AdvEx: Interactive Visual Explorations of Adversarial Evasion Attacks

by

Yuzhe You

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2023

Author's Declaration

This thesis consists of materials all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis is based on the research project I led under the supervision of Dr. Jian Zhao. As the main investigator, I am the primary system designer and builder. I contributed to designing and implementing the system, conducting the user studies, analyzing data and writing the draft manuscript. The draft was further edited by Dr. Jian Zhao and Jarvis Tse. Jarvis Tse contributed to implementing the confidence score view, designing and conducting the user studies, and analyzing data. Hanyu Xu contributed to conducting the user studies. I would also like to thank Parjanya Vyas and Qing Guo for their contributions during the early stages of this thesis.

Abstract

Adversarial machine learning (AML) focuses on studying attacks that can fool machine learning algorithms into generating incorrect outcomes as well as the defenses against worst-case attacks to strengthen the adversarial robustness of machine learning models. Specifically for image classification tasks, it is difficult to comprehend the underlying logic behind adversarial attacks due to two key challenges: 1) the attacks exploiting "non-robust" features that are not human-interpretable and 2) the perturbations applied being almost imperceptible to human eyes. We propose an interactive visualization system, AdvEx, that presents the properties and consequences of evasion attacks as well as provides data and model performance analytics on both instance and population levels. We quantitatively and qualitatively assessed AdvEx in a two-part evaluation including user studies and expert interviews. Our results show that AdvEx is effective both as an educational tool for understanding AML mechanisms and a visual analytics tool for inspecting machine learning models, which can benefit both AML learners and experienced practitioners.

Acknowledgements

I would like to thank everyone who made this thesis possible.

First of all, I cannot thank my supervisor Dr. Jian Zhao enough. He has been my greatest source of inspiration and was the one who ignited my passion for visualization research. His knowledge and encouragement have meant so much to me in helping me find my research direction and leading me toward academic excellence.

I would also like to extend my thanks to my thesis readers Dr. Daniel Vogel and Dr. Jimmy Lin. Their insightful feedback has helped me further refine this thesis by taking it to a higher level.

In addition, I would like to thank all the research partners and brilliant peers that I have worked with during my time here. It is only because of their valuable skills and endeavors that we have been able to accomplish so much together.

Finally, I want to express my gratitude to all my loved ones who have been constantly supporting me throughout my academic pursuit. They have given me the strength I need to keep pursuing my academic dreams with determination.

I am forever grateful to all the people that have inspired me, mentored me, and supported me during my graduate study. It is thanks to them that I am who I am today and have been able to reach this milestone in my academic journey.

Table of Contents

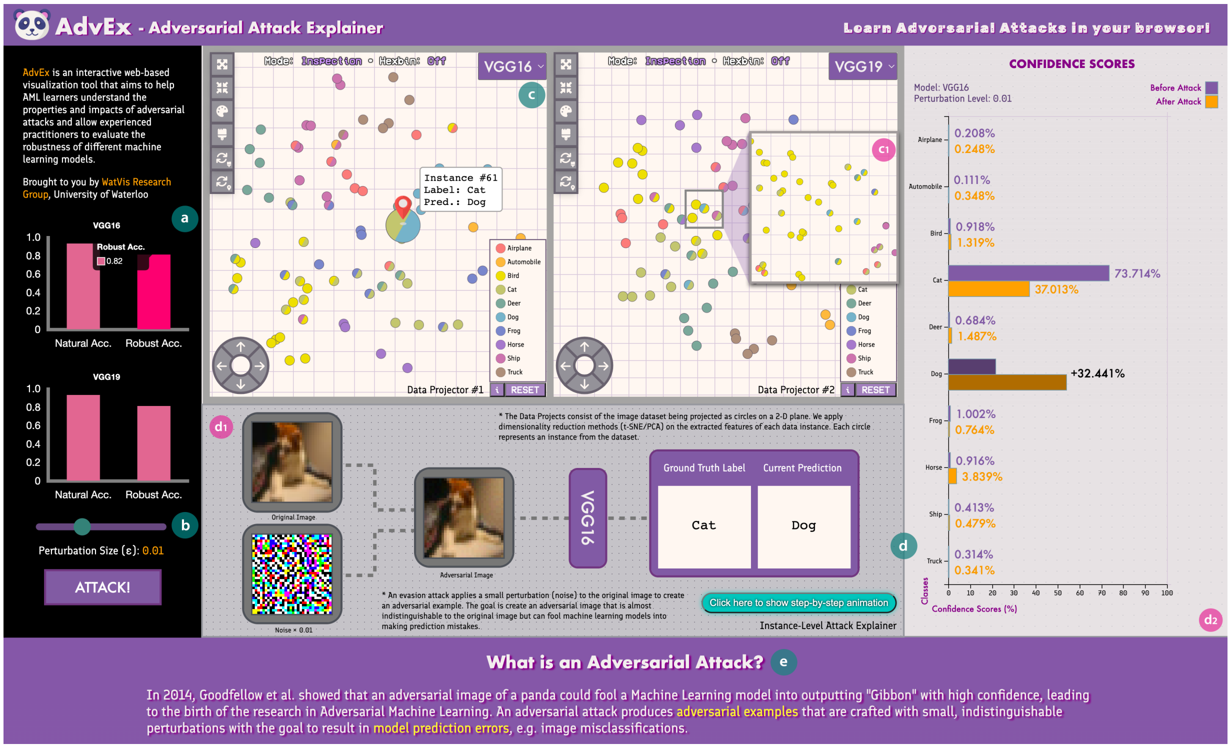
List of Figures

List of Tables

Chapter 1

Introduction

Deep learning models (e.g., neural networks) have achieved remarkable success in diverse domains that influence our life, including in safety-critical applications such as facial recognition [9] and autonomous driving [20]. Nonetheless, these models have been proven to be quite brittle to minor perturbations around the input data. In 2014, Goodfellow et al. [8] showed that an adversarial image of a panda could easily fool GoogLeNet [41] into labeling it as a gibbon with high confidence, leading to the birth of the research in *adversarial machine learning* (AML). Subsequently, Eykholt et al. [7] showed that road signs modified with physical perturbations could achieve high misclassification rates in road sign classifiers, and Lin et al. [22] demonstrated that similar methods could be employed to evade facial recognition systems. These types of adversarial attack are known as evasion attacks, which produce deceptive inputs (e.g., adversarial images) that are crafted maliciously with imperceptible perturbations to fool models into making mistakes. Although more and more students/practitioners are studying and applying machine learning, many of them are uninformed about the danger of adversarial attacks to their models due to their lack of understanding in AML. As a result, the models developed often achieve good accuracy on natural datasets but are highly susceptible to attack-perturbed inputs [40]. To help developers design or calibrate their models to be adversarial robust for real-world applications, it is essential to educate them on the concepts and risks of adversarial attacks and help them assess if their models can maintain reliable performance under these attacks. Many studies have shown that visualizations serve as an effective means of explaining machine learning concepts [45, 16, 14] and enabling model evaluations [48, 36], which has motivated our study to visualize adversarial attacks. For this work, we focus on evasion attacks in image classification, a highly active AML research path that most existing work (e.g., [8, 15, 47]) focuses on since such models are frequently used in safety-critical applications



AdvEx - Adversarial Attack Explainer

Learn Adversarial Attacks in your browser!

AdvEx is an interactive web-based visualization tool that aims to help AML learners understand the properties and impacts of adversarial attacks and allow experienced practitioners to evaluate the robustness of different machine learning models.

Brought to you by WatVis Research Group, University of Waterloo

Perturbation Size (ε): 0.01

Mode: Inspection ∘ Hexbin: Off

Instance #61
Label: Cat
Pred.: Dog

* The Data Projects consist of the image dataset being projected as circles on a 2-D plane. We apply dimensionality reduction methods (t-SNE/PCA) on the extracted features of each data instance. Each circle represents an instance from the dataset.

Ground Truth Label	Current Prediction
Cat	Dog

* An evasion attack applies a small perturbation (noise) to the original image to create an adversarial example. The goal is create an adversarial image that is almost indistinguishable to the original image but can fool machine learning models into making prediction mistakes.

Model: VGG16
Perturbation Level: 0.01

What is an Adversarial Attack?

In 2014, Goodfellow et al. showed that an adversarial image of a panda could fool a Machine Learning model into outputting "Gibbon" with high confidence, leading to the birth of the research in Adversarial Machine Learning. An adversarial attack produces adversarial examples that are crafted with small, indistinguishable perturbations with the goal to result in model prediction errors, e.g. image misclassifications.

Figure 1.1: AdvEx user interface: (a) Robustness Analyzers that display the models' prediction accuracy pre- and post-attack; (b) Perturbation Adjuster that initiates the attack sequence with specified magnitude; (c) Data Projectors that visualize data embeddings in a 2-D latent space; (d) Instance-level Attack Explainer that displays in-depth information of the highlighted instance; (e) General Information Provider that provides more background on AdvEx and AML.

[9, 20].

There are certain key challenges in understanding adversarial attacks for image classification. Firstly, Ilyas et al. [15] demonstrated that image datasets contain both "robust" features that are aligned with human perceptions and "non-robust" features that are not interpretable yet still effective for model predictions. Adversarial examples exploit those useful but "non-robust" features that are highly predictive and well-generalized over an entire dataset, but since these features are not human-interpretable, they appear as "noise" to us. Secondly, since the perturbations applied by the attacks tend to be very subtle, the resulting adversarial images can be almost indistinguishable from the original versions. Therefore, visualizations need to be deliberately designed to illustrate these attacks, including their underlying logic, their ramifications across large-scale datasets and on individual instances, and how different models behave differently under the same attack. While several related visualization systems have been proposed, existing work still possesses critical limitations. For instance, Bluff [5] visualizes how adversarial attacks confuse deep neural networks (DNNs) by displaying the features induced by the perturbations, but it does not illustrate the attack impact across a larger dataset, nor the performance of different models under the same attack. Similarly, Adversarial-Playground [28] only supports one model with very few images; it also does not demonstrate the common "imperceptible-ness" of the attacks as the perturbations applied to the images are highly visible.

To address these challenges, we parameterized the adversarial perturbations applied to data instances and illustrated the processes and impacts of adversarial attacks in the form of interactive visualizations. Specifically, we developed AdvEx, which presents the properties and consequences of adversarial attacks and provides data and model performance analytics (Figure 1.1). The goals of AdvEx are to 1) help novice *learners* understand adversarial attacks and 2) allow experienced *practitioners* to assess the robustness of their trained models. To the best of our knowledge, AdvEx is the first visualization system designed specifically to support both learning and evaluation in AML. Moreover, AdvEx visualizes attack information and model performance on both instance and population levels. By doing so, the users may not only obtain a high-level understanding of how adversarial attacks alter large-scale datasets and target various models, but also are provided with the option to conduct more detailed inspections of the way each instance is perturbed.

