EECS 442 - Final Report on ViT Implementation

John Glenn
University of Michigan
johgle@umich.edu

Zhuoyi Ai
University of Michigan
aizoey@umich.edu

Chenyi Lin
University of Michigan
chyilin@umich.edu

Abstract

This project presents a comprehensive study of the Vision Transformer (ViT), a novel approach to image classification that leverages transformer architectures traditionally used in natural language processing. The motivation for this work stems from the limitations of convolutional neural networks (CNNs), which, while effective, do not inherently understand the global context of an image. ViT, on the other hand, treats an image as a sequence of patches, allowing it to capture both local and global image features.

1. Introduction

The field of image classification has been dominated by convolutional neural networks (CNNs) for years. However, these models have inherent limitations, particularly in their ability to understand the global context of an image. This limitation motivates our exploration of a novel architecture for image classification: the Vision Transformer (ViT).

ViT applies the transformer architecture, which has been highly successful in natural language processing (NLP), to image classification. The inspiration for this approach comes from the impressive scaling of transformers in NLP. By treating an image as a sequence of patches, much like a sentence is treated as a sequence of words in NLP, ViT is able to capture both local and global image features.

Our contributions in this work are twofold. First, we introduce the application of transformer models to image classification, a significant departure from the traditional CNN-based approaches. Second, we provide a comprehensive study of ViT, including its architecture, training, and performance on standard image classification benchmarks. Our findings demonstrate the potential of ViT as a powerful tool for image classification tasks.

2. Related Work

The application of transformer models to image classification is a relatively new area of research, with several key works contributing to its development.

Section 1: Related Ideas

The original transformer model was introduced by Vaswani et al. (2017) for natural language processing tasks. The idea of applying transformers to image processing was later explored, with models like the one proposed by Cordonnier et al. (2020) using patches of images as input to the transformer. The current Vision Transformer (ViT) model builds upon these ideas, with the motivation to explore image recognition at larger scales. This involves large-scale pre-training and combining convolutional neural networks (CNNs) with self-attention mechanisms.

Recent developments include the Image GPT (iGPT), which achieved a maximal accuracy of 72% on ImageNet, and the Swin Transformer, which uses a hierarchical structure that is more suitable for processing data with a strong temporal relationship.

Section 2: Comparison of Other Models

Compared to CNN-based methods, ViT offers advantages in scenarios where global dependencies and contextual understanding are crucial. While CNNs are effective at capturing local patterns in images and can handle large-scale datasets efficiently, they lack the ability to capture global context.

Recurrent Neural Network (RNN)-based methods, on the other hand, process data sequentially, meaning they can only process one element at a time. In contrast, ViT can process all elements at once, allowing it to capture long-range dependencies and global context better than RNNs. Furthermore, RNNs suffer from issues such as vanishing or exploding gradients when dealing with long sequences, a problem that ViT does not have.

3. Method

Our approach to implementing the Vision Transformer (ViT) for image classification involves several key steps:

Step 1: Patchifying and Linear Mapping

We begin by ”sequencifying” an image. This involves breaking the image into multiple sub-images or patches.

Each patch is then mapped to a vector using a linear transformation. This process effectively converts the 2D image into a 1D sequence of vectors, each representing a patch of the image.

Step 2: Adding the Classification Token

Next, we add a special token to our sequence of vectors. This classification token plays a crucial role in capturing information about the other tokens (patches) in the sequence.

Step 3: Positional Encoding

To ensure the model understands the original spatial location of each patch in the image, we add positional encodings to the patch vectors. This step allows the model to incorporate the relative positions of the patches during the transformation process.

Step 4: The Encoder Block

The sequence of patch vectors, along with the classification token, is then passed through an encoder block. This block applies layer normalization to the tokens, followed by a multi-head self-attention mechanism. A residual connection is also added to facilitate the flow of gradients during backpropagation.

Step 5: Classification

Finally, we extract just the classification token from our sequence of vectors. This token, which has been updated to capture information about all the patches in the image, is used to obtain the final classification output.

This method allows ViT to effectively capture both local and global features in an image, leading to improved performance on image classification tasks.

4. Experiments

We used the CIFAR-10 datasset for our experiments. CIFAR-10 is a balanced dataset of 60,000 32x32 RGB images, 50,000 of which are training images and 10,000 of which are test images. There are 10 data classes consisting of: birds, airplanes, automobiles, cats, deer, dogs, frogs, horses, ships, and trucks. The data set is also fully balanced, with 6000 images per class.

