Confronting Vision Transformer to limited training resources: a comparative study

Team members[A-Z] : Jiadi Yu, Othmane Sajid, Wenhao Xu, Yan Zhang, Yuyang Xiong

Abstract

The Transformer architecture and the self-attention mechanism they leverage have revolutionized Machine Learning, especially language models. These breakthroughs, first developed in the context of Natural language processing (NLP), quickly captured the attention of researchers, which in turn adapted this architecture to other problems like image classification.

In all cases, these models have in common their significant appetite for data and computing resources, which when they are sufficiently large, propel Transformer models to achieve state-of-the art performance on a wide range of problems. Vision Transformer (ViT) is no exception to this dynamic. While extremely powerful for image classification, this model is generally trained on large-scale datasets such as ImageNet and OpenImages with millions of labeled training examples.

This raises the important question of the challenges for researchers and practitioners with limited resources. To investigate this issue, we have compared ViT's performance with diverse CNN architectures, all trained in similar conditions with a constrained training budget of 50 epochs, on the Animals With Attributes 2 (AwA2) dataset which is a small dataset in this context (37K images). Moreover, pre-trained versions of these models were fine-tuned on the dataset.

The results show that ViT was unable to compete with most of the other CNN architectures when exclusively trained on AwA2 for 50 epochs. ViT achieved 57.6% accuracy, while ResNet reached 79.3%, and EfficientNet, an astounding 90.8%. However, when pre-trained on ImageNet and fine-tuned on AwA2, ViT was able to reach 96.8% and outperform the other pre-trained CNN architectures. These results testify to its large appetite for resources, which in exchange allow it to reach state-of-the art performance. The democratization of pre-trained models is thus all the more important for the average practitioner.

(Github repo link at the end of the report)

1. Introduction

"This report presents the results of a project aimed at evaluating the performance of Vision Transformer (ViT) models on a smaller dataset and with a limited training time. ViT models have gained a reputation for their remarkable performance on various computer vision tasks, but they require significant computational resources to train. The primary goal of this project was to investigate whether ViT models can still deliver competitive results under more constrained conditions."

Recent breakthroughs in deep learning have greatly enhanced the capabilities of computer vision, as numerous models display exceptional performance in areas like image classification, object detection, and segmentation. Among these models, Vision Transformers (ViTs) have captured considerable interest because of their capacity to attain cutting-edge outcomes across an array of visual tasks.

ViT is a class of models that leverage the *Transformer* architecture, originally designed for natural language processing, to process image data by dividing it into fixed-size non-overlapping patches and then linearly embed them into a sequence of vectors. ViT models replace the traditional convolutional layers used in computer vision tasks with self-attention mechanisms, which allow them to capture global dependencies between image patches. This approach has been shown to produce superior results compared to traditional convolutional neural networks (CNNs) on several image classification benchmarks.

Despite their impressive performance, ViT models are known for their resource-intensive training requirements. They often demand vast amounts of data and extensive training time to achieve optimal results, posing challenges for researchers and practitioners with limited resources. The primary goal of this project is to investigate the performance of ViT when trained on a smaller dataset, namely the Animals with Attributes 2 (AwA2) dataset, and within a constrained training budget of 50 epochs. By assessing their efficacy under these conditions, we aim to understand the adaptability and potential applicability of ViT in resource-constrained scenarios.

To provide a comprehensive evaluation, we will compare the performance of ViTs with other popular models in the computer vision domain: AlexNet, ResNet, EfficientNet and LeNet (which is relatively small compared to other models), when trained under similar constraints. Through this comparative study, we seek to answer the following research questions:

1. Can ViT models maintain their high-performance capabilities when trained on smaller datasets and for a constrained training time?
2. How is the performance of ViT models compared to other popular computer vision models, such as AlexNet, ResNet, and CNNs, when subjected to the same constraints?

Dataset (AwA2)

Animals with Attributes 2 (AwA2) is a dataset designed for the comparative evaluation of transfer learning algorithms, such as attribute-based classification and zero-shot learning. AwA2 is a direct replacement of the original Animals with Attributes (AwA) dataset, with more images released for each category. AwA2 also provides a category-attribute matrix, which contains an 85-dimensional attribute vector (e.g. color, stripes, fur, size, and habitat) for each category. For our study, the attributes were not used; the aim was predicting the label of the animal in the photo.

The dataset includes 37,322 images with 50 different categories of animals. One particularity of this dataset is that images tend to contain other elements (noise), and sometimes exclusively portray a certain part of the animal's body (like a whale's tale). This obviously makes the classification of these images harder, but also more realistic, since these photos show animals in a wide range of contexts and positions. The following example pictures from the dataset illustrate the challenges faced.

Figure 1 : Example images from AwA2 data set

Also, it should be noted that the number of labeled examples per category is inconsistent. Among the 37,322 images, some animal categories only have a small number of labeled examples, like the mole for instance with no more than 100. Most categories tend to have more or less 500-1000 examples.