We quantitatively and qualitatively assessed AdvEx in a two-part evaluation with its two goals (i.e., as means of learning and inspecting AML) in mind. First, we performed a user study to assess the learning effects of AdvEx on novice AML learners and gathered subjective feedback from them. Second, we conducted interview sessions with AML experts to collect in-depth feedback for AdvEx as a visual analytics tool. The results of our studies show that AdvEx is highly effective both as an educational and a visual analytics tool. Additionally, our studies provide comprehensive insights into the strengths and weaknesses of AdvEx from various perspectives. In summary, our contributions with this thesis include:

- A novel interactive visualization system, AdvEx, for both novice AML learners and experienced practitioners to gain a comprehensive understanding of adversarial attacks;

- Empirical findings on how AdvEx and the designed visualizations can help users understand the underlying properties and consequences of adversarial attacks and evaluate the robustness of trained models.

Chapter 2

Background

In this chapter, we provide a summary of the related studies present in the literature. We divide these works into three categories: 1) adversarial machine learning (AML), 2) visualizations of adversarial attacks, and 3) educational visualizations for learning neural networks.

2.1 Adversarial Machine Learning

Many adversarial attacks have been proposed to work under different threat models, namely white-box and black-box attacks. A white-box attack assumes that the attacker has full access to the model's internals, while a black-box attack assumes that the attacker can only access model inputs and outputs. Fast Gradient Sign Method (FGSM) [8], Basic Iterative Method (BIM) [19], and Projected Gradient Descent (PGD) [26] are a few of the well-known white-box gradient-based attacks. At the same time, efficient black-box attacks such as Zeroth Order Optimization (ZOO) [2] and One Pixel Attack (OPA) [39] have been explored extensively in the literature as well. To counter adversarial attacks, various defense methods have been proposed to fortify the robustness of models against adversarial inputs. For instance, TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization (TRADES) [47] is the state-of-the-art method for training an adversarially robust DNN by leveraging the observed trade-off between robustness and accuracy through a regularized surrogate loss. Other examples of adversarial defenses include ensemble adversarial training [42], generative adversarial training [21], autoencoder-based input denoising [27], etc.

While our system can be generalized to all kinds of perturbation-based evasion attacks, in this thesis, we use the FGSM attack, one of the earliest and most well-known adversarial attacks [8], as an example to demonstrate the features and functionalities of AdvEx. Several prior studies have tried to understand the characteristics of the FGSM attack. Zhang et al. [49] discovered that FGSM attack may create not only 2-D adversarial images but also 3-D adversarial examples by applying the attack methodology to PointNet [34], a DNN designed for 3-D point cloud data. Crecchi et al. [3] introduced a new detector for adversarial examples and experimentally proved that examples generated by the FGSM attack could be distinguished from manifold samples using nonlinear dimensionality reduction techniques such as t-SNE [44]. Pan et al. [29] proposed a distance-based technique to identify classes susceptible to the FGSM attack and evaluated their approach with benchmark datasets such as MNIST [6], Fashion MNIST [46], and CIFAR-10 [18]. The abundance of existing works on the FGSM attack shows that this is a well-known adversarial attack and hence a good introductory example for those new to AML. For AdvEx, we utilize interactive visualizations to help users explore the characteristics of adversarial attacks like the FGSM attack. As AML is a relatively new area of machine learning research, it is crucial to raise awareness on these attacks to encourage practitioners to build safer AI applications, especially those that are safety-critical and rely heavily on model robustness.

2.2 Visualizations of Adversarial Attacks

Several visualization tools designed to illustrate adversarial attacks have been proposed in the past studies. Adversarial-Playground [28] is a web-based application that visualizes adversarial attacks by demonstrating the efficacy of common adversarial methods against a simple CNN. The tool allows its users to choose from a set of pre-defined inputs and displays clean and adversarial images side by side to illustrate the impact of an attack. Similarly, Bluff [5] visualizes and characterizes adversarial attacks on vision-based neural networks. However, instead of focusing on explaining the attack logic, it compares the activation pathways of benign and attacked images by highlighting the neurons and connections that an attack exploits to confuse the model. In addition, Lin and Soylu designed AdVis.js [23], a system that visualizes the FGSM attack by comparing the original and adversarial images side by side and displaying the heat maps of the perturbed pixels. Cao et al. [1] proposed AEVis, a visual analytics tool that analyzes adversarial attacks targeted at DNNs by extracting critical neurons and their connections and showing how adversarial examples deactivate and activate specific features to fool the models. Ma et al. [25] proposed a visual analytics framework that employs a multi-faceted visualization scheme to support the analysis of data poisoning attacks from the perspectives of models, data instances, features, and local impacts.

Nonetheless, these works still do not fulfill our design goals for visualizing adversarial attacks in several ways. For instance, Adversarial-Playground [28] and AdVis.js [23] fall short with their overly simplistic interfaces that juxtapose an adversarial image with its original, which becomes less effective when the two images look identical due to minor perturbations. Bluff [5], on the other hand, relies on visualizing the internal model logic on benign and adversarial examples, sacrificing model generalizability. Both tools fail to depict attack impact across a larger dataset, nor the performance of different models under the same attack. As for the advanced visual analytic tools, AEVis [1] lacks model comparisons and population-level visualizations, while Ma et al.'s work [25] is limited to data poisoning attacks in binary classification, lacking support for evasion attacks in multiclass classification. Both systems are also designed for experienced practitioners to perform model analysis, featuring intricate visualizations that may be challenging for AML learners to comprehend. Therefore, these tools are either too simplistic or excessively intricate for our target audience, or cannot effectively demonstrate an attack's impact on a larger dataset or how models with varying robustness exhibit different behaviors under the same attack. In contrast, for AdvEx, we aim to enable users who have little or no knowledge of AML to learn adversarial attacks on both population and instance levels, and also allow experienced practitioners to evaluate the adversarial robustness of multiple models at once.

Moreover, dimensionality reduction methods have been used extensively to understand and visualize adversarial attacks by projecting representations of data instances onto a low-dimensional space. Ma et al.'s proposed framework [25] for data poisoning attacks contains a projection view that utilizes t-SNE to display the poisoned dataset and visualize global data distributions in a scatterplot. Park et al. [31] proposed VATUN, an interactive visualization system that also uses t-SNE to create a data embedding view that interactively visualizes the impacts of adversarial attacks and data augmentations. In addition, Principal Component Analysis (PCA) [32] has been used to understand adversarial data and as a form of anomaly detector to identify adversarial examples. Panda and Roy [30] introduced a Noise-based Learning (NoL) approach for training robust DNNs and provided a simplistic visualization tool that uses PCA for adversarial dimensionality and loss surface visualization analysis. Hendrycks and Gimpel [11] incorporated PCA into adversarial image detection and visualized how adversarial images abnormally emphasize coefficients for low-ranked principal components. Inspired by these works, in AdvEx, we apply dimensionality reduction methods to project the data embeddings onto a two-dimensional plane, and use animated transitions and colors of circular glyphs on the 2-D plane to visualize how the attacks alter the models' perception of the images.

2.3 Visualizations for Learning Neural Networks

Several visualization systems specifically designed for users to learn about neural networks have been proposed as well. For instance, GAN Lab [17] is an online visualization tool designed for non-experts to learn and experiment with generative adversarial networks (GANs). The tool allows its users to interactively train GAN models on a simple dataset and visually examine each step of the training process in real time. CNN Explainer [45] enables non-experts to learn about CNNs and inspect the interplay between low-level mathematical operations and high-level model structures. Learners can input images into a CNN and observe the intermediate outputs at every layer, gaining a full understanding of the network's inner mechanisms. Another past study has proposed Summit [14], an interactive system that provides higher-level explanations of DNNs by intuitively visualizing the image features detected by the networks and how those features interact to make predictions. The tool adapts two scalable summarization techniques to create visualizations that reveal crucial neuron associations and structures that contribute to a model's predictions, providing valuable insights into DNNs' decision-making processes.

Despite focusing on visualizing common neural networks instead of adversarial attacks, all three aforementioned studies have provided us with inspirations for the design of AdvEx. Specifically, similar to GAN Lab [17] and CNN Explainer [45], AdvEx is accessible to any user with a modern browser without the need to install specialized hardware for deep learning. Motivated by GAN Lab [17]'s step-by-step training visualization, AdvEx provides step-by-step executions of the attack methodology to visualize the intricate attack process. Like CNN Explainer [45] and Summit [14], AdvEx also adapts smooth transitions across different levels of abstraction to enable a streamlined visual exploration and to serve as the link that connects different views of the visualization tool. Inspired by existing work, we aim to develop AdvEx as an interactive visualization tool for adversarial attacks with compelling visualizations and animated transitions and present the attack properties at multiple levels of detail.

Chapter 3

Design Goals

Through an extensive literature survey, we came up with the following design goals to guide the development of AdvEx:

G1 Provide visual abstraction of the attack impact at multiple levels of detail. Many existing tools that visualize adversarial attacks (e.g., [28, 5]) only focus on displaying instance-level information, such as how an attack perturbs a specific image. Though instance-level details may help demonstrate the attack logic, they are insufficient to illustrate an attack's overall impact across a larger dataset. Therefore, population-level (and subpopulation-level) information also needs to be incorporated to better visualize an attack, such as how the attack decreases the model's overall accuracy and shifts its data representations. In AdvEx, visual abstractions at multiple levels are included to provide both a population-level overview of the attack's capabilities and the options to conduct more in-depth investigations on the attacked instances.

G2 Enable visual analysis of the adversarial robustness of different models under attack. Models with different architectures and training methods exhibit varying levels of robustness against the same attack. For instance, adversarially trained models tend to perform better than naturally trained models on attack-perturbed examples [8]. However, most existing learning tools [28, 5, 23] use only one arbitrary model to illustrate the attack, often with little to no information provided on the model itself. As a single model's performance cannot represent the impact of the same attack on other models, enabling visual analysis of multiple models will help users better understand the overall consequences of the attacks. For AML learners, this provides them with the opportunity to explore how different models respond differently to the same attack. For more experienced practitioners, this allows them to get a quick sense of their models' robustness in comparison to other models.

G3 Facilitate dynamic experimentation with fluid transition between attacks with different perturbation sizes. To help users get a holistic picture of the models' performance before and after an attack, and how the attack gets more aggressive as the perturbation size increases, similar to Adversarial-Playground [28], we aim to have the users dynamically experiment with the perturbation size. Interfaces are included to enable easy manipulation of the perturbation size, and seamless animations are incorporated to transition between different sizes and vividly visualize the changes in attack impact. The combination of dynamic experimentation and fluid animations helps users quickly grasp the correlation between the perturbation size and the attack strength, and allows them to effortlessly track the travelling path of data instances in the projected latent space.

G4 Allow step-by-step execution for learning the attack process in detail. Adversarial attacks often require very minor perturbations to produce a model mistake, thus the resulting adversarial images could be almost indistinguishable from the original images. While existing work like Adversarial-Playground [28] illustrates adversarial attacks by placing the original and perturbed images side by side, it fails to emphasize the common "imperceptible-ness" of the attacks as the perturbations shown are highly visible. Therefore, to help users intuitively understand the attack logic while preserving the "imperceptible-ness" of adversarial images, we aim to include a step-by-step execution of the attack process along with a side-by-side comparison of the natural and perturbed images.

