Exploring Invariance-Based Adversarial Examples: Algorithmic Generation and Human Perception

Samuel Oberhofer[1], Martin Nocker[2], Pascal Schöttle[3]

[1] MCI The Entrepreneurial School, Innsbruck, sa.oberhofer@mci4me.at
[2] MCI The Entrepreneurial School, Innsbruck, martin.nocker@mci.edu
[3] MCI The Entrepreneurial School, Innsbruck, pascal.schoettle@mci.edu

Abstract. This paper investigates the generation and human perception of invariance-based adversarial examples (IBAEs), a class of adversarial attacks in which significant semantic changes in an image do not affect the classifier's prediction. We propose a novel method for generating IBAEs by leveraging explainable artificial intelligence (XAI) techniques to compute importance maps, which guide the insertion of patches from attack images into base images. Unlike previous approaches that rely on human gaze data, our method is more scalable and adaptable. We validated our method by generating IBAEs using the ImageNette dataset and Integrated Gradients to generate importance maps. A human study with 166 participants was conducted to assess the divergence between machine and human perception of these examples. The results demonstrate the successful generation of IBAEs and show that IBAEs can deceive time constrained human labelers, with the probability of being fooled increasing as the proportion of the attack image in the IBAE increases. We further refine the generation algorithm to increase IBAE effectiveness and contribute a dataset of 114 high-quality IBAEs, based on ImageNette, for further research. Our findings underscore the importance of aligning machine learning models with human perception for robust visual classification systems.

Keywords: adversarial machine learning, invariance-based adversarial examples

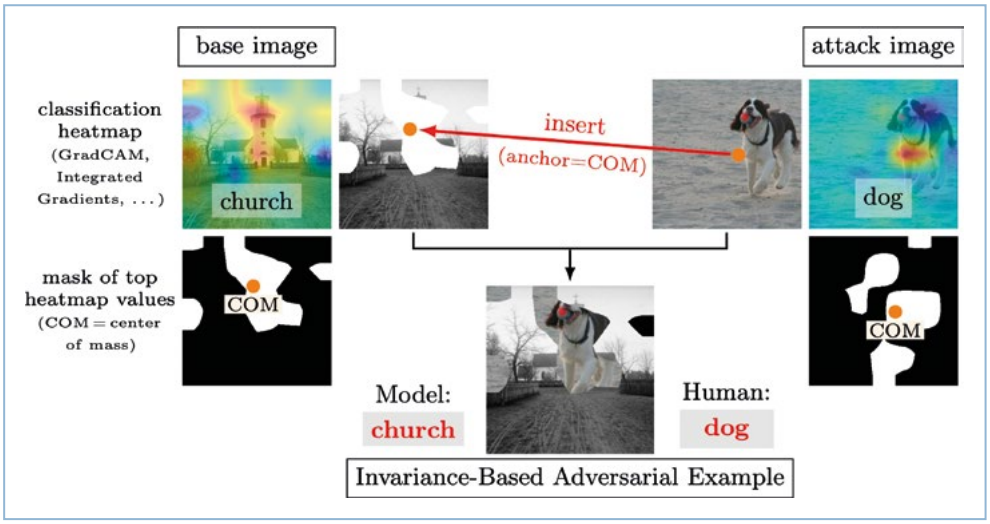

Figure 1: Overview of the IBAE generation algorithm. Important parts of a base image are exchanged by important parts of an attack image as much as possible such that the model's classification remains unchanged. The importance of image parts is determined by XAI methods, typically in the form of heatmaps.

Introduction

Machine learning (ML) algorithms are increasingly applied across diverse fields like autonomous driving (Bachute & Subhedar, 2021), finance (Ahmed et al., 2022), and medicine (Piccialli et al., 2021). In computer vision (Voulodimos et al., 2018), Convolutional Neural Networks (CNNs) and Transformers have achieved near-human performance in image classification. However, the rise of adversarial attacks underscores the need to evaluate ML models under adversarial conditions. Since 2014, researchers have demonstrated how subtle image modifications, which are imperceptible to humans, can alter classifier outputs, highlighting the importance of aligning machine predictions with human perception (Szegedy et al., 2014). Such adversarial examples can be divided into two types. Sensitivity Based Adversarial Examples (SBAE) (Tramèr et al., 2020) introduce human-imperceptible changes to an image that alter a machine classifier's output, while Invariance Based Adversarial Examples (IBAE) (Jacobsen et al., 2020) exploit a model's excessive invariance by significantly altering the image's semantics without changing its output. In this work, we focus on IBAEs and their automatic generation. Additionally, we validate the generated images in a human subject study.

