Adversarially Robust Medical Classification via Attentive Convolutional Neural Networks

Isaac Wasserman
University of Pennsylvania
isaacrw@seas.upenn.edu

Abstract

Convolutional neural network-based medical image classifiers have been shown to be especially susceptible to adversarial examples. Such instabilities are likely to be unacceptable in the future of automated diagnosis. Though statistical adversarial example detection methods have proven to be effective defense mechanisms, additional research is necessary that investigates the fundamental vulnerabilities of deep-learning-based systems and how best to build models that jointly maximize traditional and robust accuracy. This paper presents the inclusion of attention mechanisms in CNN-based medical image classifiers as a reliable and method for increasing robust accuracy without sacrifice. This method is able to increase robust accuracy by up to 16% in typical adversarial scenarios and up to 2700% in extreme cases. Additionally, the behavior in attack scenarios of vanilla CNNs and CNNs with attention is compared through the use Grad-CAM attention mapping.

1. Introduction

Deep neural networks are critical tools in the computing landscape of today, and they have achieved superhuman performance in many incredibly complex tasks such as protein folding [8] and gaming [12]. For years, they have been applied to medical diagnostic tasks and have achieved levels of performance on par with highly trained pathologists, radiologists, and other diagnosticians [16]. However, these systems are often relegated to lab settings and clinical decision support systems, as their black-box nature does not inspire the confidence necessary for life-critical environments [1] [2].

The decisions of such models are near impossible for the researchers that created them to understand, let alone the proposed clinical end-user [14]. Additionally, neural networks are incredibly vulnerable to adversarial examples, instances x whose true label is certifiably y, but the mechanics of the estimator (either inherent to the architecture or specific to the weights) result in a confident and incorrect prediction [13]. These examples can be carefully contrived instances $x' = x + p$ which are extremely similar to a natural instance x but are classified differently, $h(x) \neq h(x')$. However, adversarial examples aren't just the product of bad-actors. Recent research has demonstrated the existence of many naturally occurring instances which reliably behave adversarially when used as input to popular models [6]. Prior to this discovery, researchers and engineers building models for environments where security and bad-actors were not a concern could somewhat understandably deprioritize adversarial robustness, but with the combination of increased reliance on AI-driven systems, everything we've learned about natural adversarial examples, and the proliferation of civilian-targeted cyber-warfare, secure and reliable neural networks are more important than ever.

If the goal of medical AI is to increase accessibility to high-quality diagnostics and treatments, humans will need to abdicate their roles in some procedural medical processes where AI succeeds such as image classification and segmentation. Unfortunately, these are also the applications where a well-placed orthopedic pin or speck of dust could make all the difference between a properly and improperly diagnosed patient. For medical applications, natural- and robust-accuracy must be optimized simultaneously and with equal priority.

This paper proposes a simple architectural suggestion for medical image classification models that greatly increases robust-accuracy without sacrificing natural-accuracy, unlike previous techniques [15]. Additionally, this method only increases the parameter count of ResNet-50-based [5] architectures by less than 1.3%; therefore, training time and data requirements are not significantly affected. Later sections will carefully compare the behavioral differences between baseline and adjusted architectures to investigate the root-cause of this increased performance.

2. Related Work

2.1. Vulnerability of Medical Image Models

Paschali et al. (2018) was among the first to examine the accuracy of popular medical image classification and segmentation models against adversarial images. Their results showed that current adversarial image perturbation methods were able to decrease accuracy on popular medical classification models by up to 25% and medical segmentation models by up to 41%. Based on their limited results, they also concluded that dense blocks and skip connections were correlated with more adversarially robust segmentation and that in classification tasks deeper networks tended to be more robust [11].

Finlayson et al. (2019) also looked into the susceptibility of medical diagnosis models to adversarial image attacks. Their inquiry found that, although no methods specific to medical images had been developed, attacks methods developed for natural (i.e. non-medical) images were transferable. This work also considered possible motivations for such attacks on these models, contributing the thought that even if clinicians do not soon relinquish their role to neural networks, insurance companies are likely begin outsourcing payment and authorization decisions to AI systems [4].

