Related work

Adversarial Label Flips

Matthias Dellago & Maximilian Samsinger

April 22, 2021

1 Notation

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

2 On neural networks

Convolutional neural networks are the de facto standard in computer vision related tasks. They were first introduced in [1] to classify handwritten digits. In [2], Krizhevsky et al. demonstrated the effectiveness of deep convolutional neural networks on ImageNet dataset [3], winning the ImageNet Large Scale Visual Recognition Challenge 2012 [4]. The architecture of convolutional neural networks has since been further optimized. Residual neural networks [5] and their variants are state-of-the-art for image recognition tasks and dominate the leaderboard on websites such as `https://paperswithcode.com/task/image-classification`.

Experiments For our experiments we will consider the MNIST [6], Fashion-MNIST [7] and CIFAR-10 [8] datasets. For MNIST and Fashion-MNIST we will use the convolutional neural network described in [9], Table 1. For CIFAR-10 we may use a ResNet-18 architecture [5] with the "pre-activation" optimization [10].

3 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 [11]. Further research showed that these adversarial examples generalize over multiple datasets and architectures [12]. Even very inexpensive attacks like the Fast Gradient Sign Method (FGSM) [12] 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 $\epsilon > 0$. 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

[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 [...]" [13].

Descent (PGD) and was first introduced [14]. Their experiments demonstrated the effectiveness of such attacks and showed that convergence is achieved after only a few hundred iterations or less.

Foolbox For our experiments we will use a suite of different attacks using the FoolBox library [15]. The documentation is available at `https://foolbox.readthedocs.io/en/stable/`.

References

[1] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[2] 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.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[5] 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.

[6] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[7] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[8] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[9] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.