One-pixel Attacks Against Medical Imaging: A Conceptual Framework

Tuomo Sipola(✉)[0000-0002-2354-0400] and Tero Kokkonen[0000-0001-9988-6259]

Institute of Information Technology,
JAMK University of Applied Sciences,
Jyväskylä, Finland
{tuomo.sipola, tero.kokkonen}@jamk.fi

Abstract. This paper explores the applicability of one-pixel attacks against medical imaging. Successful attacks are threats that could cause mistrust towards artificial intelligence solutions and the healthcare system in general. Nowadays it is common to build artificial intelligence models to classify medical imaging modalities as either normal or as having indications of disease. One-pixel attack is made using an adversarial example, in which only one pixel of an image is changed so that it fools the classifying artificial intelligence model. We introduce the general idea of threats against medical systems, describe a conceptual framework that shows the idea of one-pixel attack applied to the medical imaging domain, and discuss the ramifications of this attack with future research topics.

Keywords: Adversarial examples · Artificial intelligence · Cyber security · Machine learning · Model safety · Medical imaging · Healthcare · Security

1 Introduction

Modern networked and digitalized cyber domain is an extremely complex entity that comprises of unpredictable circumstances. As a part of the critical infrastructure, the healthcare sector is one of the major domains of interest from the cyber security perspective. In healthcare, there are numerous networked systems that can be targets for cyber attacks or intrusions. Finland's cyber security strategy indicates healthcare as an area which does not produce cyber security related solutions, services or products, but the activities of which are affected by cyber security, and where possible cyber security incidents will have a significant impact [13].

The state-of-the-art target in the development of the healthcare digitalization is the smart hospital environment. As defined by the European Union Agency for Network and Information Security (ENISA) [17]: *"A smart hospital is a hospital that relies on optimised and automated processes built on an ICT environment of interconnected assets, particularly based on Internet of things (IoT), to improve existing patient care procedures and introduce new capabilities."* According to

ENISA, one capability of the smart hospital environment are devices that lead to overall smartness. There are numerous systems used in the medical domain with capability of autonomic classification or diagnosis based on machine learning (ML) or deep learning (DL) [1, 10, 15].

Medical imaging technologies such as X-rays, tomography methods and whole-slide imaging digital pathology have become more widespread in the modern medical practice [2, 6]. However, new technologies attract malicious actors who want to profit from the misuse of those technologies or otherwise reach their goals by disrupting normal operations. The medical domain is an especially lucrative target for cyber criminals because of the sensitive nature of the data. For example, in Finland a psychotherapy service and 40,000 of its customers were blackmailed causing public mistrust towards healthcare [7, 8]. This causes long-term side effects from which it might take considerable time to recover. Similar kind of mistrust could be directed to medical imaging systems. Even if such doubts are not known among the public, the experts using imaging systems might lose their trust in AI-based models, and when such models remain in use, their misdiagnoses could cause unneeded overload in the healthcare system.

In this paper, we describe a framework to conduct one-pixel attacks against medical imaging. The remainder of this paper is organized as follows. Section 2 introduces the fooling of AI models using adversarial examples. In section 3, the attack framework is described. Finally, discussion about future research topics is presented in section 4.

2 Adversarial Examples

Fooling AI models using adversarial examples is a known threat. There are many attacks against deep neural networks that analyze images, especially when the goal is image classification. Most of the known attacks are iterative and white-box type, i.e., the inner configuration of neural network models is available to the attacker. However, some defences are available: gradient masking hides the gradient so that attack methods cannot use it, robust optimization uses attacks to re-train the model to be more resistant against attacks and detection methods try to identify attacks again before the input is being passed to the actual AI model [18].

The field of medical imaging is not immune to adversarial attacks. There are examples of crafting images and patches that create unwanted results when using an AI classifier in the medical domain [5, 12, 14]. Ma et al. noted that medical deep neural network models are more vulnerable than those used for natural image detection. However, simple detectors are able to capture the majority of adversarial examples because they contain differing fundamental features [11]. Finlayson et al. demonstrated that the use of projected gradient descent (PGD), natural patches and adversarial patches is effective against fundoscopy, X-ray and dermoscopy imaging [5]. Finlayson et al. have also raised the question of when to intervene regarding these vulnerabilities in medical imaging systems. Acting early could build more resilient systems but also hinder agile development. They

describe the problem of adversarial images similar to the cat-and-mouse game of cyber defence against hacking. As a solution they suggest amending regulatory best practices, for example hash-based fingerprinting of images [4].

