Addressing Vulnerability in Medical Deep Learning through Robust Training

Josué Martínez-Martínez
Computer Science and Engineering
University of Connecticut
Storrs, USA
josue.martinez-martinez@uconn.edu

Sheida Nabavi
Computer Science and Engineering
University of Connecticut
Storrs, USA
sheida.nabavi@uconn.edu

2023 IEEE Conference on Artificial Intelligence (CAI) | 979-8-3503-3984-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/CAI54212.2023.00150

***Abstract*—Deep neural networks have been incorporated into healthcare for the purpose of diagnosing and detecting medical conditions. However, studies have shown that the vulnerability of neural networks to adversary and noise remains a pervasive problem that compromises trust of medical practitioners and accuracy in diagnosis, prognosis, and outcome prediction by such systems. In this study we show that robust training methods can help models perform more robustly against not only adversarial attacks, but also noises and calibration errors.**

***Index Terms*—Data Augmentation, Robustness, Medical Images, Deep Learning**

I. Introduction

Machine learning models used in analysing medical images must be robust to adversarial attacks and noise, meaning they should perform well even when input data is intentionally or unintentionally perturbed or corrupted. Adversarial attacks are inputs specifically crafted to cause misclassification or unexpected behavior in a model, while noise refers to random variations in input data that can also impact model performance. Model robustness is crucial in real-world deployment, where models may encounter unexpected or adversarial inputs that could lead to incorrect diagnosis or treatment, potentially causing harm to patients.

In addition to model robustness, model calibration and prediction uncertainty are equally important in medical image applications. Model calibration ensures that the outputs of a machine learning model accurately reflect the underlying probability distribution of the data, which is critical in medical imaging for informed decision making. Therefore, machine learning models for medical image applications must be robust, accurately calibrated, and provide less randomness in the output estimates to support informed decision making in clinical practice.

In this study we propose a novel framework for evaluating an AI system in medical domain along multiple dimensions of performance (e.g., clean accuracy, adversarial robustness, natural robustness, probability calibration), we compare multiple training methods along those dimensions, and we adapt robust training techniques from the natural image domain (AugMix [1] and RobustAugMix [2]) for use with medical images. Our results demonstrate that RobustAugMix not only achieves robustness against adversarial perturbations and noise, but also reduces the calibration error of the model.

II. Methodology

Our objective in this study is to examine the performance of various robust learning strategies in analyzing biomedical images when there are different sources of noise and corruptions. We used the COVID-19 chest X-Ray imagery dataset presented before by [3] to examine and compare the performance of a standard deep neural network (DNN) model, and models trained with three robust data augmentation strategies: adversarial training [4], AugMix [1] and RobustAugMix [2]. We used ResNet 18 as the backbone architecture to train the models with different training methods. We trained the standard model using the Empirical Risk Minimization (ERM) method [5], the robust model using the robust optimization method [4], the AugMix model using the Jensen-Shanon Loss [1], and RobustAugMix using the combination of Jensen-Shanon and the robust optimization [2].

To examine the performance of these models we compare their accuracy against clean, noisy, and adversarial images. In this study, we show the benefits and the trade-off of using robust learning in the medical domain. We study the accuracy of the models when images are corrupted by different types of medical images' noise. Also, we examine the trade-off in the model's output estimates. To evaluate this, we compute the calibration error of the model and the entropy of the predicted class probabilities.

III. Experiments

All the models in this study were trained for 41 epochs, with the SGD optimizer and an initial learning rate of 0.01. The robust models were trained using Projected Gradient Descent (PGD) with an epsilon value of 4 and 7 gradient descent steps [6]. Epsilon is the size of the perturbation. For the AugMix which mixes augmented images through linear interpolation, we used random flips augmentations.

A. Adversarial Attack

Adversarial training was first introduced in [7] as a defense against adversarial attacks [8]. In this study we tested the models against PGD attacks. The epsilon values for these

979-8-3503-3984-0/23/$31.00 ©2023 IEEE
DOI 10.1109/CAI54212.2023.00150

experiments goes from 1 to 5 with 20 steps. We calculated the mean adversarial accuracy of these models as well, to determine the overall robust model, considering all the epsilon values. We used the Responsible Artificial Intelligence (RAI) toolbox [9] to perform adversarial training and testing. Results are presented in table I.