Methods

To generate IBAE, we first start with a base image and an attack image. These images are taken from the ImageNette dataset (GitHub – fastai/Imagenette, 2019). The

goal is to take the "most important" patches from the attack image and insert them at the "most important" location of the base image. Unlike previous methods, which used human eye-gazing data (Merkle et al., 2024), we generate our importance data using methods from the field of Explainable Artificial Intelligence (XAI). These models give us a heatmap, which shows how important each pixel was for the classifier's final decision. We start with small patches and check whether the classifier's output has changed. Iteratively increasing the patch size yields our IBAE candidates. These candidates represent an image with the maximum amount of attack image possible, while still being classified as the base class by the image classifier. We call them candidates, as we cannot yet determine whether human labelers would assign a class other than the base class, and the misalignment of human and machine classification is the definition of a valid adversarial example. This generation process is illustrated in Figure 1. To validate these IBAE candidates, we perform a human subject study with 166 participants. The images are shown for one second, after which the participants had to choose one out of four available labels. Analyzing the participants' responses, we can identify some features of IBAEs which lead to a high probability of fooling the labelers. This knowledge can be used to tweak our generation algorithm in a way which produces highly effective IBAEs.

Results

First, we looked at how often human labelers disagreed with machine classification for the different image classes. We call an IBAE successful if human labelers assign a label other than that of the base image. The results show that images with the label "chain saw" were classified using a different label most often (76.6 %) versus the class "golf ball", which was mostly correctly identified by participants (only 19.6 % were fooled). We could also see a correlation between the amount of base image present in the final IBAEs and the proportion of participants who assigned the base label during the experiment. Especially for the classes "garbage truck" and "golf ball" the correlation coefficient was very high (0.84 and 0.88, respectively). This means that IBAEs are more likely to be successful if we are able to insert a large amount of the attack image before the classifier flips its prediction. Also, remember that we stop increasing the size of our masks when the classifier switches its prediction away from the base class. The new prediction does not necessarily need to be the label of the attack image. We frequently observed that the image classifier switched its prediction to another label completely. In our study, we saw that IBAEs images were more likely to be successful when the final prediction of the machine classifier was the same as the label of the attack (71.4 % vs. 49.0 %).

