Project: Comparing adversarial robustness of networks with PGD attacks

Kesar Murthy,
University of Central Florida
kesar@Knights.ucf.edu

Daniel Silva
University of Central Florida
danielzgsilva@Knights.ucf.edu

1 Abstract

Recent work has uncovered dangerous phenomena of neural networks, highlighting their instability and vulnerability when attacked with carefully tuned perturbations to the input space. Projected gradient descent (PGD), in particular, is an adversarial attack which finds these perturbations through constrained optimization. By optimizing the input to maximize prediction error, PGD generates quasi-imperceptible perturbations that fool neural classifiers at a high rate. In this work, we evaluate and compare the performance of two image classification networks on such adversarial examples. Our findings suggest that the ResNet architecture, perhaps as a result of skip-connections, demonstrates a stronger resilience to such attacks than its counterpart, VGG.

Keywords: ***Projected Gradient Descent, quasi-imperceptible perturbations, ResNet, skip-connections, VGG, ImageNet***

2 Introduction

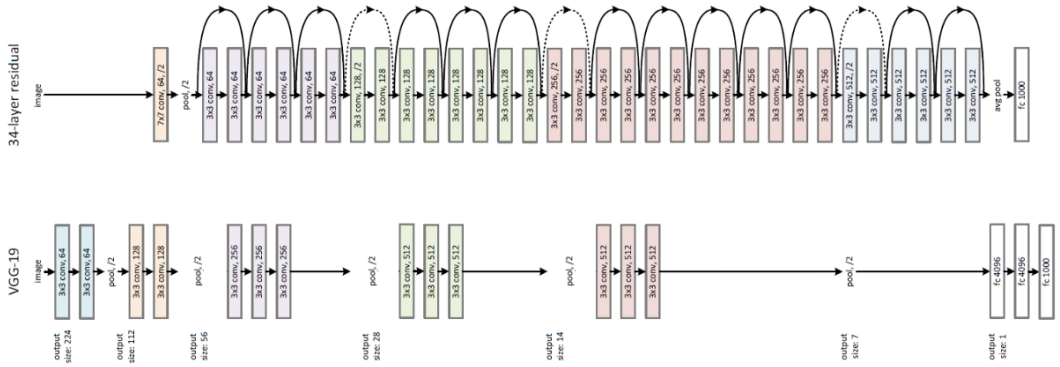
Despite the remarkable success of deep neural networks in a variety of settings such as image understanding [9] and natural language processing, there remains a number of interesting properties [5] which must be understood before these systems can be relied on for safety critical tasks. Perhaps the most notable, is their vulnerability to adversarial perturbations in their input space. More specifically, adversarial perturbations are small, carefully tuned modifications to the input which force the model to mis-predict. In the context of computer vision, these perturbations have been shown to fool state of the art image classifiers while being imperceptible to the human eye [4; 6; 7; 8]. This raises important questions surrounding the robustness and reliability of neural networks and has led to a rich research space in both adversarial attack and defense.

The majority of approaches for generating such perturbations are known as gradient based adversarial attacks. This family of attack methods exploit the same back-propagation concepts that are used in the training of neural networks. During training, the gradient of an error function is computed with respect to model weights, then used to optimize the weights to minimize error. On the contrary, gradient-based adversarial attacks consider model weights to be constant, and compute the gradient with respect to the input to optimize the input itself to maximize error.

One of the most successful approaches within this family is Projected Gradient Descent (PGD) [4], an iterative white-box attack (meaning the algorithm has access to a model's weights). PGD frames adversarial perturbation generation as a constrained optimization problem. As such, PGD attempts to find the perturbation which maximizes a model's error on a particular input, while ensuring the perturbation's magnitude is less than some value ϵ. Included to ensure the adversarial example isn't exceedingly different from the original, this magnitude constraint is applied as the L_p norm of the perturbation. The PGD algorithm begins with a random perturbation from within the L_p ball around an input. The algorithm then proceeds for a desired number of iterations, repeatedly calculating the error from the current perturbation, computing the gradient, updating the current perturbation by

Preprint. Under review.

Figure 1: Model architecture of ResNet and VGG networks



taking a step of desired size in the direction of the greatest loss, and lastly projecting the current perturbation back onto the L_p ball to satisfy the previously mentioned magnitude constraint. This iterative perturbation update procedure can be formalized as

$$\text{Initialize } \delta;$$
$$\text{Repeat for } n \text{ iterations:}$$
$$\delta = P(\delta + \alpha \nabla_\delta L(x + \delta, y, \theta))$$

where δ represents the generated perturbation, P denotes the projection onto the L_p ball, x is the original input, y is its ground truth label, θ are the static model weights, L is the model loss, ∇_δ is the loss gradient with respect to the input, and α is a desired step size.

In this work, we apply Projected Gradient Descent on ImageNet [3] images and use these adversarial examples to evaluate the robustness of ResNet [1] and VGG [2] image classification networks.

