2022 IEEE International Performance, Computing, and Communications Conference (IPCCC) | 978-1-6654-8018-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/IPCCC55026.2022.9894322

Exploring Adversarial Attacks on Neural Networks: An Explainable Approach

Justus Renkhoff *[1a], Wenkai Tan *[1a], Alvaro Velasquez[2b], William Yichen Wang[5g], Yongxin Liu[1c], Jian Wang[3d]
Shuteng Niu[4e], Lejla Begic Fazlic[6f], Guido Dartmann[6f], Houbing Song[1c]
[1]Embry-Riddle Aeronautical University, FL 32114 USA, [2]University of Colorado Boulder, CO 80309, USA
[3]University of Tennessee at Martin, TN 38237 USA, [4]Bowling Green State University, OH 43403 USA,
[5]Purdue University, IN 47907 USA, [6]Trier University of Applied Sciences, Germany
[a]{renkhofj, tanw1}@my.erau.edu, [b]alvaro.velasquez@colorado.edu, [c]{LIUY11, songh4}@erau.edu,
[d]jwang186@utm.edu, [e]sniu@bgsu.edu [f]{l.begic, g.dartmann}@umwelt-campus.de, [g] wywang@purdue.edu

***Abstract*—Deep Learning (DL) is being applied in various domains, especially in safety-critical applications such as autonomous driving. Consequently, it is of great significance to ensure the robustness of these methods and thus counteract uncertain behaviors caused by adversarial attacks. In this paper, we use gradient heatmaps to analyze the response characteristics of the VGG-16 model when the input images are mixed with adversarial noise and statistically similar Gaussian random noise. In particular, we compare the network response layer by layer to determine where errors occurred. Several interesting findings are derived. First, compared to Gaussian random noise, intentionally generated adversarial noise causes severe behavior deviation by distracting the area of concentration in the networks. Second, in many cases, adversarial examples only need to compromise a few intermediate blocks to mislead the final decision. Third, our experiments revealed that specific blocks are more vulnerable and easier to exploit by adversarial examples. Finally, we demonstrate that the layers *Block4_conv1* and *Block5_cov1* of the VGG-16 model are more susceptible to adversarial attacks. Our work could potentially provide useful insights into developing more reliable Deep Neural Network (DNN) models.**

I. Introduction

Deep learning (DL) provides unlimited possibilities for addressing various scientific problems. However, the robustness of Deep Neural Networks (DNNs) has caused many concerns; for example, some researchers revealed that DNNs, such as the VGG-16 model [1], can be misled by intentionally mutated images that are imperceptible to humans [2]–[4]. In these scenarios, the mutated pixels have pseudo-random characteristics and thus raise concerns about the uncertainty and trustworthiness of DNNs under the natural Gaussian noise of their operational environments [5], [6]. Most of the current efforts still regard DNNs as black-box models and have not yet analyzed the effect of adversarial attacks in an explainable way. In this paper, we use images from the ImageNet database [7] and then manipulate them with DLFuzz [8], which generates adversarial examples based on given seed images and tries to activate as many neurons as possible simultaneously. Grad-CAM [9] heatmaps are generated to make the decision-making procedure explainable. Wrongly classified mutated images are analyzed and compared to their origin to find out why and in which layers behavior deviations occur. We compare the response characteristics of the VGG-16 model when the input images are perturbed with adversarial and Gaussian random noise.

* Wenkai Tan and Justus Renkhoff are co-first authors.

II. Methodology

A. Adversarial Example Generation

We selected 1000 images from the dataset. With DLFuzz we were able to generate a total of 93 adversarial examples which will be used in the analysis. Based on these perturbations we calculated Gaussian random noise with the same statistical properties and applied it to the original images as well.

B. Locating Vulnerabilities

The original samples, adversarial examples, and noisy samples are analyzed using Grad-CAM. A heatmap is generated for each layer of the VGG-16 model showing the focus point of the DNN in the corresponding layer. Then, the cosine similarity is calculated between the heatmaps of the adversarial examples and those of the original images for every layer. By locating layers where the similarity between the adversarial and the original samples are particularly low, we can determine which layers react strongly to the perturbations.

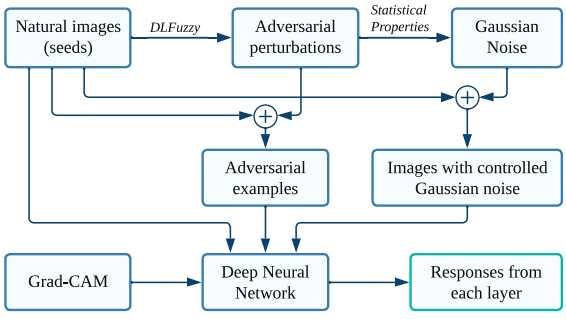


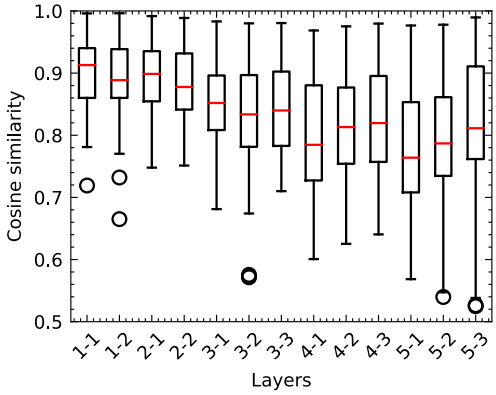
Fig. 1. Exploring neural network response against adversarial perturbations and Gaussian noise.

