ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Detecting Adversarial Examples with Inductive Venn-ABERS Predictors

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

1- Ghent University
Dept of Applied Mathematics, Computer Science and Statistics
9000 Ghent, Belgium

2- imec-IPI-Ghent University
9000 Ghent, Belgium

3- VIB Inflammation Research Center
Data Mining and Modeling for Biomedicine
9052 Ghent, Belgium

Abstract. Inductive Venn-ABERS predictors (IVAPs) are a type of probabilistic predictors with the theoretical guarantee that their predictions are perfectly calibrated. We propose to exploit this calibration property for the detection of adversarial examples in binary classification tasks. By rejecting predictions if the uncertainty of the IVAP is too high, we obtain an algorithm that is both accurate on the original test set and significantly more robust to adversarial examples. The method appears to be competitive to the state of the art in adversarial defense, both in terms of robustness as well as scalability.

1 Introduction

The reliability of machine learning techniques in adversarial settings has been the subject of much research for a number of years already [1]. Early work in this field studied how a linear classifier for spam could be tricked by carefully crafted changes in the contents of spam e-mails, without significantly altering the readability of the messages. More recently, [2] showed that deep neural networks also suffer from this problem: deep convolutional neural networks (CNNs) can be fooled into returning erroneous predictions by slightly modifying a handful of pixels in the original image. Usually, these modifications are almost imperceptible to humans, calling into question the generalization ability of CNNs. Since this work, research interest in the phenomenon of so-called *adversarial examples* has increased substantially and many attacks and defenses have been proposed. Despite this, at the time of writing only one technique is generally accepted as having any noticable effect: adversarial training, which is a form of data augmentation with adversarial samples. However, even this method currently has too limited success. The Madry defense [3], for instance, achieves less than 50% adversarial accuracy on the CIFAR-10 data set even though state-of-the-art clean accuracy is over 95%.

[*] We gratefully acknowledge the support of the NVIDIA corporation who provided us with a Titan Xp GPU to conduct our experiments under the GPU Grant Program. Jonathan Peck is sponsored by a Ph.D. fellowship of the Research Foundation - Flanders (FWO). Yvan Saeys is an ISAC Marylou Ingram Scholar.

In this work, we propose to defend machine learning models against adversarial examples using *inductive Venn-ABERS predictors* (IVAPs). This construction, first proposed in [4], uses a computationally efficient procedure to hedge the predictions of an underlying *binary* classification model which otherwise would provide only point predictions. Specifically, every prediction $\hat{y} \in \{0, 1\}$ from the model is augmented with two probabilities p_0 and p_1, with $p_0 \leq p_1$, such that $p_0 \leq \Pr[y = 1 \mid x] \leq p_1$. That is, p_0 and p_1 are bounds on the true conditional probability that the label is 1 given the input x. The size of this interval, $p_1 - p_0$, serves as a natural measure of confidence that the IVAP has in the prediction of the model. If it is too large, we flag the model's prediction as unreliable. This flag can serve as actionable information for human operators, who might then defer to other experts or more expensive classification methods. It could also be used by automated systems: for instance, a file sharing service might refuse an upload if it cannot ascertain with sufficient confidence that a file is free of malware.

The effectiveness of this defense relies on the hypothesis that adversarial examples exist because the softmax probabilities commonly used in classification models are not *calibrated*, that is, they do not accurately reflect the true underlying conditional probability distribution. This idea was also put forward, for example, in [5]. By contrast, IVAPs enjoy provable guarantees on the calibration of their probabilities regardless of the underlying machine learning model [4].

2 The IVAP Defense

We consider the typical supervised learning setup for classification. There is a measurable *object space* $\mathcal{X}$ and a measurable *label space* $\mathcal{Y}$. In this work, we only consider the case where $\mathcal{Y} = \{0, 1\}$, so the classification is binary. We let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. There is an unknown probability measure P on $\mathcal{Z}$ which we aim to estimate. In particular, we have a class of $\mathcal{X} \to \mathcal{Y}$ functions $\mathcal{H}$ and a data set $S = \{z_i = (x_i, y_i) \mid i = 1, \ldots, m\}$ of i.i.d. samples from P. Our goal is to find a function in $\mathcal{H}$ which fits the data best in the sense of minimizing the empirical risk.

Algorithm 1 shows the pseudocode of our defense. A tunable precision parameter β thresholds the width of the interval $p_1 - p_0$. Predictions are only accepted if $p_1 - p_0 \leq \beta$, otherwise a special `REJECT` label is returned. In case of acceptance, the defense uses the estimate $p_1/(1 - p_0 + p_1)$ to correct the prediction from the underlying model, as in [4].