One-pixel attacks are a known method of fooling neural network models. Changing just one pixel in the image causes the model to classify an image as being of another class label than the image is in reality. Differential evolution (see, e.g. [3]) can be used to find the optimal perturbation to change the predicted class label of an image. The one-pixel perturbation is encoded with x-y coordinates and RGB values, so that each perturbation is a vector of five elements. This kind of attack applies to different network structures and image sizes but could benefit from more advanced optimization methods [16]. There have been research concerning attack attempts against medical imaging using one-pixel attacks. Although a simplified case of pose estimation of surgical tools, Kügler et al. find adversarial examples near the decision boundary, creating vulnerable regions inside the images [9].

3 Attack Framework

A straightforward way of using an artificial intelligence (AI) solution is to classify medical images. The images are classified either normal or as having indications of disease. This information is accompanied with a score, which indicates how much the image is seen as part of its class. Attacking against medical imaging can be thought as a way of creating mistrust against the healthcare system. The basic principle can be applied in two ways, from normal to indications of disease, and vice versa. Firstly, we have a normal image as a starting point. This image is modified so that the AI model will instead predict the image as having indications of disease. Such a misdiagnosis could create unnecessary use of medical resources. It could also undermine the trust in systems using an AI model because the are producing less accurate results. Secondly, we have an image with indications of disease as a starting point. After appropriately modifying the image, the AI model will classify it as normal. This approach could lengthen the time after which the patient gets treatment. Such misclassifications could be even fatal. These factors could undermine the trust in the healthcare system.

Building and deploying an AI model using machine learning methods is usually broken into two major steps. The first one is the actual training of the model, during which the training images are used to teach the AI model to carry out the classification task as efficiently as possible within the constraints of the training. The second step is the deployment of the AI model so that it predicts or classifies completely unknown images, yielding a result: the classification and the score. If the input images are engineered to intentionally create wrong classification of the said image, we speak of adversarial examples. The image itself could look like healthy tissue; however, the engineered adversarial example could include information that fools the AI model. One such engineering attempt could be a one-pixel attack that changes only one pixel in the image to fool the AI model. This setup is schematically described in Figure 1, which indicates the train-

ing and deployment for predictive/diagnostic use. The one-pixel attack would be performed by modifying the input images the class label of which is being predicted.

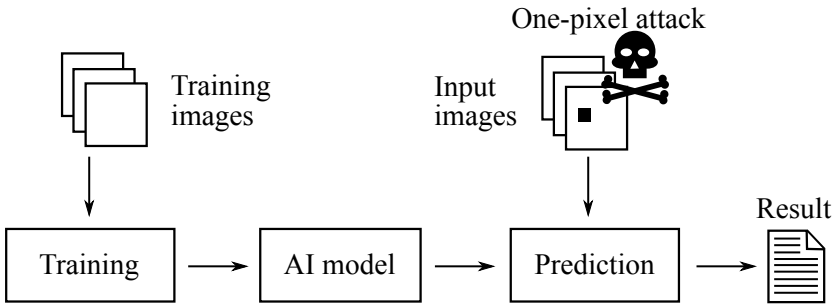

Fig. 1. A schematic presentation of the one-pixel attack against a machine learning model. Adapted from authors' previous paper [14].

Performing the one-pixel attack can be achieved by searching for these images using optimization methods. As seen in the study by Su et al. [16], differential evolution is one suitable candidate for the optimization problem. The problem of finding an adversarial example can be thought as a challenge of finding the necessary change in order to achieve a measurable goal. The goal, measured by a cost function, is to get the AI model produce wrong results. As said, the AI model usually returns a score indicating how confident it is in the classification result. The score is usually expressed in the range of $[0, 1] \in \mathbb{R}$, and it is suitable for acting as the cost function. This attack is a black-box solution because the target AI model is only needed for feeding input and querying the classification result. The inner workings of the AI model are not needed because it is only used as part of the cost function during the optimization. Figure 2 gives the basic idea behind the differential optimization process, where a population of attack images is created. This population is then used as input to the AI model, which predicts the class label and gives a probability score for it. These results are then evaluated, and images that are better at fooling the AI model are retained as the precursors for the future populations. This way of thinking is geared towards the differential evolution method, but it equally applies to many other optimization methods.

