Using the Transformer Model on Image Data

Anup Shakya
University of Memphis
ashakya@memphis.edu

Abstract

Pure transformer-based applications has been very limited in computer vision tasks. For the vision transformer to compete with the convolution-based models, it needs to be robust to commonly occurring noise and perturbations. We present a controlled study of the robustness of the Vision Transformer when subjected to different types of noise. Further, we also show how even the simplest data augmentation techniques can help the model generalize and perform better.

1 Introduction

The transformer model has been a breakthrough in NLP domain. It has become the go-to model for any NLP task like Language Modeling, Machine Translation, etc. The transformer model makes use of its parallel processing capability and thus requires comparatively lesser computation resource. This feature has widely been exploited in NLP. However, its use in Computer Vision tasks has been very limited. Convolution-based models are the dominant models for any Computer Vision task like: Image classification, segmentation, object detection, etc. The general consensus in the research community is that there is just too much reliance on the convolution-based models and we just have to live with its cons. The convolution-based models require more data, computation resource and time to train. Therefore, an alternative to the convolution-based methods is required.

In this work, we aim to study and examine the infamous Vision Transformer (ViT) model. This model uses pure transformer-based architecture for image classification task. The results of the ViT show promising prediction accuracy on many classic image classification datasets. However, to compete with the conolution-based models, ViT needs to be robust to any noise and perturbations added to the images which is very common in real-world scenario. Therefore in this work, we test the robustness of the ViT model by introducing different kinds of commonly occurring noise and corruptions to images. Furthermore, we also implement different data augmentation techniques and compare their effect in the overall performance of ViT models.

The challenge with using transformer model on image data is that the transformer models require sequence data as input. The original ViT paper (Dosovitskiy et al., 2020) solves this issue by dividing the image into patches of size n x n, and flattening it to obtain a data sequence. This is then linearly projected to embedding, followed by addition of positional encoding and an additional CLS label. In this work, We use the same technique with patch sizes 32×32 and 16×16.

When the ViT model is subjected to noisy data, the performance suffers a lot. The degree of performance loss depends on the type of noise added and the dataset itself. However, once the data augmentation techniques are applied, the ViT model generalizes better and has an improved performance for noisy data.

2 Related Work

The attention-based mechanism is very widely being followed by researchers today. The principles of the attention-based mechanism seem very intuitive that not all parts of data contain important information. Vaswani et al. (2017) introduced the Transformer with Scaled Dot-product attention and Multi-headed self attention, which became the state of the art method in many NLP tasks. The transformer based models mostly use transfer-learning where a large model is pre-trained on a large dataset and then fine-tuned to a specific task at hand. BERT (Devlin et al., 2018) uses denoising self-supervised pre-training where GPT uses lanaguage modeling as its pre-training task (Brown et al., 2020). There have been lots of work on using attention-based

mechanisms on image data. But, they only partially used the attention mechanism as a certain portion of the convolution layers. Cordonnier et al. (2019) introduced the first model that applied full self-attention on top. This model uses patches of size 2×2. Due to the small patch size, this model is usable only to small-dimension images.

There has also been interest research works done in combining CNNs with different forms of self-attention, e.g. by augmenting feature maps for image classification (Bello et al., 2019), by processing the outputs of a CNN using self-attention for tasks like: object detection (Hu et al., 2017; Carion et al., 2020), video processing (Wang et al., 2017), or image classification (Wu et al., 2020). Chen et al. (2020) proposed the image GPT (iGPT) which applies Transformer to image pixels after reducing the image resolution and color space. Dosovitskiy et al. (2020) presented the Vision Transformer (ViT) which is the first pure transformer based model that is applicable for larger dimension images as well. The results of the ViT challenges state-of-the-art benchmarks. Another work researches on the training parameters of the ViT (Steiner et al., 2021).