III. Evaluation and Discussion

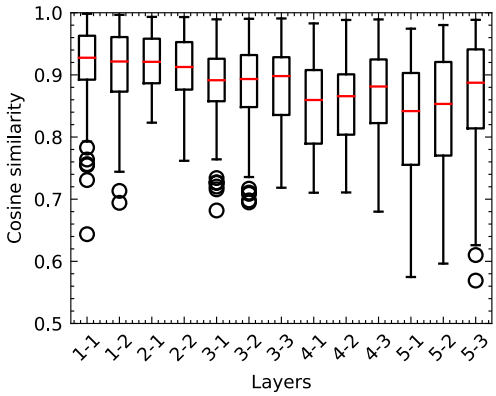
The cosine similarities between the original image's heatmaps and the adversarial sample's heatmaps have noticeable lower medians and greater variances in most layers compared to the ones of the original images and the noisy images as it can be seen in figure 2. Especially the layers Block4_Conv1 and Block5_Conv1 are considered to be comparatively easier to compromise because their response deviates significantly under adversarial attacks, as they have the two lowest medians. Gaussian noise also causes the DNN's behavior to deviate as shown in figure 2b. However, noisy images cause less significant deviation. In general, more significant behavioral deviations can be observed in deeper layers of the network, as shown in figures 2a and 2b. We found that the VGG-16 network can be misled by compromising only a few intermediate layers. Perturbations based on Gaussian noise can cause behavioral drift, but usually do not mislead the classifier. Using this, we define a threshold for each layer. This threshold is equal to the median of the cosine similarity between the heatmap of the original image and the one based on Gaussian noise in the corresponding layer. A layer is said to be compromised if the cosine similarity between the heatmap of the adversarial example and the heatmap of the original image is lower than the threshold. With DLFuzz's standard level of perturbation, we calculated a 40% chance of compromising a layer. This probability increases significantly and reaches 80% when we increase the level of perturbation by a factor of 2. We observed that there are adversarial examples that can cause misclassifications by compromising only a small number of layers.

IV. Conclusion

We generated adversarial examples using DLFuzz. Through Grad-CAM, we were able to analyze the decision-making procedure of the VGG-16 network layer by layer. With our approach, we were able to show that compared to Gaussian random noise, intentionally generated adversarial perturbations cause more severe behavioral deviations. Furthermore, we were able to show that, only a few intermediate layers of a DNN need to be compromised in order to manipulate the final decision. Finally, we demonstrated that, the layers $Block4\_conv1$ and $Block5\_cov1$ of the VGG-16 model are more susceptible to adversarial attacks.



(a)



(b)

Fig. 2. Comparison of behavioral deviations under: (a) adversarial examples and (b) random noise for the VGG-16 convolutional layers (x-axis 1-1 represents the Block1_Conv1 layer in VGG-16).

Acknowledgment

This research was supported in part by the Air Force Research Laboratory Information Directorate, through the Air Force Office of Scientific Research Summer Faculty Fellowship Program®, Contract Numbers FA8750-15-3-6003, FA9550-15-0001 and FA9550-20-F-0005. This research was also partially supported by the National Science Foundation under Grant No. 2150213.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] K. Pei, Y. Cao, J. Yang, and S. Jana, "Deepxplore: Automated whitebox testing of deep learning systems," in *Proceedings of the 26th Symposium on Operating Systems Principles*, ser. SOSP '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1–18. [Online]. Available: https://doi.org/10.1145/3132747.3132785

[3] D. J. Miller, Z. Xiang, and G. Kesidis, "Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks," *Proceedings of the IEEE*, vol. 108, no. 3, pp. 402–433, 2020.

[4] B. Luo, Y. Liu, L. Wei, and Q. Xu, "Towards imperceptible and robust adversarial example attacks against neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[5] C. Englert, P. Galler, P. Harris, and M. Spannowsky, "Machine learning uncertainties with adversarial neural networks," *The European Physical Journal C*, vol. 79, no. 1, pp. 1–10, 2019.

[6] A. Ignatiev, "Towards trustable explainable ai." in *IJCAI*, 2020, pp. 5154–5158.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[8] J. Guo, Y. Zhao, H. Song, and Y. Jiang, "Coverage guided differential adversarial testing of deep learning systems," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 2, pp. 933–942, 2021.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.