Our task is image classification, which allows for a simple method to evaluate the model by running the model on the test set. Overall learning can be compared against the random guess starting accuracy of 10% and to a lesser extent, the accuracy atained with the model in the original Vision Transformer (Alexey Dosovitskiy et al.) paper, accounting for the differences in dataset size and compute power available.

Our hyper parameters include:

- Patch Size
- Number of Transformer Layers
- Number of Multi-Head Attention Heads
- Embedded Dimension Size
- MLP Dimensional Ratio with Embedded Dimension

We ran many different experiments with different sets of hyper parameters. These experiments will be categorized in the table.

P-Size	Layers	Heads	Embed	MLP	Epochs	LR	Training Acc	Test Acc
16	8	4	256	4	30	1e-04	90.00%	50.00%
16	4	2	128	4	30	1e-04	62.00%	52.00%
16	4	2	128	4	40	1e-04	70.00%	52.50%
8	4	2	128	4	30	1e-04	62.00%	53.00%
8	6	4	128	4	30	1e-04	77.00%	52.00%
4	16	8	1024	4	40	1e-04	78.30%	55.40%
4	16	8	1024	4	100	1e-04	99.00%	52.86%
8	6	4	256	8	30	1e-04	92.47%	53.42%
4	8	8	512	8	30	1e-04	93.68%	**57.57**%

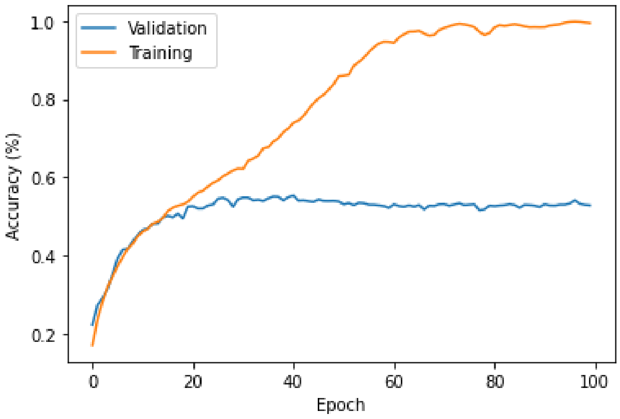
As it can be seen, most models ran in a range of 50-60% testing accuracy. This performance isn't great, but it also isn't unexpected. The Vision Transformer architecture lacks the inherent helpful inductive bias that CNN based model has with locality and translation. The positional embeddings work to try to minimize this but it doesn't amount to much for a model trained on such a small dataset, with limited computation resources.

The models in the original Vision Transformer paper performed quite well on CIFAR-10, but those models had a large number of parameters, as well as the benefit of being pretrained on very large datasets before being fine tuned. Their paper showed that being trained on large datasets allows the ViT architecture to perform well against other traditional CNN based models.

Our largest model with an embedding dimension of 1024 performed slightly above the average at its peak accuracy at 40 epochs, before losing accuracy as it continued training to 100 epochs.

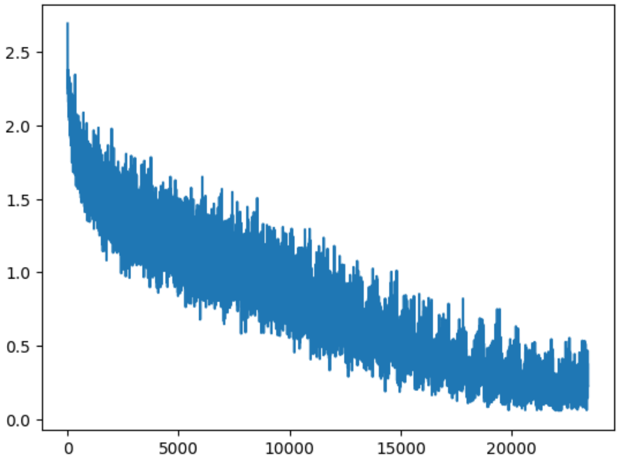
Our best performing model was a mid range architecture with 512 embedded dimensions, scoring two and a half percentage points above our largest model.

