Project 1: Adversarial Attacks on ResNet-34 and VGG-16

Zach Schickler, Blake Wyatt
zachschickler@knights.ucf.edu, blakewyatt@knights.ucf.edu

This report reflects the work for Project 1 for the Spring 2021 semester of CAP6412.

February 17, 2021

ResNet-34 and VGG-16 are two pretrained deep convolutional network models used in image classification. Despite their success under natural data, they are weak to adversarial examples. While both models have a relatively low error rate on the ImageNet validation data, their error rate goes up significantly under PGD attacks for both I_2 and I_{inf}. While this is true, the L_∞ attacks proved to be much more effective than the L_2 attacks. Furthermore, the VGG-16 model was slightly better at resisting attacks than the ResNet-16 model.

INTRODUCTION

In this paper, the two models that will be the focus of all experiments are VGG16 and ResNet-34. VGG16 is a convolutional neural network model that is used for image classification. It was designed by Simonyan and Zisserman as an improvement over AlexNet. It was submitted to the Large Scale Visual Recognition Challenge 2014. It reduced the large kernels of AlexNet into multiple smaller kernels as a way of optimization.
ResNet can be considered the successor to VGG. It won first place in the following challenge year. ResNet performs much better than previous options because it solved a problem that other deep networks had. With an increasing network depth, accuracy would plateau and then drop. ResNet fixes this problem by optimizing a mapping function that represents the non-linear layers. Like VGG, this model deals with image classification. Both models are trained on the ImageNet dataset.

EXPERIMENTS

The experiments are split into two sections. One section pertains to trials run with the VGG-16 Batch Normalization model, and one section is on ResNet-34 trials. For each model, there were 18 trials. 9 of these trials were performed using the PGD_{L_2} adversarial method, while the other 9 were performed using the PGD_{L_∞} adversarial method. In each of these sets of 9 trials, the ϵ was varied from 2 to 10, and the number of iterations was set to be $2 \times \epsilon$. The step size was always constant at 1. This made for a total of 36 trials. All experiments were done inside of a Jupyter Notebook. For each trial, we recorded the robust accuracy, L_2 and L_∞ distance.

1 RESULTS

VGG-16 BN	L_2 Attack		
	Robust Acc. %	L_2 Dist.	L_∞ Dist.
2	40.30%	0.01	0.11px
3	39.78%	0.01	0.16px
4	39.44%	0.02	0.21px
5	38.96%	0.02	0.27px
6	38.46%	0.02	0.32px
7	38.00%	0.03	0.37px
8	37.62%	0.03	0.43px
9	37.11%	0.03	0.49px
10	36.63%	0.03	0.56px

Table 1: *L_2 attack on the VGG-16 model.*

VGG-16 BN	L_∞ Attack		
	Robust Acc. %	L_2 Dist.	L_∞ Dist.
2	0.04%	2.54	2.00px
3	0.04%	3.38	3.00px
4	0.00%	4.42	4.00px
5	0.00%	5.19	5.00px
6	0.00%	6.02	6.00px
7	0.00%	6.99	7.00px
8	0.02%	7.04	8.00px
9	0.02%	8.25	9.00px
10	0.02%	9.12	10.00px

Table 2: *L_∞ attack on the VGG-16 model.*

ResNet-34	L_2 Attack		
	Robust Acc. %	L_2 Dist.	L_∞ Dist.
2	51.18%	0.01	0.09px
3	50.50%	0.01	0.13px
4	49.88%	0.02	0.18px
5	49.28%	0.02	0.23px
6	48.46%	0.02	0.27px
7	47.64%	0.03	0.32px
8	46.96%	0.03	0.36px
9	46.26%	0.04	0.40px
10	45.72%	0.04	0.45px

Table 3: *L_2 attack on the ResNet-34 model.*

As can be seen in Tables 1-4, there is a significant difference between the robustness accuracy of the L_∞ and L_2 attacks. The L_∞ attacks are extremely effective on these two networks, in fact we assumed there was an error in the code, whereas the L_2 attacks are effective, however, not nearly as effective as L_∞. Although, L_2 attacks do seem to have a major strong point in small L_∞ and L_2 distances. L_∞ distance represents the largest perturbation in an adversarial image and so, having a small L_∞ means that there are less large visible differences between adversarial images and original images. L_2 distance

ResNet-34	L_∞ Attack		
	Robust Acc. %	L_2 Dist.	L_∞ Dist.
2	0.04%	2.46	2.00px
3	0.00%	3.26	3.00px
4	0.00%	4.17	4.00px
5	0.00%	4.93	5.00px
6	0.00%	5.68	6.00px
7	0.00%	6.23	7.00px
8	0.00	6.95	8.00px
9	0.00	7.29	9.00px
10	0.00	7.94	10.00px

Table 4: *L_∞ attack on the ResNet-34 model.*

represents the perturbation distances across the entire image. An adversarial image as a whole, will appear more like the original.

Although there isn't much of a difference, VGG-16 BN seems to outperform ResNet-34. It is able to achieve slightly better metrics in every category. It seems as though ResNet-34 is inherently more robust than VGG-16. In both models, as epsilon grows, and the perturbation limit is increased, accuracy improves.

CONCLUSION

From the evidence presented, L_2 attacks minimize both the perturbation of images and their robustness. L_∞ attacks minimize the robustness and maximize the perturbations within the limit of epsilon. As epsilon is increased, more perturbations are allowed and greater robustness is able to be achieved. Overall, it appears there is little to no protection inherent in VGG-16 and ResNet-34 against adversarial examples. For such important models this is extremely important to note and consider when choosing a model for sensitive or security reasons.

REFEREENCES

https://medium.com/@14prakash/understanding-and-implementing-architectures-of-resnet-and-resnext-for-state-of-the-art-image-cf51669e1624

https://neurohive.io/en/popular-networks/vgg16/

http://www.image-net.org/challenges/LSVRC/2014/results

https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035?gi=daa66c36c207

https://arxiv.org/abs/1706.06083

https://academictorrents.com/details/5d6d0df7ed81efd49ca99ea4737e0ae5e3a5f2e5

https://github.com/Harry24k/adversarial-attacks-pytorch

http://dl.caffe.berkeleyvision.org/caffe_ilsvrc12.tar.gz