Seminar thesis

Adversarial Label Flips (sexier title?)

Matthias Dellago & Maximilian Samsinger

June 12, 2021

Abstract

sell it to sb looking for something good to read. only 4 sentences or so.

1 Introduction

What is the open question? How do we solve it.

ie. what is our contribution. (we show that these confusion matrix are surprisingly nonrandom)
use forward references i.e. "we elaborate on this in sectrion 4".
Give an example for a confusion matrix straight away.

Existence of adversarial examples Demonstrated that attacking deep neural networks are susceptible to attacks [1]. They actually coined the term "adversarial examples".

2 Background and related work

move it after the results so that the reader, can first get the interesting stuff, and then get the background?

2.1 Attacks

Fast gradient sign method [2] developed the fast gradient sign method. They are the guys with the panda image.

Projected gradient descent The projected gradient descent, which is basically iterated FGSM, was first shown in [3]. Their experiments suggest that these attacks converge, i.e. they find a local maxima. This may require some restarts.

Carlini-Wagner attack

Foolbox A Python library with lots of attacks [4]. They include the attacks above.

2.2 Neural networks (Necessary)

Is this section necessary? It seems that whoever is interested in our results, easily already knows this.

First introduced in [5]. The authors of [6] demonstrated the effectiveness of deep convolutional neural networks on ImageNet.

ResNets Paradigm shift in deep learning. In [7] they developed Residual Networks to train very deep neural networks. We will probably use ResNet18. If we do, we probably also cite [8] for the "pre-activation" optimization. This is just a better architecture obtained by having BatchNorm-ReLU-Weights blocks instead of Weights-BatchNorm-ReLU blocks.

3 Methods

Reference to our github.

3.1 Datasets

MNIST, Fashion MNIST, CIFAR-10

3.2

We probably use `https://arxiv.org/pdf/1608.04644.pdf` Table 1 as a neural network for MNIST & Fashion-MNIST.

4 Experiments

5 Results

pictures, pictures, and maybe a graph or two

6 Discussion

Symmetry of matrices -¿ maybe find a way to quantify symmetry?
-¿ NN can recognise "similarity"

Attractor classes -¿ manage with an extra "noise"-class or so?

In Figures X, Y and Z one can observe that adversarial examples computed with large perturbation budgets ϵ are misclassified as "8", "TODO" and "frog" for MNIST, Fashion-MNIST and CIFAR-10 respectively. In order to shed light onto this phenomenon we generate and classify 10000 white noise images sampled from a uniform distribution on the input domain. Figure A shows that these randomly generated images are also, most commonly, classified as "8", "TODO" and "frog" respectively. This result suggests that the neural networks in question have a default output for low probability images with respect to distribution of the input domain, which in turn affects adversarial examples computed with large perturbation budgets.

7 Conclusion

What was the main idea.

8 Contribution Statement

References

[1] 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.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[4] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

[5] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.