serves as the backbone of this paper with its popularity in recent years. First, the ViT breaks up an input image into equal-sized patches. These patches are then "flattened" into number vectors and attached with a positional embedding, since the model can only accept arrays of numbers. The purpose of the positional encoding is to help the model understand the relative position of each patch within the image. Finally, the model calculates the likelihood of each class in the final layer and outputs the most likely label for the image. Architecture of the ViT is shown in figure 1. Despite Convolutional Neural Networks (CNN) being the de-facto model for imaging tasks, vision transformers have slightly outperformed CNNs due to two main features; the attention mechanism and positional encoding [4]. Attention in Deep Learning directs a machine to general locations within an image to extract important information about the contents of the image. As mentioned before, positional encodings teach the model about how the image patches combine and are a more efficient solution compared to passing the whole image into the model. This architecture will serve as the backbone for the experiments that are conducted and presented in this paper.

2.2 Transfer Learning

In the past, there has been similar work done concerning the classification of galaxy morphology using deep learning. Domínguez Sánchez et al. apply transfer learning for galaxy classification using three datasets: The Sloan Deep Sky Survey (SDSS), the Dark Energy Survey (DES), and the Dark Energy Camera Legacy Survey (DECaLS) [5]. The data contained images with one of three labels: Smooth/Disk, Edge-on, and Bar Sign. Consisting of three different types of hidden layers, Sanchez and her team created a standard deep neural network architecture to investigate the transferability of one image survey to another. To demonstrate the effects of transfer learning, models were trained and evaluated under unique and specific conditions. One model was trained on SDSS data and immediately transferred to DES data, another model was trained on SDSS data with some model fine-tuning, a model trained on SDSS data with added layers for DES data, and a model trained on DES data from scratch. Ultimately, the team achieved results of >90% accuracy using this strategy but required more than 300k prepared galaxy classifications in the process.

Tonkes et al. also investigates the application of transfer learning by comparing state-of-the-art CNN architectures with Vision Transformers in the artistic domain [6]. By pre-training these models on ImageNet, a very popular benchmark dataset consisting of images of common objects, Tonkes' evaluation examines how well these architectures perform on data outside of these common objects; in this case, art. Also, his work includes investigations into transformer

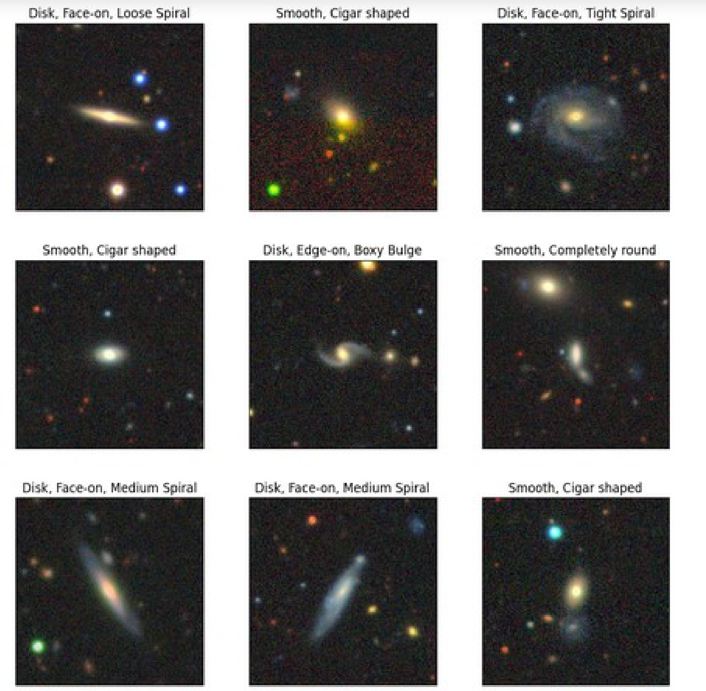
Disk, Face-on, Loose Spiral
Smooth, Cigar shaped
Disk, Face-on, Tight Spiral
Smooth, Cigar shaped
Disk, Edge-on, Boxy Bulge
Smooth, Completely round
Disk, Face-on, Medium Spiral
Disk, Face-on, Medium Spiral
Smooth, Cigar shaped

