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 activation 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 [20] and gaming [31]. 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 [43]. 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 [3] [4].

The decisions of such models are near impossible for the researchers that created them to understand, let alone the proposed clinical end-user [36]. 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 [35]. 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 [16]. 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 [39]. Additionally, this method only increases the parameter count of ResNet-50-based [15] 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 [27].

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 [12].

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 [24]. This correlation between sharp loss landscapes and more vulnerable models is outlined by Madry et al. (2017), which attributes this sharpness to overparametrization [25]. 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 [24].

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 [25]. 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) [39]. 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 [18].

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% [10].

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 [26].

Despite the multitude of NAS algorithms developed recently [10] [26] [22] [17], 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) [14]. Models utilizing this family of architectures commonly match or exceed the performance of convolutional neural networks [11] [37] [40] [23].

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 [33]. 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 [11] [15].

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 [33].

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 [14], and recently it has been confirmed that this property is also responsible for the architecture's superior robustness [44]. 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 [2].

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 [2]. 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% [42]. 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 pretrained CNNs and found that attention increased weighted average AUC for all but VGG16 [8].

Based on the findings above, it appears possible that the use of attention is a minimal architectural feature that challenges the findings of Tsipras et al. (2018) [39] by improving clean- and robust-accuracy simultaneously.

3. Method

3.1. Datasets

Benchmark medical image classification tasks were chosen based on their frequent use in similar studies, such as Ma and Niu et al. (2021) [24] and Finlayson et al. (2019) [12]. These tasks are diabetic retinopathy detection from fundoscopy images, pneumothorax detection from chest x-rays, and skin lesion classification from dermatoscopy images. Fundoscopy images were sourced from the popular Kaggle diabetic retinopathy detection competition [21]. Chest x-rays used were from the ChestX-Ray14 dataset [41]. Unlike Finlayson et al. (2019) [12] and Ma and Niu et al. [24] (2021), dermatoscopy images were not sourced directly from ISIC [19] and were instead taken from the HAM10000 dataset [38] which draws from the ISIC archive; this choice was made to more closely resemble the setup of Datta et al. (2021) [8].

3.2. Preprocessing

All images were resized to the ImageNet standard 224×224 pixels [9] and preprocessed according to the specifications of the TensorFlow [1] implementation of ResNet-50 [15], which include subtracting 123.68, 116.779, and 103.939 from the red, green, and blue channels respectively [5].

The fundoscopy and chest x-ray datasets were distilled from their original multiclass classification form into binary classification datasets. For the fundoscopy images, the "No DR" and "Mild DR" classes were considered negative while the "Moderate," "Severe," and "Proliferative DR" classes were considered positive. The chest x-ray dataset originally included 14 non-exclusive classes, but for the purposes of this study, all images labeled as having a pneumothorax present were labeled positive, while other images were simply labeled negative. No changes were made to the classes of the dermatoscopy dataset.

In an attempt to solve for the relatively large range in image contrast of the grayscale chest x-rays, all images were put through a gaussian filtering pipeline that maximized the prominence of edges and shadowed areas. Images are first min-max normalized to the range $[0, 255]$. A gaussian blur

is applied to a copy of this image; the kernel size of the filter is computed based on parameters $\sigma_x = 20,\ \sigma_y = 0$. This blurred image blur(x) is subsequently added to the original normalized image x according to the following formula $x' = 4x - 4(\text{blur}(x)) + 128$.

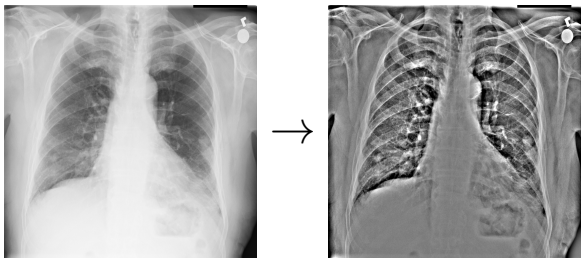


Figure 1. Visualization of the chest x-ray image preprocessing pipeline

This pipeline was initially used for the fundoscopy images as well, but early results revealed that this type of processing was not well suited to the task.