3 Experiments

The experiments were carried out using the ImageNet validation dataset. More information regarding this dataset is provided in Section 4. The ImageNet images have an input size of 224x224. We normalize the images with their mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. For validation, we use a batch size of 24 and and shuffle the input data. The experiments were carried out on UCF Newton clusters.

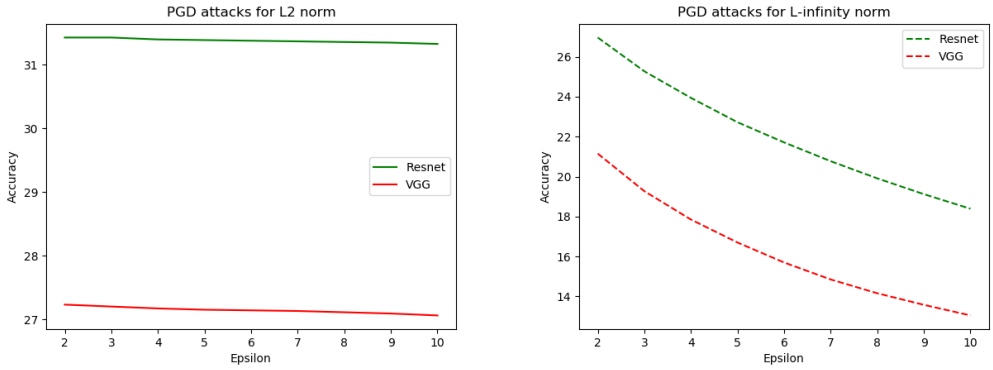
We obtain the results for ResNet-34 and VGG-16 by leveraging pre-trained models provided by Pytorch. The model architecture for each of these models are described in Figure 1, with the addition of batch normalization for each of the layers. The baseline Top-1 accuracy for the ResNet-34 and VGG-16 are 26.70 and 26.63 respectively.

We compare the performance of the two models using several variations of the PGD attack. Specifically, we vary ϵ from 2 to 10, which controls the magnitude of the perturbation. We then set the number of iterations n to be twice the value of ϵ. We also apply PGD using both L_2 and L_∞ norms, and use a static step size α of 1. The results for our experiments are shown in Table 1, with the VGG model performing slightly worse than the ResNet model across all PGD variations. A more detailed analysis of these results is provided in section 5.

4 Dataset

The Imagenet [3] dataset was released for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition in 2012. The training data comprises of 1 million images pertaining to 1000 classes. In our experiment, we use the validation set comprising of 50,000 images with 50 images per class.

Figure 2: Comparing the robustness of ResNet and VGG on PGD adversarial attacks



5 Results

As noted above, the results are obtained after carrying out 2ϵ number of PGD iterations. The test accuracy for each experiment is described in Table 1 and visualized in Figure 2. As we can see, the ResNet model shows minimal variation in accuracy with L_2 norm, and a steep decline in accuracy with L_∞ norm as epsilon and number of iterations increase. The same trends also follow for the VGG model.

Some interesting findings which can be gathered from these results include:

- According to the paper, the resulting adversarial error plateaus after a significant number of PGD iterations. In our implementation, although we perform a relatively low number of iterations, we did not experience this plateau, especially with L_∞ norm. Of course, in our experiments, iteration count is linearly increased with epsilon, which dictates the strength of the perturbation. Therefore, it is slightly unclear whether the accuracy decrease shown in Figure 2 is due to the increase in epsilon or in iteration number.

- The paper also mentions that the L_∞ norm is more robust than the L_2 norm for a standard model. We notice the same behaviour for both the VGG and ResNet models, with $L\infty$ resulting in a higher error rate than its counterpart for the same PGD parameters. Moreover, the error rate does not improve with increased epsilon values or iteration counts in the case of the L_2 norm, as it does with L_∞.

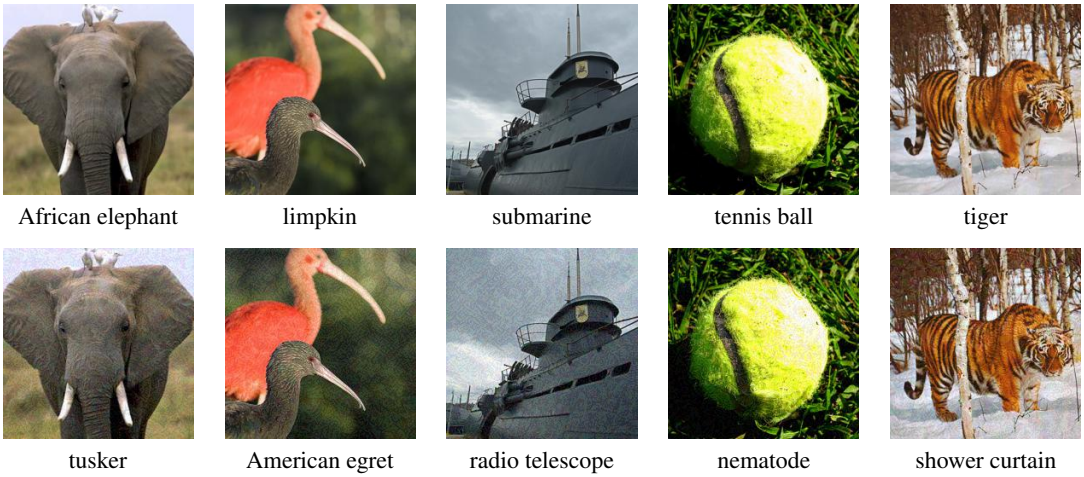
- The ResNet-34 model displays a higher accuracy on adversarial examples than VGG across all PGD parameter variations, despite the two models having a nearly equal baseline accuracy. This is a major finding, suggesting that the ResNet architecture has a higher degree of adversarial robustness than VGG. It is also interesting to note that the relationship between epsilon, iterations, and classification accuracy stays constant across the models. This can be observed in Figure 2, where the models have a static accuracy difference when these parameters are varied.