Ma and Niu et al. (2021) expanded on the works discussed above, reaching the fascinating conclusion that models designed for medical images were, in fact, more vulnerable to adversarial image attacks. Key to this conclusion was their finding that the medical image models they tested had significantly sharper loss landscapes than their natural image counterparts [9]. This correlation between sharp loss landscapes and more vulnerable models is outlined by Madry et al. (2017), which attributes this sharpness to overparametrization [10]. Ma and Niu et al. (2021) echoes this concern of overcomplexity and also suspects that the salience of intricate textures in medical images may also contribute to their compatibility with adversarial attacks. Additionally, they find that while medical image models are easily fooled by adversarial images, adversarially perturbed medical images are more easily detected than their natural image counterparts; they attribute this property to the tendency of popular attacks to place perturbations outside of the image's salient region [9].

2.2. Adversarially Robust Architectures

Training adversarially robust neural networks is an extremely active area of research. Current best-practices involve adversarial training, in which adversarial examples are included in the training set [10]. However, this method serves neither to remedy nor understand the underlying vulnerabilities of neural networks and often comes at the cost of training time and clean accuracy (i.e. accuracy on non-adversarial inputs) [15]. For this reason, it is imperative that the research community continues to investigate architectural modifications that yield greater adversarial robustness.

In an extensive grid search of CNN architectures for predicting CIFAR-10, Huang et al. (2021) found that while the total number of parameters does not seem to be correlated to robustness, reductions in the depth or width of deeper layers does seem to produce more robust models. However, this relationship was not consistent, and the authors were unable to articulate why it appeared [7].

Dong et al. (2020) developed an NAS (neural architecture search) algorithm called "Adversarially Robust Neural Architecture Search with Confidence Learning" (RACL) which was shown to produce models that significantly outperformed state-of-the-art models and models produced by other NAS algorithms in terms of both clean- and robust-accuracy. Key to the algorithm's success is its method of approximating and minimizing the lipschitz constant of the resultant model. When combined with adversarial training, RACL scored between 0.43% and 3.17% higher than the next best model and had the second-best clean-accuracy of all those tested. Without adversarial training, RACL had the highest clean-accuracy and the highest accuracies against MIM and PGD attacks by 10.64% and 359.52% respectively; however, a standard DenseNet-121 outperformed it against FGSM attacks by 21.77% [3].

Mok et al. (2021) furthered this work to develop an NAS for finding robust architectures. Called AdvRush, their algorithm prioritizes the smoothness of input loss landscape and uses a novel technique to simultaneously evaluate the robustness of all candidate architectures against perturbations in a two-dimensional projection of the feature-space. In practice the resultant model outperformed state-of-the-art architectures and other NAS algorithms (including RACL) on CIFAR-10 under FGSM, PGD, APGD, and AutoAttack by 2.79%, 3.25%, 2.82%, and 3.07% respectively; and it additionally outperformed all other models in clean-accuracy by 1.57%. However, the model was significantly more complex than those generated by DARTS and RACL, utilizing 17% more parameters.

References

[1] Kanadpriya Basu, Ritwik Sinha, Aihui Ong, and Treena Basu. Artificial intelligence: How is it changing medical sciences and its future? *Indian J. Dermatol.*, 65(5):365–370, Sept. 2020. 1

[2] Hannah Bleher and Matthias Braun. Diffused responsibility: attributions of responsibility in the use of ai-driven clinical decision support systems. *AI and Ethics*, Jan 2022. 1

[3] Minjing Dong, Yanxi Li, Yunhe Wang, and Chang Xu. Adversarially Robust Neural Architectures. *arXiv*, 2020. 2

[4] Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[6] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, June 2021. 1

[7] Hanxun Huang, Yisen Wang, Sarah Monazam Erfani, Quanquan Gu, James Bailey, and Xingjun Ma. Exploring Architectural Ingredients of Adversarially Robust Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34 of *arXiv*, pages 5545–5559. Curran Associates, Inc., 2021. 2

[8] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021. 1

[9] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, 2021. 2

[10] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv*, 2017. 2

[11] Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. *Lecture Notes in Computer Science*, pages 493–501, 2018. 2

[12] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, Dec 2020. 1

[13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 1

[14] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, 2021. 1

[15] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness May Be at Odds with Accuracy. *arXiv*, 2018. 1, 2

[16] Hang Yu, Laurence T. Yang, Qingchen Zhang, David Armstrong, and M. Jamal Deen. Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. *Neurocomputing*, 444:92–110, 2021. 1