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 [13] and gaming [20]. 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 [28]. 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 [2] [3].

The decisions of such models are near impossible for the researchers that created them to understand, let alone the proposed clinical end-user [23]. 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 [22]. 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 [10]. 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 [25]. Additionally, this method only increases the parameter count of ResNet-50-based [9] 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 [19].

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 [7].

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 [16]. This correlation between sharp loss landscapes and more vulnerable models is outlined by Madry et al. (2017), which attributes this sharpness to overparametrization [17]. 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 [16].

2.2. Finding 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 [17]. 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) [25]. 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 [12].

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% [5].

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 at least 2.79%, 3.25%, 2.82%, and 3.07% respectively; and it additionally outperformed all other models in clean-accuracy by at least 1.57%. However, the model was significantly more complex than those generated by DARTS and RACL, utilizing 17% more parameters [18].

Despite the multitude of NAS algorithms developed recently [5] [18] [14] [11], previous research has not discovered reliable methods or guidelines for building robust architectures. Instead, these algorithms are merely able to efficiently search a finite set of architectures for the most robust option in a given task.

2.3. Vision Transformers

Recent advances in natural language processing have offered the computer vision research community a new option for image classification, the vision transformer (ViT) [8]. Models utilizing this family of architectures commonly match or exceed the performance of convolutional neural networks [6] [24] [26] [15].

Shao et al. (2021) found these architectures to be more adversarially robust than their CNN counterparts. When compared to various versions of ResNet, ShuffleNet, MobileNet, and VGG16, ViT-S/16 was up to 46% more robust to PGD attacks and 44% more robust to AutoAttack [21]. In fact, for attack radii less than 0.01, the most robust CNN was less robust than the least robust vision transformer. All the while, the clean accuracy of the transformers was, on average, 7% higher than that of the CNNs. Additionally, ViT-S/16 (the most robust overall) has just 22 million trainable parameters, compared to the 25 million parameters of ResNet-50 [6] [9].

Based on their analysis, Shao et al. (2021) found that ViTs tend to learn more robust, high-level features, allowing them to ignore the high frequency perturbations of many attack methods. They also found that in architectures that combined transformer and convolutional blocks, a high proportion of transformer blocks correlated to higher robust-accuracy [21].

2.4. CNNs with Attention

2.4.1 Adversarial Robustness

Since their inception, it has been understood that the superior clean-accuracy of vision transformers is related to their reliance on attention [8], and recently it has been confirmed that this property is also responsible for the architecture's superior robustness [29]. Zhou et al. (2022) found that the self-attention of vision transformers tends to promote the saliency of more meaningful (non-spurious) clusters of image regions.

Working from this knowledge, it is natural to wonder whether attention mechanisms can offer additional adversarial robustness to CNNs. Agrawal et al. (2022) concluded that this was not necessarily the case, finding that while their attentive CNNs had slightly superior robustness to PGD attacks on the CIFAR-100 dataset, it fell behind ResNet-50 on CIFAR-10 and Fashion MNIST. Based on these results, they suspected that the adversarial robustness of attentive models on a given dataset may correlate to the number of classes [1].

2.4.2 Clean Accuracy

Agrawal et al. (2022) also found that the attentive model had slightly lower clean-accuracy compared to the vanilla CNNs on CIFAR-10 and CIFAR-100 but slightly higher clean-accuracy on Fashion MNIST [1]. However, this finding is inconsistent with early research on the use of CNNs with attention for fine-grained classification datasets. For example, the attention-based model developed by Xiao et al. (2015) outperformed equally supervised vanilla-CNNs by 19% [27]. As small details are incredibly salient in medical image datasets, research on fine-grained classification is especially relevant. The ability of attention mechanisms to improve the accuracy of medical image classification CNNs was confirmed by Datta et al. (2021) which appended a minimal soft-attention mechanism to five off-the-shelf pre-trained CNNs and found that attention increased weighted average AUC for all but VGG16 [4].

Based on the findings above, it is possible that the use of attention is a minimal architectural feature that challenges the findings of Tsipras et al. (2018) [25] by improving clean- and robust-accuracy simultaneously.

References

[1] Prachi Agrawal, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. Impact of Attention on Adversarial Robustness of Image Classification Models. *2021 IEEE International Conference on Big Data (Big Data)*, 00:3013–3019, 2021. 3

[2] 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

[3] 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

[4] Soumyya Kanti Datta, Mohammad Abuzar Shaikh, Sargur N Srihari, and Mingchen Gao. Soft-Attention Improves Skin Cancer Classification Performance. *arXiv*, 2021. 3

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

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*, 2020. 2, 3

[7] 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

[8] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2022. 2, 3

[9] 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, 3

[10] 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

[11] Ramtin Hosseini, Xingyi Yang, and Pengtao Xie. DSRNA: Differentiable Search of Robust Neural Architectures. *arXiv*, 2020. 2

[12] 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

[13] 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

[14] Jia Liu and Yaochu Jin. Multi-objective search of robust neural architectures against multiple types of adversarial attacks. *Neurocomputing*, 453:73–84, 2021. 2

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 00:9992–10002, 2021. 2

[16] 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

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

[18] Jisoo Mok, Byunggook Na, Hyeokjun Choe, and Sungroh Yoon. AdvRush: Searching for Adversarially Robust Neural Architectures. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 00:12302–12312, 2021. 2

[19] 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

[20] 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

[21] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the Adversarial Robustness of Vision Transformers. *arXiv*, 2021. 3

[22] 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

[23] 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

[24] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv*, 2020. 2

[25] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness May Be at Odds with Accuracy. *arXiv*, 2018. 1, 2, 3

[26] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 00:548–558, 2021. 2

[27] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The Application of Two-level Attention Models in Deep Convolutional Neural Network for Fine-grained Image Classification. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 842–850, 2015. 3

[28] 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

[29] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M. Alvarez. Understanding the robustness in vision transformers. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27378–27394. PMLR, 17–23 Jul 2022. 3