Model Fooling Attacks Against Medical Imaging: A Short Survey

Tuomo Sipola, Samir Puuska and Tero Kokkonen

Abstract This study aims to find a list of methods to fool artificial neural networks used in medical imaging. We collected a short list of publications related to machine learning model fooling to see if these methods have been used in the medical imaging domain. Specifically, we focused our interest to pathological whole slide images used to study human tissues. While useful, machine learning models such as deep neural networks can be fooled by quite simple attacks involving purposefully engineered images. Such attacks pose a threat to many domains, including the one we focused on since there have been some studies describing such threats.

Key words: deep learning, artificial neural networks, adversarial images, machine learning, medical imaging

Introduction

Artificial Intelligence (AI) based solutions, especially deep learning based on neural networks, are widely used in the medical domain. For example, AI is used for helping and automatizing cancer diagnosis based on image data. The benefit of this approach is to relieve experts to work on more important tasks while automated systems can inspect images and give initial recommendations.

If an attacker can fool the AI processing, ramifications can be devastating. Such attacks may result in incorrect treatment procedures, causing extreme circumstances with a worst case scenario of losing human lives. In addition, wrong diagnoses could

Tuomo Sipola
JAMK University of Applied Sciences, Jyväskylä, Finland, e-mail: tuomo.sipola@jamk.fi

Samir Puuska
JAMK University of Applied Sciences, Jyväskylä, Finland, e-mail: samir.puuska@jamk.fi

Tero Kokkonen
JAMK University of Applied Sciences, Jyväskylä, Finland, e-mail: tero.kokkonen@jamk.fi

© 2020. Licensed under the Creative Commons BY-NC 4.0 license. This is the authors' postprint version of the article. The original article appeared as: Tuomo Sipola, Samir Puuska, and Tero Kokkonen. Model Fooling Attacks Against Medical Imaging: A Short Survey. Information & Security: An International Journal 46(2): 215-224, 2020. https://doi.org/10.11610/isij.4615

undermine the public trust in medical professionals. This paper presents a short survey of model fooling attacks against neural networks in the medical domain.

Fooling neural networks is an important subject because machine learning models are widely used in medicine for automating many processes and for helping with diagnosis. For example, Rai et al. proposed a convolutional neural network for healthcare assistant [29] while Rastgar-Jazi and Fernando used neural networks for detecting heart abnormalities from electrocardiogram (ECG) data [30]. Similarly, authors of [31] used neural networks for prediction and prevention of heart attacks from ECG data while Murugesan and Sukanesh used neural networks for detecting brain tumors in electroencephalograms (EEG) signals [24]. Syam and Marapareddy discussed three different scenarios of classification problems, where one is skin lesion (cancer) classification from images [34]. As can be seen, these machine learning solutions are useful for many medical applications.

Effectiveness of neural network based deep learning is based on the used algorithm and learning data. If learning dataset is inadequate or contains incorrect information, results will be inaccurate. Similarly, if there are known weaknesses in the used algorithm, they can be compromised. In that sense, AI components can be attacked and fooled to behave incorrectly. As an example of a weakness, Afifi and Brown explore how white balance of photography impact the performance of deep neural networks [2], while authors of [21] generated adversarial noise for fooling the neural networks. Gu et al. discussed about gradient shielding method for understanding the vulnerabilities in neural networks [13]. MOEA-APGA is an algorithm for achieving targeted attacks against neural networks [9], and another similar algorithm is called DeepFool implemented for computing perturbations that fool neural networks [23]. In a medical domain, Chuquicusma et.al. studied about fooling radiologists for lung cancer diagnosis [8]. As can be seen, many such attack vectors exist.

As a powerful machine learning method, deep learning has also been applied to images related to pathology, for example, trying to classify images of cancer whole slide images (WSI). Serag et al. present an overview of the application of artificial intelligence for pathology and tissue analytics [32]. As another example, convolutional neural networks have been used for nuclear segmentation, which is an important part of tissue cancer grading [18]. Deliberately produced wrong segmenatation could result in wrong diagnoses. Pre-trained convolutional neural networks have been compared to training from scratch using the Kimia Path24 dataset, with results indicating that pre-trained networks are quite competitive [15]. Using such pre-trained models creates a possibility of hidden attacks trained into the model or abusing known deficiencies of such models.

