Calibrated Multi-Probabilistic Prediction as a Defense against Adversarial Attacks*

Jonathan Peck[1,2], Bart Goossens[3], and Yvan Saeys[1,2]

[1] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, 9000, Belgium
[2] Data Mining and Modeling for Biomedicine, VIB Inflammation Research Center, Ghent, 9052, Belgium
[3] Department of Telecommunications and Information Processing, IMEC/Ghent University, Ghent, 9000, Belgium

Machine learning techniques have made great progress in recent years, obtaining state of the art performance in areas such as natural language processing [3] as well as image and speech recognition [2]. However, the theoretical properties of the deep neural networks responsible for this success remain poorly understood. At present, there is no theory which can satisfactorily explain the success of deep learning and many open questions remain [6]. A peculiar example of this lack of theoretical understanding is the existence of so-called *adversarial perturbations* [1]. These are small modifications to the inputs of a model which can drastically change its output, even though the alterations are completely insignificant.

In this work, we propose a novel defense against adversarial manipulation which aims to scale to realistic problems and provide non-trivial robustness. It is based on methods from *conformal prediction* and therefore enjoys frequentist guarantees of validity [4]. Empirical evaluations as well as theoretical results also support the idea that our defense can be scaled to realistic models. We evaluate our method against existing (oblivious) adversarial attacks as well as a white-box attack specifically designed to fool the MultIVAP. We find that these attacks have limited success when the norms of the perturbations are reasonably constrained.

The basic construction of the MultIVAP is as follows. Given any machine learning classifier, we use the *inductive Venn-ABERS predictor* algorithm by Vovk et al. [5] in a one-vs-all manner in order to obtain pairs of probabilities $(p_0^{(1)}, p_1^{(1)}), \ldots, (p_0^{(K)}, p_1^{(K)})$, one for each class. Intuitively, the pair $(p_0^{(i)}, p_1^{(i)})$ form lower and upper bounds on the probability that the given sample belongs to class i. These probabilities are then processed into a multi-probabilistic prediction by solving a mixed integer linear program (MILP). The output of the MultIVAP is the solution to this optimization problem, which consists of a vector of bits $(\alpha_1, \ldots, \alpha_K)$. Here, α_i indicates that we can accept the label i for

* We thank the NVIDIA Corporation for the donation of a Titan Xp GPU with which we were able to carry out our experiments. Jonathan Peck is sponsored by a fellowship of the Research Foundation Flanders (FWO). Yvan Saeys is an ISAC Marylou Ingram scholar.

the given input at the ε significance level, where $\varepsilon \in [0, 1]$ is a user-specified parameter.

Task	η	ε	Accuracy (baseline)	Adversarial error
Fashion-MNIST	0.3	24.20%	94.22% (93.84%)	18.96%
CIFAR-10	0.03	20.77%	83.36% (81.51%)	27.55%
Asirra	0.03	41.86%	88.56% (89.04%)	47.57%
SVHN	0.03	25.23%	96.81% (96.40%)	9.85%

Table 1: Results of the MultIVAPs on the adversarial white-box attack.

Table 1 shows experimental results when we evaluate the MultIVAP on four different image recognition tasks. For each task, we report several metrics: η, the ℓ_∞ norm bound on the magnitude of the perturbations; ε, the significance level at which these results were obtained; accuracy of the MultIVAP and accuracy of the underlying model; the adversarial error of the MultIVAP. Note that on three out of four tasks, the MultIVAP increases the accuracy of the classifier. Also, the adversarial error of the MultIVAP is significantly lower than that of unprotected machine learning classifiers evaluated against adaptive white-box attacks (these are almost invariably close to 100%). The computational overhead we incurred with this construction was roughly linear in the number of classes of the task.

We conclude that the MultIVAP is a computationally efficient procedure for protecting multi-class classifiers against adversarial perturbations. We make our code available at `https://github.com/saeyslab/multivap`.

References

[1] Battista Biggio and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning". In: *Pattern Recognition* 84 (2018), pp. 317–331.

[2] Jonathan Shen, Patrick Nguyen, Yonghui Wu, Zhifeng Chen, Mia X Chen, Ye Jia, Anjuli Kannan, Tara Sainath, Yuan Cao, Chung-Cheng Chiu, et al. "Lingvo: a modular and scalable framework for sequence-to-sequence modeling". In: *arXiv preprint arXiv:1902.08295* (2019).

[3] David So, Quoc Le, and Chen Liang. "The Evolved Transformer". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, June 2019, pp. 5877–5886.

[4] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

[5] Vladimir Vovk, Ivan Petej, and Valentina Fedorova. "Large-scale probabilistic predictors with and without guarantees of validity". In: *Advances in Neural Information Processing Systems*. 2015, pp. 892–900.

[6] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization". In: *arXiv preprint arXiv:1611.03530* (2016).