Related work

Adversarial Label Flips

Matthias Dellago & Maximilian Samsinger

April 22, 2021

1 Notation

We denote neural network classifiers with $f_\theta \colon \mathcal{X} \to \mathcal{Y}, x \mapsto y$ with trainable parameter θ, where $\mathcal{X}$ are a set of images with corresponding labels (classes) $\mathcal{Y}$. The parameter θ are optimized by minimizing a training objective $(\theta, x, y) \mapsto J(\theta, x, y)$ with respect to θ.

2 On adversarial attacks

Deep neural networks have been shown to be vulnerable to tiny, maliciously crafted perturbations applied to otherwise benign inputs. These so-called "adversarial examples" were first introduced in [1]. Further research [2] showed that these adversarial examples generalize over multiple dataset and architectures. Even very inexpensive attacks like the Fast Gradient Sign Method (FGSM) [2] can be used to fool neural networks. FGSM requires white-box access to the targeted neural networks architecture and its weights. Adversarial examples are computed by performing a gradient ascent step with respect to the sign of the gradient

$$\mathrm{FGSM}_\epsilon(x) = x + \epsilon \operatorname{sign}(\nabla_x J(\theta, x, y))$$

given a step size ϵ. FGSM is an L^∞-bounded attack[1], i.e. $\|x - \mathrm{FGSM}_\epsilon(x)\|_\infty \leq \epsilon$, meaning that each pixel value of a benign image x may only be perturbed by up to ϵ. Stronger attacks can be computed by repeatedly applying FGSM with smaller step sizes. This type of attack is known as Projected Gradient Descent (PGD) and was first introduced [4]. Their experiments demonstrated the effectiveness of such attacks and showed that convergence is achieved after only a few hundred iterations.

Foolbox A Python library with lots of attacks [5]. They include the attacks above.

3 On neural networks

3.1 Neural networks

First introduced in [6]. The authors of [7] demonstrated the effectiveness of deep convolutional neural networks on ImageNet.

ResNets Paradigm shift in deep learning. In [8] they developed Residual Networks to train very deep neural networks. We will probably use ResNet18. If we do, we probably also cite [9] for the "pre-activation" optimization. This is

[1] While bounds with respect to an L^p norm are commonly used in the machine learning literature, we are aware that they are "[...] neither necessary nor sufficient for perceptual similarity [...]" [3].

just a better architecture obtained by having BatchNorm-ReLU-Weights blocks instead of Weights-BatchNorm-ReLU blocks.

4 Methods

4.1 Datasets

MNIST, Fashion MNIST, CIFAR-10

4.2

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

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] Mahmood Sharif, Lujo Bauer, and Michael K Reiter. On the suitability of lp-norms for creating and preventing adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1605–1613, 2018.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.