B. Corruptions

Miscalibration of imaging equipment can lead to noisy images. For this reason the models should be robust to these corruptions and avoid misclasification. We considered common imaging noise: Gaussian, salt and pepper (S&P), and Speckle. Gaussian noise is due to randomness and has a normal distribution. S&P noise is created by replacing random pixels of the image with dark and bright values that are at the extreme ends. Speckle noise is created when coherent light interacts with a rough surface, producing a random pattern of bright and dark spots that can degrade image quality. We varied the Gaussian and Speckle noise by changing the variance between 0.1 to 1, and we varied the S&P noise by changing the amount between values 0.1 to 1. The models' accuracy are presented in Table I.

It is evident that standard models in the medical domain does not guarantee any robustness against these adversarial attacks. Additionally, training the model using a data augmentation technique like AugMix, which was primarily designed for natural images, can reduce the model's robustness against adversarial attacks and noise. However, the Robust and RobustAugMix models perform better compared to the Standard and AugMix models against adversarial perturbations, Gaussian noise and Speckle noise .

TABLE I
ROBUSTNESS EVALUATION IN TERM OF ACCURACY IN PERCENTAGE

Methods	Clean	Adversarial	Gaussian	S&P	Speckle
Standard	95.80	22.24	72.78	**72.00**	86.51
Robust	95.96	**89.83**	82.91	61.98	**95.91**
AugMix	94.25	12.41	21.27	26.07	44.35
RobustAugMix	**96.27**	82.10	**83.56**	69.75	94.86

C. Calibration Error and Output Randomness

To measure the calibration error, we used the root mean square calibration error (RMSCE) metric [10]. And to measure the output estimate randomness we calculated the entropy of the classification results. The results are shown in Table II RMSCE quantifies the accuracy of the predicted class probabilities and entropy measuring the randomness and spread of the predicted class probabilities. A low RMSCE and low entropy indicating a well-calibrated and certain model. It has been reported that AugMix can reduce the calibration error [1], and adversarial robust models have more stable predictions [11], [12]. As shown in Table II, the combination of robust learning and AugMix reduced the calibration error from 7.14% to 2.32%. However, the Robust model provides the most certain output estimates.

TABLE II
MODEL CALIBRATION AND OUTPUT RANDOMNES

Methods	Calibration Error	Output Entropy
Standard	7.14%	2.96
Robust	5.61%	**2.81**
AugMix	6.11%	4.47
RobustAugMix	**2.32%**	4.09

IV. CONCLUSION

This study explored various training approaches and proposed a framework for evaluating the performance of DNNs in the medical domain. We show that adversarial training can make DNNs more robust against not only adversarial perturbation, but also other kind of noise and corruptions, and results in more calibrated and certain outputs. A well-calibrated and robust model can aid doctors in making informed decisions and decreasing the risk of misdiagnosis or mistreatment. It can also contribute to the development of more accurate and trustable medical decision support systems and enhance overall care quality.

ACKNOWLEDGMENT

We thank Olivia Brown (MIT Lincoln Laboratory) for her feedback and helpful comments on this work. This material is based upon work supported by the program Team-TERRA at the University of Connecticut, which is supported by the National Science Foundation under Grant DGE-2022036.

REFERENCES

[1] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *arXiv preprint arXiv:1912.02781*, 2019.

[2] J. Martinez-Martinez and O. Brown, "Robustaugmix: Joint optimization of natural and adversarial robustness," in *NeurIPS ML Safety Workshop*.

[3] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," *arXiv 2006.11988*, 2020.

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[5] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[6] J. Goodwin, O. Brown, and V. Helus, "Fast training of deep neural networks robust to adversarial perturbations," in *2020 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–7, 2020.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[9] R. Soklaski, J. Goodwin, O. M. Brown, M. Yee, and J. Matterer, "Tools and practices for responsible AI engineering," *CoRR*, vol. abs/2201.05647, 2022.

[10] A. Kumar, P. S. Liang, and T. Ma, "Verified uncertainty calibration," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[11] J. Roberts and T. Tsiligkaridis, "Ultrasound diagnosis of covid-19: Robustness and explainability," *arXiv preprint arXiv:2012.01145*, 2020.

[12] M. Z. Joel, S. Umrao, E. Chang, R. Choi, D. X. Yang, J. S. Duncan, A. Omuro, R. Herbst, H. M. Krumholz, and S. Aneja, "Using adversarial images to assess the robustness of deep learning models trained on diagnostic images in oncology," *JCO Clinical Cancer Informatics*, vol. 6, p. e2100170, 2022.