3.3. Network Architecture

Four models were trained for each dataset, a ResNet-50 with and without a soft attention block and an Inception-ResNetV2 with and without soft attention. The architecture of each model was based on the TensorFlow [1] implementations of ResNet-50 [7] and InceptionResNetV2 [6].

ResNet-50 was instantiated with the top block included, initial weights based on ImageNet [9] pretraining, and class count set to 1000. The final three layers of the network were removed. In the models without attention, these final layers were replaced with a 2D global average pooling layer followed by a fully-connected layer with softmax activation. In the models with attention, the last three layers were replaced with a soft attention block implemented according to the specifications of Datta et al. (2021). This block produces a 7×7 feature map which is 2×2 max pooled and concatenated with a 2×2 max pooled version of the input to the attention block. This concatenated output is put through ReLU activation, 50% dropout, and global average pooling before being handed off to the fully connected prediction head with softmax activation [8].

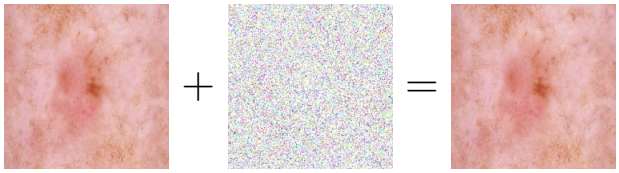
InceptionResNetV2 [34] was instantiated with the top block included, initial weights based on ImageNet, and classifier activation as softmax. In the models without attention, the final 28 layers were replaced with a ReLU activation and 50% dropout whose output was flattened and sent to the fully connected prediction head with softmax activation. In the models with attention, these layers were replaced by the soft attention block described above. This output is 2×2 max pooled and concatenated with a max pooled version of the input to the attention block before being sent through ReLU activation and the softmax prediction head.

In each model, the width of the prediction head was equal to the number of classes in the dataset (seven for dermatoscopy and two for fundoscopy and chest x-ray).

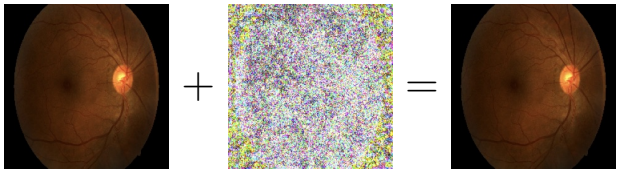
All models used the Adam optimizer with $\eta = 0.01$ and $\epsilon = 0.1$ and minimized categorical cross-entropy during training. During training classes were weighted based on their inverse frequency relative to the other classes. Each model was trained for a maximum of 300 epochs with early stopping (patience $= 40$, minimum delta $= 0.001$) causing most to stop after 60-90 epochs. In the dermatoscopy task, models only saw 10% of the dataset during each epoch; this was done to increase the odds of reproducing the results of Datta et al. (2021) [8].

3.4. Evaluation and Analysis of Robustness

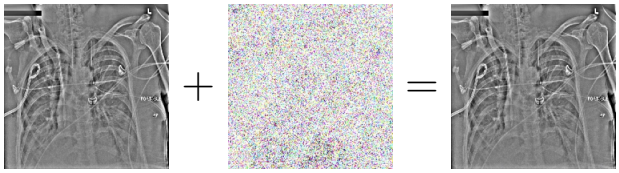
After training, the models' clean- and robust-accuracy was evaluated using the FoolBox [29] [28] library's implementation of l_∞ projected gradient descent attacks. Each set of models was tested at increasing epsilons (perturbation radii). Attacks of these perturbation radii were created for each image in the test sets. Unweighted accuracy was calculated for each perturbation radius.



(a) An example of $\epsilon = 0.01$ perturbation on the dermatoscopy dataset. 0.01 is the maximum perturbation radius used for this dataset.



(b) An example of $\epsilon = 0.32$ perturbation on the fundoscopy dataset. 0.32 is the maximum perturbation radius used for this dataset.



(c) An example of $\epsilon = 0.32$ perturbation on the chest x-ray dataset. 0.32 is the maximum perturbation radius used for this dataset.