Figure 3 showcases the working principle of differential evolution in this scenario. The process is started by giving it an input image and information towards which class label we want the AI model to be fooled. The logic is the same as with the earlier images; however, this is a more detailed view of the differential evolution process when searching for adversarial examples. This process takes

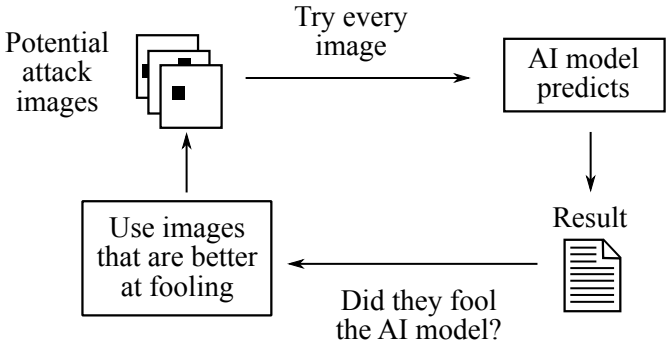

Fig. 2. Basic idea behind the optimization procedure.

one image as a source for its input population, which is initialized based on random or search space spanning one-pixel permutations. In other words, each image in the population will be based on the same source image but have one pixel changed to another by the permutation. The evolutionary process is used to change the images. Then the effectiveness of these attack images is evaluated by using the black-box AI model. This, in turn, makes it possible to select the best images that confuse the AI model. If any image in the population fooled the AI model with acceptable certainty, we can stop and declare that an adversarial image has been found. If no acceptable images can be found, and the optimization does not converge, the search should be stopped.

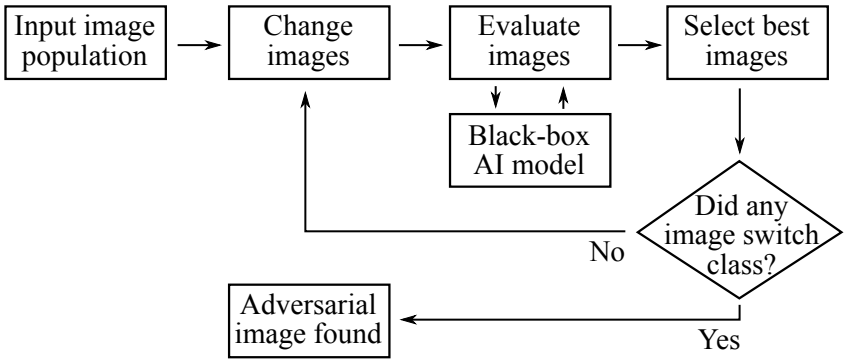

Fig. 3. Block diagram of the procedure of finding an adversarial image using differential evolution.

4 Discussion

The integrity and robustness of medical systems needs to be tested and hardened against known attacks. Furthermore, deeper inspection of robust behavior of machine learning systems will benefit the systems in the medical domain. Such inspection could be directed at least towards two directions. First of them are theoretical bounds of machine learning systems that warrant more detailed mathematical analysis. Understanding the behavior of AI models and the boundaries of their inaccurate behavior would help create more trustworthy solutions. Secondly, employing robustness strategies during training could harden the AI models against adversarial examples that misuse the theoretical bounds. Bringing these new mitigations using theoretical bounds and defensive robustness strategies into production will be a challenge; however, this ultimately ensures that the professionals and the public trusts in these efficient tools that make the healthcare process faster and more accurate.

One-pixel attack is a decent example of an attack against automatic analysis and diagnosis in medical domain, especially when the pixel is not noticeably prominent. By affecting merely one pixel of an image under analysis, the diagnosis can be incorrect, which can lead to improper treatment. Since the logic of the attack is well understood, it is possible to create uncertainty with a proper attack vector to insert the image into the diagnosis pipeline. The latest real-life attacks have demonstrated that there is a desire to conduct cyber attacks against medical systems, and furthermore, medical systems are seen as valuable targets.