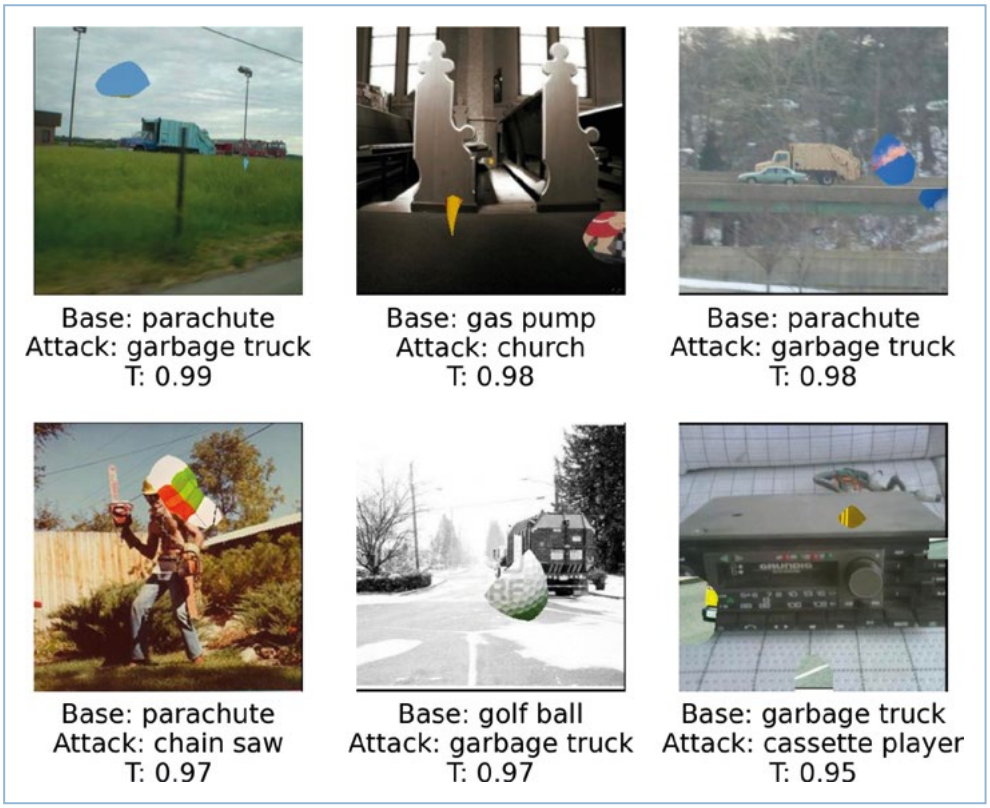

Figure 2: IBAEs with almost no pixels from the base image left. *T* is the threshold at which the model flipped its prediction.

Discussion

Our results show that the generated IBAEs can reliably lead to differing classifications by humans and machines. Our generation algorithm can be adapted to only look for IBAEs with a high amount of the attack image inserted, since our results show that such examples are more likely to succeed. Looking at the results we obtained from the user study, we conducted another round of IBAE generation where we selected images with a high amount of attack image and where we only included images with the final label as the attack label. This resulted in a dataset of 114 IBAEs, all consisting of at least 80 % attack image. Based on our findings, we expect this dataset to contain a substantial number of high-quality IBAEs where the majority of humans are expected to assign the attack label when classifying them under a time constraint.

Figure 2 shows some examples from this dataset. The dataset is anonymously available and will be made publicly available upon acceptance for further research in the field of adversarial examples. In conclusion, we demonstrate the efficacy of IBAEs

and validate them in a human subject study. With the release of a high-quality IBAE dataset, we contribute to the ongoing research on adversarial attacks providing implications for building more robust and secure computer vision systems.

References

Ahmed, S., Alshater, M. M., Ammari, A. E., & Hammami, H. (2022). Artificial intelligence and machine learning in finance: A bibliometric review. *Research in International Business and Finance*, *61*, 101646. https://doi.org/10.1016/j.ribaf.2022.101646

Bachute, M. R., & Subhedar, J. M. (2021). Autonomous driving architectures: Insights of machine learning and deep learning algorithms. *Machine Learning with Applications*, *6*(8), 100164. https://doi.org/10.1016/j.mlwa.2021.100164

GitHub – fastai/imagenette: A smaller subset of 10 easily classified classes from Imagenet, and a little more French, 2019. https://github.com/fastai/imagenette/tree/master (retrieval date 01/12/2025)

Jacobsen, J.-H., Behrmann, J., Zemel, R., & Bethge, M. (2020). Excessive invariance causes adversarial vulnerability. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA.*

Merkle, F., Sirbu, M. R., Nocker, M., & Schöttle, P. (2024). Generating invariance-ased adversarial examples: Bringing humans back into the loop. In G. L. Foresti, A. Fusiello, & E. Hancock (Eds.), *Image analysis and processing – ICIAP 2023 workshops. Proceedings, part II* (pp. 15–27). Cham: Springer. https://doi.org/10.1007/978-3-031-51023-6_2

Piccialli, F., Somma, V. D., Giampaolo, F., Cuomo, S., & Fortino, G. (2021). A survey on deep learning in medicine: Why, how and when? *Information Fusion*, *66*, 111–137. https://doi.org/10.1016/j.inffus.2020.09.006

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. arXiv:1312.6199. https://doi.org/10.48550/arXiv.1312.6199

Tramèr, F., Behrmann, J., Carlini, N., Papernot, N., & Jacobsen, J.-H. (2020). Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *Proceedings of the 37th International Conference on Machine Learning, ICML*.

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, *2018*, e7068349. https://doi.org/10.1155/2018/7068349