# Project 2: Adversarial Attacks on Vision Transformers

Blake Wyatt
blakewyatt@knights.ucf.edu

**This report reflects the work for Project 2 for the Spring 2021 semester of CAP6412.**

**With the recent entrance of Vision Transformers (ViT) into the state-of-the-art image classification space, it has become ever more critical that the extent of their adversarial nature is understood. This paper covers two existing ViT networks. One provides no defense and the other implements an input image resizing defense. Projected Gradient Descent (PGD) and a new attack introduced in this paper which we term Interpolated Projected Gradient Descent (IPGD) are used to evaluate the two ViT networks. We empirically show both PGD and IPGD are effective methods of attack. In fact, IPGD has comparable, perhaps better, performance to that of PGD itself on the defended ViT model.**

## 1 Introduction

After mass adoption of neural networks, it became apparent and important that they were vulnerable to adversarial attacks. Thus rose an influx of research pertaining to the attack and defense of neural networks. From this research came the discovery of attacks laying the foundation for adversarial research we know today such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and the Charles-Wagner attack. However, most of this research was performed on the effective and widely-used network architectures of the time such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Recently, transformers have proven themselves in the image classification space. There was prior work in natural language processing (NLP), however, their image classification performance was not as performant. Now that their image classification performance has become comparable to state-of-the-art, they are being applied to real life image classification tasks. It is therefore necessary to evaluate the adversarialness of vision transformers (ViTs) and do research into defending them.

Two ViT models are provided to us. Both are based on the Vision Transformer recently published by Google. The first is a simple reimplementation with no particular defense to adversarial attacks. The second adds a simple defense to the reimplementation. That defense is to simply resize all input from 32x32 to 16x16 with a bilinear interpolation and compare the prediction of the 32x32 input to the prediction of the 16x16 input. If the predictions are the same, the ViT model believes the prediction is correct and uses it as output.
To evaluate these models, we attack them with PGD and a new attack which we term Interpolated Projected Gradient Descent (IPGD). For the undefended ViT model, we compare its baseline accuracy to the robustness accuracy of an attack with PGD. Since we are constrained to only 100 images of CIFAR10 for this paper, the time it takes to attack is not an issue and PGD is well-known for being an effective attack against undefended networks. We find that our assumptions are correct. PGD is very effective against the undefended ViT model which leads us to believe it could likely also be effective on the defended ViT model. For the defended ViT model, we compare its baseline accuracy to the robustness accuracy of an attack with PGD and an attack with IPGD. Surprisingly, despite the defense, PGD also performs well on the defended ViT model. In fact, the robustness accuracy of PGD on the undefended ViT model and the robustness accuracy of PGD on the defended ViT model are nearly the same. IPGD builds upon the foundation left by PGD and attempts to be more performant than PGD itself. In the end, we prove IPGD is at least as performant as PGD on our 100 images of CIFAR10 and could be more performant if evaluated against the rest of the dataset.

Interpolated Projected Gradient Descent (IPGD) is the primary contribution of this paper. It is specifically designed to attack the resize defense of our defended ViT model. IPGD bilinearly interpolates the input from 32x32 to 16x16 precisely like the defended ViT model then attacks the defended ViT model with PGD. The resulting perturbations are noted and a second optimization algorithm is performed which adds perturbations to the original 32x32 input image based on the pixel values of the perturbed, interpolated 16x16 image. This way, when the 32x32 input is bilinearly interpolated, the interpolation creates the perturbations necessary for the image to be misclassified. Unexpectedly, the perturbed 32x32 image also succeeds at causing the network to misclassify it. Combined, the overall prediction of the model is a misclassification and the IPGD attack succeeds.

We find that PGD is effective at attacking both the undefended and defended ViT models and that IPGD is effective at attacking the defended ViT model. Empirical evidence on the 100 CIFAR10 images indicates IPGD is at least as effective PGD when attacking the defended ViT model. Testing on the rest of the CIFAR10 dataset could very well prove IPGD is more effective than PGD on the defended ViT model.