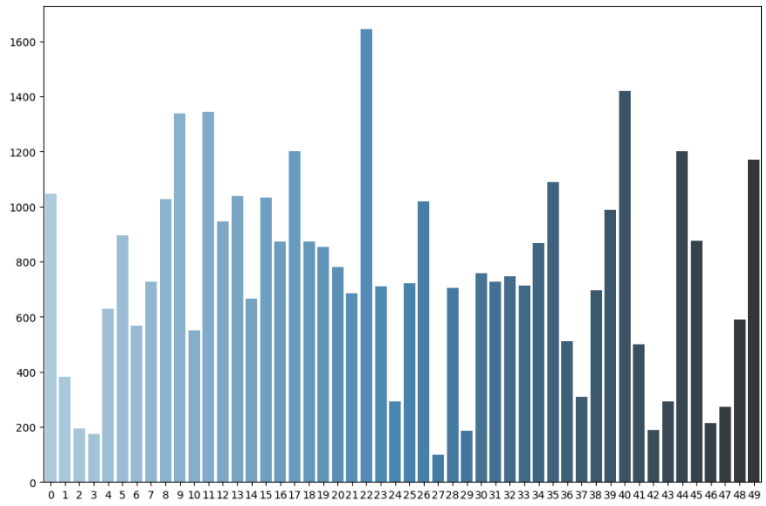


Figure 2 : Distribution of the number of labeled examples per animal category

2. Literature Review

Since 2017, when the Transformer architecture was introduced by Vaswani et al. in their now famous paper “Attention is all you need “ (Vaswani et al., 2017), many papers have been published to propose adaptations of this architecture for solving other types of problems. For instance, Dosovitskiy et al. have introduced in 2021 the Vision Transformer model in their paper titled "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale" (Dosovitskiy et al., 2021). Since then, various successful experiments have been performed with ViT. Almost all of them have used large-scale datasets and computing resources to achieve cutting-edge performance. For instance, the original ViT paper by Dosovitskiy et al. trained their model on ImageNet-1k dataset, which consists of 1.28 million images. Other high-profile studies have trained on even much larger datasets: Chen et al. (2020) trained on JFT-300M (300M images) and later on JFT-2.5B (2.5B images) in 2021.

To our knowledge, the challenge of training ViT on much smaller datasets has not been discussed sufficiently. These high-profile studies fail to address the challenges for researchers and practitioners with limited resources. And so, our comparative study goal is to contribute to filling this gap. The results obtained further highlight ViT’s appetite for data and computing resources and the need of addressing the issue of Transformer models’ use by organizations with a limited access to resources.

3. Methodology

In addition to ViT, four different CNN architectures were trained: LeNet, EfficientNet, AlexNet and Resnet. In total, there were three rounds of training. For each of model, we trained three different versions:

- One version without data augmentation;
- One version with data augmentation;
- One version pre-trained and fine-tuned on AwA2 dataset.

Also, as discussed, one important factor in this comparative study is to reproduce the same training conditions as to allow comparability. And so, all models were trained under a constrained training budget of 50 epochs maximum. They were trained in similar hardware, namely on the Google Colab Pro+ cloud servers (GPU Nvidia Tesla T4/A100-SXM4-40GB and RTX3090).

It should also be noted that images were reshaped into 224x224 pixels and normalized with the mean and standard deviation of the dataset. Data augmentation techniques were applied in the second round of training for all models (Random rotations and flips, color jittering and random cropping). In the third and final round of training, pre-trained versions of the models were fine-tuned on the dataset.

4. Results

At first, a Vanilla version of ViT was trained on the dataset for 50 epochs. It was subject to significant overfitting. The model's high capacity immediately translated into overfitting on the training examples and a lack of generalization to the unseen test data, which is reflected by its poor accuracy of merely 18.7% and by text-book overfitted curves, as illustrated in Figure 3.

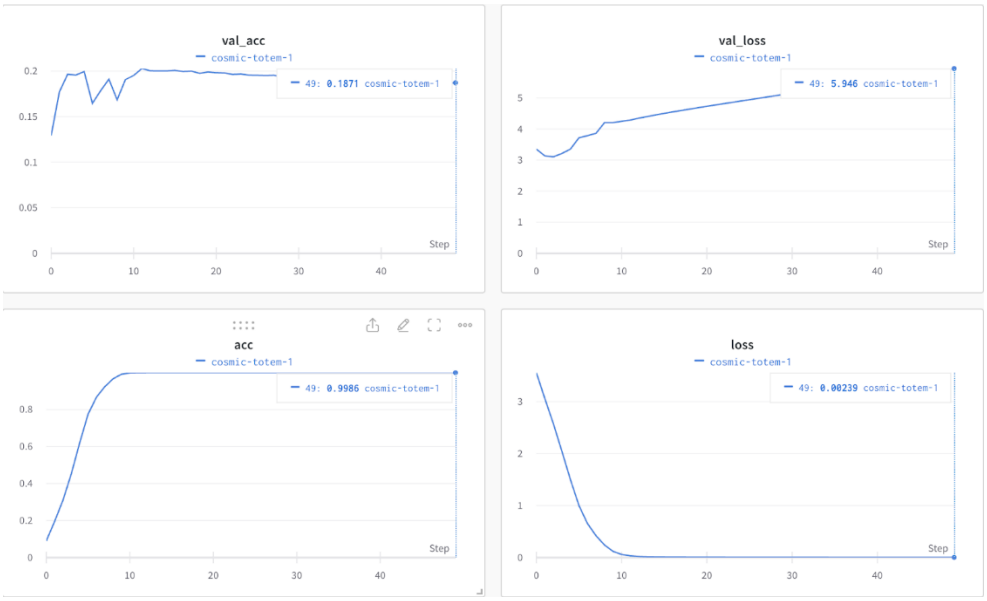


Figure 3. Learning curves of ViT without Regularization (overfitting)

And so, we added regularization techniques to remedy overfitting. In addition to dropout and weight decay, Data augmentation techniques were introduced to ViT's training. They were also applied to the other models for comparison purposes in this second round of model fitting. These techniques proved extremely effective, as ViT was no longer overfitting (see figure 3) and saw its test accuracy improve by almost 40 percentage points, from 18.7% to 57.6%.