Fig. 2. One image from each of the nine class of the Galaxy10 dataset. Note that we used images from nine classes instead ten classes, as we found one class have significantly lower number of images.

architectures, unlike past research into this domain which has extensively explored transfer learning with CNNs. The dataset consisted of digitized artwork that falls into three categories: Type, Material, and Artist. To classify these images, eight different architectures (four CNNs, and four ViTs) were implemented to test how transferable each model is when extracting prominent features from images. Each of these models was tested with an on-the-shelf approach (OTS) and a fine-tuned approach (FT). Similar to Domínguez Sánchez et al., the OTS approach consisted of directly applying the model to a new dataset without any fine-tuning of the weights in the model while the FT approach altered the model to fit the new dataset. Ultimately, the Swin Transformer (a variant of the ViT) after fine-tuning provided the best results overall with an average of 92% accuracy for each art classification, followed closely by the ConvNext CNN Architecture. Our paper is similar to the work done in this paper but will be different in two ways. First, we use a different dataset (Galaxy10 dataset) from what Tonkes [6] used. Secondly, we evaluate the required dataset size where the pre-trained ViT model demonstrates similar results to that of the scratch ViT model using 100% of the dataset.

3 Dataset

The data used to fine-tune the ViT architectures was the Galaxy10 dataset, which consists of 17,736 images across 10 galaxy classifications [7]. This dataset is a subset of the Galaxy10 DECals and Galaxy Zoo datasets. The Galaxy Zoo Data Release 2 (GZ DR2) had volunteers classify $\sim 270k$ images from the Sloan Digital Sky Survey (SDSS) across 10 broad classes [8]. From there, the GZ DR2 combined with the DESI Legacy Imaging Surveys (DECals) to create the Galaxy10 DECals dataset. resulting in 441k unique galactic images in one dataset. For the purposes of this paper, we utilize a smaller version of the Galaxy10 DECals dataset, which contains $\sim 18k$ images across the same 10 galaxy classes as GZ DR2 [7]. Some example images from the dataset are displayed in figure 2. Due to the disproportionate size of the *Disk, Edge-on, Rounded Bulge* class with only 300 images, we delete this class from our current dataset. The modified version of this dataset can be seen in figure 3 below. Each image has a shape of 256 × 256 pixels and contains 3 color channels. The Galaxy10 dataset is publicly available online through the *astroNN* Python library.

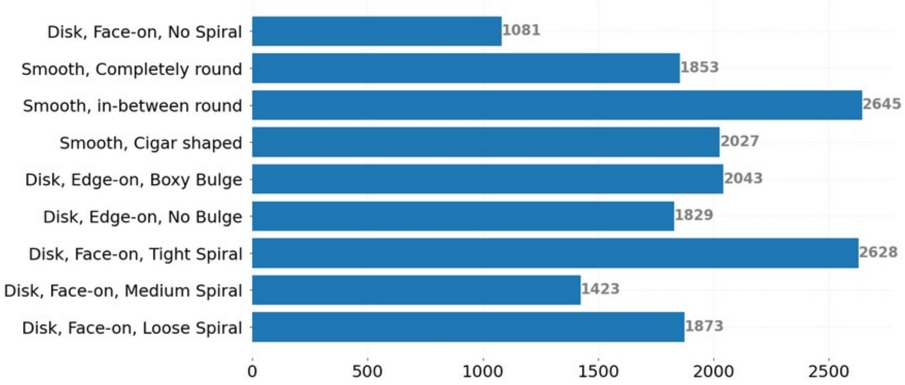

Fig. 3. Distribution of the modified Galaxy10 DECals dataset for each classification label.

3.1 Fine-Tuning Preparation

Using a pre-trained network and applying it to other domains has yielded better performing models than training from scratch. However, when fine-tuning a deep network, the optimal setup varies between applications [4]. For this paper, we create a data pipeline that performs various transforms on the images for the ViT models. These transforms include resizing images to 224 × 224, converting the image to a tensor, as well as normalizing them with a mean of $[0.485, 0.456, 0.406]$ and a standard deviation of $[0.229, 0.224, 0.225]$ [9]. Images are then loaded into training and validation Torch dataloaders with an 80/20 split and batch sizes of 32.