In this study, we collected a list of relevant research papers concerning medical imaging and attacks against neural networks. We queried the publicly available Google Scholar database to identify publications relevant to deep neural network fooling, deep neural networks in medical imaging and deep neural networks fooled in that domain. The results of this short survey should be useful for anyone trying to understand the vulnerabilities of neural networks in specific domains. Moreover, the use of them in medical imaging raises the question of reliability and robustness when targeted by such attacks. As can be seen, targeted attack against neural networks

in medical domain is a realistic scenario. From the attacker perspective, medical domain can be considered as a valuable target because of the critical ramifications of possible attack. In addition, there are known vulnerabilities with neural networks that are highly used in medical domain.

Below, we present the medical imaging domain and then discuss about machine learning regarding that domain. Next sections describe the state of fooling deep neural networks and how it has been applied to the medical domain. Finally, we present our concluding thoughts.

Short Introduction to Whole Slide Images in Cancer Diagnosis

Quick and affordable laboratory cancer diagnosis methods are of great importance. One of the well-established methods is light microscopy with a stain, such as haematoxylin and eosin (H&E). The H&E stain makes various tissue components visible, allowing diagnosis based on e.g. their morphological features. [14]

The advent of digital pathology and whole-slide imaging (WSI) have provided a computerized way to analyze and share the results of light microscopy. By digitizing the tissue images, a variety of automated methods can be used to perform image analysis, annotation, and workflow improvements. Turning glass slides to a digital format requires a slide scanner, which digitizes the slide using specialized format that allows e.g. various zoom levels and metadata to be stored in one data file. This data can then be easily shared, further processed using a variety of tools, and even easily used in teaching in a virtualized microscopy environment. [1]

Distinguishing between benign and malignant tumors is essential for accurate prognosis. One of the features that separate the two is differentiation and anaplasia. In general, benign tumors consist of cells that resemble the tissue where they originated from. They retain much of the functionality and morphology of their non-transformed counterparts but may invade surrounding tissue. Malignant tumors, on the other hand, lose their resemblance to their normal counterparts and become undifferentiated (anaplastic). This change results in noticeable change in cell morphology, and it is possible to observe this using light microscopy and stains. These observable changes include variations in size and shape, nuclear abnormalities and atypical mitoses. Assigning a value to this differentiation is called grading. The criteria and schemes are dependent on the type of tumor. [19]

For breast cancer, observing mitoses has been shown to be a good predictor for tumor development and prognosis. In order to proliferate, tumor cells need to overcome various limitations that prevent ordinary cells from dividing indefinitely. These mutations may result in increased cell-cycle activity, and even majorly affect the mitosis process itself by causing atypical-looking cell divisions which may be visually observed using light microscopy. [27] Figure 1 shows an example of breast cancer WSI from Al-Janabi et al. [4].

Detecting abnormal morphology and quantifying the number of various cell features is a good candidate for automatization via machine learning and computer

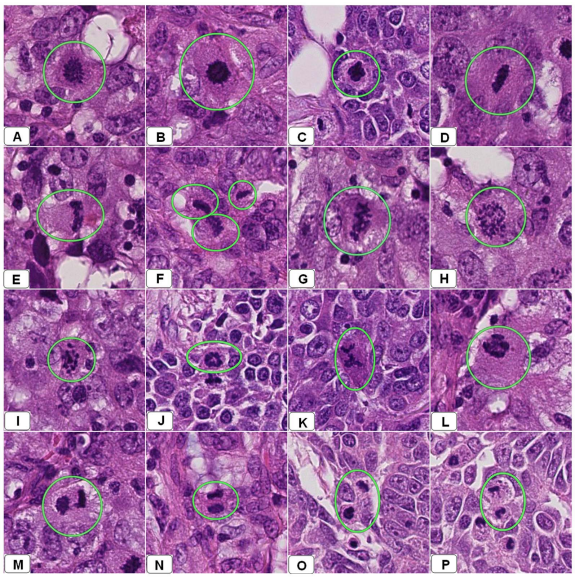

Fig. 1 Example of WSI showing several breast resections with infiltrative ductal carcinoma. Figure courtesy of Al-Janabi et al. [4], distributed under the terms of the Creative Commons Attribution License.

vision methods. WSI with sufficient quality can be automatically annotated. Digital pathology is expected to improve convenience and quality of the process. Nam et al. provide an introduction to digital pathology aimed at healthcare professionals [25]. Furthermore, Komura and Ishikawa made a short review of machine learning methods for histopathological image analysis, listing seven whole slide image datasets and 21 hand annotated histopathological datasets [16].