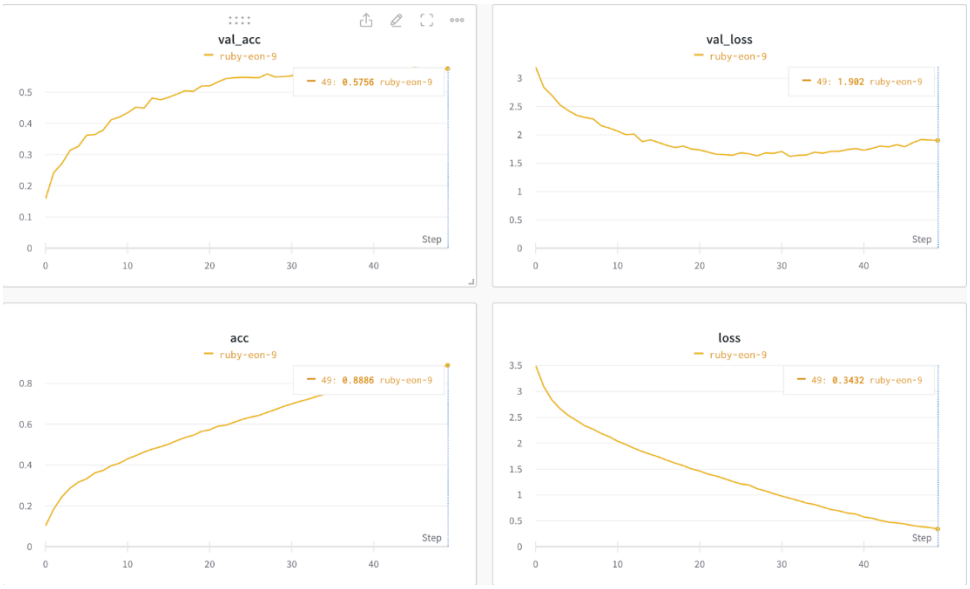


Figure 3. Learning curves of ViT with Regularization (no more overfitting)

The other models also exhibited vastly improved test accuracies thanks to data augmentation, with their respective accuracies showing gains ranging from 8.9 points for the lowest improvement (AlexNet) to 25.9 points for the highest (LeNet).

In fine, compared to the other models, ViT with its 57.6% accuracy achieved mid-table performance, beating AlexNet (38.5%) and LeNet (51%), but was still far behind ResNet (79.3%) and EfficientNet (90.8%). The latter proved to be particularly efficient for this dataset. The following tables show the respective accuracies (on test data) after 50 epochs for each version of each model, as well as their Precision, Recall and F1 score.

Model	**Not pre-trained and w/o data augmentation (Accuracy %)**	**Not pre-trained and with data augmentation (Accuracy %)**	**Pre-trained and fine-tuned with data augmentation (Accuracy %)**
LeNet	25.08	50.96	**N/A**
EfficientNet	67.01	90.82	96.26
AlexNet	29.62	38.48	75.37
ResNet	65.26	79.28	85.48
ViT	***18.71***	***57.56***	***96.76***

Table 1. Accuracy on test data for each version of each model (after 50 epochs)

Model (not pre-trained)	**Accuracy %**	**F1-score (macro avg)**	**Precision (macro avg)**	**Recall (macro avg)**
Without data augmentation				
LeNet	25.08	0.21	0.25	0.21
EfficientNet	67.01	0.61	0.67	0.61
AlexNet	29	0.23	0.23	0.23

ResNet	65.26	0.57	0.59	0.57
ViT	***18.71***	***0.16***	***0.17***	***0.16***
With data augmentation				
LeNet	50.96	0.49	0.51	0.51
EfficientNet	90.82	0.88	0.91	0.87
AlexNet	38	0.30	0.36	0.31
ResNet	79.28	0.73	0.74	0.73
ViT	***57.56***	***0.52***	***0.56***	***0.53***

Table 2. Evaluation metrics of non pre-trained models (after 50 epochs)

Model (pre-trained)	**Accuracy (%)**	**F1-score (macro avg)**	**Precision (macro avg)**	**Recall (macro avg)**
With data augmentation				
EfficientNet	96.26	0.95	0.96	0.95
AlexNet	**N/A**	**N/A**	**N/A**	**N/A**
ResNet	85.48	0.81	0.82	0.80
ViT	***96.76***	***0.95***	***0.96***	***0.95***

Table 3. Evaluation metrics of pre-trained models (after 50 epochs)

The 57.6% mid-table accuracy of ViT can be explained among other things by its lack of inductive bias, absence of translation invariance, and inability to capture local features of more efficient convolutional neural networks like EfficientNet. These lacks are usually compensated with more training resources, as ViT is a very high-capacity model and is a lot slower in its learning curve so to say. ViT clearly suffered here from the constrained resources used in its training.

Moreover, this performance of ViT can be further explained by its inability to clearly differentiate some animals with similar features. For instance, Tiger Vs Leopard, Antelope Vs Deer, Dolphin Vs Whale, Moose vs Buffalo and Otter Vs Seal among others resulted in frequent misclassifications of test examples. It should also be noted that some other notable misclassifications were more surprising from the perspective of the Human eye (e.g., German Shepherd Vs Chihuahua, Cow Vs Horse, Sheep Vs Horse, Grizzly Bear Vs Lion, etc.). Some of these might be explained in part by the inconsistent number of labeled examples per animal category. Figure 4 shows the confusion matrix of ViT classification of test examples.

