Overcoming Adversarial Attacks for Human-in-the-Loop Applications

Ryan McCoppin [1] **Marla Kennedy** [2] **Platon Lukyanenko** [1] **Sean Kennedy** [3]

Abstract

Including human analysis has the potential to positively affect the robustness of Deep Neural Networks and is relatively unexplored in the Adversarial Machine Learning literature. Neural network visual explanation maps have been shown to be prone to adversarial attacks. Further research is needed in order to select robust visualizations of explanations for the image analyst to evaluate a given model. These factors greatly impact Human-In-The-Loop (HITL) evaluation tools due to their reliance on adversarial images, including explanation maps and measurements of robustness. We believe models of human visual attention may improve interpretability and robustness of human-machine imagery analysis systems. Our challenge remains, how can HITL evaluation be robust in this adversarial landscape?

1. Introduction

The adversarial advent has greatly impacted the machine learning landscape. Adversarial images can degrade a model's ability to perform its task. In addition to degrading ML model performance, these perturbations are often crafted to evade detection by image analysts. Additional data can be included in image analysis applications to aid the analyst; including robustness metrics and explanation maps. Even so, adversarial attacks have managed to corrupt or evade many of these additional tools. HITL defenses are needed in order to thwart attacks targeting the human element of computer vision. We present the need for further research in HITL applications and human observation in the adversarial domain.

[1]CAE Inc. at Air Force Research Laboratory, Wright-Patterson AFB, USA [2]Leidos Inc. at Air Force Research Laboratory, Wright-Patterson AFB, USA [3]Warfighter Interactions & Readiness Division, Air Force Research Laboratory, Wright-Patterson AFB, USA. Correspondence to: Ryan McCoppin <ryan.mccoppin.1.ctr@afrl.af.mil>, Sean Kennedy <sean.kennedy.10@afrl.af.mil>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

2. Related Work

2.1. Attacking Active Learning

Humans are involved throughout the learning pipeline, including in curating training data. Bespoke training data is unlabeled and is labeled at substantial expense. Active learning involves more efficient labeling methods and is largely separated from the field of adversarial machine learning. While several studies have looked at misleading active learning querying (Miller et al., 2017) there is much to explore in how active learning is affected by adversarial attacks. Specifically, how can a model query and labeling be enacted in an adversarial environment without poisoning the model.

2.2. Manipulating Model Explanations

Model utility hinges on user trust, which requires reasonable reliability measures (e.g. confidence intervals) and understanding why it made a particular decision. Past work has developed attacks that can modify classifier explanations without changing the classifier prediction. (Dombrowski et al., 2019). The relationship between model outputs and trust is complicated, and a recent report found that user trust following attacks is poorly studied (Xiong et al., 2022). Explanation maps which are expected to reveal adversarial images (Ye et al., 2020) may be manipulated to disguise adversarial attacks and model biases. This poses obstacles not only for model trust but also for HITL evaluation (Figure 1). While solutions have been put forward which provide robustness toward manipulation (Dombrowski et al., 2022), the issue remains when analysts are looking at unfamiliar classes or using traditional explanatory techniques.

3. Opportunity: Human-In-The-Loop Studies

3.1. Human Vision and Human Vision Models

Because humans are not fooled by adversarial images in the same way as deep networks, humans and human vision models may contribute to adversarial robustness. Human attention models can be used to predict where humans will look when viewing a scene. Task-specific attention models can be built with gaze tracking data or crowd-sourced using interfaces that direct users to highlight (or deblur) areas of an image that are critical to their classification decision

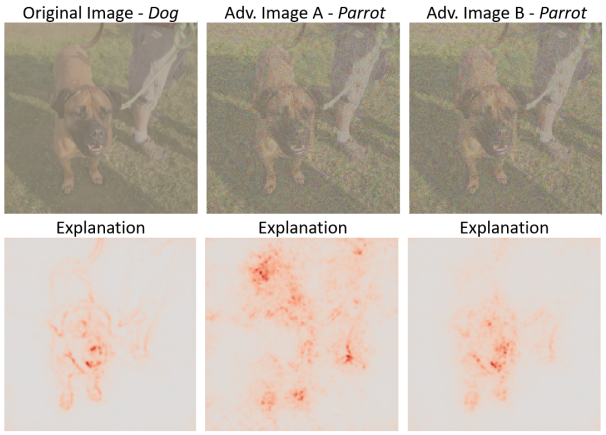

Figure 1. Adversarial images can have manipulated explanations. Image A - adversarial image; explanation reveals target class. Image B - adversarial image; manipulated explanation hides target class. Manipulation based on (Dombrowski et al., 2019)

(Linsley et al., 2017).