G5 Integrate beginner-friendly user interface design for AML learners. As one of our goals is to develop an interactive system that can introduce users to AML, we want to make sure AdvEx is easy to understand and accessible to learners who are unfamiliar with this topic. To accomplish this goal, we aim to make AdvEx welcoming by accompanying our visualizations with beginner-friendly designs and ensuring that AdvEx is not too overwhelming for learners to digest. Additionally, AdvEx focuses more on the visual exploration of the attack process instead of excessively emphasizing the architecture or internal logic of the models.

Chapter 4

AdvEx

With the above design goals in mind, we developed AdvEx. In this chapter, we provide an overview of AdvEx and then describe our backend system as well as each interface component in great detail.

4.1 System Overview

As briefly discussed, we designed AdvEx as a web-based interactive visualization system with two primary goals in mind: 1) to help AML learners understand the properties and impacts of adversarial attacks and 2) to allow experienced practitioners to evaluate the robustness of different machine learning models. As depicted in Figure 4.1, AdvEx consists of two system modules: A) a *backend pipeline* (see section 4.3) and B) a *frontend user interface* (see section 4.4).

In the backend pipeline, an *Attacker* module fetches the image dataset and converts the data into numeric matrices normalized between 0 and 1. The Attacker conducts attacks on the converted dataset to create adversarial instances of the original data. The original instances, along with their adversarial counterparts, are both fed into the selected models to obtain information such as image embeddings, confidence scores, model prediction accuracy, etc. Additionally, an *Embedding Projector* is responsible for extracting each model's embedding vectors by removing its final output layer and retrieving outputs of the backbone only. Dimensionality reduction methods such as t-SNE [44] and PCA [32] are applied to the extracted embeddings to project the data representations onto a 2-D space. The output information is relayed to the frontend components for visual display.

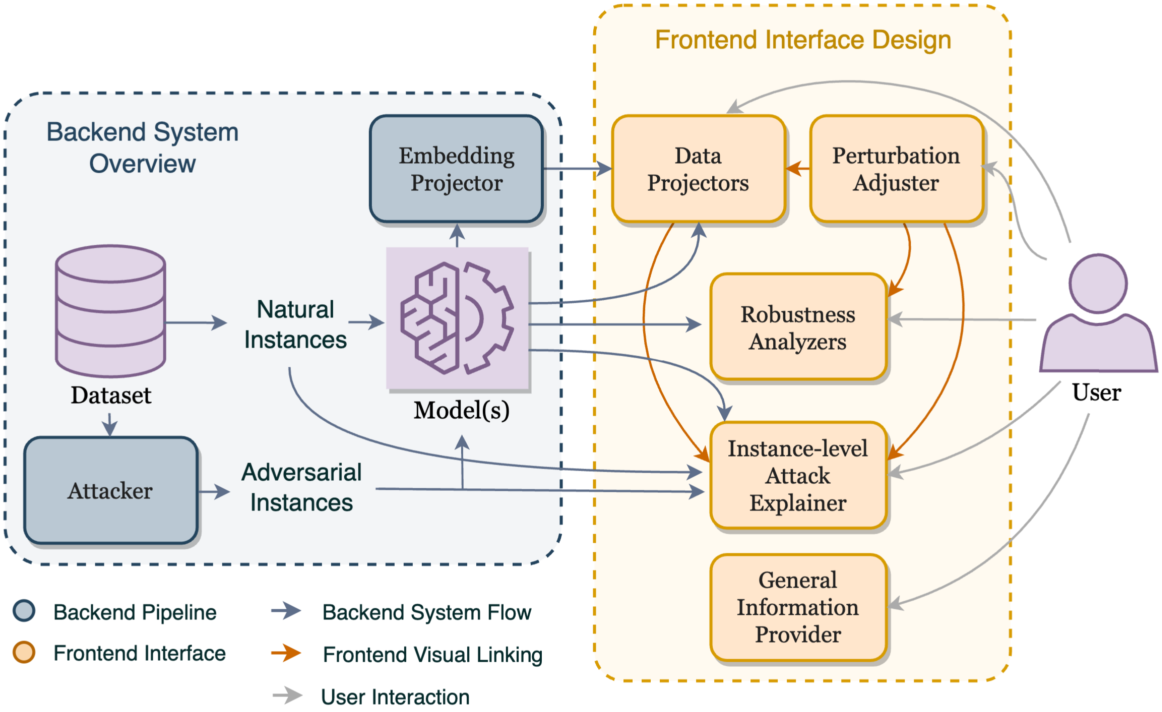


Figure 4.1: The schematic diagram depicting the system architecture.

The frontend visual interface can be divided into the following components: 1) *Data Projectors* (Figure 1.1c), 2) *Instance-level Attack Explainer* (Figure 1.1d), 3) *Robustness Analyzers* (Figure 1.1a), 4) *Perturbation Adjuster* (Figure 1.1b), and 5) *General Information Provider* (Figure 1.1e) + interactive tutorials. The Robustness Analyzers are a pair of interactive bar charts that quickly evaluate the models' overall performance under the current attack (**G1**) and provide comparisons of the models' robust accuracy to their natural accuracy (**G2**). The Data Projectors utilize coordinates provided by the Embedding Projector to visualize data in 2-D interactive scatterplots, enabling explorations of how the attack alters the data representations (**G1**) with side-by-side model comparisons (**G2**). The Instance-Level Attack Explainer reflects more details regarding a specific instance (**G1**) and includes a confidence score view and a step-by-step attack execution feature that further illustrates the attack process (**G4**). The Perturbation Adjuster allows the user to select the desired attack strength and is directly linked to the three aforementioned components to initiate animated sequences and simulate the attack (**G3**). Lastly, combined with interactive tutorials, the General Information Provider guides the user through the navigation of the interface and provides additional background on AML (**G5**).

4.2 Dataset and Models

In this thesis, we use the CIFAR-10 dataset for demonstrating AdvEx, but our system can be employed with any image classification datasets. The CIFAR-10 dataset [18] consists of 60,000 32×32 colored images from 10 different classes (50,000 training data and 10,000 testing data), with 6,000 images per class. We choose this dataset due its popularity of being used in machine learning research to evaluate the natural accuracy and adversarial robustness of image classification models (e.g., [47, 4, 12]).

In addition, AdvEx supports a variety of image classification models and allows the user to compare two models side by side (**G2**). For instance, the user may want to compare the robustness of CNNs with the same architecture but different numbers of convolutional layers. Alternatively, they can investigate how a model trained adversarially may outperform a model trained naturally under attack. For this thesis, we loaded two pairs of models for our studies: 1) VGG-16 vs. VGG-19 [37], and 2) ResNet-34 [10] trained naturally vs. trained adversarially with TRADES [47].

4.3 Backend Pipeline

In this section, we describe how the backend processes and analyzes the data in AdvEx, including how it generates the adversarial examples and prepares the data instances and model outputs for frontend display.

4.3.1 Attacker Module

The backend "Attacker" module produces adversarial examples of the original dataset by conducting adversarial attacks on the targeted models. It first retrieves the dataset and normalizes all images' pixel values between 0 and 1, then conducts the attack on the data instances to create the adversarial images. Here we use the FGSM attack [8] as an example for demonstrating our system, but AdvEx can be easily employed with various other types of evasion attacks.

We choose the FGSM attack due to its notoriety for creating the very first adversarial image, namely the panda image from [8] that is well-known among AML researchers. The attack is commonly used as a baseline method to evaluate the robustness of machine learning models and the effectiveness of adversarial training methods (e.g., [26, 49, 35]). In addition, compared to other forms of evasion attacks, the FGSM attack is relatively simple

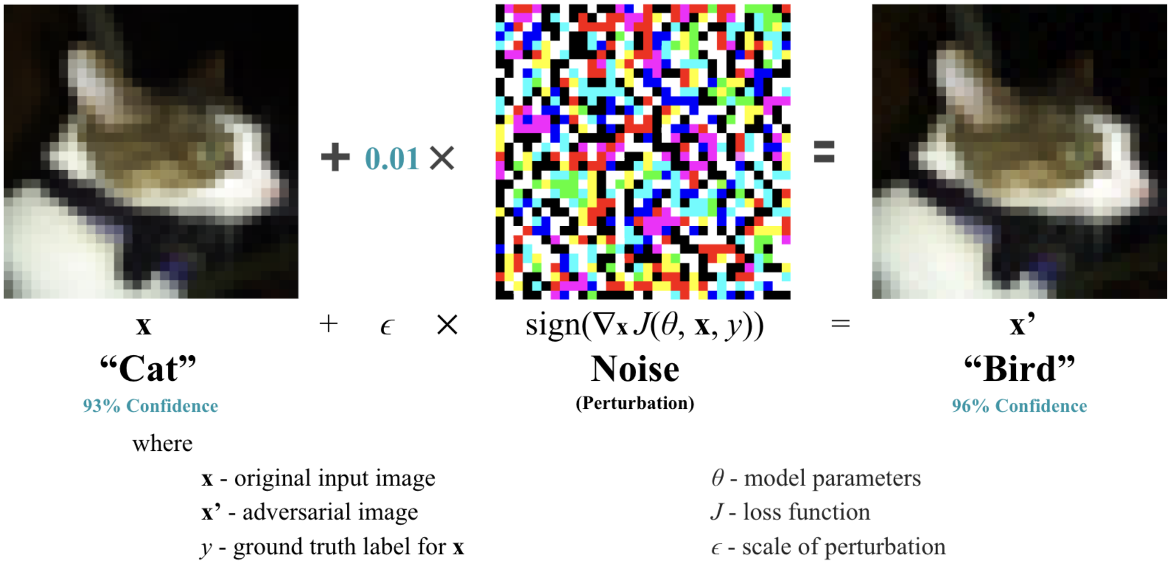


Figure 4.2: A demonstration of fast adversarial example generation applied to ResNet-34 on the CIFAR-10 dataset with the FGSM attack. By applying an imperceptibly small perturbation in the direction of the sign of the back-propagated gradients to maximize loss, we can fool ResNet-34 into misclassifying the input.

in logic and is often used as the introductory attack in AML courses or tutorials. Though simple in logic, the attack has been proven to be extremely effective [8] (Figure 4.2):

$$\mathbf{x}' = \mathbf{x} + \epsilon \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y)). \tag{4.1}$$

The attack adjusts the input image by taking a step towards the sign of the back-propagated gradients for each pixel to maximize $J(\theta, \mathbf{x}, y)$, where $\mathbf{x}$ is the original input image, $\mathbf{x}'$ is the generated adversarial image, y is the ground truth label associated with $\mathbf{x}$, θ are the model parameters, J refers to the loss function utilized by the targeted model, and ϵ refers to the scale of the perturbation [8].