3 Methodology

The Vision Transformer requires the input to be a sequence. For this, we use the same approach used in the original ViT paper (Dosovitskiy et al., 2020) where an image is divided into patches. We experiment on patches of size 32×32 and 16×16. For the design of the internal transformer, we only use the encoder part with minimum modification on the original Transformer architecture. The implementation is heavily inspired from the open-source code released by Google Research team. Following will be the methods used for this project:

- Download a pre-trained model variation.
- Fine-tune to a small dataset
- Test on noisy dataset (and report the robustness)
- Apply augmentation techniques to the fine-tuned models
- Compare and report the effect of the augmentation techniques

3.1 Model Variants

The Vision Transformer has different variants that differ in the number of self-attention heads, layers, hidden dimension size and number of parameters. The variants consist of Base (ViT-B), Large (ViT-L) and Huge (ViT-H) models. The Huge model requires at least 21 GB in GPU memory to load. Therefore, we do not use this variant in this work. The model variants used for experimentation are:

- ViT-B_32
- ViT-B_16
- ViT-L_32
- ViT-L_16

The number at the end of the model variant names refer to the patch size used in that variant. ViT-B_32 extracts patches of size 32×32 and ViT-L_16 extracts patches of size 16×16 from the input images.

3.2 Fine-Tuning Parameters

The pre-trained ViT model needs to be fine-tuned to a specific dataset at hand. For this fine-tuning, the Vision Transformer provides the following tuning parameters: batch size, learning rate, training steps, warm-up steps, accumulation steps, gradient normal clipping, etc. In this work, we use a batch size of 512, base learning rate of 0.03 with cosine decay, 300 training steps, 20 warm-up steps and 8 accumulation steps. We borrowed many of these hyper-parameters from the original ViT model. However, this setting performed very well for all the experiments discussed in Section 4 as well.

4 Experiments

A series of experiments were carried out to compute and compare the performance of the different variations of the transformer model on different types of commonly occurring corruptions in image data.

4.1 Datasets

This work primarily focuses on two very popular image classification datasets: CIFAR-10 and MNIST. CIFAR-10 consists images of birds, cars, cats, etc and MNIST consists images of hand written digits. Both of these datasets have 10 labels with 50k training images, 10k test images. We have

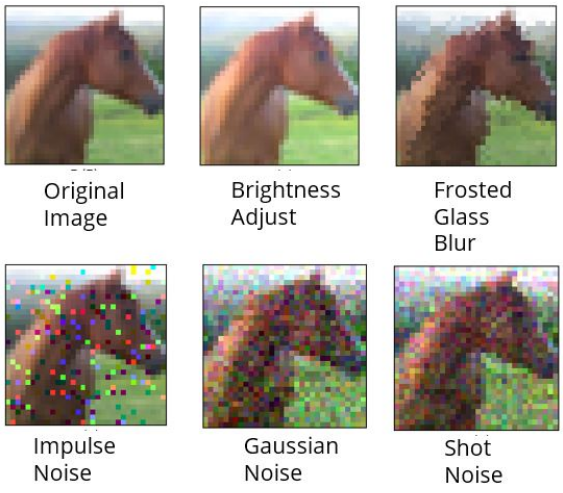

Figure 1: Different types of commonly occurring corruptions in image data added to the image of a horse from the CIFAR-10 dataset.

used the API provided by TensorFlow to download and use these datasets. TensorFlow additionally provides corrupted versions of these datasets, where different commonly occurring corruptions are added to the test images. We have used the corrupted test sets as a benchmark for analyzing the robustness of the ViT model, and for comparing the effect of the data augmentation techniques on those models.

We have used the following corruptions on the test images: Brightness Adjustment, Impulse Noise, Frosted Glass Blur, Gaussian Noise and Shot Noise. Figure 1 shows how the images look once the mentioned corruptions are added.

4.2 Data Augmentation

We tested the models with only few out of many available data augmentation techniques. Those include the following: Random Flip, Random Rotation, Gaussian Noise addition, Contrast Adjustment and a random combination of these techniques. We will discuss how each of these techniques effected the model performance in Section 5.

