Ph.D. thesis proposal: Input space exploration for the security of neural-network models.

- **Keywords:** Neural-networks, machine learning models, security, black-box interaction, decision boundaries.
- **Location:** Inria Research Center, Rennes, France.
- **Ph.D advisors:** Teddy Furon (teddy.furon@inria.fr) & Erwan Le Merrer (erwan.le-merrer@inria.fr)
- **Start:** September 2020.
- **Requirement:** Due to the topic being funded by the French AID (Agence pour l'Innovation de la Défense), only EU students can be considered for this grant.

Context: The topic of the PhD thesis pertains to the security of deep neural networks. It considers the *black-box* scenario where a classifier has been learned (supervised or un-supervised training) and deployed to predict classes of input observations. An attacker aims at deluding its inference by adding small perturbations to the observations in order to trigger wrong predictions.

The classifier partitions the observation space into regions associated to each class. The creation of decision boundaries by modern deep learning models are still not well understood [10, 13, 3]. This causes security issues when those models are placed in production: The literature on adversarial examples [11] in the *white-box* scenario shows that perturbations of very small amplitude can succeed to delude the network. Yet, that perturbation is not random ; on the contrary it is crafted thanks to the full knowledge of the classifier.

In a *black-box* scenario [8], the attacker knows nothing about the internals of the network to challenge. To compensate for this lack of knowledge, the attacker is given a limited access to the sealed classifier: he/she can submit observations and see their predictions. The complexity of an attack is usually measured by the number of calls necessary to delude to the classifier.

Attacks: A first strategy is to steal knowledge from the black-box. The attacker bombards the classifier with observations, records their outputs, and then learns in a supervised way a 'proxy' network mimicking the behavior of the black-box. Once the proxy network trained, the attacker is back to a *white-box* attack: he/she mounts an attack against its own white-box proxy, conjecturing

the attack will transfer to the unknown network [12]. The cost is a huge number of calls to the black box in order to build a similar enough proxy, especially when the observations lie in a high dimensional space.

The Phd thesis will investigate strategies for sampling observations [14] in order to ease the learning of the proxy. There is definitively a connection with active learning [15]: the attacker chooses which observation is the most amenable of increasing the accuracy of the proxy. This yields a complex trade-off between the number of samples/calls, the accuracy of the proxy, and the probability of deluding the black box classifier.

The concept of black-box attack has existed since the 2000's in Digital Watermarking under the name *oracle attack* [2, 4, 5]. It is surprising that the Machine Learning community never takes advantage of this know-how. There are subtle differences, and addressing them will enrich both fields. For example, attacks in watermarking consume less calls to the oracle, but they are memory-less: a new attack must be initiated from scratch for each observation. The attacker starts adding an extreme perturbation so that the observation is pushed into another class region. Then, a line search between the original and the perturbated observations finds the nearest point on the boundary of the class regions. Another mechanism locally estimates this boundary by a tangent hyper-plane. This reveals the direction where the distortion from the original observation decreases while staying close to the boundary. These two mechanisms are iterated with the hope to converge to the nearest point inducing a wrong prediction.

The Phd thesis will investigate the power of these attacks in the context of deep neural network classification. This kind of attacks being not sustainable for hacking a lot of observations; a hybrid strategy will be devised to gain knowledge on the black-box while forging adversarial samples.

Defenses: There are two tracks of defense in the literature. The first tries to understand the fundamental reasons where the vulnerability of networks is stemming from [10, 13]. Some are inherited from the data, others from the training procedure. The representational dimension of the input space is often too big. The observations typically lie on narrower manifolds. The classifier however takes a decision whatever the input. This suggests that the boundaries extrapolated outside the data manifolds are doubtful, and the classifier should give up predicting in these regions of the space. This raises the issue of confidence measurement in the classifier prediction [7]. As for training, gradient penalization and adversarial training [9] are promising procedures to robustify the networks. Small perturbations can snow-ball into triggering a wrong prediction because the network function has large amplitude gradient at some points. Penalizing their norm in the training allows to smooth the network function. Adversarial training has the same goal by flattening the network function over balls centered on training data.

The second track comes from the watermarking community, where mechanisms detecting that an oracle attack is ongoing have been invented [1]. Defending follows: shutting down the black-box, delaying the outputs to slow down the attack, or randomizing the output to confuse the attacker [6].

References

[1] M. Barni, P. Comesana-Alfaro, F. Pérez-González, and B. Tondi. Are you threatening me? Towards smart detectors in watermarking. In *Media Watermarking, Security, and Forensics*, volume 9028, 2014.

[2] P. Comesaña, L. Pérez-Freire, and F. Pérez-González. The blind Newton sensitivity attack. *IEE Proc. Information Security*, 153, 2006.

[3] E. Dohmatob. Generalized No Free Lunch Theorem for Adversarial Robustness. *arXiv:1810.04065*, 2018.

[4] J. W. Earl. Tangential sensitivity analysis of watermarks using prior information. In *Security, Steganography and Watermarking of Multimedia Contents IX*, 2007.

[5] M. El Choubassi and P. Moulin. Noniterative algorithms for sensitivity analysis attacks. *IEEE Transactions on Information Forensics and Security*, 2(2):113–126, June 2007.

[6] M. El Choubassi and P. Moulin. On reliability and security of randomized detectors against sensitivity analysis attacks. *IEEE Transactions on Information Forensics and Security*, 4(3):273–283, Sep. 2009.

[7] A. Jaouen and E. Le Merrer. zoNNscan: boundary-entropy index for zone inspection of neural models. *CoRR*, abs/1808.06797, 2018.

[8] Y. Kilcher and T. Hofmann. The best defense is a good offense: Countering black box attacks by predicting slightly wrong labels. *arXiv:1711.05475*, 2017.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *Int. Conf. on Learning Representations*, 2017.

[10] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2924–2932. Curran Associates, Inc., 2014.

[11] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *CVPR*, 2017.

[12] S. J. Oh, M. Augustin, B. Schiele, and M. Fritz. Whitening black-box neural networks. *arXiv:1711.01768*, 2017.

[13] G. Ortiz-Jimenez, A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard. Hold me tight! influence of discriminative features on deep network boundaries, 2020.

[14] D. G. L. R, M. Pedergnana, and S. G. García. Smart sampling and incremental function learning for very large high dimensional data. *Neural Networks*, 78:75 – 87, 2016. Special Issue on Neural Network Learning in Big Data.

[15] S. Rane and A. Brito. A version space perspective on differentially private pool-based active learning. In *Proc. of IEEE Workshop on Information Forensics and Security (WIFS)*, December 2019.