4 Experiment Details

Each trial was run using the Vision Transformer architecture as provided in the Torchvision [11] package in PyTorch [10]. Utilizing a pre-trained network, we then analyze the effectiveness of the network when using it just as a feature extractor versus training the same model from scratch. We also change the fine-tuning dataset to determine the dataset size where the pre-trained model yields similar results to the model with no initialized weights.

4.1 Model Selection & Pre-training

Within the Torchvision PyTorch package, there are many models and pre-trained weights that are pubicly available [11]. Out of the various Vision Transformer variants, we chose the **vit_b_16** model since this is the exact architecture that was proposed in Dosovitskiy et al.'s *An Image is Worth* 16×16 *Words: Transformers For Recognition At Scale*. The *b* within the name refers to the "Base" version of the model and *16* refers to the 16×16 input patch size. With authenticity in mind, we chose the **ViT_B_16_Weights.IMAGENET1K_V1** weights to use for our pre-trained model because these weights were retrieved from a model trained exclusively on ImageNet. Specifically, the **vit_b_16** was run on ImageNet images for 300 epochs using the AdamW optimizer with a learning rate of 0.003. For this paper, we also use *AdamW* as our optimizer while using the *CrossEntropyLoss* loss function and a learning rate of 0.0001.

4.2 Pre-Trained ViT as a Feature Extractor

We use the pre-trained network as a feature extractor and only modify the classification layer. To accomplish this, the last layer in the Sequential block of the pre-trained ViTs is changed to have a different output shape. Specifically, the *out_features* within the Linear layer was edited from 1000 classes for ImageNet to 9 classes for the Galaxy10 dataset [12]. We also prevent the weights within every layer of the ViT from being updated except for the last Linear layer. Finally, we conduct nine different trials: one trial trained a ViT with no weights using 100% of the dataset while the other eight used pre-trained ViTs. Of these eight models, the dataset sizes that were used for fine-tuning were 40%, 50%, 60%, 70%, 80%, 90%, 95%, and 100%. We present our findings in Section 5.

4.3 Accuracy and Loss Calculation

We use the standard equations of calculating accuracy and loss throughout the trials. The accuracy calculations utilized the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The loss, L_δ, is defined as

$$L_\delta = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

where the number of classes, M, must be $M > 2$. Here, we have $M = 9$ since we have nine classes within the Galaxy10 dataset.

4.4 Hardware and Software

Both experiments were performed on a NVIDIA GeForce RTX 3060 TI GPU. To activate this GPU, PyTorch, an open-source machine learning framework [10], is used for all trials. The Vision Transformer architecture and its pre-trained weights were taken from the Torchvision library within PyTorch [11]. We use Matplotlib to visualize our results [13].

Table 1. Comparing the results of training from scratch (ViT reported in [3]) to classification after fine-tuning a pre-trained ViT with varying dataset sizes. The number in the trial name represents the percentage of the Galaxy10 dataset that the respective ViT was fine-tuned on. The best results are highlighted in bold. The colored cells represent the models trained on 100% of the dataset.

Scheme	Trial	Validation Accuracy	Validation Loss
Train from scratch	base_vit[3]	69.49%	0.9018
Fine-tuning	ft_vit_40	67.13%	0.92
Fine-tuning	ft_vit_50	69.60%	0.8878
Fine-tuning	ft_vit_60	71.28%	0.8541
Fine-tuning	ft_vit_70	71.75%	0.8129
Fine-tuning	ft_vit_80	73.33%	0.7735
Fine-tuning	ft_vit_90	73.33%	0.7797
Fine-tuning	ft_vit_95	75.20%	0.7548
Fine-tuning	ft_vit_100	**75.53%**	**0.7225**

5 Results

We now present the main findings of our study and report results for the eight trials as seen in Table 1. It is important to note that for both experiments, an early stopping approach was used to avoid the models from overfitting. To do this, the lowest loss value, *min_loss_value*, is set to the first value that the model outputs (oftentimes the largest value throughout the training). We also initialize a counter variable to 0, a threshold variable(δ), and a patience variable. These variables are explained further in the following paragraph.

After each epoch, *min_loss_value* is compared to the current epoch; if the current value is smaller, then *min_loss_value* would become the current value.

Table 2. Comparing the time duration for training from scratch versus fine-tuning a pre-trained ViT with varying dataset sizes. The colored cells represent the models trained on 100% of the dataset.