4.3 Hardware Configuration

The computational resource is always a constraint when working with large deep learning models. Although the transformer models require very less computation power compared to their convolution-based counterparts, the resource constraint is still an issue for a student practitioner. We used the following hardware configurations for all the experiments performed in this work.

CPU: Intel Core i9-9980HK CPU @ 2.40GHz x 16
GPU: Quadro RTX 5000 16 GB
RAM: 64 GB

This hardware configuration only allowed us to run the Large (ViT-L) variant of the vision transformer. For running the largest (ViT-H) variant, we would need to use cloud-based solutions like Google Cloud or AWS, which is not affordable for a student practitioner.

4.4 Setup

Task	Dataset	Data Statistics
Pre-Train	ImageNet-21k	14M images, 21k classes
Fine-Tune	CIFAR-10	50k train images, 10 classes
	MNIST	50k train images, 10 classes
Test	CIFAR-10	10k test images, 10 classes
	CIFAR-10 + Noise	10k test images, 10 classes
	MNIST	10k test images, 10 classes
	MNIST + Noise	10k test images, 10 classes
Augment	CIFAR-10 + Aug	50k train + Aug
	MNIST + Aug	50k train + Aug

Table 1: Experimental setups for pre-training, fine-tuning, augmentation and testing tasks

Table 1 shows the different setups for the experiments performed in this work. The original plan with this work was to pre-train at least the smallest Vision Transformer model. However, that was not possible due to computation constraints. So in this work, we use a pre-trained model that was trained on the ImageNet-21k dataset for all the experiments. We fine-tune on the standard datasets: CIFAR-10 and MNIST, and test the performance on the noiseless as well as noisy data. Then, we apply the augmentation techniques discussed above and evaluate the effect of these techniques.

5 Results

5.1 Robustness

The initial results on the test data shows that the performance of the vision transformer models suffer a lot when noise is added. This shows that although these models can show a good performance on test data, they are not very robust to noise additions and perturbations which is quite common in real-world scenario. Table 2 shows the details on how the models under-perform with different corruptions on the CIFAR-10 dataset. We can clearly see that the impulse noise hurt the performance the most.

Model	ViT-B32	ViT-B16	ViT-L32	ViT-L16
No Noise	98.88	99.02	99.06	99.13
Brightness Adjust	97.37	97.4	97.5	98.3
Frosted Glass Blur	85.2	85.4	87.13	88.78
Impulse Noise	55.85	74.16	62.17	80.92
Gaussian Noise	82.22	82.9	85.47	88.04
Shot Noise	75.92	76.93	71.32	83.5

Table 2: Initial test results on CIFAR-10 after fine-tuning. Performance accuracy shown in %.

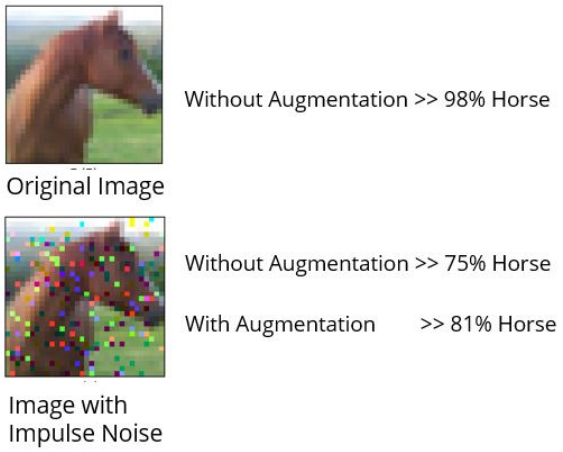

Figure 2: Example of performance improvement after data augmentation

Model		ViT-B32	ViT-B16	ViT-L32	ViT-L16
No Noise	Before	97.1	97.3	95.9	96.91
	After	97.1	97.5	97.4	98.1
Frosted Glass Blur	Before	59.1	66.6	59.6	71.7
	After	64.95	68.1	65.6	74.6
Shot Noise	Before	88.7	90.8	86	90.1
	After	91.7	91.7	91.4	93.6
Motion Blur	Before	74.1	79.6	69.6	74.6
	After	74.6	79.9	69.9	77.82

Table 3: Performance results for before and after using Flip + Rotation + Gaussian Noise augmentation technique on the MNIST dataset. Performance accuracy shown in %.