Figure 3 shows a comparison of original images in the top row and the perturbed images at the bottom. Since most of the image labels in the ImageNet dataset have a super-category that is also a class among the 1000 labels, (for example there exists a class 'Bookcase' and a subsequent class 'Library', which are similar in many ways and fall under a super-category 'books') we observe that many of the mis-classified adversarial images are still classified with a label in the same super-category as the ground truth. We hypothesize that this is due to the nature of our perturbations lying in the same sample space, right outside the class' decision boundary and yet under the same super-category. Some examples of this include the African Elephant image being classified as a Tusker upon perturbation. Similarly, the image of the limpkin is classified as an American Egret. However, we also see results which do not follow this trend, such as the tennis ball being classified as a nematode and the tiger being classified as a shower curtain upon perturbation.

Table 1: Experimental results of PGD on ResNet-34 and VGG models. The baseline classification accuracy on the ImageNet validation set for each model is included in the far right. Accuracy on PGD generated images with varying norms, epsilon, and number of iterations are in the table body.

		Epsilon									Baseline
Model	Norm	2	3	4	5	6	7	8	9	10	
VGG	L_2	27.2	27.2	27.3	27.2	27.1	27.1	27.1	27.1	27.1	73.4
	L_∞	21.2	19.4	17.9	16.7	15.7	14.9	14.2	13.6	13.1	
ResNet	L_2	31.4	31.4	31.4	31.4	31.4	31.4	31.4	31.4	31.3	73.3
	L_∞	26.9	25.3	23.9	22.7	21.7	20.8	19.9	19.1	18.4	

Figure 3: Visualizing perturbations generated with PGD using the ResNet-34 model. Parameters of PGD are $\epsilon = 10$, $n = 20$, norm = L_∞. The top row are original images while the bottom row are perturbed. Below each image we include the predicted label.



6 Conclusion

- Which architecture is more vulnerable to adversarial attacks?
  Based on the results obtained from testing both models on PGD-generated adversarial examples, the VGG-16 [2] network with batch normalization is more vulnerable to adversarial perturbations than its ResNet-34 [1] counterpart. We draw this conclusion after observing that the ResNet model consistently performs better than VGG for all epsilon values, iteration counts, and across L_∞ and L_2 norms.
- Is the difference statistically significant?
  The results indicate that the difference in classification accuracy between models is statistically significant. To show this, we use a paired T-test on the accuracy results from each of our experiments. Here, we consider the corresponding accuracies for both models on each PGD parameter combination as our observation pairs. In the paired T-test, upon calculating the difference between observation pairs, the hypothesis is that the mean difference is zero. Therefore, we can prove the difference in adversarial robustness of our models is statistically significant by rejecting this hypothesis. For the L_2 norm, a paired T-test on the accuracies of both models results in a T value of 179.6, and a p-value of 0.0001. For the L_∞ norm, the result is a T value of 72.5 and a p-value of 0.0001. Therefore, we can say with high confidence that the difference in PGD-based adversarial robustness for VGG and ResNet is statistically significant.
- If it is, what is your justification?
  We attribute the difference in performance to the skip-connections present in the ResNet model. The residual blocks have skip-connections that are summed up at the end of each block. This would mean that the gradients being updated have a 'highway' path during back-propagation, perhaps allowing weights to learn a more robust representation. Moreover,

in a forward pass, the skip connections allow information from earlier layers to be better represented in deeper layers. This may allow the network to rely more strongly on the basic, general features in earlier layers, rather than complex and perhaps over-fitted features in deep layers, enabling the model to be more robust to these small perturbations. The paper [10] explains in depth how the skip-connections matter and how they affect the adversaries.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[2] Karen Simonyan, Andrew Zisserman. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition.

[3] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu. (2019). Towards Deep Learning Models Resistant to Adversarial Attacks.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In International Conference on Learning Representations (ICLR), 2014.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR), 2015.

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In Symposium on Security and Privacy (SP), 2017.

[8] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In Computer Vision and Pattern Recognition (CVPR), 2016

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (NeurIPS), 2012.

[10] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, Xingjun Ma.Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets International Conference on Learning Representations (ICLR) 2020.