Trial	Total Epochs	Average Epoch Duration	Total Training Time
base_vit	26	302 sec.	131 min.
ft_vit_40	16	109 sec.	29 min.
ft_vit_50	23	123 sec.	47 min.
ft_vit_60	19	147 sec.	47 min.
ft_vit_70	22	161 sec.	59 min.
ft_vit_80	27	178 sec.	80 min.
ft_vit_90	42	190 sec.	133 min.
ft_vit_95	36	198 sec.	119 min.
ft_vit_100	50	206 sec.	172 min.

If the current epoch had a loss greater than $min\_loss\_value + \delta$, then we increment the counter. Once the counter is equivalent to the patience, then the model stops training. To acquire the best results without overfitting, we set $\delta = 0.03$ and $patience = 3$.

5.1 Training from Scratch vs. Fine-Tuning

In this section, we analyze the performance of both the ViT baseline as well as the ViT that was fine-tuned on 100% of the Galaxy10 dataset. When it comes to galaxy morphology classification, we can see that the baseline ViT and fine-tuned ViT achieved final accuracies of $\sim 69\%$ and $\sim 76\%$, respectively. After the first three epochs of this trial, the fine-tuned ViT (*ft_vit*) had an average accuracy increase of 5% compared to the baseline ViT (*base_vit*) as seen in figure 4. This result was not unexpected, as the pre-trained weights from ImageNet allowed *ft_vit* to recognize patterns within the galaxy images more accurately than training *base_vit* with no initialized weights. Additionally, *ft_vit* had an average training time of 201 seconds/epoch whereas *base_vit* trained at 302 seconds/epoch as seen in table 2, resulting in a 31% speed-up from the fine-tuned ViT to the base ViT. It is important to note that *base_vit* began to show signs of overfitting after 25 epochs while *ft_vit* began to overfit at 50 epochs. This experiment clearly confirms the benefits of transfer learning and its ability to increase accuracy, reduce training time, and delay overfitting.

5.2 Fine-tuning with Varying Dataset Sizes

Now that we have compared the fine-tuned ViT with the base ViT, we report our findings on the second part of this paper: evaluating the dataset size required for the pretrained model. While it is clear in figure 4 that pre-trained weights demonstrate a considerable boost in performance over training from scratch, this begs the question: At what dataset size does a fine-tuned model yield similar results to a baseline model using 100% of the dataset? We approach this problem

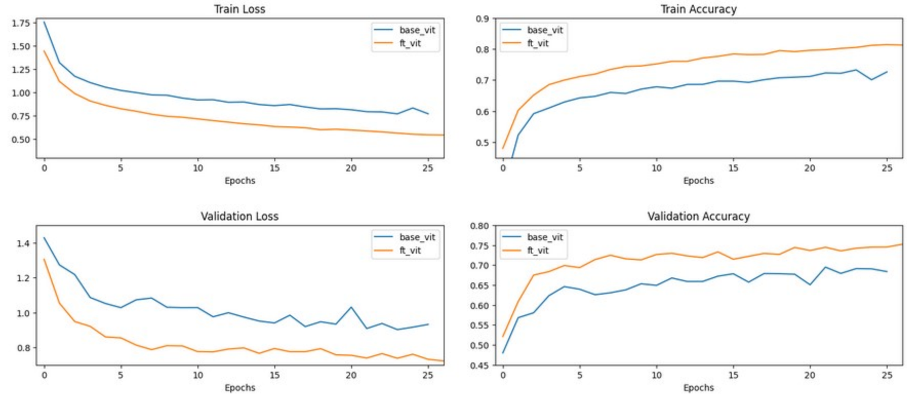

Fig. 4. Loss and accuracy curves of the fine-tuned model (*ft_vit* in orange) and the base model (*base_vit* in blue)

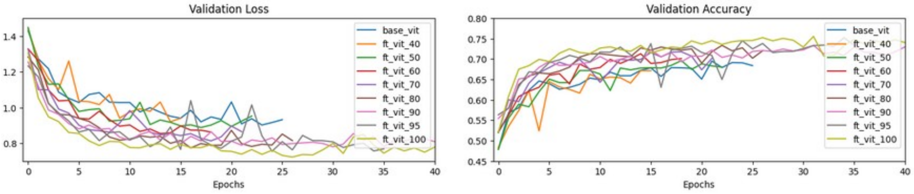

Fig. 5. The validation loss and accuracy curves obtained from both experiments. The number next to the model name represents the dataset size percentage that the model was fine-tuned on.

by creating seven additional trials to those proposed in Training from Scratch vs. Fine-Tuning.