2.1 Experiments

We conduct experiments on the T-shirts and trousers classes from Fashion-MNIST [6], the cats vs dogs data set [7] and a subset of the CIFAR-10 dataset [8] which contains only images of airplanes or automobiles. For cats vs dogs, we resized all the samples to 64×64 pixels as the images come in various sizes. For all three sets, we rescaled the inputs to the $[0, 1]$ interval and randomly split the sets into disjoint training, calibration, validation and test data. The training

Algorithm 1: Detecting adversarial manipulation with IVAPs.

```
Data: precision $\beta \in [0,1]$, bag of examples $\wr z_1, \dots, z_n \wr \subseteq \mathcal{Z}$, object $x \in \mathcal{X}$, learning
      algorithm $A$
Result: An element of the set $\mathcal{Y} \cup \{$REJECT$\}$.
Use an IVAP on the hypothesis output by $A$ to compute $p_0, p_1$ for $x$.
if $p_1 - p_0 \leq \beta$ then
    Set $p \leftarrow \frac{p_1}{1 - p_0 + p_1}$.
    return 1 if $p > 0.5$ else 0
else
    return REJECT
end
```

Task	Clean Accuracy	Adversarial Accuracy
Cats vs dogs	89.87%	1.8%
Airplanes vs automobiles	96.69%	3.75%
T-shirts vs trousers	99.75%	3.43%

Table 1: Summary of performance indicators for the unprotected CNNs.

data was used to train a standard CNN for 500 (cats vs dogs) and 50 (T-shirts vs trousers, airplanes vs automobiles) epochs using the Adam optimizer [9]. The calibration data set was used to calibrate the IVAP and the validation data was used to tune the precision parameter β of the defense. The attacks we employed were projected gradient descent with random restarts [3], DeepFool [10], local search [11], the single pixel attack [12], NewtonFool [13], fast gradient sign [14] and the momentum iterative method [15]. All of these attacks were used to augment both the validation set as well as the test set, for tuning the precision parameter and subsequently evaluating the performance of the models. The training and calibration data sets contained no adversarials. The attacks were unbounded, meaning that there was no limit imposed on the norm of the generated perturbations. We used the implementations provided by the Foolbox library [16] and left all parameters at their default values. The implementation of the IVAP we used was due to Toccaceli.[1]

Results. Table 1 shows the results for the unprotected CNNs. The cut-off for the optimal β was determined by maximizing the difference $\text{TPR} - \text{FPR}$, known as *Youden's index* [17], on a data set consisting of adversarials for the underlying model generated on the validation set along with the validation set itself. In our setting, a sample is considered *positive* if it is accepted by the detector and *negative* if it is flagged. The thresholds are 0.01 for cats vs dogs and airplanes vs automobiles and 0.07 for T-shirts vs trousers. With these values of β, we obtain the results shown in table 2. The *Clean* row shows results on the clean test set; the *Adversarial* row shows results for the adversarials generated on the test set for the underlying model alone and the *Adapted* row shows the performance of the defense on the adversarial examples generated by our suite of attacks when

[1] `https://github.com/ptocca/VennABERS`

Task	Attack	Size	Accuracy	TPR	FPR
Cats vs dogs	Clean	3,988	72.94%	75.9%	51.15%
	Adversarial	22,599	52.56%	71.25%	61.14%
	Adapted	17,187	59.27%	100.0%	40.82%
	Custom ℓ_∞	3,145	1.69%	100.0%	100.0%
Airplanes vs automobiles	Clean	1,600	74.38%	73.21%	3.7%
	Adversarial	8,256	58.14%	0.0%	0.0%
	Adapted	6,586	93.99%	0.0%	5.18%
	Custom ℓ_∞	1,600	0.0%	0.0%	100.0%
T-shirts vs trousers	Clean	1,600	96.88%	96.86%	2.56%
	Adversarial	7,076	50.99%	0.0%	0.03%
	Adapted	5,923	77.53%	0.0%	22.01%
	Custom ℓ_∞	1,600	0.0%	0.0%	100.0%

Table 2: Summary of performance indicators for the IVAP defense on the different tasks.

the IVAP is taken into account. The defense suffers a reduction in clean accuracy each time. However, its accuracy on adversarial examples is always much higher than the accuracy of the unprotected model.