The next step of the continuing research is to research the feasibility and effectiveness of the attack in a real-life scenario with a real dataset and machine learning algorithms. As the concept is quite evident and its targets abundant, studying the feasibility and effectiveness of the attack seems to be a potential way forward.

Acknowledgments. This research is funded by the Regional Council of Central Finland/Council of Tampere Region and European Regional Development Fund as part of the Health Care Cyber Range (HCCR) project of JAMK University of Applied Sciences Institute of Information Technology. The authors would like to thank Ms. Tuula Kotikoski for proofreading the manuscript.

References

1. Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., Greenspan, H.: Chest pathology detection using deep learning with non-medical training. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), pp. 294–297 (2015). DOI 10.1109/ISBI.2015.7163871
2. Doi, K.: Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics **31**(4–5), 198–211 (2007). DOI 10.1016/j.compmedimag.2007.02.002
3. Feoktistov, V.: Differential evolution. Springer (2006). DOI 10.1007/978-0-387-36896-2

4. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019)
5. Finlayson, S.G., Chung, H.W., Kohane, I.S., Beam, A.L.: Adversarial attacks against medical deep learning systems. arXiv e-print (2019)
6. Ghaznavi, F., Evans, A., Madabhushi, A., Feldman, M.: Digital imaging in pathology: whole-slide imaging and beyond. Annual Review of Pathology: Mechanisms of Disease **8**, 331–359 (2013)
7. Huhtanen, J.: Potilaiden tietoja vietiin psykoterapiakeskuksen tietomurrossa, yritys kertoo joutuneensa kiristyksen uhriksi. Helsingin Sanomat (2020). URL https://www.hs.fi/kotimaa/art-2000006676407.html
8. Kleinman, Z.: Therapy patients blackmailed for cash after clinic data breach. BBC News (2020). URL https://www.bbc.com/news/technology-54692120
9. Kügler, D., Distergoft, A., Kuijper, A., Mukhopadhyay, A.: Exploring adversarial examples. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, pp. 70–78. Springer (2018)
10. Latif, J., Xiao, C., Imran, A., Tu, S.: Medical imaging using machine learning and deep learning algorithms: A review. In: 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp. 1–5 (2019). DOI 10.1109/ICOMET.2019.8673502
11. Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F.: Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition **110**, 107,332 (2020). DOI 10.1016/j.patcog.2020.107332
12. Paschali, M., Conjeti, S., Navarro, F., Navab, N.: Generalizability vs. robustness: adversarial examples for medical imaging. arXiv e-prints (2018)
13. Secretariat of the Security Committee: Finland's Cyber security Strategy, Government Resolution 3.10.2019 (2019). URL https://turvallisuuskomitea.fi/wp-content/uploads/2019/10/Kyberturvallisuusstrategia_A4_ENG_WEB_031019.pdf
14. Sipola, T., Puuska, S., Kokkonen, T.: Model fooling attacks against medical imaging: A short survey. Information & Security: An International Journal **46**(2), 215–224 (2020). DOI 10.11610/isij.4615
15. Soumik, M.F.I., Hossain, M.A.: Brain tumor classification with inception network based deep learning model using transfer learning. In: 2020 IEEE Region 10 Symposium (TENSYMP), pp. 1018–1021 (2020). DOI 10.1109/TENSYMP50017.2020.9230618
16. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation **23**(5), 828–841 (2019). DOI 10.1109/TEVC.2019.2890858
17. The European Union Agency for Network and Information Security (ENISA): Smart Hospitals, Security and Resilience for Smart Health Service and Infrastructures. Tech. rep. (2016). DOI 10.2824/28801
18. Xu, H., Ma, Y., Liu, H.C., Deb, D., Liu, H., Tang, J.L., Jain, A.K.: Adversarial attacks and defenses in images, graphs and text: A review. International Journal of Automation and Computing **17**(2), 151–178 (2020). DOI 10.1007/s11633-019-1211-x