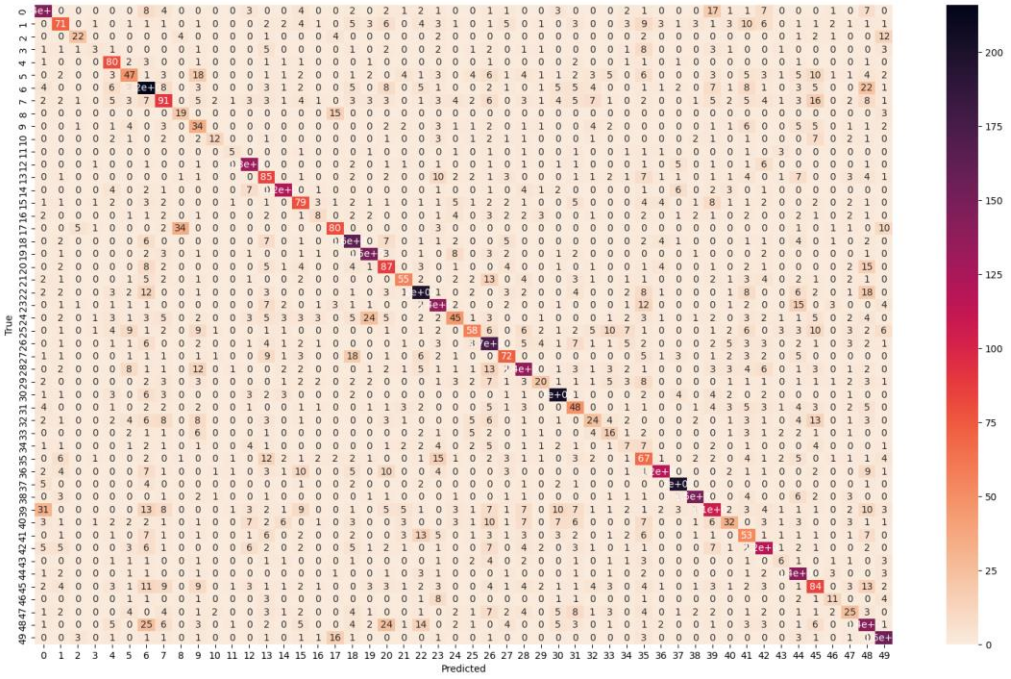


Figure 4. Confusion matrix of test data classification (ViT, 57.6% accuracy, F1 avg=0.52)

A third and final round of training involved this time pre-trained versions of each model. As seen in table 1 although ViT had achieved at first only mid-table performance, when it was this time pre-trained on ImageNet and fine-tuned for this dataset, it was able to achieve the best performance among all models, with 96.8% accuracy (even though EfficientNet pre-trained was close behind).

To briefly summarize what the results entail for our research questions, it appears that:

(1) ViT models fail to maintain their high-performance capabilities when trained on smaller datasets and for a constrained training time.

(2) Compared to the four CNN architectures trained in this study, ViT without pre-training only managed to achieve mid-table performance (i.e., it beat two models out of four, but was far behind EfficientNet).

However, when pre-trained and fine-tuned, ViT was able to outperform all the other models and thus reclaim its status of state-of-the art classifier and demonstrated once again the impressive results it can attain when training resources are not an issue. Indirectly, our study also highlighted the importance of regularization, as shown by the significant accuracy gains for all models after data augmentation was introduced in the training pipelines. The latter should indeed always be part of the routine of any practitioner working with limited training data and especially for high-capacity models.

5. Discussion

The iteration of technology is lightning fast. Even with a small dataset that is AwA2, the performance of ViT (after regularization and not pretrained) is better than that of older architectures like LeNet and AlexNet under similar training conditions. However, as we highlight throughout this comparative study, without pre-training, ViT is likely to exhibit lower performance than other more performant CNN architectures, as it was not able to maintain its high-performance capabilities when trained on smaller datasets and for a constrained training time due to its lack of inductive bias.

Indeed, good-performing CNN models demonstrate significant translation invariance and local feature capture due to the shared local perceptual field and weights of the convolution and pooling layers. This allows CNNs to perform very well in image recognition tasks, generally without the need for large-scale amounts of training data. This is an undeniable advantage they hold over Transformer models like ViT, and thus, CNNs remain more accessible to ML practitioners with scarce resources. For instance, in this study, EfficientNet exhibited an exceptional performance compared to the other models. It was by far the highest performing model in the training round without pre-training, as it reached an astounding 90.8% accuracy, followed by ResNet in second position (79.3%) and ViT in third (56.6%). Even with pretraining, where ViT managed to climb to the first position with 96.7% accuracy, EfficientNet was close behind with 96.3%, a difference that could be summarized to a rounding error. As its name suggests, EfficientNet is able to use fewer parameters and a lot less computation than most other comparable networks, a feature achieved thanks to its coordinated *compound scaling* of depth, width and resolution of the network.

In light of the results of this study, we remain of the opinion that the days in which CNN architectures will be abandoned in favor of Transformer models is not anytime soon and might in fact never come. Even though Transformer models like ViT are the hype of the moment, their vast appetite for resources remains a non-negligible deterrent for ML practitioners who are not part of organizations with virtually unlimited resources and CNNs like EfficientNet continue to offer them a robust and viable alternative. Also, this study has the merit of further highlighting that the importance of the democratization of ML models cannot be understated. Indeed, with the advent of powerful models with exponentially growing capacity (e.g., Chat-GPT iterations), it seems that the importance of computing power in the field of AI will continue to grow rapidly, and that average practitioners are at risk of becoming obsolete. In this context, open-sourcing pre-trained models allows the aforementioned ML practitioners to tackle more challenging problems despite their limited resources.