Figure 2. Perturbation examples

Small samples of the resultant adversarial images were saved and used to analyze model behavior using gradient weighted class activation mapping (Grad-CAM) [32] local-

ization maps. These maps were generated using the Grad-CAM implementation of Datta et al. (2021) [8] and provide attention-map-esque representations of spatial importance for neural networks regardless of whether they utilize attention mechanisms or not. For each model, these maps were generated at each epsilon for 16 sample images with and without perturbations applied. Maps for a given starting image were compared to their perturbed counterpart as well as against other models and perturbation radii.

4. Results

4.1. ResNet-50 Models

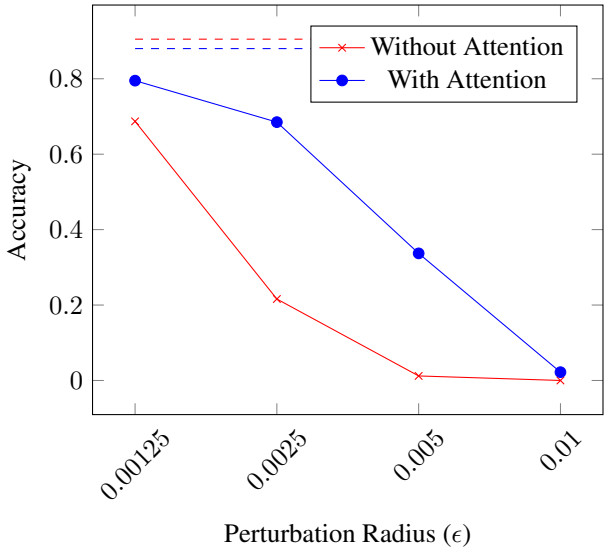


Figure 3. Clean- and robust-accuracy of ResNet-50 models for skin lesion classification. Dashed lines represent clean-accuracy.

For the task of skin lesion classification, the models with attention are clearly superior, in terms of robustness (Fig. 3). Although the model without attention has slightly higher clean-accuracy (0.905 vs. 0.88), even the slightest perturbation ($\epsilon = 0.00125$) is able to reduce its accuracy by 24% and put the model with attention in the lead. By the time $\epsilon = 0.005$, the model is worse than random, and by the time $\epsilon = 0.01$, the model is rendered completely useless. It is worth noting that perturbation of this size (Fig. 2a) is far from being human perceptible. Meanwhile, at this perturbation radius, the model with attention remains better than random selection.

The models for diabetic retinopathy detection tell a slightly different story (Fig. 4). These models were, overall, much more robust, requiring a perturbation of at least $\epsilon = 0.01$ to reduce accuracy by 3%. As in the skin lesion classification models, the model without attention has a slightly higher clean accuracy (0.832 vs. 0.818). It retains this very small lead until $\epsilon = 0.02$, at which point, the

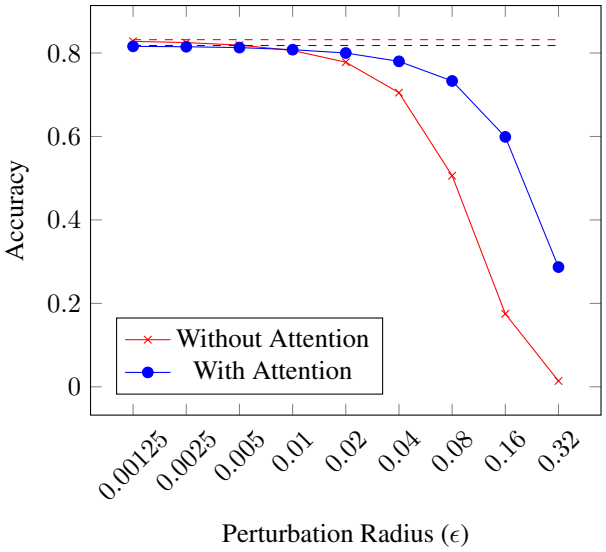


Figure 4. Clean- and robust-accuracy of ResNet-50 models for diabetic retinopathy detection. Dashed lines represent clean-accuracy.