To modify the dataset for fine-tuning, we created a function that implemented the train_test_split function from the Sci-Kit Learn library [14]. After splitting the original dataset by our desired dataset size, each ViT was trained using the early stopping technique described earlier. The results of these trials are shown in figure 5. Upon examination, it is clear that ViT_80 − 100 perform very well compared to their counterparts as they reach an accuracy of $\sim 72\% - 75\%$. The highest accuracies and lowest loss values for each trial in figure 5 can be found in figures 6 and 7 below.

Additionally, we can see that as the dataset size decreases, the overall training time decreases. This is clearly shown in table 2 where the trials with smaller dataset sizes have lower epochs and training times than those of higher dataset sizes. In response to the question proposed at the beginning of this section, we can conclude that a pre-trained model only requires around 50% of the images and 40% of the time used to train the baseline model in order to get similar results when classifying galaxy morphology on the Galaxy10 dataset using the Vision Transformer.

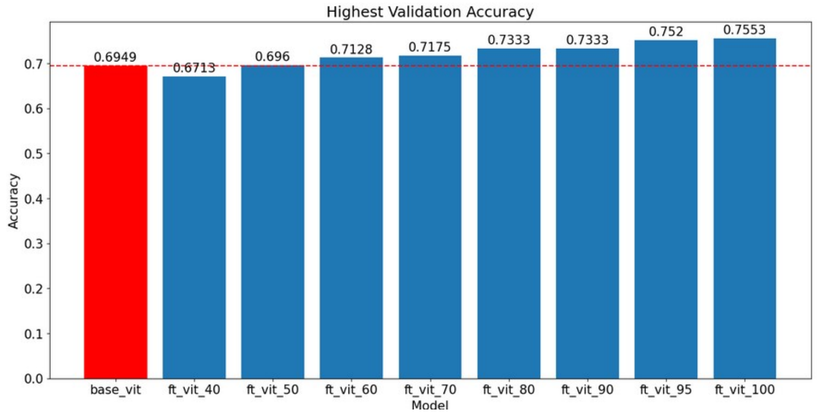

Fig. 6. Visualization of the highest validation accuracy for each trial. The red dotted-line represents the best accuracy from the base ViT with no pre-training.

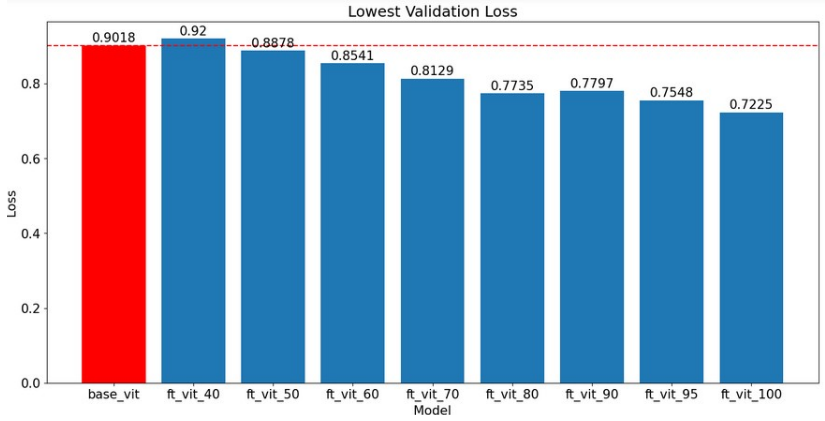

Fig. 7. Visualization of the lowest validation loss for each trial. The red dotted-line represents the best loss from the base ViT with no pre-training.

6 Conclusion

In this paper, we assess the transferability of the Vision Transformer (ViT) for galaxy morphology classification. We evaluate scratch and pre-trained versions of the ViT with two main experiments: comparing a fine-tuned ViT with a baseline

ViT using the same dataset size and comparing multiple pre-trained ViTs with varying dataset sizes against the baseline ViT. Based on the experimental results, we can see that the pre-trained and fine-tuned ViTs were able to outperform the scratch ViT until the dataset percentage was below 50%. Additionally, we can see the differences in training duration from scratch to pre-trained models. Specifically, we see a 32% decrease in training time when comparing *base_vit* and *ft_vit_100*, hence demonstrating the implications on time saved when using pre-trained models. When applying computer vision to problems within industry, a reduction of 50% in dataset size and 32% in training time allows for businesses to spend less money and time on data acquisition and preprocessing.

