Adversarial Machine Learning for Reliable AI

A.Milani, V. Poggioni, A.E.Baia, G.Biondi, V.Franzoni, and M.Baioletti

Dip. Matematica e Informatica, Università di Perugia, Knowledge and Information Technology Lab. (KITLab)
{alfredo.milani,valentina.poggioni}@unipg.it

Abstract. The diffusion and the wide use of deep learning methods for AI systems posed significant security and privacy issues. Moreover, with the increasing popularity of online social networks, privacy-preserving systems for photos and images have received considerable attention. In this research we have developed a nested evolutionary algorithm able to craft very effective hard-label black-box image adversarial attacks based on the composition of Instagram inspired image filters. The system proposed produces natural looking artifacts-free images and produces adversarial attacks that cannot be distinguished from any other filtered images produced extensively every day, especially on social media platforms.

Keywords: Adversarial Machine Learning · Privacy Protection · Evolutionary computation · Multiobjective optimization

1 Research Description

The diffusion and the wide use of deep learning methods for AI systems posed significant security and privacy issues. From the security point of view, adversarial attacks showed that deep learning models can be easily fooled [14,18,17,7,8,20,13,9,15] while, from a privacy point of view, it has been shown that information can be easily extracted from dataset and learned model [2,19,12]. It has also been shown that attacking methods based on adversarial samples can be used for privacy preserving purposes [1,11,10,22,16]: in this case data are intentionally modified in order to avoid unauthorized information extraction by fooling the unauthorized software.

Attack reliability is strictly connected to its applicability in real-world scenarios and to its ability to bypass potential defense mechanisms; this is why the hard-label black-box setting (also called decision-based attack) and the transferability property gained increasing attention.

Besides the categorization in white-box and black-box methods, attacks can be classified as *restricted* or *unrestricted*, considering the amount of modifications they apply to the images in order to fool the systems. Restricted attacks generally use a L_p-norm distance to bound the modifications. The attacks are crafted with the aim of minimizing the differences between the original image and the adversarial one, even if it means having visible (more or less) artifacts. On the other hand, unrestricted attacks employ large and visible perturbations while

keeping the images realistic, natural looking and non-suspicious. The idea is to obtain images that can admit great differences from the original one but, beyond a direct comparison with the original one, they cannot be distinguished from any other real (maybe filtered) image.

In this research we are studying systems to craft reliable, effective and highly transferable attacks, also in the universal approach, by composing image filters. We decided to focus on generating non-targeted unrestricted adversarial attacks in a hard-label black-box scenario, since the limited knowledge and ability of the attacker is more similar to a real-world scenario, making the attack more challenging but with a wider applicability.

We proposed a gradient-free method based on nested evolutionary algorithms and multi-objective optimization that, given a set of commonly-used image filters, finds an optimal sequence of them that, when applied to an image is hardly detectable and causes the classifier to misclassify the image. To find a successful adversarial filter configuration, a population of candidate solutions is evolved and the solution quality is assessed by means of a fitness value.

This kind of attack offers multiple benefits: *(i)* it is naturally robust to detecting methods able to find noise, injected patterns and irregularities in the high frequencies image, *(ii)* it is naturally robust to masking-gradients (and their numerical estimate) defending methods. *(iii)* it produces natural looking artifacts-free images. *(iv)* it produces adversarial attacks that cannot be distinguished from any other filtered images produced extensively every day, especially on social media platforms. *(iv)* it requires a very low number of queries, that can also be limited by the algorithm itself, to find an attack.

Starting from the core algorithm, we proposed several versions of the system, with different aims and different characteristics:

- Universal approach: in this case the objective is to find *only one* filter composition able to fool the classifier for *almost all* the data points available in the dataset. In other words, the system is able to find an image-agnostic *universal perturbation* having a high attack success rate for many images. [3]
- Multi-network approach: it implements a multi-network training to increase the attack transferability; the fitness function is used to implement a sort of generalization ability induced by attacking more models simultaneously. In this case our goal is to find *one* adversarial perturbation that can fool *many* deep learning models avoiding the natural overfitting trap we can fall into when the attacks are crafted by using a unique reference model. This objective is much harder to achieve than attacking a single model but it guarantees complete (100%) transferability of the attack towards the reference networks [6].
- Multi-objective approach: the fitness function is exploited to optimize the, attacks combining more objectives, for example attack success rate with detection rate, as described in [3], or attack success rate with image quality assessment, as described in [4,5].

2 Applications developed / under development

- AGV: Evolutionary adversarial attacks for classification systems *(released on github)*.
- AGV-emotions: Adversarial attacks for face based emotion recognition systems *(released on github)*.
- AGV-object recognition: Adversarial attacks for object recognition systems *(under development)*.

3 Challenges/Prospects

This research focuses on a subsector of machine learning with a lot of perspectives ranging from the robustness/reliability of ML models to privacy preserving issues, passing through the security issues of smart city systems that are often based on image analysis and object recognition/classification.

