Vision Transformers and their Adversarial Robustness

Abhirama Subramanyam
penamakuri.1@iitj.ac.in

Kanika Choudhary
choudhary.10@iitj.ac.in

February 16, 2021

1 Introduction

Transformers [Vaswani et al. (2017)] are known for the state of the art results on many downstream tasks in the natural language processing domain, like translation, question answering, etc. Researchers have recently extended the transformers to the vision community, one such model is vision transformer [Dosovitskiy et al. (2020)]. Despite state of the art performances on various complex tasks, existing deep learning methods are highly susceptible to adversarial attacks i.e., adding small imperceptible noise to the image will make the deep learning model to give false predictions of the image. The existence of adversarial examples has put some breaks over the huge success of deep learning models. Through this project we aim to validate the robustness of vision transformers.

2 Motivation

Our motivation in selecting this problem statement lies in our impression that self attention models act as good regularizers when it comes to images. As in self attention module when it computes attention over multiple patches of the input image, suppose a pixel or an attack has happened in the particular patch of the image, we anticipated that the attention weights are robust as they are many clean patches, that might drive the result to correct classification.

3 Attack Details

We have adopted the no-box attack framework proposed in [Li, Guo, and Chen (2020)]. Considering the vision transformer model as the target model. We assume complete black box setting, along with almost zero access to query the target model. The only information we consider that is provided to us to come up with an attack is 20 examples per class, 10 classes in total, summing up to 200 examples. Objective is to train an autoencoder [Xichen (n.d.)] using this data, and use the trained autoencoder to perform attack and generate perturbed test images. We have trained an auto encoder (a simple version to that of what

is mentioned in the paper) on these 200 samples. [A sample of reconstructed images and input images is shown below Fig.2].

3.1 Auto encoder configuration

- Epochs - 100
- Optimizer - Adam
- Learning rate - 0.001
- Loss - Binary Cross Entropy

```
ConvAutoencoder(
  (conv1): Conv2d(1, 16, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (conv2): Conv2d(16, 4, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (pool): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  (t_conv1): ConvTranspose2d(4, 14, kernel_size=(2, 2), stride=(2, 2))
  (t_conv2): ConvTranspose2d(14, 1, kernel_size=(2, 2), stride=(2, 2))
)
```

Figure 1: AE architecturea

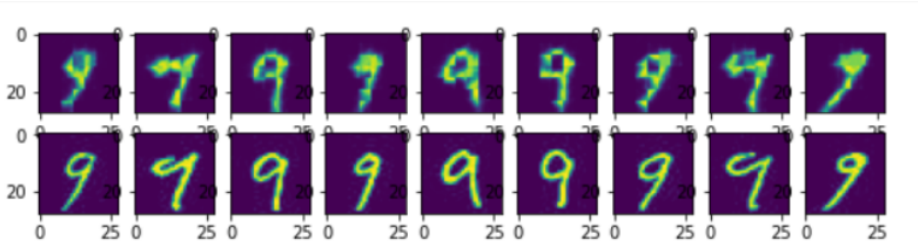

Figure 2: AE reconstructed vs input

We have performed two attacks on autoencoder, with varying arguments which are given below,

1. FGSM - epsilons[0.1,0.2,0.3 and 0.4]
2. PGD - [epsilon 0.3, alpha-0.1, iterations=30]

4 Vision Transformer details

- Image: 28*28
- patch size: 7*7

- no. of patches: 4
- epochs - 25
- Self Attention Layers (Depth) - 6
- Attention heads in one self attention layer - 8
- Loss - Negative Log Likelihood (Categorical Cross entropy)
- Optimizer - Adam
- learning rate - 0.003

5 Dataset Details

We have trained our vision transformer [Wang (n.d.)] on tiny-imagenet, cifar-10 and mnist. Owing to space and time constraints, we are sticking to mnist dataset.

1. Train data for Vision Transformer - 6000 images from each class, 10 classes in total.
2. Test data for Vision Transformer (and also to generate perturbed images) - 200 images from each class, 10 classes in total, summing upto 2000 test samples.

[Transformer has overfitted over tiny-imagenet, as the self attention layers are fewer than what is required - train accuracy is around - 90% and test accuracy is around (7-12)%]

6 Results

We have used the attacked images on (i) linear model and (ii) vision transformer to benchmark the drop in accuracy.

Attack	**Epsilon**	**Liner Model Accuracy**	**Vision Transformer Accuracy**
No Attack (clean)	0	95.8%	97.2
FGSM	0.1	91.1%	95.7%
FGSM	0.2	78.9%	78.2%
FGSM	0.3	58.4%	53.2%
FGSM	0.4	41%	34.2%
PGD	0.3 ($\alpha = 0.1, iter = 30$)	41.7%	29.9%

7 Summary

1. Vision Transformers are robust at small epsilons, but not robust as epsilon increases accuracy drops drastically. (Inferred from the reference table above)
2. To train a vision transformer on large data with more classes (here - tiny imagenet), it should have more depth.
3. We have also tried authors implementation for no box attack, but it is taking lot of time to generate the perturbed sample due to the complex autoencoder. We have visualized few examples generated in (i) cifar-10 , (ii)tiny-imagenet. Generated perturbations using the attack, on cifar-10, results didn't look great.

[saved models/data etc. are not included in the submission, will be shared on request]

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Li, Q., Guo, Y., & Chen, H. (2020). Practical no-box adversarial attacks against dnns. *Advances in Neural Information Processing Systems*, *33*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wang, P. (n.d.). *Vision transformer - pytorch.* Retrieved from `https://github.com/lucidrains/vit-pytorch`

Xichen, A. (n.d.). *Pytorch playground.* Retrieved from `https://github.com/aaron-xichen/pytorch-playground`