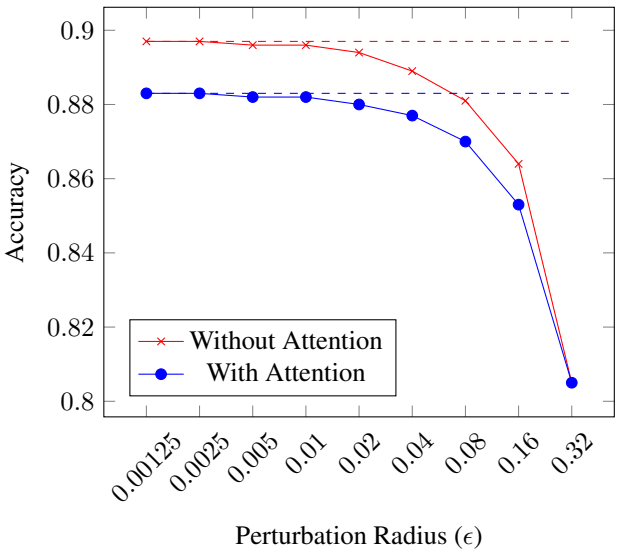


Figure 5. Clean- and robust-accuracy of ResNet-50 models for pneumothorax detection. Dashed lines represent clean-accuracy.

accuracy of both models begins falling, but the model with attention does so a bit less dramatically. At the maximum perturbation radius tested ($\epsilon = 0.32$) (Fig. 2b), the accuracy of the model with attention is over 20 times higher than that of the model without.

The pneumothorax detection models, collectively, were even more robust than the diabetic retinopathy detection models (Fig. 5). After perturbations of $\epsilon = 0.32$ (Fig. 2c), the accuracy of both models met at 0.805. Prior to this, the accuracy of the model without attention hovered 1-2%

above that of the model with attention. However, based on the trajectory of the models' accuracy with respect to ϵ, perturbation radii higher than 0.32 would likely result in the model with attention leapfrogging the model without.

4.2. InceptionResNetV2 Models

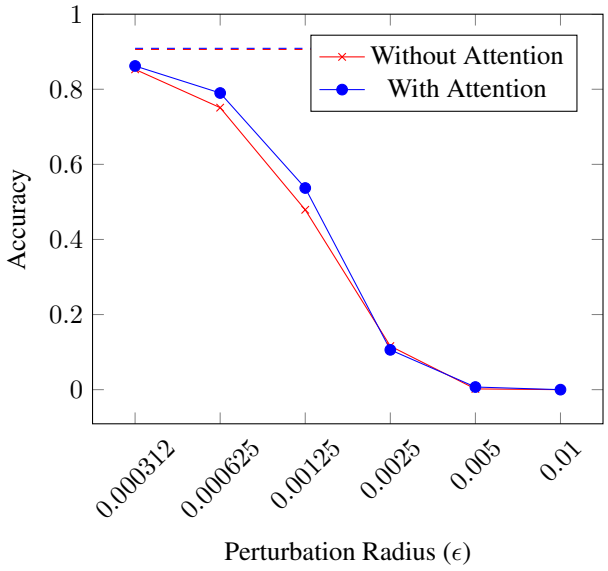


Figure 6. Clean- and robust-accuracy of InceptionResNetV2 models for skin lesion classification. Dashed lines represent clean-accuracy.

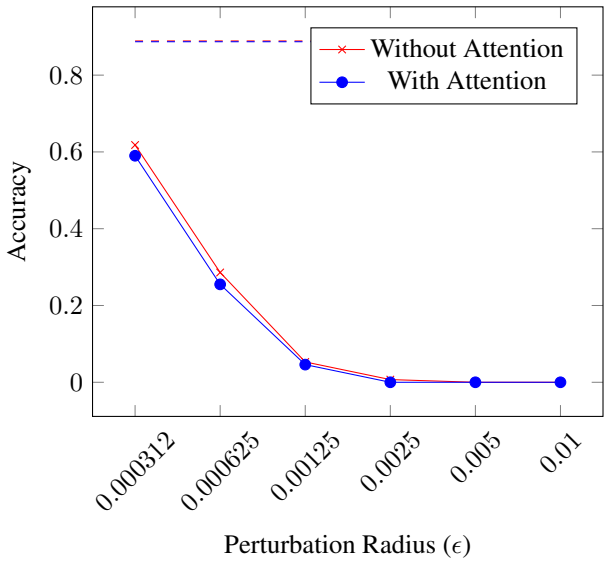


Figure 7. Clean- and robust-accuracy of InceptionResNetV2 models for diabetic retinopathy detection. Dashed lines represent clean-accuracy.