An extension of this project that we are investigating move towards the building of adversarial attacks to systems based on smart sensors and in general IoT devices. The use of smart sensors increases the functionality of the devices and undoubtedly makes possible the concept of smart infrastructure; however, these sensors can also be used as vehicles to launch attacks on the devices and applications, undermining the infrastructure security and reliability. Attackers can use the sensors to convey malicious information in order to mislead the applications, disrupt the decision making process and lead the system to take different decisions. These sensor-based threats, fully-fledged adversarial attacks, can expose the smart infrastructure to significant risks also with catastrophic effects. Moreover smart sensors and IoT based systems are not in general protected with respect these kind of threats, even in the case of infrastructures implementing data security and privacy layers. Understanding these threats is necessary for researchers to design reliable solutions to increment the infrastructure security and resilience. Because machine learning and deep learning systems have been extensively used and applied in many IoT systems, the security implications of smart infrastructers need to be studied also following an adversarial machine learning approach.

From an algorithmic point of view we are working on a targeted version of the system where the goal is not simply cause the system to make any error, but a specific targeted error. Moreover we are studying on how use the multi-objective approach to optimize the trade-off between the performance of a target task and an associated privacy budget (as formalized in [21]).

References

1. Arcelli, D., Baia, A.E.B., Milani, A., Poggioni, V.: Enhance while protecting: privacy preserving image filtering. In: In Proc of IEEE/WIC/ACM Int. Conf. on Web Intelligence (WI-IAT '21) (2021)

2. Bae, H., Jang, J., Jung, D., Jang, H., Ha, H., Lee, H., Yoon, S.: Security and privacy issues in deep learning. arXiv preprint arXiv:1807.11655 (2018)
3. Baia, A.E., Di Bari, G., Poggioni, V.: Effective universal unrestricted adversarial attacks using a moe approach. In: In Proc of EvoAPPS 2021 (2021)
4. Baia, A.E.B., Biondi, G., Franzoni, V., Milani, A., Poggioni, V.: Lie to me: shield your emotions from prying software. Submitted to Sensors, MDPI (2022)
5. Baia, A.E.B., Milani, A., Poggioni, V.: Combining attack success rate and detection rate for effective universal adversarial attacks. In: In Proc of ESANN 2021 (2021)
6. Baia, A.E.B., Milani, A., Poggioni, V.: One for many: an instagram inspired black-box adversarial attack. In: Submitted to ICLR 2022 (2022)
7. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. 2017 IEEE Symposium on Security and Privacy (SP) pp. 39–57 (2017)
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. CoRR **abs/1412.6572** (2015)
9. Hayes, J., Danezis, G.: Learning universal adversarial perturbations with generative models. In: 2018 IEEE Security and Privacy Workshops (SPW). pp. 43–49. IEEE (2018)
10. Liu, B., Ding, M., Zhu, T., Xiang, Y., Zhou, W.: Using adversarial noises to protect privacy in deep learning era. 2018 IEEE Global Communications Conference (GLOBECOM) pp. 1–6 (2018)
11. Liu, Y., Zhang, W., Yu, N.: Protecting privacy in shared photos via adversarial examples based stealth. Secur. Commun. Networks pp. 1–15 (2017)
12. Mireshghallah, F., Taram, M., Vepakomma, P., Singh, A., Raskar, R., Esmaeilzadeh, H.: Privacy in deep learning: A survey. arXiv preprint arXiv:2004.12254 (2020)
13. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 86–94 (2017)
14. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2574–2582 (2016)
15. Mopuri, K.R., Ganeshan, A., Babu, R.V.: Generalizable data-free objective for crafting universal adversarial perturbations. IEEE transactions on pattern analysis and machine intelligence **41**(10), 2452–2465 (2018)
16. Sánchez-Matilla, R., Li, C., Shamsabadi, A.S., Mazzon, R., Cavallaro, A.: Exploiting vulnerabilities of deep neural networks for privacy protection. IEEE Transactions on Multimedia **22**, 1862–1873 (2020)
17. Shahin Shamsabadi, A., Sanchez-Matilla, R., Cavallaro, A.: Colorfool: Semantic adversarial colorization. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2020)
18. Shamsabadi, A.S., Oh, C., Cavallaro, A.: Edgefool: an adversarial image enhancement filter. In Proc of ICASSP 2020 (2020)
19. Shokri, R., Shmatikov, V.: Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC 2015. pp. 1310–1321 (2015)
20. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. CoRR **abs/1312.6199** (2014)
21. Wu, Z., Wang, Z., Wang, Z., Jin, H.: Towards privacy-preserving visual recognition via adversarial training: A pilot study. In: Proc of ECCV 2018 (2018)
22. Xue, M., Sun, S., Wu, Z., C.H., Wang, J., Liu, W.: Socialguard: An adversarial example based privacy-preserving technique for social images. ArXiv **abs/2011.13560** (2020)