Disagreements between human attention models and model explanations may indicate manipulated images, low salient classes or faulty models. Taking user gaze into account could also improve model value by reducing user workload, simplifying training, or improving user performance (Matzen et al., 2016).

3.2. Prototype Humans-in-the-Loop Tools

Active and Interactive machine learning methods allow users to label data within an interface. This can be used to investigate attacks on active/interactive labeling, user interfaces, and be used to train models from user responses. One of our prototype HITL tools examines how users can contribute to adversarial or poisoned image detection. Users assign 'cards' that contain images and metadata to 'poisoned' or 'benign' categories, as seen in Figure 2. Visual explanations of a classifier's decision, obtained via Grad-CAM (Selvaraju et al., 2017), provide the user with regions of an image the ML model focuses on during classification. As user annotated data is collected, a poison detection ML model is updated via active learning. With this tool, we aim to explore:

- How do adversarial detection models compare to analyst detection capability?
- What explanations do users find useful and convincing?
- Do more robust explanations improve human performance?
- Which attacks are really invisible? What tools can make them more visible?

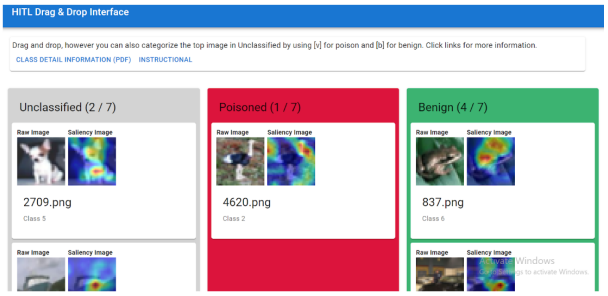

Figure 2. A HITL labeling interface with explanation maps

4. Challenges

There are many challenges in HITL research concerning adversarial robustness that have yet to find answers. These concerns include visualizing perturbations, which explanation maps are robust and beneficial, the reliability of active learning query methods and which additional metrics may be beneficial to the analyst.

Generally, users are unable to detect adversarial images and thus rely on "explanation maps" to provide visibility. These maps can also be adversarial distorted (Dombrowski et al., 2019). If explanation maps are not always reliable or truthful, what information can be provided to the user in order to detect adversarial images and improve model robustness? In addition, is it possible to gain a robustness to these distortions in a similar manner to adversarial training? We believe human vision models may provide a key in detecting and interpreting adversarial images. However, recent research has shown that deep models of human attention are also vulnerable to adversarial attack (Che et al., 2020). Can we improve the adversarial robustness of these models and combine human and machine attention to identify adversarial images? Research has already begun in adversarial robustness, but not in how it relates to HITL evaluation and how analysts are able to use more robust explanations.

References

Che, Z., Borji, A., Zhai, G., Ling, S., Li, J., and Le Callet, P. A new ensemble adversarial attack powered by long-term gradient memories. 34:3405–3413, Apr. 2020.

Dombrowski, A. K., Alber, M., Anders, C., Ackermann, M., Müller, K. R., and Kessel, P. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems 32*. Elsevier, 2019.

Dombrowski, A. K., Anders, C. J., Müller, K. R., and Kessel, P. Towards robust explanations for deep neural networks. In *Pattern Recognition*. Elsevier, 2022.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Computer Science Department, University of Toronto, Toronto, Canada, 2009.

Linsley, D., Eberhardt, S., Sharma, T., Gupta, P., and Serre, T. What are the visual features underlying human versus machine vision? In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 2706–2714, 2017.

Matzen, L., Haass, M., Tran, J., and McNamara, L. Using eye tracking metrics and visual saliency maps to assess image utility. *Electronic Imaging*, 2016:1–8, 02 2016.

Miller, D. J., Hu, X., Qiu, Z., and Kesidis, G. Adversarial learning: A critical review and active learning study. In *27th International Workshop on Machine Learning for Signal Processing*, pp. 1–6. IEEE, 2017.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. In *International Journal of Computer Vision (IJCV)*. Springer, 2015.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Xiong, P., Buffett, S., Iqbal, S., Lamontagne, P., Mamun, M., and Molyneaux, H. Towards a robust and trustworthy machine learning system development: An engineering perspective. In *Journal of Information Security and Applications*, pp. 65. Elsevier, 2022.

Ye, D., Chen, C., Liu, C., Wang, H., and Jiang, S. Detection defense against adversarial attacks with saliency map. In *International Journal of Intelligent Systems*. Wiley Online Library, 2020.