Detecting Attacks with Conformal Test Martingales

Genghua Dong GENGHUA@KTH.SE
Roman Bresson BRESSON@KTH.SE
Henrik Boström BOSTROMH@KTH.SE
KTH Royal Institute of Technology, Stockholm, Sweden

Editor: Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

Keywords: conformal test martingales, adversarial examples, whitebox attacks

1. Introduction

Several techniques for attacking a target model by providing adversarial examples (AEs) have been developed in the past, where the AEs are original examples perturbed in a way that is inconspicuous to humans, misleading a model into an incorrect prediction. Techniques for generating attacks by AEs fall into two categories; *whitebox* and *blackbox* (Wiyatno et al., 2019), depending on whether the attacker has full access to the target model or not. Techniques of the former category, targeting DNNs, include FGSM (Goodfellow et al., 2014), PGD (Madry et al., 2017) and CW (Carlini and Wagner, 2017b).

Detecting whether a model is under attack is an important first step towards mitigating or avoiding adversarial attempts to change the outcome of the model. In order to test whether a model is under attack, i.e., whether the distribution of examples has shifted due to the presence of AEs, we propose to use conformal test martingales (Vovk, 2021). In this work, we will investigate whether attacks on a sequence of images can be detected by conformal test martingales. Specifically, we focus on *whitebox* attacks, as they are generally considered to be more powerful due to their full access to models (Wang et al., 2022; Ebrahimi et al., 2018; Carlini and Wagner, 2017a).

2. Empirical Investigation

We consider image sequences, randomly sampled with replacement from the STL-10 dataset (Coates et al., 2011). Each test sequence consists of a reference segment of length $T_{ref} = 50$ and a simulated attack segment of length $T_{atk} = 500$. To simulate attacks, each image in the attack segment is randomly replaced by an AE with probability $r \in [0, 1]$ (the attack rate). In each test sequence, all AEs are generated using a fixed attack method, selected from FGSM, PGD, and CW. Following (Madry et al., 2017), the standard attack parameters for CIFAR-10 (Krizhevsky et al., 2009) are adopted. We use FGSM with $\epsilon = 0.03$, and PGD with $\epsilon = 0.03$, $\alpha = 0.01$, and 7 steps. For CW attacks, we adopt the parameters $c = 5$, $\kappa = 0$ with 50 steps and 0.01 learning rate. This configuration is slightly stronger than the TorchAttacks (Kim, 2020) default, which uses $c = 1$, in alignment with common practice for generating more effective adversarial examples.

As prior studies, e.g., (Yun et al., 2025), (Huang et al., 2019), indicate that white-box attacks, such as FGSM, can cause non-trivial perturbations in the outputs of intermediate

layers within DNNs, we compute conformal test martingales based on embeddings of images from DNNs. Therefore, we adopt the method proposed in (Dong et al., 2025) to extract embeddings from a finetuned ResNet-18 model based on the STL-10 dataset. To increase sensitivity of detection, we further apply embedding augmentation via gradient-based attention (Selvaraju et al., 2020), tuned by $\gamma = 1$, and local intrinsic dimension (LID) (Ma et al., 2018) residual, tuned by $\beta = 1$. Thus, we obtain embeddings $\mathbf{Z}^{(l)} = [\mathbf{z}_1^{(l)}, \ldots, \mathbf{z}_T^{(l)}] \in \mathbb{R}^{T \times D}$ from a convolutional layer, where D is the dimension of the embedding and l is the convolutional layer index.

We conduct conformal testing following the procedure in (Vovk, 2021). First, a sequence of p-values is computed for each dimension d. Then, conformal test martingales are computed from the p-values using the Sleeper/Drifter algorithm (Vovk et al., 2022), with the parameters $M = 50$ and $r = 0.2$, yielding a martingale matrix $S^{(l)} \in \mathbb{R}^{T \times D}$. We further average the martingales across dimensions to obtain a martingale $\bar{S}^{(l)} \in \mathbb{R}^T$ per layer. Finally, we average the martingales across all convolutional layers to obtain the martingale used for detecting attacks. An attack is detected if the final martingale exceeds a predefined threshold of 100, which limits the false positive rate to at most 1%.

Our experiments indicate that attacks by FGSM, PGD and CW can be detected at an attack rate of $r = 0.4$. For illustration, Figure 1 shows the martingale calculated by the Sleeper/Drifter algorithm, where AEs are generated by FGSM.

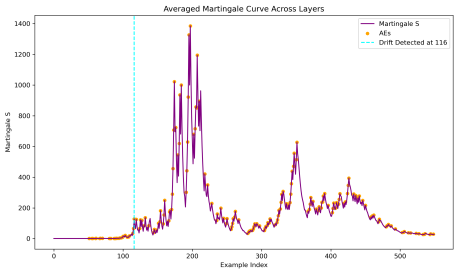

Figure 1: Global martingale across all convolutional layers. AEs are generated by FGSM.

3. Concluding Remarks

The presented investigation of using conformal test martingales based on augmented image embeddings shows that *whitebox* attacks to some extent can be detected, although at fairly high attack rates. Future work will focus on adaptive selection of key parameters, both in embedding augmentation and in the Sleeper/Drifter algorithm. We will also explore alternative inputs for computing p-values, e.g., based on cosine distances between embeddings in the reference and attack segments, to improve detection at lower attack rates. Another direction is to investigate which subsets of convolutional layers are more sensitive to attacks. At last, we will evaluate the effectiveness of our approach in conjunction with additional attack methods, e.g., *backdoor* attacks and *blackbox* attacks.

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 3–14, 2017a.

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, 2017b.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

Genghua Dong, Henrik Boström, Michalis Vazirgiannis, and Roman Bresson. Obtaining example-based explanations from deep neural networks. In *International Symposium on Intelligent Data Analysis*. Springer, 2025. Accepted, to appear.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. On adversarial examples for character-level neural machine translation. *arXiv preprint arXiv:1806.09030*, 2018.

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

Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4733–4742, 2019.

Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

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

Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

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.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020.

Vladimir Vovk. Conformal testing in a binary model situation. In *Conformal and Probabilistic Prediction and Applications*, pages 131–150. PMLR, 2021.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World (2nd ed.)*. Springer, 2022.

Yixiang Wang, Jiqiang Liu, Xiaolin Chang, Ricardo J Rodríguez, and Jianhua Wang. Di-aa: An interpretable white-box attack for fooling deep neural networks. *Information Sciences*, 610:14–32, 2022.

Rey Reza Wiyatno, Anqi Xu, Ousmane Dia, and Archy De Berker. Adversarial examples in modern machine learning: A review. *arXiv preprint arXiv:1911.05268*, 2019.

Sanggeon Yun, Ryozo Masukawa, Hyunwoo Oh, Nathaniel D Bastian, and Mohsen Imani. A few large shifts: Layer-inconsistency based minimal overhead adversarial example detection. *arXiv preprint arXiv:2505.12586*, 2025.