Adversarial defense suggestion

Aurélien Fouque

La Rochelle, France
19th June 2020

As Machine Learning is becoming more and more present in our society, there is a growing demand in innovative fields such as robustness [1, 2, 3] and comprehensibility [4, 5, 6]. This summary focuses on the selection of one publication dealing with robustness, especially the adversarial defense notion. The title of the selected article is *Encryption inspired adversarial defense for visual classification*, written by AprilPyone *et al.*, whose first version was released on 16th May 2020 [7].

An adversarial defense is meant to make algorithms more robust to a kind of attack called *adversarial example*, expression coined by Szegedy *et al.* [1]. In our study, an adversarial example is a modified input, which is an image in this case. The difference between the original image and the modified one cannot be perceived by the human eye. The adversarial example is then supposed to make the algorithm classify the image into either a wrong or a targeted class. Various attack strategies were developed in the past, for instance *Fast Gradient Sign Method* by Goodfellow *et al.* [2].

The central contribution of the selected publication consists in transforming the images. This transformation, also known as an encryption method, is based on *shuffling block-wise pixels*. The way the shuffling method is made by the provider remains secret for the attacker. Here is how the defense works. The model is trained with encrypted images. When the user sends some test images to the provider, these images happen to be intercepted and modified by an attacker. The test images are also encrypted and they are finally tested by the provider. The classification must be right – what does the picture represent – and the accuracy – how confident the algorithm is – must be as high as possible, whether the images are attacked or clean.

In the article, the neural network used is ResNet18 from by He *et al.* [8] and the training set comes from CIFAR-10 dataset [9], which is a collection of 60,000 labelled colored images. A threat model is developed in order to test the proposed defense. The authors take into account the advice of Carlini *et al.*, who published a checklist to avoid common mistakes when testing the robustness of a defense [3]. Two attacks are contained in the threat model. The first one, a non-adaptative attack, is called *Projected Gradient Descent* (PGD) from Madry *et al.* [10]. It is considered to be one of the strongest attack under the infinity norm. The second one, an adaptative attack, is performed using *Backward Pass Differentiable Approximation* (BPDA) from Athalye *et al.* [11]. The results of this defense were compared with other defenses visibles on RobustML Catalog[1]. The other defenses are *Latent Adversarial Training* (LAT) from Kumari *et al.* [12], *Adversarial Training* (AT) from Madry *et al.* [10], and *Thermometer Encoding* (TE) from Buckman *et al.* [13].

This article has been chosen because it is very recent compared to the last defense published in the catalog. Its defense is based on a quite simple principle which might be hard to crack. Having said that, here are what we consider as possible pros and cons. Firstly, apart from being a simple principle, the method achieved a very good accuracy: 91% on clean images and 61% minimum on attacked images. It outperforms the state-of-the-art defenses with respectively 90% and 53% maximum. Secondly, about the accuracy of the proposed method under a PGD attack, it may be interesting to try more block sizes $M \times M$. For instance, the block size can take these values: $M \in \{3;5;6;7;9;10\}$, to refine the accuracy table, especially around the best result $M = 4$.

Then, concerning the secret key, it might be broken by finding the right size of the block-wise pixels and the pattern of the permutation vector. The maximal number of permutations is $n!$, where

[1] https://www.robust-ml.org/
Today, the last version of the most recent published defense is from 30th October 2019.

the number of pixel $n = M^2$. For $M = 4$, we have then 2.09×10^{13} possibilities, which seems already quite good. As an example, $M = 5$ would give 1.55×10^{25} possibilities. The robustness could be even more improved by not only using squares as blocks – rectangles – or use different shapes. As an additional check or to generalize this method, a next step would to be try other datasets such as CIFAR-100 [9], MNIST [14], SHVN [15], or ImageNet [16] datasets. The defense is based on infinity norm. Other norms may be tested, for example 1 or 2, in order to check their influence. A last remark is that the defense is implemented on a white box. A white box is an algorithm where the mechanisms are accessible to everyone. A next step would be to implement it on gray boxes or even black boxes, where the settings are kept secret, which is more interesting for companies.

All things considered, this defense may be a new efficient tool for the company. A regular check is required, in order to see whether the defense will be published on the catalog and whether it remains robust. Indeed, the aforementioned catalog contains on the one hand the published defenses, with their accuracy claim, and on the other hand the attempts to break the defenses with their accuracy analysis. Definitely not robust defenses, that is to say 0% accuracy analysis, occurred several times. We are thus looking forward to seeing how long the selected encryption method will survive! One can also recommend an hybrid method, combining this encryption method with an another defense method, to make the algorithm even more robust.

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. 2013, arXiv:1312.6199 [cs.CV].

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

[3] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. 2019, arXiv:1902.06705 [cs.LG].

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 2016, arXiv:1602.04938 [cs.LG].

[5] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841, 2018. pdf.

[6] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. pdf.

[7] MaungMaung AprilPyone and Hitoshi Kiya. Encryption inspired adversarial defense for visual classification. 2020, arXiv:2005.07998 [cs.LG].

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2015, arXiv:1512.03385 [cs.CV].

[9] Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, April 2009. pdf.

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

[11] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. 2018, arXiv:1802.00420 [cs.LG].

[12] Nupur Kumari, Mayank Singh, Abhishek Sinha, Harshitha Machiraju, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Harnessing the vulnerability of latent layers in adversarially trained models. 2019, arXiv:1905.05186 [cs.LG].

[13] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, 2018. pdf.

[14] Y Le Cun, L.D Jackel, B Boser, J.S Denker, H.P Graf, I Guyon, D Henderson, R.E Howard, and W Hubbard. Handwritten digit recognition: applications of neural network chips and automatic learning. *IEEE Communications Magazine*, 27(11):41–46, 1989.

[15] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. *Neural Information Processing Systems*, 01 2011. pdf.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Computer Vision and Pattern Recognition*, 2009. pdf.