Model		ViT-B32	ViT-B16	ViT-L32	ViT-L16
Brightness Adjustment	Before	97.33	97.4	97.55	98.3
	After	97.43	98.2	98	98.65
Frosted Glass Blur	Before	85.2	85.4	87.13	88.78
	After	86.75	86.8	88.2	91.92
Impulse Noise	Before	55.85	74.16	62.17	80.92
	After	61.5	75.12	66.25	82.16
Gaussian Noise	Before	82.22	82.9	85.47	88.04
	After	84.56	84	88.54	90.21
Shot Noise	Before	75.92	76.93	71.32	83.5
	After	79.51	78.2	84.1	85.83

Table 4: Performance results for before and after using Flip + Rotation + Gaussian Noise + Contrast Adjustment augmentation technique on the CIFAR-10 dataset. Performance accuracy shown in %.

We can see in Figure 1 that impulse noise adds random RGB pixels on the images. This was too much for the model to handle and we can safely say that the models might be overfit.

Similar sets of experiments on the MNIST dataset (not included in the report) shows that the model performance suffered the most with frosted glass blur effect whereas impulse noise didn't effect it too much. This might be due to the fact that MNIST dataset contains black-and-white images so randomly added RGB pixels were easy to pick out.

5.2 Augmentation Efficacy

This sub-section presents the model performance on different image corruptions when subjected to a series of augmentation techniques. Please note that we have not presented the results for all the experiments in this report. We experimented with different combinations of augmentation techniques mentioned in Section 4.2. Comparing the pre-augmentation and post-augmentation model performance, we find that all the techniques help the models improve. Table 3 and 4 show the effect of using combined augmentation techniques. We can see the improvement in prediction accuracy for all the variants of ViT. Figure 2 shows an example of an image from the CIFAR-10 dataset. The original ViT model (without augmentation) classifies the original image as horse with 98% confidence. This confidence drops to 75% when the image is subjected to impulse noise. With the data augmented model, the prediction confidence bumps up to 81%.

One very prominent feature that we observed during the various experiments with the augmentation techniques is that the combination of these methods seem to work better than the individual methods. This might be due to the fact that the combined augmentation methods provide several different variations of the same image which results in better generalization. We also observed that the larger variants of ViT did better when data augmentation was implemented. This can be explained by the fact that larger models might be able to better learn the representation of the images from its different variations provided by the combined augmentation techniques due to their complex (superior) architecture.

6 Conclusion

The vision transformer model is still behind the traditional convolution-based models. It will take lots

of further research and work for the vision transformer to compete with CNN. However, the path is very bright. In this work, we saw that the pure vision transformer models are not so robust to corruptions and perturbations in images. But, with the addition of simple augmentation techniques, the overall performance of the model improved. The final takeaway from this work is that larger models perform better with data augmentation and combined augmentation techniques work better than individual techniques.

References

Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. 2019. Attention augmented convolutional networks. *CoRR*, abs/1904.09925.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. *CoRR*, abs/2005.12872.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR.

Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. 2019. On the relationship between self-attention and convolutional layers. *CoRR*, abs/1911.03584.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. 2017. Relation networks for object detection. *CoRR*, abs/1711.11575.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. 2021. How to train your vit? data, augmentation, and regularization in vision transformers. *CoRR*, abs/2106.10270.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. 2017. Non-local neural networks. *CoRR*, abs/1711.07971.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. 2020. Visual transformers: Token-based image representation and processing for computer vision. *CoRR*, abs/2006.03677.