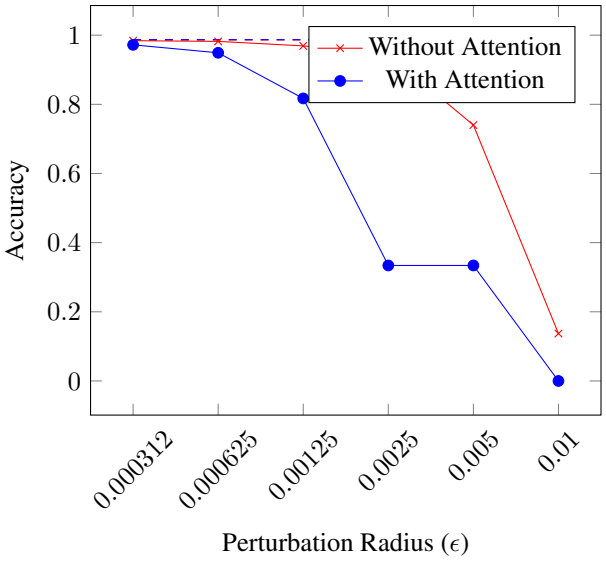


Figure 8. Clean- and robust-accuracy of InceptionResNetV2 models for pneumothorax detection. Dashed lines represent clean-accuracy.

Experiments with the InceptionResNetV2 architecture paint a slightly different picture. In all tasks, clean-accuracy for the baseline and attentive models was within 1%. This result is, somewhat, at odds with those of Datta et al. (2021) [8] which found soft attention to improve accuracy of InceptionResNetV2 (on HAM10000 [38]) by 3%.

For the skin lesion classification and diabetic retinopathy detection tasks, the baseline and attentive architectures performed near-identically in adversarial scenarios. The diabetic retinopathy task slightly favored the baseline, whereas skin lesion classification slightly favored the attentive model; however, at most, there was a 6% discrepancy between their accuracies ($\epsilon = 0.00125$).

In the pneumothorax detection task, the model with attention performed significantly worse under adversarial scenarios. The model reached a level of accuracy worse than random selection after the perturbation radius was pushed beyond 0.00125, while the baseline remained usable with a perturbation radius of 0.005.

5. Discussion

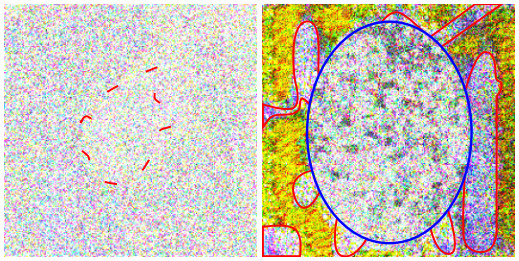
5.1. Analysis of Perturbations and Activation Maps

In an attempt to better understand the principles and behaviors that led the CNNs with attention to perform better in both clean and adversarial scenarios, perturbation difference maps and Grad-CAM activation maps were generated for a select sample of images on each model. Individually, these difference maps and activation maps were unremarkable. However, a number of dataset specific patterns were

noticed throughout the images used for this analysis.

5.1.1 ResNet-50 Models

While imperceptible in the final images, the perturbations for dermatoscopic and fundoscopic images were found to carry easily perceptible information about the source image, when the changes were scaled to a range of $[0, 255]$. In these cases, the adversarial "noise" contained the shape of the lesion or eye. For dermatoscopic images, this shape was in the form of a densely perturbed ring enclosing a sparsely perturbed center (Fig. 9a). This phenomenon was most visible when $\epsilon = 0.00125$. For the fundoscopic images, this shape was represented by a sparsely perturbed ellipse. In the attacks generated for the model with attention, the shape of this ellipse occasionally diverged from the true shape of the eye (Fig. 9b).