Fooling Deep Neural Networks

Deep neural networks and deep learning in general refers to a field of study, were complex concepts are learned from simpler representations by creating an interconnected network of activation functions and weights [11]. Due to their nature, these networks may contain flaws which make them susceptible to various classes of errors. These imperfections may be used maliciously to force the network into making an erroneous prediction. There have been several successful attempts at creating methods to fool deep neural networks.

One approach is to give an adversarial images as input to the classifier. Nguyen, Yosinski and Clune used an evolutionary algorithm optimization method to generate unrecognizable images to the human eye. Those images fooled a neural network to classify them as an object with high certainty, even though it should not have. They describe these images as costly exploits that could be used against deep neural networks [26]. Moosavi-Dezfooli, Fawzi and Frossard propose the DeepFool algorithm that efficiently generates adversarial images and quantifies the robustness of image

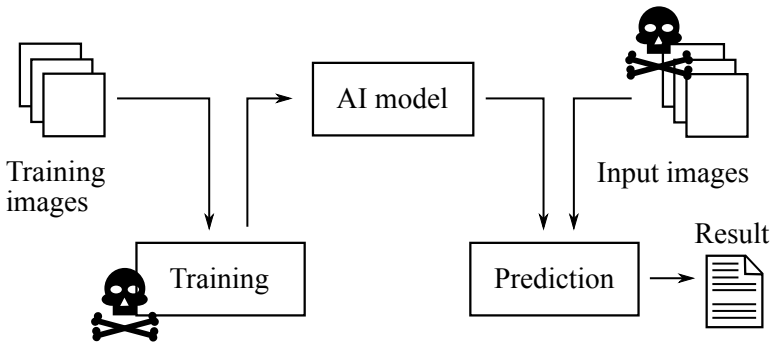

Fig. 2 A schematic picture of AI model fooling. The input images could be altered using adversarial images, patches or one pixel attacks. In addition, the training process itself could be tampered with.

classifiers [23]. An adversarial image is wrongly classified as something else than what the image clearly represents to the human eye. In the paper, a slight perturbation was added to an image of an animal to misclassify a whale as a turtle. Furthermore, it is possible to create adversarial 2D images robust to noise, distortion and affine transformations, and even adversarial 3D printed objects (a turtle) [6].

Adversarial patches are images that can be placed inside another image to fool a neural network classifier. Brown et al. have shown the effectiveness of such images [7]. It is easy to see that inserting such patches to medical images could yield similar results, resulting in a false classification.

There have been research about changing only one pixel of an image to cause it to be classified as another object [33]. It is remarkable that a change of color in one pixel could fool the neural network. A move towards a more theoretical understanding of one pixel-attacks and incorrect mapping to low dimensional manifold has also been proposed. This makes it easy to find localized areas where one pixel attacks should be more effective. [17]

Backdoored images can be created when attacking the learning stage of a neural network. These malign models can be deployed to production, and the fault is only revealed when the bad image is given as an input, resulting in wrong classification. Outsourced training opens the possibility of creating backdoored neural networks that behave badly on input specified by the attacker [12]. In a similar scheme, called poisoning attack, artificially poisoned data being sent to a model gradually change the model to conform to the attacker's goals. Yang et al. used an autoencoder (instead of the more traditional direct gradient method) to generate poisoned input data for deep neural networks [37].

The evident vulnerability of neural networks against several types of attacks is alarming because these methods are being proposed in several real world domains. See [3] for a survey of adversarial attacks against deep learning in computer vision. The authors not only list several attacks but also include defences. They conclude that there is a threat against safety and security critical applications.

Figure 2 shows a schematic presentation of the possible attack routes described above. In this paper we have identified two parts of the AI process, which could be targeted.