Custom white-box attack. The *Custom* ℓ_∞ row shows the results for our defense when run on adversarial examples generated on the test set using a custom attack we constructed specifically to fool the IVAP. This attack works by solving the following optimization problem:

$$\min_{\delta \in [-1,1]^d} \|\delta\| + c(s(x+\delta) - s_i)^2 \text{ subject to } x + \delta \in [0,1]^d.$$

Here, δ is the adversarial perturbation, $\|\cdot\|$ is the norm of choice (ℓ_∞ in this case), x is the original clean sample, c is a parameter, d is the data dimensionality, $s(x)$ returns the logit score assigned to x by the CNN and s_i is a target score we want our adversarial to achieve. This target score is determined from the calibration set, where a sample is picked that belongs to the target class. The idea is that whenever a new sample has the same score as an old sample in the calibration set, the result of applying algorithm 1 to it will be the same as applying the algorithm to the old sample because of the way the IVAP computes the probabilities. The constant c can be determined via binary search, as in the Carlini & Wagner attack [18]. Figure 1 shows the empirical cumulative distribution of the adversarial distortion introduced by our custom attack. We see that, although the attack is highly successful in circumventing our defense, the adversarials it produces have sufficiently high ℓ_∞ distortion so as to be clearly distinguishable from the originals. Note also that neither the IVAP nor the underlying CNN were exposed to these adversarials during training or calibration.

Comparison to Madry et al. We also compare our defense to the one proposed in [3], which is considered state of the art at the time of this writing.[2] We trained

[2] Code taken from `https://github.com/MadryLab/mnist_challenge`

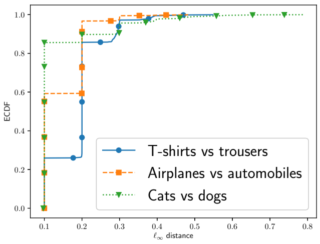

Fig. 1: Empirical cumulative distributions of the adversarial distortion produced by our custom ℓ_∞ white-box attack.

the same underlying CNNs for the same number of epochs with PGD adversarial training using the parameter settings Madry et al. recommend for CIFAR-10 (on cats vs dogs and airplanes vs automobiles) or MNIST (for T-shirts vs trousers). On the cats vs dogs data set, we achieved 54.94% clean and 3.5% adversarial accuracy; for airplanes vs automobiles, we have 82.06% clean and 29.19% adversarial accuracy and for T-shirts vs trousers we obtain 97.12% clean accuracy with 85% adversarial accuracy. By contrast, on the PGD adversarials that fool the Madry models, our defense obtains 61.64% (cats vs dogs), 45.81% (airplanes vs automobiles) and 72.75% (T-shirts vs trousers) accuracy respectively.

We conclude that our IVAP construction achieves higher robustness on cats vs dogs and airplanes vs automobiles. On T-shirts vs trousers it appears the Madry defense outperforms ours, which is not very surprising: the Madry defense is known to be highly robust on MNIST-like tasks but much less so on more difficult data sets. We believe that Madry outperforms us on "toy" data sets such as MNIST and Fashion-MNIST but cannot scale to more difficult tasks such as cats vs dogs or CIFAR-10. The results we present here indicate that IVAPs are a viable defense on more realistic classification tasks where Madry fails.

3 Conclusion

We have proposed using inductive Venn-ABERS predictors to protect machine learning models against adversarial manipulation of input data in the case of binary classification. Our defense uses the width of the uncertainty interval produced by the IVAP as a measure of confidence in the prediction of the underlying model, where the prediction is rejected in case this interval is too wide. The acceptable width is a hyperparameter of the algorithm which can be estimated using a validation set. The resulting algorithm is much less vulnerable to adversarial examples and appears to be competitive to the defense proposed by [3], which is state-of-the-art at the time of this writing. Avenues for future work include (1) generalizing the IVAP defense to multiclass and regression problems; (2) increasing clean and adversarial accuracy. The performance of the IVAP defense is still not ideal at this stage since clean accuracy is noticeably reduced. However, we believe these findings represent a significant step forward. As such, we suggest that the community further look into methods from the field of conformal prediction in order to achieve adversarial robustness at scale. To our knowledge, we are the first to apply these methods to this problem, although the idea has been mentioned elsewhere already [19, Section 9.3, p133].

References

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

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

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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

[5] Harm de Vries, Roland Memisevic, and Aaron Courville. Deep learning vector quantization. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016.

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

[7] Jeremy Elson, John JD Douceur, Jon Howell, and Jared Saul. Asirra: a CAPTCHA that exploits interest-aligned manual image categorization. In *Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS)*. ACM, 2007.

[8] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[11] Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial perturbations for deep networks. *arXiv preprint arXiv:1612.06299*, 2016.

[12] Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi. One pixel attack for fooling deep neural networks. *arXiv preprint arXiv:1710.08864*, 2017.

[13] Uyeong Jang, Xi Wu, and Somesh Jha. Objective metrics and gradient descent algorithms for adversarial examples in machine learning. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pages 262–277. ACM, 2017.

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR, abs/1412.6572*, 2015.

[15] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. *arXiv preprint*, 2018.

[16] 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.

[17] William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

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

[19] Yevgeniy Vorobeychik and Murat Kantarcioglu. Adversarial machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3):1–169, 2018.