We utilize FGSM attack with L^∞ norm to generate adversarial examples. Also known as the Chebyshev distance, the L^∞ distance is commonly adapted by adversarial attacks to generate perturbed images by measuring the maximum pixel difference between two images. For example, if $\mathbf{x}$ is the original image input, and $\mathbf{x}' = \mathbf{x} + \mathbf{n}$ is the adversarial output where $\mathbf{n}$ is equivalent to $\epsilon \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y))$, then the L^∞ distance between $\mathbf{x}$ and

$\mathbf{x}'$ is computed as the following:

$$||\mathbf{n}||_\infty = \max_i |n_i|. \tag{4.2}$$

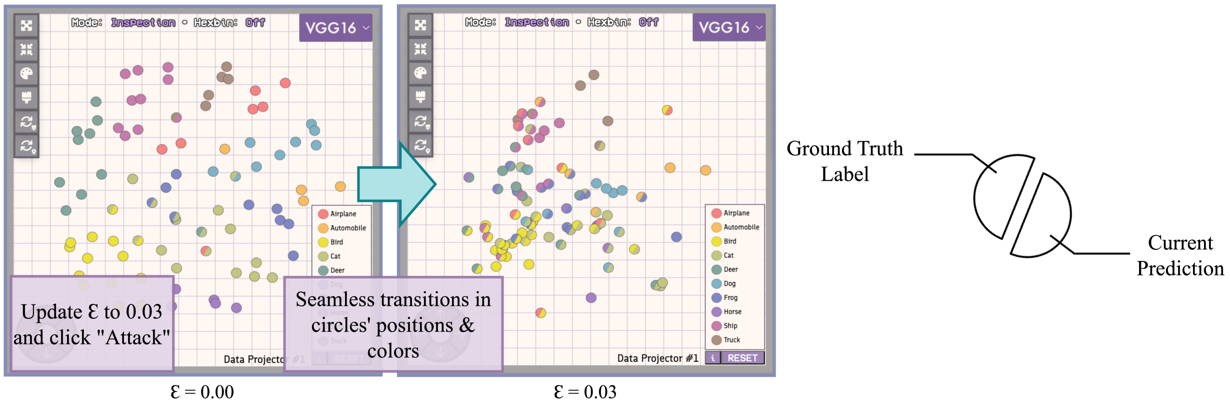
The Attacker module performs attacks on the selected models respectively with perturbation sizes ϵ of 0.00, 0.01, 0.02, and 0.03. This is achieved by first feeding the natural images into the targeted models to obtain the gradients of the loss function w.r.t the input pixels, where the required gradients can be computed efficiently using backpropagation [8]. Next, the pixel values of the input image are adjusted by taking a step in the direction of the gradients, i.e., $\text{sign}(\nabla_\mathbf{x} J(\theta, \mathbf{x}, y))$, which produces an adversarial image by maximizing the loss value. Both natural and adversarial images are inputted into the models for classification and embedding extraction.

4.3.2 Embedding Projector

The Embedding Projector is a backend module tasked with the followings: 1) processing the embeddings produced by the models and 2) analyzing the information of the extracted features and preserving it in a low-dimensional representation. The goal is to unveil important patterns among the embeddings and transform them into a format that can be readily fetched for frontend rendering. The module temporarily detaches the final output layer of a model to obtain the image embeddings and further reduces the embeddings' dimensions by applying the user's choice of dimensionality reduction method to later visualize them in a 2-D space. For instance, in the case of t-SNE, the module analyzes the features of the instances by constructing a lower-dimensional probability distribution that represents the similarities between the objects in the high-dimensional space. If PCA is selected, the module preserves the most significant variability in the embeddings while reducing the number of features. The resulting outputs are scaled to be used as the x- and y-coordinates of the instances in scatterplots and are stored as tabular data accessed by AdvEx's frontend Data Projectors.

4.4 Frontend User Interface

Within this section, we introduce all frontend components of AdvEx and describe how they work in detail. We demonstrate our approach on VGG models that were pre-trained on the CIFAR-10 dataset [33]. Specifically, we select VGG-16 and VGG-19 models to introduce AdvEx.



Update ε to 0.03 and click "Attack"

Seamless transitions in circles' positions & colors

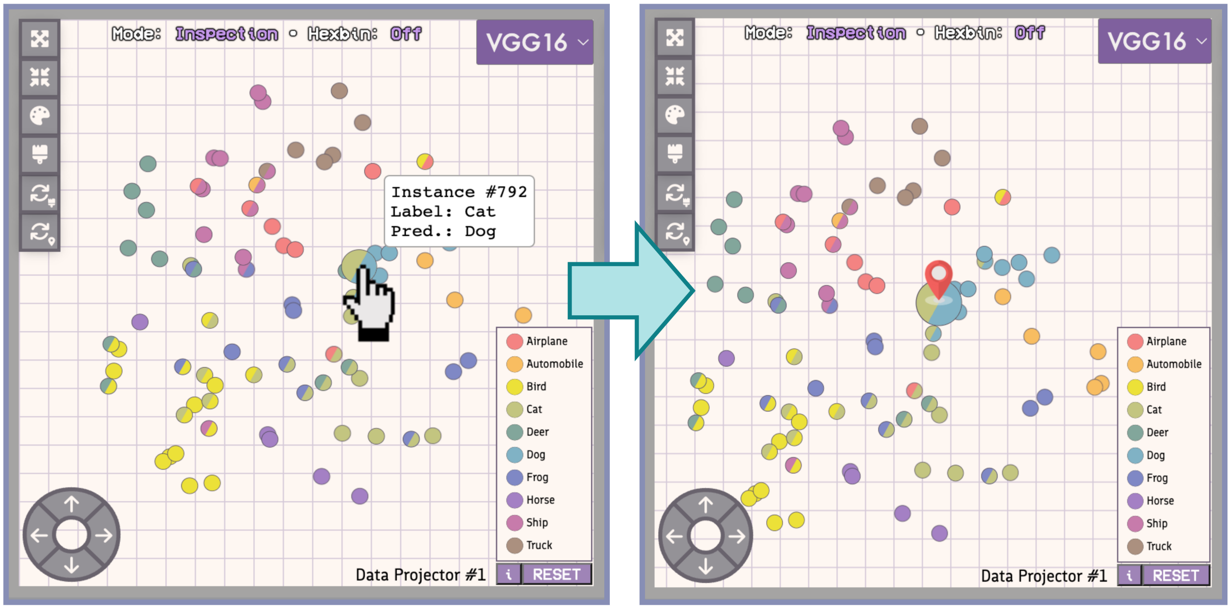
(a) Animated Attack Visualization

(b) Color Encoding

Figure 4.3: (a) As the perturbation size changes, the circles in the Data Projectors dynamically shift in color while traveling around the plane. (b) Each circle is divided into two halves to visually indicate its label and prediction.

4.4.1 Data Projectors

The Data Projectors represent dimensionality reduction overviews of the dataset and consist of two scatterplots where the image embeddings are projected as circles on a 2-D plane. Each circle corresponds to a data instance and is sliced into two halves: the color of the left half represents the instance's ground truth label, while the color of the right half represents its current prediction (Figure 4.3b). The spatial positions of the circles encode the relationships between them in the original high-dimensional space (e.g., similarities, variance, local and global structure). Taking inspiration from nanocubes [24], we use a combination of binned aggregation and hierarchical clustering with multiple zoom levels to preserve data scalability (Figure 1.1c1). Our approach allows the user to interactively explore data sources with large numbers of instances while maintaining the global data structures without high-performance devices. When an attack is conducted on the dataset with new specified magnitude, the Data Projectors visualize the attack with an animated sequence (Figure 4.3a) that emphasizes each circle's change in position and color (**G3**). For example, if a circle travels to a different coordinate, this means that how the model perceives the instance's features has been altered by the attack. Moreover, if the class "airplane" is assigned with the color red and the class "automobile" is assigned with orange, then a red circle that transitions into a circle with its left half colored red and right half colored orange means that this is an image of an airplane that is incorrectly labeled as an automobile as a result of the attack. When the user hovers over each circle, a brief


Mode: Inspection ○ Hexbin: Off

Instance #792
Label: Cat
Pred.: Dog

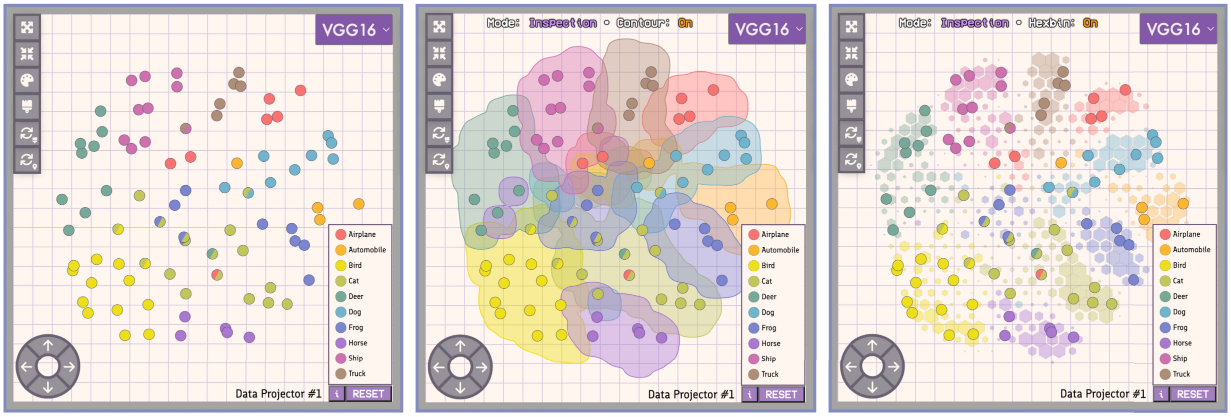
(a) Hovered

(b) Clicked

Figure 4.4: (a) When a circle of interest is hovered over, a tooltip is revealed to display quick information regarding the instance. (b) Upon clicking, the circle gets highlighted and repositioned to the center of the plane.

summary including the instance's original label and the current prediction is displayed in a tooltip (Figure 4.4a). To further enhance the user's navigation experience, the following functionalities are incorporated:

- **Inspection mode.** If multiple images share very similar features, the instances may get projected on top of each other due to small differences between their coordinates. To prevent this inconvenience and enable effortless exploration of the embedding views, we allow the user to freely zoom in and out and drag around the scatterplots with their cursor while dynamically maintaining the radius of the projected circles, allowing them to inspect every individual instance. Clicking on each circle will highlight the instance by enlarging its radius and placing a pin on it, then moving the circle to the center of the 2-D plane via panning the entire scatterplot (Figure 4.4b). A series of buttons are also provided for the user to instantly restore each scatterplot's scale and position, with a guidance button that provides quick instructions on navigating the projectors (**G5**).
- **Selection mode.** In addition, we allow the user to enter the "selection mode" (Fig-


Mode: Inspection ○ Contour: On

Mode: Inspection ○ Hexbin: On

(a) Scatterplot

(b) Scatterplot + Density Contours

(c) Scatterplot + Hexbin Map

Figure 4.5: We explored a variety of visual encodings and aggregating features for the Data Projectors. We chose binned aggregation with multiple zoom levels, with an optional hexbin toggle to display the overall distribution (c). This preserves data scalability and displays global data structure without the need for high-performance devices.

ure 4.8) by toggling the brush button in each scatterplot. Under this mode, the user can highlight a specific subset of the dataset, including a subset of size 1, by specifying a selected region via a pointing gesture, i.e., clicking and dragging the cursor. As a result, only the colors of the circles selected by the brush will be displayed, and all other circles will be grayed out. This allows the user to easily track the movements of specific subgroups/instances between different perturbation levels, thus adding an additional subpopulation-level display (**G1**). When a group or instance is highlighted in one Data Projector, the same group or instance is also highlighted in the other projector.