Fooling Deep Learning in Medical Imaging and Pathology

Although a rather new concern, the vulnerabilities intrinsic to neural network solutions have been identified by the medical community. For evident health reasons the accuracy and robustness of methodology in the medical domain is very important. Tizhoosh and Pantanowitz list challenges and opportunities related to artificial intelligence and digital pathology. One of the challenges concerns adversarial attacks and the shakiness of deep decisions made by neural networks [36]. This fundamental lack of robustness could be one avenue of future research.

Adversarial examples in medical imaging can change the behavior of classifiers and segmentation, illustrating the lack of robustness in the neural network models. Such approach can also be used for model evaluation [28]. Vulnerability during segmentation could lead to wrong representation of reality during the following stages of diagnosis. Again, the less understood and erratic boundaries of classification are a concern that enable an attack vector.

Deep learning networks classifying X-ray images are also vulnerable to attacks [35]. Being perhaps the most familiar scenario to the public, X-ray image processing is a natural target for automation. However, these kinds of perturbation attacks show that the models can be fooled.

Finlayson et al. successfully use adversarial attacks against medical imaging in three domains: fundoscopy, chest X-ray images and dermoscopy. They also present a risk model for the machine learning pipeline [10]. Patch attacks and projected gradient descent both seem to work against real world images, reducing the reliability of the classifier. However, neural networks can be made robust against perturbation attacks by exploiting the structure of the optimization task [20].

Not all uses of these methods are harmful. It is possible to use existing medical imaging data to generate more training data and to tackle uneven class balance using various methods, including generative adversarial networks [22]. The methods described above can also be used for beneficial inpainting of missing areas in biomedical imaging. Armanious et al. used generative adversarial networks to inpaint missing areas or incomplete medical images [5].

The identified fooling methods are listed in Table 1. As can be seen, some of the fooling methods have been used in the medical domain. It should be noted that training process tampering is probably more difficult to execute in practice.

Table 1 Fooling methods against deep neural networks and those in the medical domain

Method	References	Medical domain
Adversarial images	[26],[23],[6]	[28],[35],[10],[22],[5]
Adversarial patches	[7]	[10]
One pixel attack	[33],[17]	
Training process tampering	[12],[37]	

Conclusion

Although modern neural networks have proven useful for detecting cancerous cell growth, it is possible to mislead these algorithms. There have been research exploits against deep learning methods, even in the field of pathology. Such exploits include specifically engineered adversarial images, adversarial patches put on actual images, one pixel attacks and attacks focusing on fooling the training process. The scientific studies this short survey inspected include all those attacks. Even medical imaging is not safe from them, which promotes further study of the underlying causes and robustness problems stemming from the structure of neural networks. The expert opinions from the medical community will also broaden the undersanding of the effect of these types of attacks.

Acknowledgements This research is partially funded by the Cyber Security Network of Competence Centres for Europe (CyberSec4Europe) project of the Horizon 2020 SU-ICT-03-2018 program.

References

1. Aeffner, F., Zarella, M.D., Buchbinder, N., Bui, M.M., Goodman, M.R., Hartman, D.J., Lujan, G.M., Molani, M.A., Parwani, A.V., Lillard, K., et al.: Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. Journal of pathology informatics **10** (2019)
2. Afifi, M., Brown, M.S.: What else can fool deep learning? addressing color constancy errors on deep neural network performance. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 243–252 (2019)
3. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access **6**, 14410–14430 (2018)
4. Al-Janabi, S., van Slooten, H.J., Visser, M., Van Der Ploeg, T., Van Diest, P.J., Jiwa, M.: Evaluation of mitotic activity index in breast cancer using whole slide digital images. PloS one **8**(12) (2013)
5. Armanious, K., Mecky, Y., Gatidis, S., Yang, B.: Adversarial inpainting of medical image modalities. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3267–3271. IEEE (2019)
6. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: Proceedings of the 35th International Conference on Machine Learning, vol. 80, pp. 284–293. PMLR, Stockholm, Sweden (2018)

7. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial Patch. arXiv e-prints (2018)
8. Chuquicusma, M.J., Hussein, S., Burt, J., Bagci, U.: How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), pp. 240–244. IEEE (2018)
9. Deng, Y., Zhang, C., Wang, X.: A multi-objective examples generation approach to fool the deep neural networks in the black-box scenario. In: 2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC), pp. 92–99. IEEE (2019)
10. Finlayson, S.G., Chung, H.W., Kohane, I.S., Beam, A.L.: Adversarial attacks against medical deep learning systems. arXiv e-print (2019)
11. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016). http://www.deeplearningbook.org
12. Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S.: Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access **7**, 47230–47244 (2019)
13. Gu, Z., Hu, W., Zhang, C., Lu, H., Yin, L., Wang, L.: Gradient shielding: Towards understanding vulnerability of deep neural networks. IEEE Transactions on Network Science and Engineering (2020)
14. Junqueira, L.C.U., Carneiro, J.: Basic Histology Text & Atlas. McGraw-Hill Professional (2005)
15. Kieffer, B., Babaie, M., Kalra, S., Tizhoosh, H.R.: Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–6. IEEE (2017)
16. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. Computational and structural biotechnology journal **16**, 34–42 (2018)
17. 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)
18. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE transactions on medical imaging **36**(7), 1550–1560 (2017)
19. Kumar, V., Abbas, A.K., Aster, J.C.: Robbins basic pathology, 10th edn. Elsevier, Philadelphia, USA, Saunders (2017)
20. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv e-prints arXiv:1706.06083 (2019)
21. Mihajlović, M., Popović, N.: Fooling a neural network with common adversarial noise. In: 2018 19th IEEE Mediterranean Electrotechnical Conference (MELECON), pp. 293–296. IEEE (2018)
22. Mikołajczyk, A., Grochowski, M.: Data augmentation for improving deep learning in image classification problem. In: 2018 international interdisciplinary PhD workshop (IIPhDW), pp. 117–122. IEEE (2018)
23. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582. IEEE (2016)
24. Murugesan, M., Sukanesh, R.: Automated detection of brain tumor in eeg signals using artificial neural networks. In: 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies, pp. 284–288. IEEE (2009)
25. Nam, S., Chong, Y., Jung, C.K., Kwak, T.Y., Lee, J.Y., Park, J., Rho, M.J., Go, H.: Introduction to digital pathology and computer-aided pathology. The Korean Journal of Pathology (2020)
26. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 427–436. IEEE (2015)
27. Ohashi, R., Namimatsu, S., Sakatani, T., Naito, Z., Takei, H., Shimizu, A.: Prognostic utility of atypical mitoses in patients with breast cancer: A comparative study with ki67 and phosphohistone h3. Journal of surgical oncology **118**(3), 557–567 (2018)

28. Paschali, M., Conjeti, S., Navarro, F., Navab, N.: Generalizability vs. robustness: adversarial examples for medical imaging. arXiv e-prints (2018)
29. Rai, S., Raut, A., Savaliya, A., Shankarmani, R.: Darwin: convolutional neural network based intelligent health assistant. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1367–1371. IEEE (2018)
30. Rastgar-Jazi, M., Fernando, X.: Detection of heart abnormalities via artificial neural network: An application of self learning algorithms. In: 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), pp. 66–69. IEEE (2017)
31. Ravish, D., Shanthi, K., Shenoy, N.R., Nisargh, S.: Heart function monitoring, prediction and prevention of heart attacks: Using artificial neural networks. In: 2014 International Conference on Contemporary Computing and Informatics (IC3I), pp. 1–6. IEEE (2014)
32. Serag, A., Ion-Margineanu, A., Qureshi, H., McMillan, R., Saint Martin, M.J., Diamond, J., O'Reilly, P., Hamilton, P.: Translational ai and deep learning in diagnostic pathology. Frontiers in Medicine **6** (2019)
33. 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)
34. Syam, R., Maraparedy, R.: Application of deep neural networks in the field of information security and healthcare. In: 2019 SoutheastCon, pp. 1–5. IEEE (2019)
35. Taghanaki, S.A., Das, A., Hamarneh, G.: Vulnerability analysis of chest x-ray image classification against adversarial attacks. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, pp. 87–94. Springer (2018)
36. Tizhoosh, H.R., Pantanowitz, L.: Artificial intelligence and digital pathology: challenges and opportunities. Journal of pathology informatics **9** (2018)
37. Yang, C., Wu, Q., Li, H., Chen, Y.: Generative poisoning attack method against neural networks. arXiv e-prints (2017)