While we focused on the classic ViT architecture for this study, we aim to experiment with other models, such as CNN variants and the latest Transformer architectures like SWin Transformers (SWin) and Data-Efficient Image Transformers (DEiT), for the future. As the world shifts its focus towards space exploration, galaxy classification will allow us to better understand the development of distant galaxies. Specifically, we will gain more insight into the origins of the Universe based on a galaxy's shape and composition, hence solving questions that have gone unanswered for thousands of years. This research is a step in the direction for classifying galaxy morphology where data is scarce.

ACKNOWLEDGEMENTS

This research is conducted with the support of the University of Hartford through the Vincent Coffin Grant (ID:398131).

References

1. Buta, Ronald J. "Galaxy morphology." arXiv preprint arXiv:1102.0550 (2011).
2. Robertson, Brant E. "Galaxy formation and reionization: Key unknowns and expected breakthroughs by the James Webb space telescope." Annual Review of Astronomy and Astrophysics 60 (2022): 121-158.
3. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
4. Zhao, Yucheng, et al. "A battle of network structures: An empirical study of CNN, transformer, and mlp." arXiv preprint arXiv:2108.13002 (2021).
5. H Domínguez Sánchez, M Huertas-Company, M Bernardi, S Kaviraj, J L Fischer, T M C Abbott, F B Abdalla, J Annis, S Avila, D Brooks, E Buckley-Geer, A Carnero Rosell, M Carrasco Kind, J Carretero, C E Cunha, C B D'Andrea, L N da Costa, C Davis, J De Vicente, P Doel, A E Evrard, P Fosalba, J Frieman, J García-Bellido, E Gaztanaga, D W Gerdes, D Gruen, R A Gruendl, J Gschwend, G Gutierrez, W G Hartley, D L Hollowood, K Honscheid, B Hoyle, D J James, K Kuehn, N Kuropatkin, O Lahav, M A G Maia, M March, P Melchior, F Menanteau, R Miquel, B Nord, A A Plazas, E Sanchez, V Scarpine, R Schindler, M Schubnell, M Smith, R C Smith, M Soares-Santos, F Sobreira, E Suchyta, M E C Swanson, G Tarle, D Thomas, A R Walker, J Zuntz, Transfer learning for galaxy morphology from one survey to another, Monthly Notices of the Royal Astronomical Society, Volume 484, Issue 1, March 2019, Pages 93–100, https://doi.org/10.1093/mnras/sty3497.. 3009-3012.

6. Tonkes, Vincent, and Matthia Sabatelli. "How Well Do Vision Transformers (VTs) Transfer To The Non-Natural Image Domain? An Empirical Study Involving Art Classification." arXiv preprint arXiv:2208.04693 (2022).
7. Henry Leung, Jo Bovy (2020), Galaxy10 DECals. Retrieved from https://github.com/henrysky/Galaxy10
8. Walmsley, Mike, et al. "Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314,000 galaxies." Monthly Notices of the Royal Astronomical Society 509.3 (2022): 3966-3988.
9. Liu, Yahui, et al. "Efficient training of visual transformers with small datasets." Advances in Neural Information Processing Systems 34 (2021): 23818-23830.
10. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., . . . Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32 (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
11. Falbel D (2022). torchvision: Models, Datasets and Transformations for Images. Available at https://torchvision.mlverse.org and https://github.com/mlverse/torchvision.
12. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009 , pp. 248-255.
13. Hunter, J. D.. "Matplotlib: A 2D graphics environment". Computing In Science & Engineering 9. 3(2007): 90–95.
14. Lars Buitinck, et al. "API design for machine learning software: experiences from the scikit-learn project." ECML PKDD Workshop: Languages for Data Mining and Machine Learning. 2013.
15. Touvron, Hugo, et al. "Three things everyone should know about vision transformers." Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV. Cham: Springer Nature Switzerland, 2022.