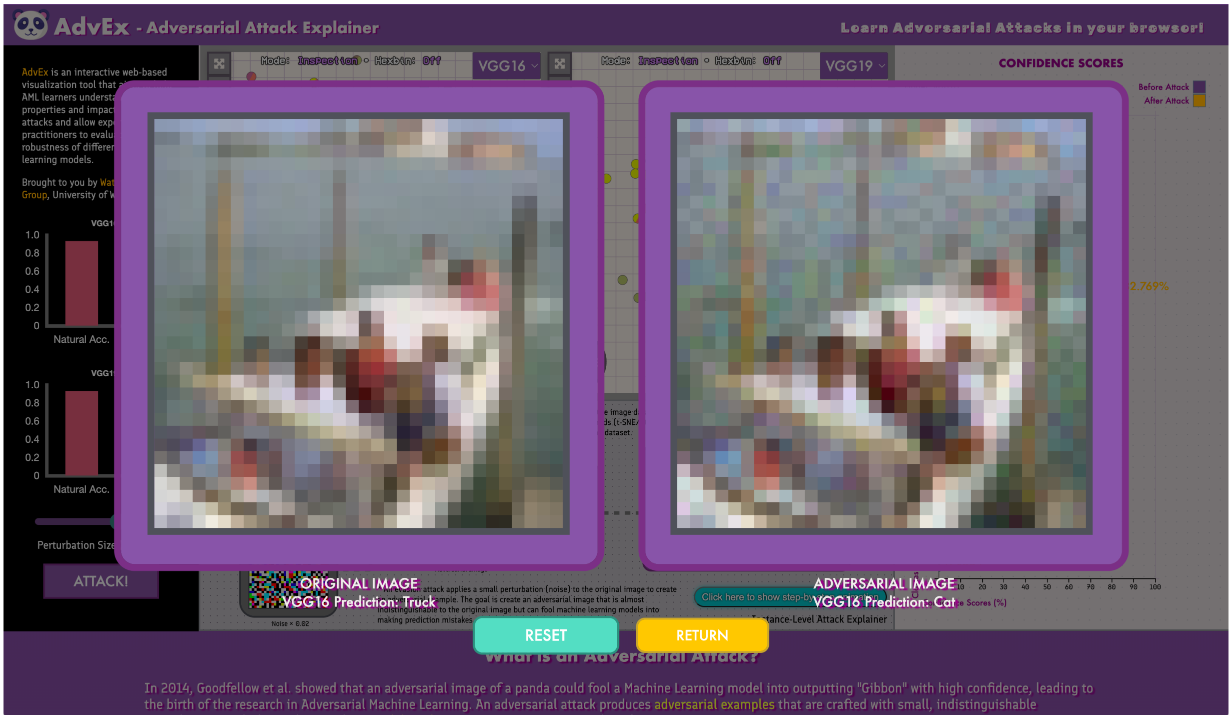
- **Hexagonal binning toggle.** To help the user keep track of the global data structure when navigating, we offer an additional feature that enables the user to toggle the hexagonal binning map (Figure 4.5c) for each projector. The hexbin map displays the general trend of the instance clusterings based on their predictions, allowing the user to observe the high-level distribution of circles and quickly identify groups of images predicted as the same class (**G1**). Our approach also allows the user to observe the overall structure of the entire dataset even when the projectors are only displaying a subset at higher zoom levels.

In summary, the Data Projectors are animated embedding views of the image dataset. They illustrate the relationship between instances via spatiality and the impact of the adversarial attack across the population via animated transitions (**G1, G3**). Since the set of adversarial examples generated would differ depending on the attack method chosen and the specific model being attacked, to demonstrate the attack impact on different models, we also include visualizations of two specific models side by side (**G2**). Through the Data Projectors, the user can intuitively observe how the data representations and the resulting predictions of the images differ as the perturbation size changes.

4.4.2 Instance-level Attack Explainer

While the Data Projectors visualize population-level properties and impacts of the attack, the Instance-level Attack Explainer displays more in-depth information regarding each perturbed input. Specifically, the Instance-level Attack Explainer provides details on the underlying logic of the adversarial attack and visualizes instance-level information such as the applied noise, the input image, the confidence scores, and more (**G1**). When the user wishes to see more instance-level details or how the attack perturbs a specific image, they may click on the circle that corresponds to the entry, and both the bottom panel known as the "general view" and the right panel known as the "interactive confidence score view" are updated immediately. Precisely, the Instance-level Attack Explainer can be divided into the following components:

- **General view.** The general view (Figure 1.1d1) is updated at the bottom whenever the user clicks on any circle from either projector. It displays information about the selected instance such as the original image, the applied noise, the resulting adversarial image, the targeted model, the ground truth label and the current prediction. A combination of animations are used to visualize the generation process of the adversarial image. For instance, a repeated animated sequence shows the original image and the generated perturbation slowly moving towards each other with reduced transparency and stacking on top of each other, then gradually fading into the final perturbed image to visualize the attack result. The dashed lines connecting the images are also animated to continuously move from the original image and the visualized "noise" to the resulting adversarial image to intuitively illustrate the general flow of the attack.
- **Side-by-side image inspection.** To inspect the images more closely, the user may click on the image thumbnails shown in the general view to see enlarged versions. A comparison mode (Figure 4.6) is also provided if the user wishes to investigate the original and adversarial images side by side and observe the exact pixel differences.



In 2014, Goodfellow et al. showed that an adversarial image of a panda could fool a Machine Learning model into outputting "Gibbon" with high confidence, leading to the birth of the research in Adversarial Machine Learning. An adversarial attack produces adversarial examples that are crafted with small, indistinguishable

Figure 4.6: By clicking on the image thumbnails in the attack explainer, the user can view an individual image in an enlarged format or enter comparison mode, which displays the original and adversarial images side by side. This way, the user can identify the subtle perturbations applied to create the adversarial image.

- **Interactive confidence score view.** An interactive grouped bar chart (Figure 1.1d2) displays the model's confidence scores across all classes for the selected entry before and after an attack. The confidence scores for each class pre- and post-attack are grouped together to allow easy comparison between the two values. Hovering over each pair of them will display their exact difference in percentage.

- **Step-by-step execution view.** The step-by-step execution view (Figure 4.9) delves into the details of the underlying attack logic and how the "noise" is generated. To trigger this view, the user clicks on the button located at the bottom right corner of the general view. This initiates a series of engaging step-by-step animated sequences in which the explanations of the attack process appear consecutively (**G4**). For instance, explanation #2 (Figure 4.9-2) will not appear until the user clicks the play button next to explanation #1 (Figure 4.9-1), which only pops up when explanation #1 has finished playing. In addition, we provide the following features: 1) a "show all" button to skip the step-by-step animations and display all explanations at once, 2) a "replay" button that replays all animations, and 3) an optional toggle that allows the user to substitute the default image with the currently selected instance as an example for the view's demonstration.

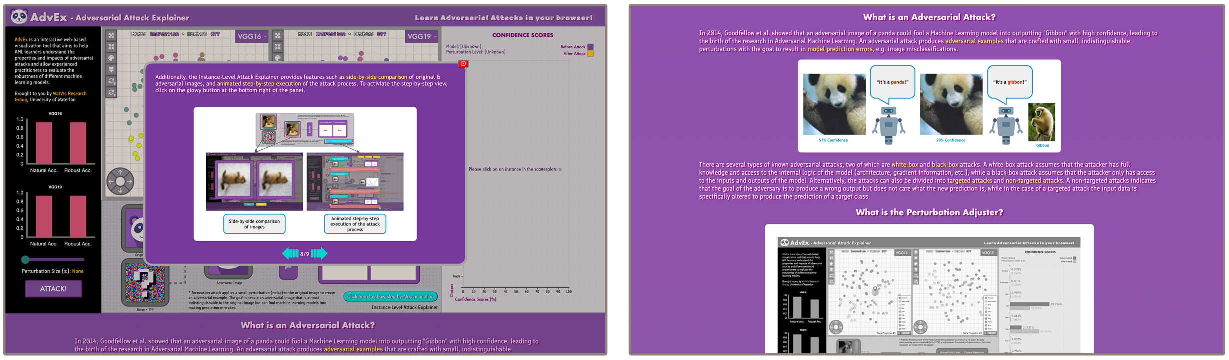
In short, the Instance-level Attack Explainer visualizes the attack on an instance level (**G1**) and intuitively shows that the perturbed image is a result of the original image and the "noise" combined, with a detailed step-by-step execution view to help the user further understand the attack logic (**G4**). Additionally, through the confidence score view, the user can easily observe how the corresponding model's confidence scores change before and after an attack.

4.4.3 Robustness Analyzers

The Robustness Analyzers are a pair of small interactive bars charts (each with 2 bars) on the left-most panel that show the natural and robust accuracy of the two models displayed (**G1, G2**). The natural accuracy refers to the model's prediction accuracy on the original dataset, while the robust accuracy refers to the model's accuracy on the current adversarial dataset. The right bars of the Robustness Analzyers transition up and down to reflect the models' changes in adversarial robustness. With the Robustness Analyzers, the user can easily compare 1) the robustness of a model to its natural accuracy and 2) the performance of one model to the other. Consequently, the user can intuitively understand the concept of model robustness against adversarial attacks.

4.4.4 Perturbation Adjuster

The Perturbation Adjuster refers to a slider and an attack button below the Robustness Analyzers. The user can adjust the slider horizontally to choose the perturbation size of interest (i.e., None, 0.01, 0.02, & 0.03). After a size is selected, the user clicks the attack button to initiate the animated attack sequence (**G3**). An example sequence is that when the button is clicked with a new size specified, the circles of both Data Projectors transition to new coordinates with potential changes in colors, and the right bars of the Robustness Analyzers adjust their heights seamlessly based on the models' new robust accuracy. Through the Perturbation Adjuster, the user can effortlessly adjust the attack strength and observe how the consequences of the attack get more drastic as the perturbation size increases.



AdvEx - Adversarial Attack Explainer
Learn Adversarial Attacks in your browser!
Mode: Inspection ○ Hexbin: Off
VGG16
VGG19
CONFIDENCE SCORES
Model: [Unknown]
Perturbation Level: [Unknown]
Before Attack
After Attack
VGG16
Natural Acc.
Robust Acc.
VGG19
Perturbation Size (ε): None
ATTACK!
Side-by-side comparison of images
Animated step-by-step execution of the attack process
8/9
Please click on an instance in the scatterplots
Adversarial Image
Noise × ???
Click here to show step-by-step animation
Instance-Level Attack Explainer
Truck
Classes
Confidence Scores (%)
"It's a panda!"
"It's a gibbon!"
57% Confidence
99% Confidence
Gibbon

AdvEx is an interactive web-based visualization tool that aims to help AML learners understand the properties and impacts of adversarial attacks and allow experienced practitioners to evaluate the robustness of different machine learning models.