(a) Shape of skin lesion found in adversarial "noise"

(b) Divergence from shape of eye

Figure 9. Shape information carried over into perturbations

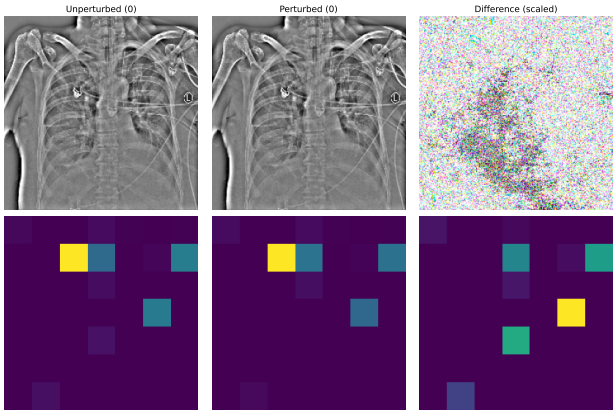


Figure 10. Comparison of activation maps (with attention) for perturbed and unperturbed chest x-ray

In the attacks tailored to the fundoscopic image models, perturbations introduced dark splotchy patterns to the eye area (Fig. 12). Based on diagnostic literature, these spots could be "attempts" by the attack to emulate cotton-wool

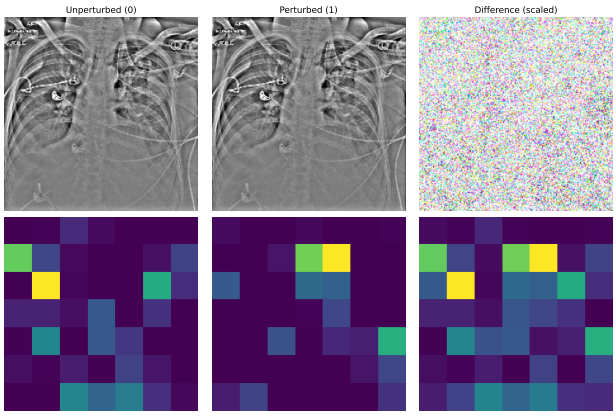


Figure 11. Comparison of activation maps (without attention) for perturbed and unperturbed chest x-ray

spots or hemorrhages, key indicators of diabetic retinopathy [13]. Similar patterns appear in the perturbations for the chest x-ray model with attention (Fig. 10) (these spots are larger and less speckled), though these do not appear similar to the key indicators of pneumothoraces (air in pleural space, misplaced lung edge, less distinct lung markings, etc.) [30], it remains unclear whether the attack could be emulating these features indirectly. If so, this behavior and its perceptibility in the difference maps could be responsible for medical adversarial images being so easily detected, as observed by Ma and Niu et al. (2021) [24].

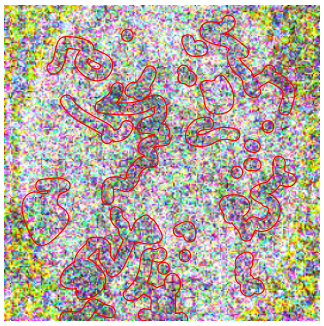


Figure 12. Dark splotches in perturbation of fundoscopic images

While the activation maps for the dermatoscopic and fundoscopic image models were invariant to perturbation, the chest x-ray model without attention occasionally produced very different activation maps under adversarial attacks (Fig. 11). However, this was only true under successful attacks (of which there were few).

For all three datasets, models with attention produced activation maps that were very different from their baseline counterparts; in most cases, the regions of highest activation did not match across models. Additionally, the activation maps for the chest x-ray model with attention were

highly focused, typically having a single region of significance (Fig. 11).

5.1.2 InceptionResNetV2 Models

Similar to the ResNet-50-based models, perturbations of dermatoscopic and fundoscopic images carried perceptible information about the source image. Across all three datasets, the perturbations for attentive models were generally higher contrast; in other words, the difference in the level of perturbation between the most- and least-perturbed pixels was greater. Like the ResNet-50 models, the perturbations targeting fundoscopic and chest x-ray images produced intensely perturbed splotches. However, unlike the ResNet-50 models, the splotches on the x-rays appear in both the baseline and attentive models and are less widespread (Fig. 14).