6. Annex

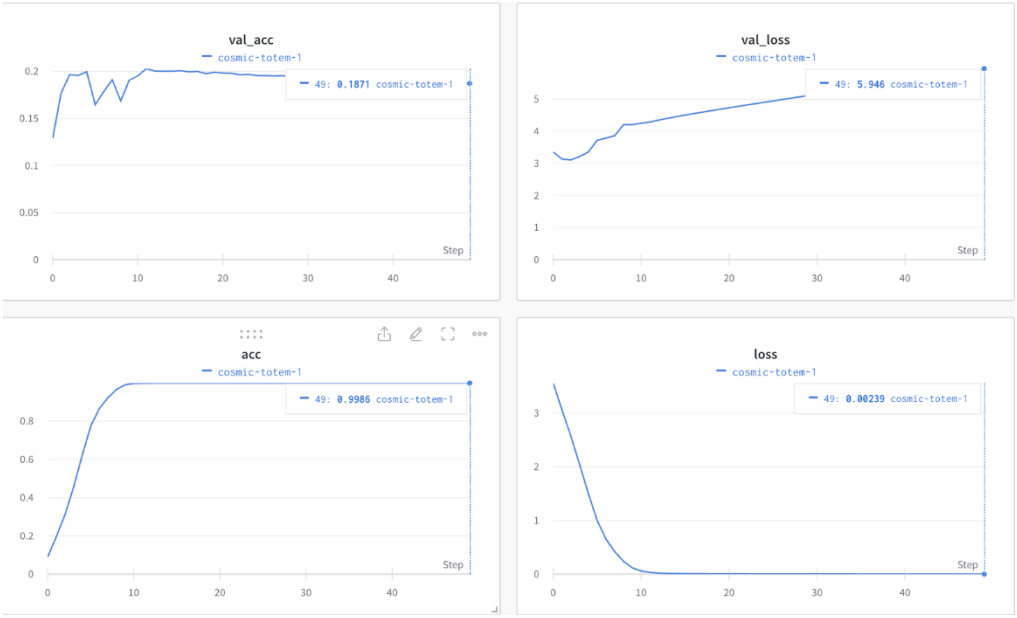


Figure 5. Non pre-trained ViT model without Data Augmentation

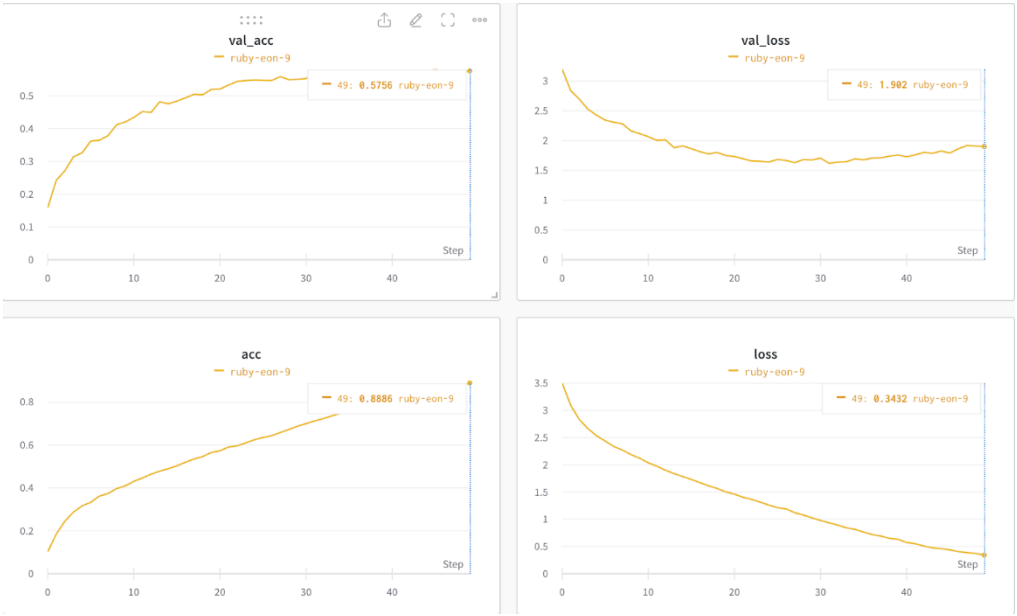


Figure 6. Non pre-trained ViT model with Data Augmentation

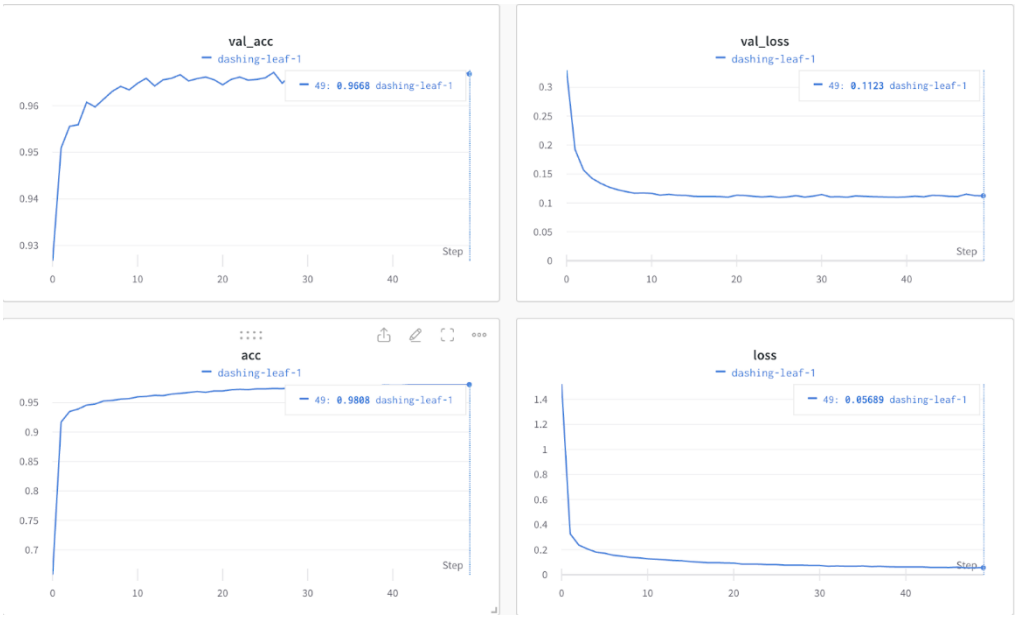


Figure 7. Pre-trained ViT model with Data Augmentation

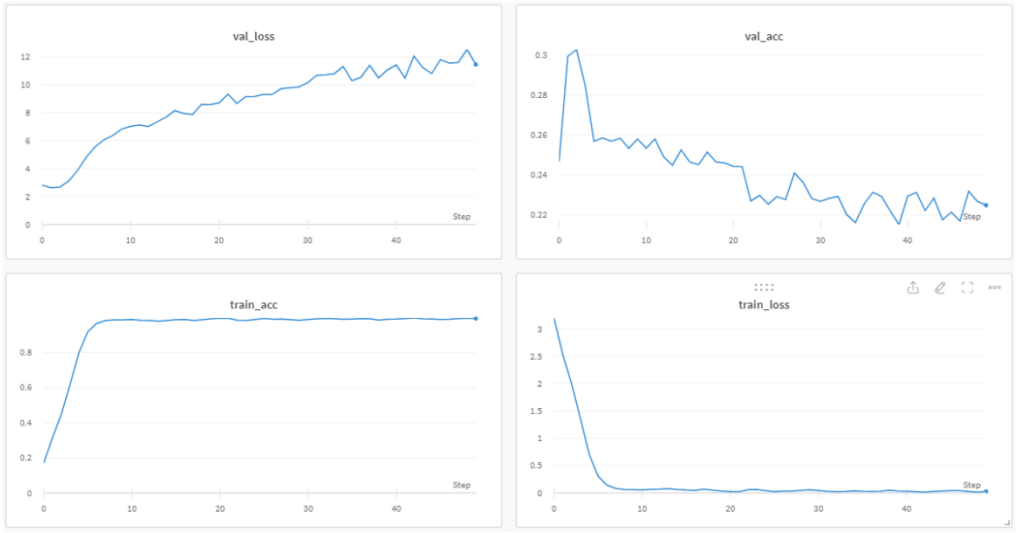


Figure 8.. LeNet (not pre-trained) model without Data Augmentation

Figure 9. LeNet (not pre-trained) model with Data Augmentation

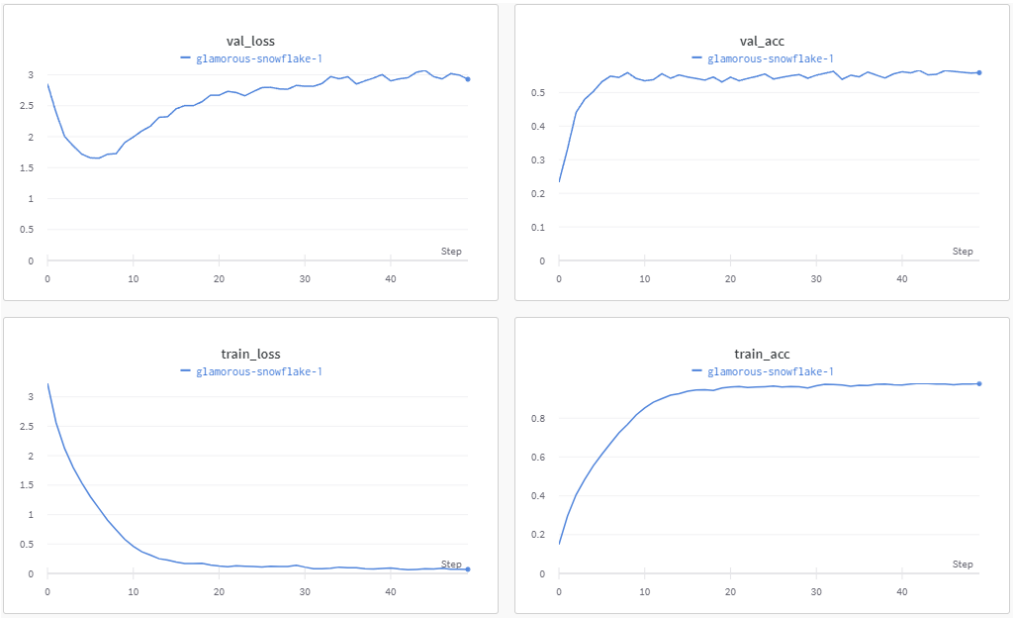


Figure 10. EfficientNet (not pre-trained) model without Data Augmentation

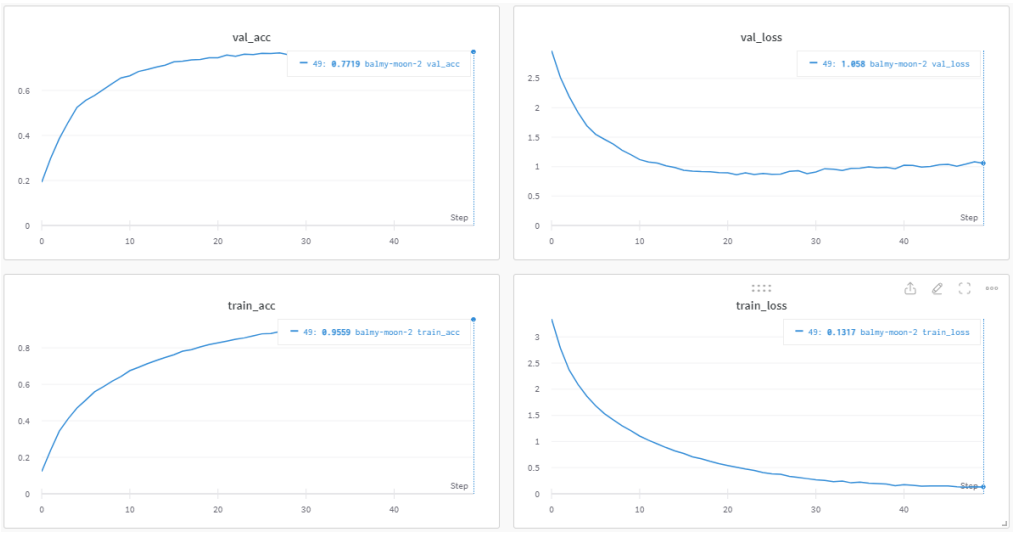


Figure 11. EfficientNet (not-pretrained) model without Data Augmentation

Figure 12. Pre-trained EfficientNet model with Data Augmentation

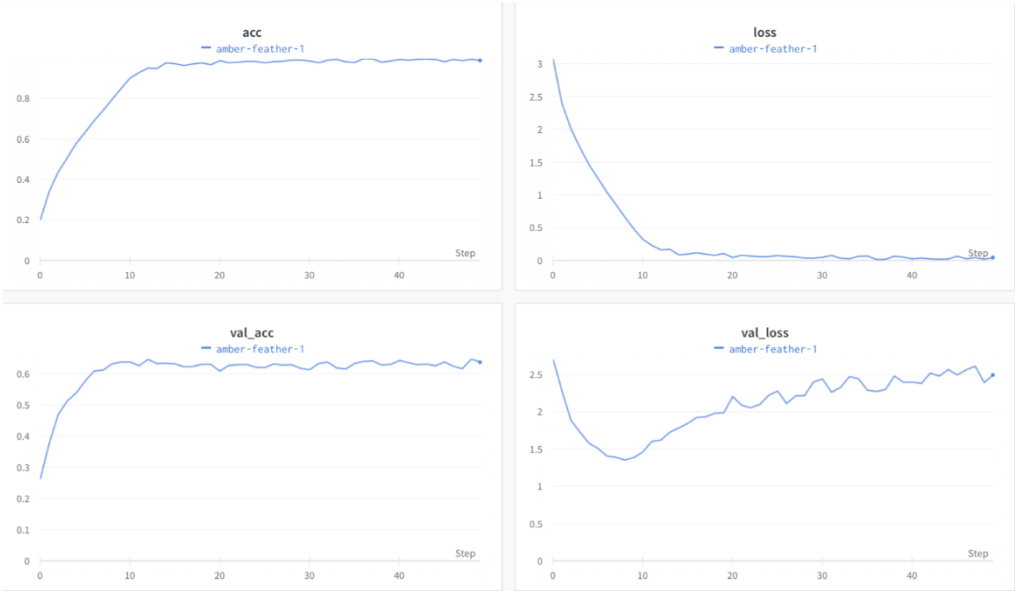


Figure 13. Non pre-trained ResNet model without Data Augmentation

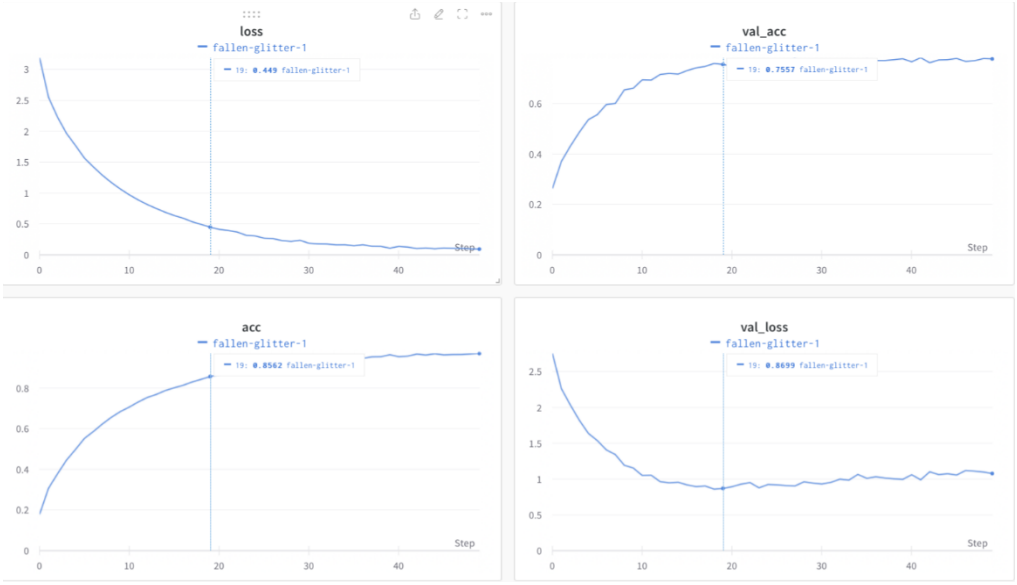


Figure 14. Non pre-trained ResNet model with Data Augmentation

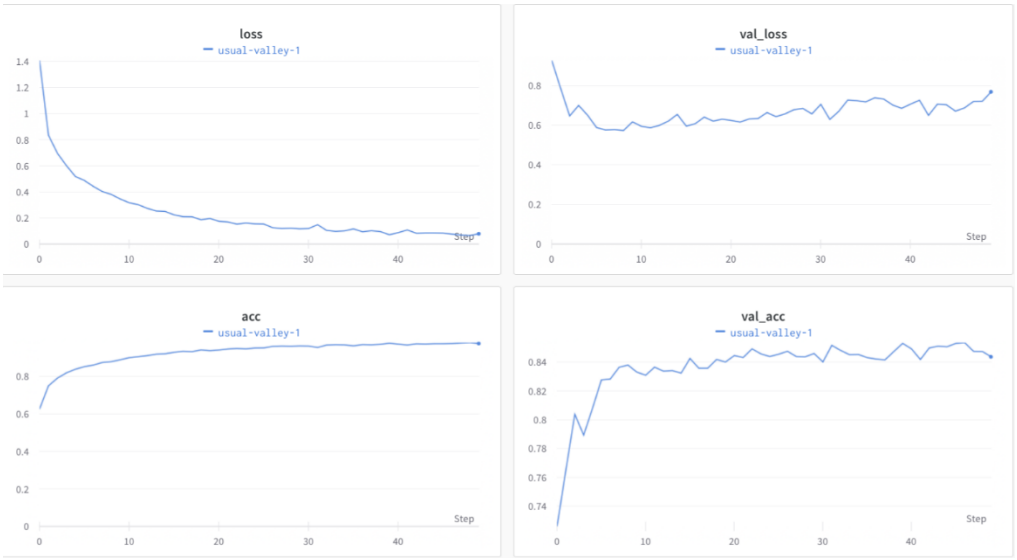
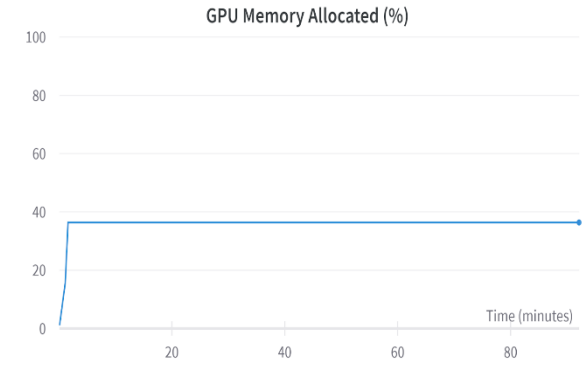
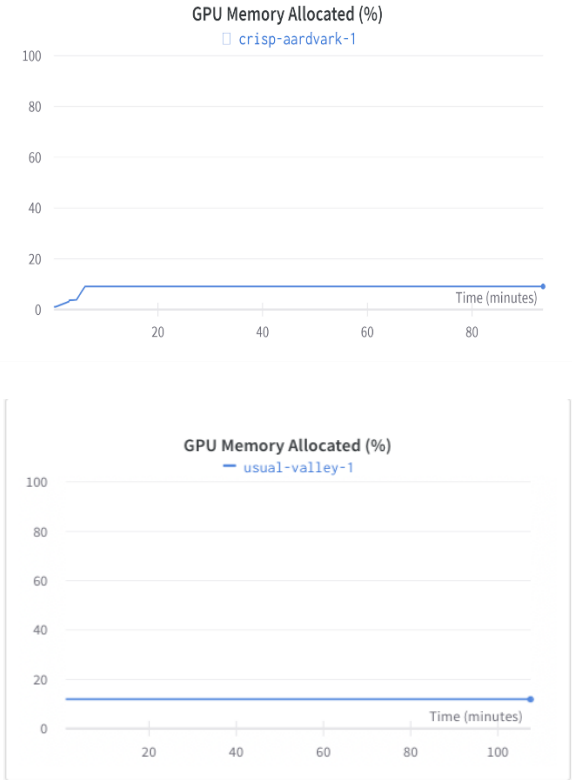