Brought to you by WatVis Research Group, University of Waterloo

Additionally, the Instance-Level Attack Explainer provides features such as side-by-side comparison of original & adversarial images, and animated step-by-step execution of the attack process. To activiate the step-by-step view, click on the glowy button at the bottom right of the panel.

* An evasion attack applies a small perturbation (noise) to the original image to create an adversarial example. The goal is create an adversarial image that is almost indistinguishable to the original image but can fool machine learning models into making prediction mistakes.

What is an Adversarial Attack?

In 2014, Goodfellow et al. showed that an adversarial image of a panda could fool a Machine Learning model into outputting "Gibbon" with high confidence, leading to the birth of the research in Adversarial Machine Learning. An adversarial attack produces adversarial examples that are crafted with small, indistinguishable perturbations with the goal to result in model prediction errors, e.g. image misclassifications.

There are several types of known adversarial attacks, two of which are white-box and black-box attacks. A white-box attack assumes that the attacker has full knowledge and access to the internal logic of the model (architecture, gradient information, etc.), while a black-box attack assumes that the attacker only has access to the inputs and outputs of the model. Alternatively, the attacks can also be divided into targeted attacks and non-targeted attacks. A non-targeted attacks indicates that the goal of the adversary is to produce a wrong output but does not care what the new prediction is, while in the case of a targeted attack the input data is specifically altered to produce the prediction of a target class.

What is the Perturbation Adjuster?

(a) Opening Tutorial

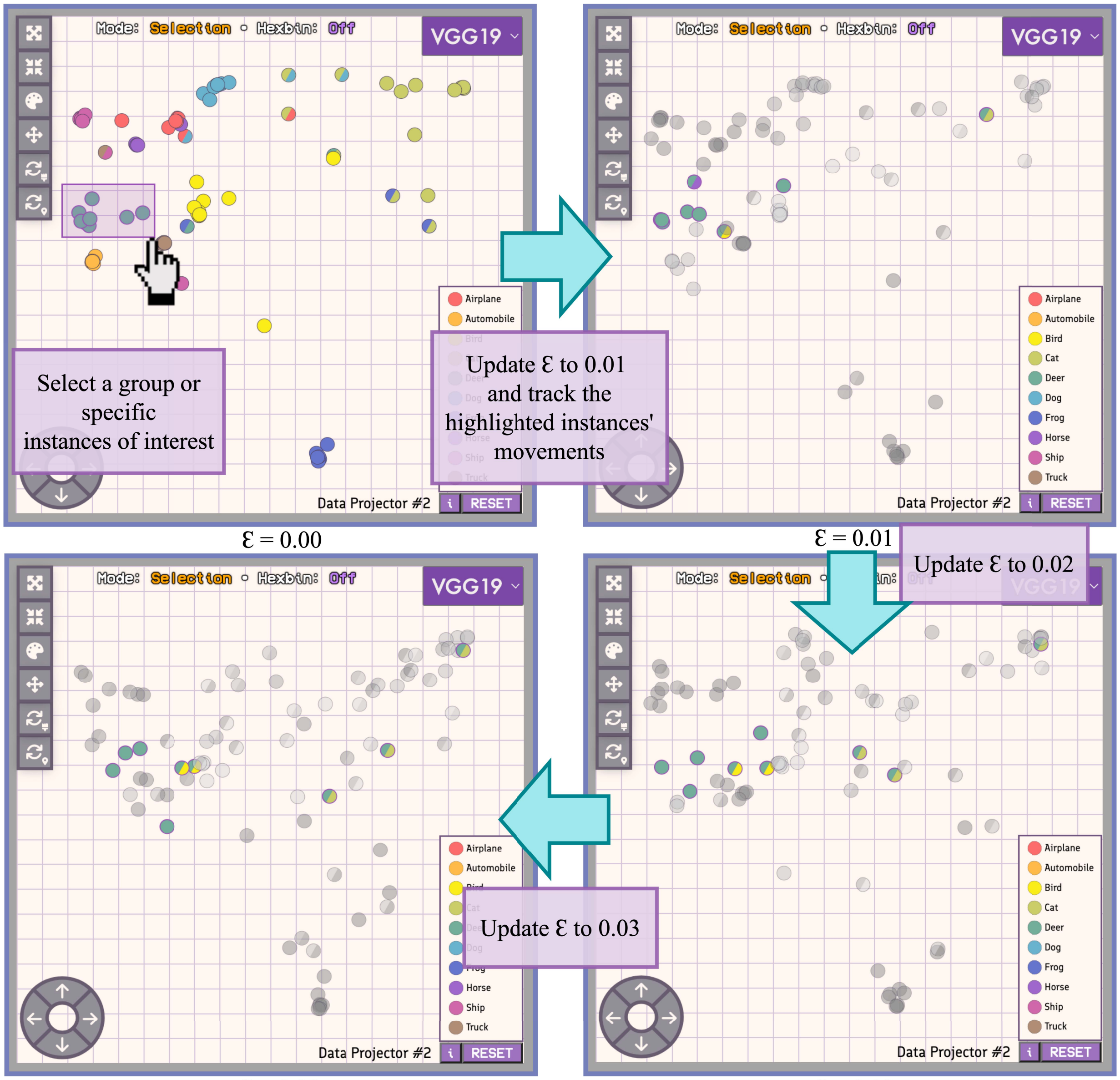
(b) General Information Provider

Figure 4.7: Screenshots of (a) the overlay tutorial shown when AdvEx is first launched and (b) the General Information Provider below the interactive components.

4.4.5 Interactive Tutorials + General Information Provider

To help the users pick up AdvEx more easily, we incorporated an interactive tutorial system into AdvEx (**G5**). When the application is first launched, the user is greeted with a welcome screen followed by an overlay tutorial (Figure 4.7a) that introduces each component of AdvEx's interface and highlights its key features. Moreover, during the interaction, if any of the Data Projectors' buttons is hovered over, a tooltip appears to provide a brief guide on what the button does. If the user has not interacted with the projectors' hexbin or brush toggles, or triggered the step-by-step execution view after 10 minutes of interaction, an animated arrow pointing at the untriggered button appears to encourage the user to explore the functionality.

Furthermore, if the user wishes to learn more about our work and the research of AML, they may read the information placed beneath the interactive components (Figure 4.7b), which provides more in-depth explanations of both AdvEx and AML. By including interactive tutorials and reading materials, the user will not only pick up our tool faster, but also gain detailed and accurate knowledge of adversarial attacks in addition to perceiving them through interactive visualizations.



Select a group or specific instances of interest

Update ℰ to 0.01 and track the highlighted instances' movements

Update ℰ to 0.02

Update ℰ to 0.03

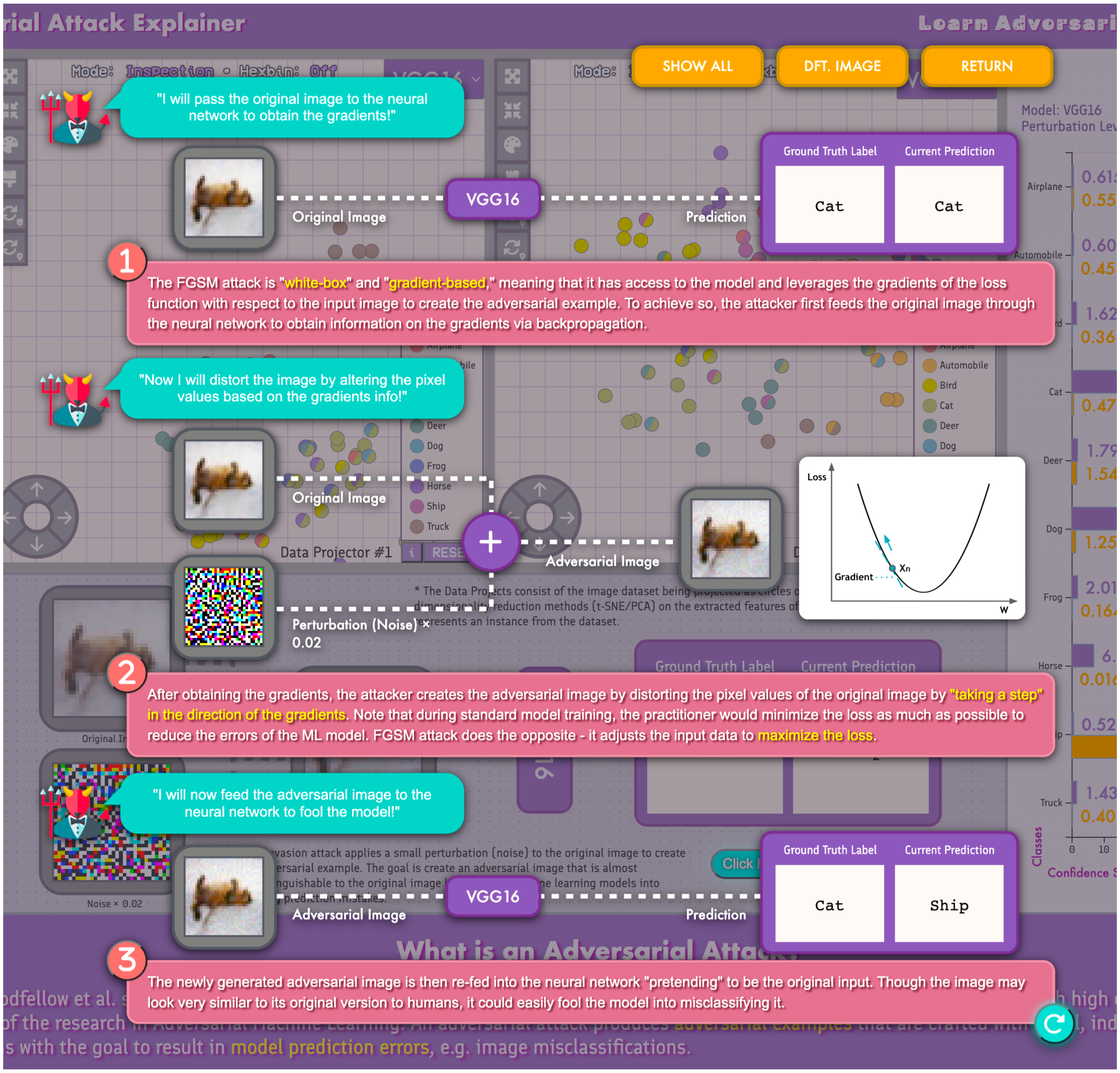
ℰ = 0.00

ℰ = 0.01

ℰ = 0.03

ℰ = 0.02

Figure 4.8: A user is highlighting and tracking a specific class from the dataset with our "selection mode." Under this mode, the user can evaluate the model performance on a specific subset of the dataset.



"I will pass the original image to the neural network to obtain the gradients!"

1 The FGSM attack is "white-box" and "gradient-based," meaning that it has access to the model and leverages the gradients of the loss function with respect to the input image to create the adversarial example. To achieve so, the attacker first feeds the original image through the neural network to obtain information on the gradients via backpropagation.