The activation maps for all skin lesion classification and diabetic retinopathy detection models were greatly impacted by perturbation, often being nearly inverted, indicating that the model is focusing on the incorrect regions of the image. Perturbations appeared to have less of an effect on the activation maps of the pneumothorax detection models. Like the ResNet-50 models, the activation maps of the attentive models shared no similarity with those of the baselines. Also, similar to its ResNet-50 counterpart, the attention maps of the chest x-ray models were significantly more focused than those of the other models (Fig. 14).

5.2. Conclusion

The inclusion of a soft-attention block was able to improve the robustness of ResNet-50 on two of the three medical classification tasks tested, while only slightly decreasing clean-accuracy. Additionally, previous experiments have shown the ability of this architectural modification to increase clean accuracy [8]. Though soft-attention was unable to significantly improve the robust accuracy of the InceptionResNetV2 models, its inclusion was not detrimental to clean- or robust-accuracy in two of three tasks. Further investigation is necessary to understand why soft-attention was detrimental to the performance of pneumothorax detection models.

These results suggest that, in most cases, the inclusion of a soft-attention block is beneficial (or at least not harmful) to the overall performance of medical image classification architectures. This modification has potential to improve accuracy and even greater potential to improve model robustness.

Additionally, the observations above regarding perturbation behavior suggest that PGD may be inadvertently emulating cotton-wool spots, hemorrhages, and lung boundaries. The fact that these patterns are so easily visible in the perturbations may lend further insight into why adversarial images targeting medical classifiers are so easily detected.

Future research into the susceptibility of medical image classifiers to adversarial images may wish to investigate this hypothesis by creating a modified PGD attack algorithm which maximizes the perceived randomness of the perturbation, as this strategy could lead to less detectable attacks.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 3, 4

[2] 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

[3] 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

[4] 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

[5] François Chollet et al. imagenet_utils.py, 2016. 3

[6] François Chollet et al. inception_resnet_v2.py, 2017. 4

[7] François Chollet et al. resnet.py, 2018. 4

[8] Soumyya Kanti Datta, Mohammad Abuzar Shaikh, Sargur N Srihari, and Mingchen Gao. Soft-Attention Improves Skin Cancer Classification Performance. *arXiv*, 2021. 3, 4, 5, 6, 8

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3, 4

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

[11] 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

[12] 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, 3

[13] Adam A Gerstenblith, Michael P Rabinowitz, Behin I Barahimi, Christopher M Fecarotta, Mark A Friedberg, and Christopher J Rapuano. *Wills Eye Manual*. Lippincott Williams & Wilkins, 2012. 7

[14] 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

[15] 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

[16] 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

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

[18] 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

[19] ISIC. The international skin imaging collaboration. 3

[20] 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

[21] Kaggle. Diabetic Retinopathy Detection, 02 2015. 3

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

[23] 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

[24] 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, 3, 7

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

[26] 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

[27] 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

[28] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. 4

[29] Jonas Rauber, Roland Zimmermann, Matthias Bethge, and Wieland Brendel. Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax. *Journal of Open Source Software*, 5(53):2607, 2020. 4

[30] Mark Rodrigues and Zeshan Qureshi. *The Unofficial Guide to Radiology*. Zeshan Qureshi, 2014. 7

[31] 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

[32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv*, 2016. 4

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

[34] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv*, 2016. 4

[35] 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

[36] 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

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

[38] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):180161, 2018. 3, 6

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

[40] 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

[41] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 3462–3471, 2017. 3

[42] 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

[43] 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

[44] 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

A. Additional Figures

Figure 13. Perturbations and attention maps of skin lesion classification, diabetic retinopathy detection, and pneumothorax detection models using ResNet-50 architecture at $\epsilon = 0.01$, $\epsilon = 0.32$, and $\epsilon = 0.32$ respectively.

Figure 14. Perturbations and attention maps of skin lesion classification, diabetic retinopathy detection, and pneumothorax detection models using InceptionResNetV2 architecture at $\epsilon = 0.00125$.