Figure 15. pre-trained ResNet model with Data Augmentation





- Fig 16.1, 16.2, 16.3, 16.4 GPU Memory Allocated rate of ViT, EffNet, LeNet and ResNet

7. References

papers

- Dosovitskiy, Alexey et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” *ArXiv* abs/2010.11929 (2020): n. pag.
- Vaswani, Ashish et al. “Attention is All you Need.” *ArXiv* abs/1706.03762 (2017): n. pag.
- Alamri, Faisal and Anjan Dutta. “Multi-Head Self-Attention via Vision Transformer for Zero-Shot Learning.” *ArXiv* abs/2108.00045 (2021): n. pag.
- He, Kaiming et al. “Deep Residual Learning for Image Recognition.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 770-778.

blogs

- Nikolas Adaloglou. “How Transformers work in deep learning and NLP: an intuitive introduction.” https://theaisummer.com/transformer/
- Nikolas Adaloglou. “How the Vision Transformer (ViT) works in 10 minutes: an image is worth 16x16 words.” https://theaisummer.com/vision-transformer/
- Konstantinos Poulinakis. “Are Transformers replacing CNNs in Object Detection?” https://www.picsellia.com/post/are-transformers-replacing-cnns-in-object-detection
- Francesco Zuppichini. “Implementing Vision Transformer (ViT) in PyTorch.” https://towardsdatascience.com/implementing-visualttransformer-in-pytorch-184f9f16f632
- Shipra Saxena. “Introduction to The Architecture of Alexnet.” https://www.analyticsvidhya.com/blog/2021/03/introduction-to-the-architecture-of-alexnet/
- Arjun Sarkar. “Understanding EfficientNet — The most powerful CNN architecture.” https://medium.com/mlearning-ai/understanding-efficientnet-the-most-powerful-cnn-architecture-eaeb40386fad
- EDUCBA. “PyTorch DataLoader.” https://www.educba.com/pytorch-dataloader/
- Vihar Kurama. “A Comprehensive Guide to the DataLoader Class and Abstractions in PyTorch.” https://blog.paperspace.com/dataloaders-abstractions-pytorch/
- Afshine Amidi and Shervine Amidi. “A detailed example of how to generate your data in parallel with PyTorch.”

https://stanford.edu/~shervine/blog/pytorch-how-to-generate-data-parallel

- PyTorch. “DATASETS & DATALOADERS.” https://pytorch.org/tutorials/beginner/basics/data_tutorial.html
- Andy Lo. “Weight Decay and Its Peculiar Effects.” https://towardsdatascience.com/weight-decay-and-its-peculiar-effects-66e0aee3e7b8

8. Contributions of the team members[A-Z]

Jiadi Yu: Trained and analysed LeNet and EfficientNet models
Tensorised labels in Dataloader

Othmane Sajid : Dataloaders
Development of ViT model
Helped other members debug
Optimized regularization techniques
Coordination

Wenhao Xu: Research and understanding of ViT
Trained and tested ViT model
Analyzed ViT model results
Fine-tuned pre-trained ViT model on target dataset
Help other members debug

Yan Zhang: Construction of AlexNet
Help other members debug

Yuyang Xiong: Trained and analyzed ResNet models
Helped with training AlexNet models

We also would like to sincerely thank our teacher Alex Hernandez-Garcia and our mentor Victor Schmidt for their availability and precious help in improving this research throughout the semester.

9. Github repo link

https://github.com/K-kiron/animal-detect

Please note that our primary development environment was Google Colab. And so, our code files for the different models are in the format of notebooks (ipynb). This has the advantage of allowing the reader to appreciate (sequentially) the output of the code snippets displayed in the notebooks without the need of running them himself or herself.