"Now I will distort the image by altering the pixel values based on the gradients info!"

2 After obtaining the gradients, the attacker creates the adversarial image by distorting the pixel values of the original image by "taking a step" in the direction of the gradients. Note that during standard model training, the practitioner would minimize the loss as much as possible to reduce the errors of the ML model. FGSM attack does the opposite - it adjusts the input data to maximize the loss.

"I will now feed the adversarial image to the neural network to fool the model!"

3 The newly generated adversarial image is then re-fed into the neural network "pretending" to be the original input. Though the image may look very similar to its original version to humans, it could easily fool the model into misclassifying it.

Figure 4.9: An example of the step-by-step execution view for introducing the FGSM attack. We provide detailed and animated explanations of the attack process to AML learners. An optional toggle allows the user to substitute the default image with the currently selected instance as an example to illustrate the underlying attack logic.

Chapter 5

User Study with Novice Learners

To evaluate how AdvEx can help novice AML learners understand adversarial attacks, we conducted a user study with participants who had basic knowledge in machine learning but were unfamiliar with AML. We aimed to investigate two aspects of AdvEx as an educational tool: **(S1)** whether AdvEx is effective for helping learners understand the concepts and impacts of adversarial attacks, and **(S2)** whether users enjoy using AdvEx to learn about adversarial attacks. We did not conduct a comparative study due to existing AML learning tools presenting either a small subset of what AdvEx can inform users with or fundamentally different information. For example, Adversarial-Playground [28] and AdVis.js [23] only provide side-by-side comparisons of natural and adversarial examples with limited functionalities and interactions. Bluff [5], on the other hand, focuses on visualizing the internal pathways formed by neurons and their connections under adversarial attacks. Thus, there is no suitable baseline for our study for a meaningful and fair comparison.

5.1 Study Setup

Participants and Apparatus. Based on self-reported qualifications from a pre-questionnaire, we recruited 12 participants (10 men, 2 women; aged 21∼31) from a local university. They came from various areas of study such as Computer Science, Transportation Engineering, and Data Science. All reported having a background in machine learning but were unfamiliar with AML. Specifically, on a 7-point Likert scale (self-rated; 1="Novice", 7="Expert"), we recruited participants that satisfied all the following constraints: machine learning experience ≥ 2, AML experience ≤ 2, completion of ≥ 1 machine learning project, completion of ≤ 1 AML project. Their median machine learning experience was 4 (IQR $= 2$), and their median AML experience was 1 (IQR $= 0.25$). Their median number of machine learning projects completed was 2.5 (IQR $= 2.25$), while their median number of AML projects completed was 0 (IQR $= 0$). The participants interacted with AdvEx on provided laptops in-person.

Task and Procedure. For this study, we loaded AdvEx with the CIFAR-10 testing dataset perturbed by the FGSM attack in varying degrees to investigate the participants' learning of the properties and impacts of the FGSM attack. Prior to interacting with AdvEx, we asked the participants to complete a pre-quiz that consisted of 9 questions to assess their background in general machine learning and knowledge in AML. These questions included 4 checker questions on basic machine learning and 5 questions that would be taught by AdvEx
4 Checker Qs
5 Taught Qs
. The checker questions were to ensure participants' self-reported expertise aligned with their background and to assess their attention during the study. After the pre-quiz, we presented a PDF file that contained basic background on adversarial attacks (e.g., what an adversarial attack is, what an FGSM attack is) to help the participants gain the minimum required knowledge to learn AML using AdvEx. They were provided with 5 minutes to read through the PDF file, but had the freedom to revisit the PDF file later during their interaction with AdvEx. After, we presented AdvEx and provided the participants with 30 minutes to interact with AdvEx freely. We instructed the participants to use AdvEx to learn about the FGSM attack as much as they could, and informed them that there would be a follow-up post-quiz to assess how much they had learnt. Next, we asked the participants to complete a 7-point Likert scale post-questionnaire (6 questions), which collected their opinions on the learning and usability aspects of AdvEx. We then asked them to complete the post-quiz (19 questions), which comprised of the 9 original questions from the pre-quiz, along with 10 new questions that were taught by AdvEx
4 Checker Qs
5 Old Taught Qs
10 New Taught Qs
. We ended the user study with a qualitative interview that asked for their thoughts and opinions on AdvEx, such as how they liked AdvEx overall, and how they liked each individual component of AdvEx.

The user study took about one hour and the participants received $15 for their effort. They were informed that the top 3 performers of both the pre-quiz and post-quiz would be awarded an additional $10.

Table 5.1: The results of the paired t-tests and the quiz averages of our participants (filtered & all). Our results show that AdvEx has a strong learning effect on both filtered and all participants. *OQ (“old questions”): 9 questions from the pre-quiz that are also included in the post-quiz. †NQ (“new questions”): 10 questions that are newly added in the post-quiz. ‡Average of total quiz (checkers + taught) scores.

	Paired T-Tests					**Quiz Averages**			
	Pre-quiz vs. Post-quiz	Pre-quiz vs. Post-quiz OQ*	Pre-quiz vs. Post-quiz NQ†	Pre-quiz Taught vs. Post-quiz Taught	Pre-quiz Checkers vs. Post-quiz Checkers	Pre-Quiz Checkers	Pre-Quiz Taught	Post-Quiz Checkers	Post-Quiz Taught
Filtered (10 Participants)	t = -5.264, p = 0.00052	t = -6.128, p = 0.00017	t = -4.229, p = 0.00221	t = -6.482, p = 0.00011	t = 1.0, p = 0.34344	85% (σ = 21.08%)	50% (σ = 17%)	82.5% (σ = 26.48%)	93.33% (σ = 6.28%)
						65.56% (σ = 16.93%)‡		91.05% (σ = 7.46%)‡	
All (12 Participants)	t = -6.225, p = 0.00006	t = -6.661, p = 0.00004	t = -5.197, p = 0.0003	t = -5.88, p = 0.00011	t = -0.561, p = 0.5863	75% (σ = 30.15%)	51.67% (σ = 15.86%)	77.08% (σ = 27.09%)	90% (σ = 10.05%)
						62.04% (σ = 17.38%)‡		87.28% (σ = 11.32%)‡	

5.2 Results and Analysis: Task Performance

Out of 12 participants, we removed 2 whose pre-quiz checker scores were below 50%. On average, the 10 remaining participants spent 3.97 minutes ($\sigma = 0.07$) on the pre-quiz, 16.17 minutes ($\sigma = 0.21$) on their interaction with AdvEx, and 5.17 minutes ($\sigma = 0.10$) on the post-quiz. Before interacting with AdvEx, the participants had an average pre-quiz score of 65.56% ($\sigma = 16.93\%$), and 50% ($\sigma = 17\%$) if exclude the checker questions. After, the participants earned an average post-quiz score of 91.05% ($\sigma = 7.46\%$), and 93.33% ($\sigma = 6.28\%$) if exclude the checker questions. The difference between the mean pre-quiz and post-quiz scores clearly indicates AdvEx's effectiveness in enabling learning.

Further, a paired t-test shows a significant difference between the participants' overall pre-quiz and post-quiz performance ($t = -5.264, p = 0.00052$); the difference is also significant when the checker questions are excluded ($t = -6.482, p = 0.00011$). Both results indicate a strong performance improvement after the participants' interaction with AdvEx. A third paired t-test shows a significant difference between their performance on the same 9 questions in the pre-quiz and post-quiz ($t = -6.128, p = 0.00017$). This shows that the participants have successfully learnt the answers to the questions that were originally included in the pre-quiz. Similarly, a significant difference can be observed between the participants' pre-quiz performance and their performance on the 10 newly added questions in the post-quiz ($t = -4.229, p = 0.00221$). This reveals that the participants have picked up additional knowledge that was not mentioned in the pre-quiz during their interaction with AdvEx. Lastly, another paired t-test was performed between their performance on the same checker questions in the pre-quiz and post-quiz and no significant difference was

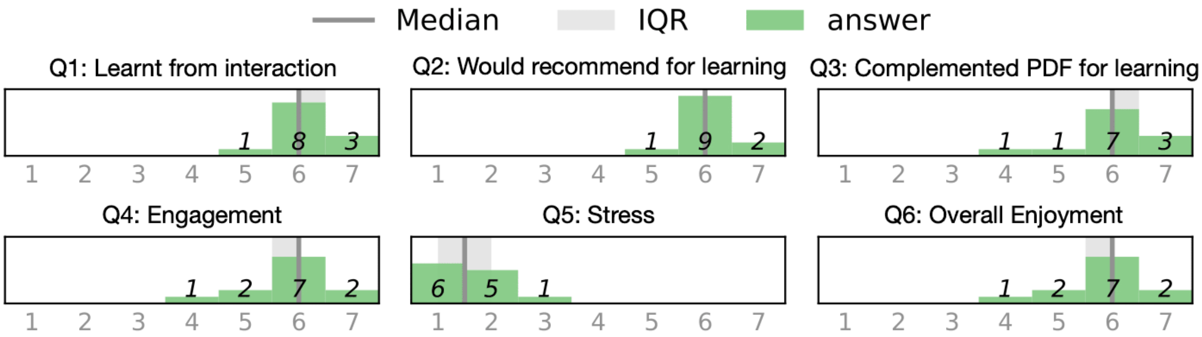


Figure 5.1: Participants' questionnaire ratings (1 = "strongly disagree"; 7 = "strongly agree") on the learning and usability aspects of AdvEx.

found ($t = 1.0, p = 0.34344$). Combined with the fact that the 10 qualified participants all scored at least 50% on the pre-quiz checker questions, this indicates that our participants were consistent with their responses to the checker questions and did not randomly choose their answers.

We repeated our statistical tests on all 12 participants with the 2 unqualified participants included and our results still indicate a strong learning effect (Table 5.1). This suggests that even if the participants did not possess basic machine learning knowledge, they could still learn effectively by interacting with AdvEx. The results of all our paired t-tests and the quiz averages of the participants are shown in Table 5.1.

5.3 Results and Analysis: Participants' Feedback

To further investigate **S1** and **S2**, we analyzed the participants' post-questionnaire responses (Figure 5.1; 7-point Likert scale with 1 = "Strongly Disagree" and 7 = "Strongly Agree") and their qualitative feedback from the semi-structured interviews.