## 2 Projected Gradient Descent (PGD)

As one of the most well-regarded adversarial attacks, Projected Gradient Descent (PGD) is often one's first choice

for attacking an undefended network. It is effective and simple. Although, it is not very performant, in this paper, we are constrained to evaluating on a 100 CIFAR10 images so running PGD does not take much time. The algorithm is mathematically described below.

To explain PGD more simply, the algorithm operates as follows. PGD randomly initializes its perturbations $\delta$ then takes a number of steps $N$. For each step, two operations are performed. First, the adversarial input $x_i + \delta$ is predicted for a classifier $f_\theta$, the loss $\ell$ is calculated, the sign of the gradient is found, the step size $\alpha$ scales the sign to take step by a certain amount in either direction, and finally the step is applied to the perturbations. Second, the the perturbations are constrained within the boundaries of epsilon $\epsilon$. After the steps are complete, the perturbations are applied to the input, creating an adversarial example.

## 3 Vision Transformer Resize Defense

The defended ViT model resizes all input with a bilinear interpolation from 32x32 to 16x16. Internally, the ViT model ignores its original fixed image size requirements of a single size and allows a variable size of input. Any input that is 16x16 is interleaved in patches to make up for the missing 32x32 data.

Two uses of the model are made per input. The first use inputs the interpolated 16x16 image. The second use inputs the original 32x32 image. The predictions of both are then compared. If they both result in the same prediction, the defense does not believe the input to be adversarial. Thus, the way to attack the defense is to make adversarial perturbations which cause the model to mispredict in two situations. First the model must mispredict through the 16x16 bilinear interpolation and second the model must mispredict without the bilinear interpolation at the normal 32x32 input.

## 4 Interpolated Projected Gradient Descent (IPGD) and The Discovery Process

To break through the ViT Resize Defense, I theorized two different attacks could work. One attack would attempt to increase the average perturbation size. Perhaps with a larger perturbation size the perturbations would survive being interpolated. The second attack would attempt PGD after interpolating the image to 16x16.

To increase the average perturbation size, I looked into PGD and realized a larger step size should do this, however, simply increasing the step size is not creating a new attack and so would not fit the requirements of the project and increasing the step size, was not increasing the pixel values to the maximum or minimum epsilon. After investigating further, my code was simply reporting tensor values, not pixel values (or rather RGB values) and there was no problem.

Interpolated Projected Gradient Descent (IPGD) went through multiple iterations. The first, which we will call IPGD1, involved simply interpolating the input from 32x32 to 16x16, using PGD, and scaling it back to 32x32. At this point, I realized interpolating the image, is also effectively perturbing the image itself and breaks the epsilon perturbation requirements of the project. I reasoned if we can't use the interpolated image, perhaps we can calculate the perturbations on the interpolated image and apply them to the original image instead. In this attack, which we will call IPGD2, the 32x32 input is bilinearly interpolated to 16x16, run through PGD, the perturbations are scaled by two to get a 32x32 size, and applied to the original input image. In the next version of IPGD (IPGD3), the 32x32 input was once again bilinearly interpolated to 16x16, however, it was scaled by two, passed to PGD, and the resulting perturbations were applied to the original input image. None of the above iterations were effective, however, I came up with one more idea for the final iteration IPGD4, or IPGD itself.

If we could not have an adversarial image which is bilinearly interpolated and perturbed, perhaps we could have an adversarial image which will be perturbed after being bilinearly interpolated. Thus, a new optimization algorithm was created. In this algorithm, the input image is bilinearly interpolated from 32x32 to 16x16 and used to attack the defended ViT model with PGD. The perturbations for that 16x16 bilinearly interpolated image which we will call the PGD output are calculated and passed to the next part of the algorithm where the original 32x32 input image is iteratively adjusted to result in the PGD output when it is bilinearly interpolated. It is done by adjusting the perturbations of the 32x32 input image based on how close the pixel values of the 16x16 bilinearly interpolated are to the PGD output. The result is an adversarial input which when bilinearly interpolated by the defended ViT model, will have the right perturbations necessary to cause the defended ViT model to mispredict it.