The model has a very hard time generalizing to the testing set and starts to stagnate on test set accuracy gains per epoch at around 50% test set accuracy. Which can be seen below

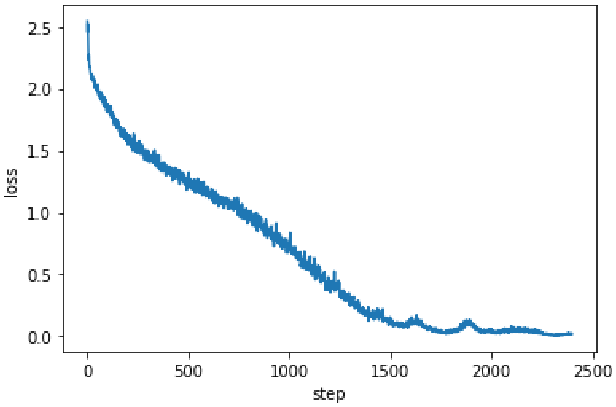


This pattern was present in every experiment, which is why even those in the table with a low training accuracy perform fairly well on the test set. The graph itself shows how it effected our largest model, which also trained for the longest period of time. It can be seen that despite stagnating, the model didn't experience much over fitting as training accuracy increased.

The loss curves all followed a general trend, excluding our worst models which had a fair bit of noise. The figure below shows the shape of our standard loss decay.



A slightly more interesting loss curve is the one for our largest model, which has experiences a bit of slow down around 500 steps.



Some ablation testing was done to try to see the effect of individual hyperparameters to see if there were any significant outliers. A table of those experiments can be seen below.

P-Size	Layers	Heads	Embed	MLP	Epochs	LR	Training Acc	Test Acc
32	1	1	32	1	30	1e-04	46.83%	44.22%
4	1	1	32	1	30	1e-04	38.77%	39.32%
8	1	1	32	1	30	1e-04	39.10%	38.59%
16	1	1	32	1	30	1e-04	41.48%	41.27%
16	1	2	32	1	30	1e-04	44.32%	44.48%
16	1	4	32	1	30	1e-04	45.80%	45.04%
16	4	4	32	1	30	1e-04	52.34%	49.77%
16	4	4	128	4	30	1e-05	44.90%	45.24%
16	4	4	128	2	30	1e-05	45.49%	45.53%

These experiments didn't show any major difference with any single hyper parameter being changed. It does however show that there is a minimum complexity of the model required to start to get the testing accuracy range of 50-60% we saw in the main table, and that without this minimum complexity, the model even struggles to gain train accuracy.

5. Review and Conclusion

What we learned, what we think was lacking, and possible solutions

Key Takeaways

The primary conclusion from this project is a deeper understanding of the Vision Transformer (ViT) model's capabilities and limitations. Despite the simplicity of our structure and the limited complexity of our model, we were able to gain valuable insights into the workings of ViT. However, our choice of a relatively smaller dataset and low-resolution images (32x32 pixels) presented a significant challenge. The model was unable to capture sufficient feature points on most patches, leading to a lack of clarity in the images.

Reproducibility and Comparison Against Baselines

In terms of reproducibility, the project demonstrated that implementing a ViT with limited resources is feasible. However, the quality of the output is heavily dependent on the size and resolution of the training dataset. Compared to baseline methods, our approach may have limitations due to the reduced model complexity and smaller dataset. However, it provides a foundation for understanding and applying ViT in more resource-intensive settings.

Future Work

With access to better hardware and more time, several improvements could be made. First, using a larger dataset with higher resolution images (up to 128x128 pixels) would allow the model to capture more feature points, improving the clarity of the output. Second, leveraging pre-trained models could provide a better starting point for training, potentially improving the model's performance. Finally, fine-tuning the model could further enhance its accuracy and adaptability to new data.

This project serves as an initial exploration into the application of Vision Transformers. The insights gained from this work can guide future research and development in this area, paving the way for more complex and robust implementations of Vision Transformers. It highlights the importance of dataset size and image resolution in model performance, providing a valuable lesson for future projects. With continued research and development, Vision Transformers have the potential to significantly advance the field of image recognition and related applications.

References

1 Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv (Cornell University).* https://arxiv.org/pdf/2010.11929

2 Pulfer, B. (2023, October 17). Vision Transformers from Scratch (PyTorch): A step-by-step guide. *Medium.* https://medium.com/mlearning-ai/vision-transformers-from-scratch-pytorch-a-step-by-step-guide-96c3313c2e0c

3 DeepFindr. (2023, July 4). *Vision Transformer Quick Guide - Theory and Code in (almost) 15 min* [Video]. YouTube. https://www.youtube.com/watch?v=j3VNqtJUoz0