The overall output of the defended ViT model compares the prediction of 16x16 input to the prediction of the 32x32 input. While the 16x16 input can now successfully cause the defended ViT model to mispredict, we would normally have to account for the 32x32 input, however, it seems as though a property of IPGD causes the 32x32 input to mispredict as well and so the overall output of the defended ViT model is mispredicted resulting in a successful attack of the defended ViT model.

## 5 Experiments

To reiterate, we empirically prove PGD successfully attacks the undefended and defended ViT model and we empirically prove IPGD successfully attacks the defended ViT model.

Thus we evaluate our attacks with a baseline accuracy for the defended ViT model and with a baseline accuracy for the undefended ViT model. After which we calculate the robustness accuracies for each attack on each defense, for all relevant parameters.

The ViT models are pretrained on CIFAR10 and as specified by the project requirements, we will be evaluating these ViT model on the first 100 CIFAR10 images from the testing set. There are two stages to this project: stage A and stage B. Stage A is an attack on the undefended ViT model at a maximum epsilon value of six. In fact, we can achieve 0% robustness accuracy at epsilon 4 with PGD for stage A so that requirement is satisfied. Stage B requires we design a new attack and maintain a maximum epsilon of 10 while attacking the defended ViT model. In fact, we can achieve 0% robustness accuracy at epsilon 4 with PGD and our new attack algorithm IPGD while attacking the defended ViT model.

All code is written with PyTorch and normalized as specified in given Project Notes. The CIFAR10 input is normalized by its mean, standard deviation, and from a range of -1 to 1 with 0 as its center. The presented algorithms normalize epsilon and the step size as specified and adjust the upper and lower limits of the algorithms to the normalized input.

All step sizes were one or two, totals steps are $2\epsilon$, and epsilon is kept between three and ten. With IPGD, comes the introduction of new parameters. IPGD has a second step size and a second number of total steps. Both are constrained the same as the other step sizes and steps.

## 6 Results

| Undef ViT | | | |
|---|---|---|---|
| | Robust Acc. % | $L_2$ Dist. | $L_\infty$ Dist. |
| Baseline | 83.00% | N/A | N/A |
| PGD | 0.00% | 1.02 | 6 |
| PGD | 0.00% | 0.76 | 4 |
| PGD | 1.00% | 0.58 | 3 |

**Table 1:** *Undefended ViT model attack results. Attack type is on the left and epsilon can be found by the $L_\infty$ distance.*

| Def ViT | | | |
|---|---|---|---|
| | Robust Acc. % | $L_2$ Dist. | $L_\infty$ Dist. |
| Baseline | 61.00% | N/A | N/A |
| PGD | 0.00% | 1.51 | 10 |
| PGD | 0.00% | 0.75 | 4 |
| PGD | 2.00% | 0.57 | 3 |
| IPGD | 0.00% | 1.63 | 10 |
| IPGD | 0.00% | 0.78 | 4 |
| IPGD | 0.00% | 0.59 | 3 |

**Table 2:** *Defended ViT model attack results. Attack type is on the left and epsilon can be found by the $L_\infty$ distance.*

As can be seen in Tables 1 and 2, IPGD and PGD perform similarly well on the defended ViT model with the same percent of robustness accuracy. If there were more CIFAR10 images to test on, perhaps, IPGD is even more performant than PGD. As for the undefended ViT model, PGD clearly performs extraordinarily well and all of them only have problems when reaching epsilon three.

## CONCLUSION

From the evidence presented, it is clear PGD and IPGD are effective against undefended and defended versions of the ViT defense. Clearly, simulating an bilinear interpolation and then optimizing for it to apply to an adversarial input is effective and perhaps with more research, more efficient methods of optimizing can be found.