For Q1, all participants agreed that they had learnt about adversarial attacks through interacting with AdvEx (MD = 6, IQR = 0.5) and gave a positive rating (≥ 5). P3 stated that *"AdvEx teaches all aspects of adversarial attacks thoroughly,"* and P8 commented that *"The clear explanations of AdvEx make the learning process much easier."* The participants also thought that AdvEx's visualizations were highly informative. *"The visualizations really show me that I can have malicious inputs to my models that are indistinguishable to my eyes. The step-by-step execution view effectively helps me understand the underlying attack logic."* -P10

Similarly, for Q2, all participants stated that they would recommend AdvEx to others for learning AML (MD = 6, IQR = 0). P2 thought that *"*AdvEx *serves as a valuable educational tool for illustrating the attacks,"* and P5 believed that *"*AdvEx *is great for beginners and it can teach them a lot about the attack process."* To further strengthen AdvEx as a learning tool, P5 suggested visualizing the internal attack process in more detail. *"For learners to gain a more in-depth understanding of the attack process, maybe visualize how the adversarial inputs modify the gradient information to alter the model outputs."* -P5

The ratings of Q3 indicate that AdvEx complemented the provided PDF file well for learning (MD = 6, IQR = 0.5). Some participants believed that AdvEx could be used in conjunction with text-based documents, such as textbooks, to help understand the attacks. *"*AdvEx *can help demonstrate and reinforce what people may have read about adversarial attacks, like from the PDF."* -P1 Other participants felt that AdvEx was sufficient on its own. *"I don't think* AdvEx *needs any additional complementary materials. The visualizations are enough to thoroughly explain the attack logic."* -P4

Eleven out of 12 participants gave a rating ≥ 5 on AdvEx's engagement in Q4 (MD = 6, IQR = 0.5). They applauded AdvEx for its highly interactive interfaces and enjoyed dynamically experimenting with the perturbation size. *"The interface is highly engaging. I enjoy changing the noise level and observing how the resulting adversarial image differs."* -P1 P5, similarly, stated that *"It is fun to see all the points moving around in the Data Projectors when I adjust the slider."* However, P12 rated AdvEx's engagement a 4 and said: *"In general, the application is good. But as a programmer, I feel like I should be able to get more involved and write custom code directly."*

For Q5, all participants agreed that it was not stressful to interact with AdvEx (MD = 1.5, IQR = 1). This was likely because AdvEx had an interactive tutorial system that provided guidance on AdvEx's functionalities, along with the General Information Provider that offered further assistance. Moreover, everything AdvEx visualized (e.g., 2-D latent space, confidence scores) were familiar to learners who knew machine learning, thus making AdvEx intuitive to learn with. *"Using* AdvEx *is very simple. The visualizations are quite straightforward."* -P7

In general, the participants rated AdvEx's enjoyment positively in Q6 (MD = 6, IQR = 0.5). They offered different reasons for why they enjoyed AdvEx. P9 and P10 claimed that AdvEx's visually appealing interfaces and its vivid animations made their interactions entertaining. P3, P8 & P10 emphasized the amount of knowledge they gained from AdvEx and found the learning experience fruitful. P4, P6 & P11, on the other hand, applauded AdvEx for its high level of interactivity. *"I enjoy* AdvEx *because I can do a lot with it.*

I can investigate different examples, try out different noise levels, visualize the confidence scores, and observe how the embedding distribution changes." -P4

Chapter 6

Interview Study with Experienced Experts

To explore whether and how AdvEx can be used for inspecting the robustness of models, we conducted an interview study with AML experts, which helped us collect in-depth qualitative feedback on AdvEx as a visual analytics tool.

6.1 Study Setup

In this study, the experts were asked to utilize AdvEx to evaluate four different models (VGG-16, VGG-19, naturally trained ResNet-34, & adversarially trained ResNet-34) on the CIFAR-10 testing dataset under the FGSM attack in a free-form analysis session.

We recruited three AML experts (all men): an AML researcher (E1), an industry data scientist who has a background in AML (E2), and an AI researcher who has expertise in AML (E3). Each study session began with a 5-minute introduction of the project background and the key features of AdvEx. We then presented a task scenario where the experts were asked to use AdvEx to *"inspect how the FGSM attack alters the input images to affect the models' performance,"* and *"compare the robustness of the models against the attack."* The experts could inspect the models with AdvEx for as long as they liked. A list of recommended tasks was provided to the experts to courage them to interact with each component of AdvEx; they were also told that they could explore the tool freely without completing those tasks as long as insights were gathered. We employed the think-aloud protocol and an experimenter was responsible for providing help and answering questions regarding the user interface, who also observed the experts' interactions and took notes. Next, a semi-structured interview (≈20 minutes) was carried out to gain a better understanding of their thoughts on AdvEx in light of the think-aloud feedback and observation gathered previously.

6.2 Results and Analysis

All three experts successfully performed the analysis and expressed a positive sentiment toward AdvEx. We conducted a thematic analysis on the unstructured feedback gathered from the experts during the free-form analysis, as well as the qualitative data provided to us during the semi-structured interviews. We came up with five systematic themes in light of our design goals and adopted a deductive approach to identify patterns of them in our data.

Visualizations of attack impacts. All experts pointed out that AdvEx's visualizations quickly helped them understand the overall impact of the attack. *"The Data Projectors allow me to quickly identify which instances are misclassified and which are not."* -E1 *"The Robustness Analyzers are straightforward and provide direct comparisons of the natural and robust accuracy of each model."* -E2 They also found the subpopulation-level and instance-level visualizations highly useful. E1 pointed out that the selection mode allowed him to observe the trajectories of all points from a specific class, thus easily seeing how the distribution of a single class differed before and after an attack. E2 explained that the comparison mode allowed him to see the exact differences between natural and adversarial images and discovered that *"When an image has a simple, single-color background like a blue sky, it is considerably easier to notice the applied perturbation compared to images with more complex backgrounds."* The above observations confirm that AdvEx can effectively visualize the attack impact at multiple levels of detail (**G1**). One limitation brought up was that when the perturbation size was above 0, it was challenging to differentiate between instances that were misclassified prior to the attack and those that were misclassified as a result of the attack. E1 suggested including the original prediction of the image along with its ground truth label and current prediction in the attack explainer.

Evaluation of model robustness. The experts applauded AdvEx for helping them quickly identify the bottlenecks of their models. During the free-form analysis, E1 pointed at VGG-19's Data Projector and commented *"Here I can see a lot of red dots being misclassified, so I know this is the class VGG-19 tends to underperform on."* The feature of comparing two models side by side was also frequently brought up as a highlight of AdvEx. E3 claimed that the model comparison feature was the most engaging part of his experience. He observed that compared to a naturally trained ResNet, the perturbations generated for an adversarially trained ResNet tend to contain more defined shapes resembling the original image, indicating that an adversarially trained model relies more on human-interpretable features for classification. *"*AdvEx *can actually help me invent new adversarial training methods, because it helps me gain a lot of insights on how adversarially trained models work."* -E3 These comments indicate that AdvEx can effectively enable visual analysis and comparison of different models under attack (**G2**). Moreover, the experts liked how comprehensive AdvEx was in terms of its model evaluations. E1 stated *"When I used Streamlit [38], I did not think about evaluating the embedding distributions. After using* AdvEx*, I realize integrating it into my workflow would be very useful."* E3 made a similar comment and explained *"People usually only focus on accuracy, but that is not the whole story. To me, these other metrics displayed by* AdvEx *are just as important."* Both E1 and E2 suggested that it may be even better for AdvEx to support comparing the same model under attacks with different perturbation sizes side by side.

Dynamic experimentation with fluid animations. The experts enjoyed dynamically experimenting with the perturbation size and observing the corresponding changes. E1 was amused to see that a bird image was first misclassified as a horse by the naturally trained ResNet-34 under minor perturbations, but as he increased the perturbation size, the instance started to be misclassified as a cat instead. They also appreciated how AdvEx was integrated with animations to vividly demonstrate the attacks. *"I think* AdvEx *is engaging because it has subtle and inviting animations. Unlike other tools, it is more interactive. It captures my attention."* -E3 The above observations indicate that AdvEx achieved **G3**. However, the experts suggested allowing users to input custom perturbation sizes directly in addition to adjusting the perturbation size with a slider. E1 explained *"While I think the Perturbation Slider is great for AML learners, it would be more convenient for experienced practitioners to directly input a custom number for the perturbation size."*

Value as an educational tool for learners. All experts agreed that AdvEx would be a great educational tool for learners to understand adversarial attacks. *"The visualizations of* AdvEx *could help learners quickly grasp the general logic and impacts of adversarial attacks, as they demonstrate key knowledge such as the underlying attack logic, the input and output images, and how the goal of an adversarial attack is to lower model accuracy."* -E1 *"The visuals of* AdvEx *are the most compelling [...] The tool is extremely valuable for beginners as it visualizes the images, the confidence scores, and other metrics instead of just the accuracy to help them understand how adversarial attacks work."* -E3 They also thought the step-by-step execution view would be very clear and informative for AML learners, confirming that AdvEx met **G4**. In addition, the experts believed that

AdvEx's interfaces, including the step-by-step execution view, would make the learning experience highly engaging. E1 commented *"I like the pace of* AdvEx*, it is like playing a game. I feel full of achievements when I interact with this software. Learning and evaluating models would be very easy for AML learners as there are not too many tedious formulas."*

Usability & beginner-friendly design. All experts believed that AdvEx was very easy to pick up. E3 liked how the beginning tutorial highlighted specific areas of the interface, which helped him easily understand the purpose and functionalities of each interface component. E1 thought learners less experienced with AML could also pick up AdvEx effortlessly. He commented *"*AdvEx *is very beginner-friendly to AML learners with a machine learning background as everything visualized are things people with machine learning knowledge already familiar with."* These comments indicate that AdvEx was successfully integrated with a beginner-friendly design (**G5**). They also thought AdvEx was highly convenient and accessible, as *"*AdvEx *is very complete and I don't need to make any adjustments or write any code [...] The automatic configuration of the visualizations is incredibly convenient."*-E1 E3, on the other hand, highlighted the zoomable binned aggregation feature of AdvEx and commented *"This feature can effectively accommodate different users' available computational power and enable smooth exploration of large-scale data for everyone."*

Chapter 7

Usage Scenario

In this chapter, we describe a hypothetical scenario that illustrates how AdvEx can be utilized by machine learning experts to learn about the properties and impacts of adversarial attacks and perform model analysis. We assume a hypothetical user called Zoey, who is an AI researcher concerned with the adversarial robustness of her trained models against potential attacks. Specifically, she wants to investigate the performance of her ResNet-34 models on the CIFAR-10 dataset under the FGSM attack, and how training the models naturally and adversarially have different outcomes on the models' defense. For simplicity, we will denote the naturally trained ResNet-34 as *ResNet*, and the adversarially trained ResNet-34 as *ResNet**. Zoey loads the CIFAR-10 testing data along with the two ResNet models into AdvEx. She launches AdvEx's interface and is greeted by a welcome screen followed by an overlay tutorial. Zoey reads through the tutorial and understands the key components and functionalities of AdvEx.

Upon entering AdvEx's main interface, Zoey immediately notices the two Data Projectors side by side, each representing the image embeddings of one model. The left projector shows ResNet's embeddings, while the right projector shows ResNet*'s. Zoey uses her cursor to pan and zoom the projectors, and as she zooms in, more sampled instances from the dataset are displayed. Zoey first observes from the Data Projectors that the embeddings of ResNet have more distinct clusterings compared to ResNet*'s (Figure 7.1a). She also notices from the Robustness Analyzers that when no attack is conducted, ResNet* $(\text{ACC} = 85\%)$ has a lower natural accuracy than ResNet $(\text{ACC} = 93\%)$ (Figure 7.1b). Zoey decides to experiment with the Perturbation Adjuster by setting the perturbation size ϵ to values greater than zero. The circles in ResNet's projector start to travel around rapidly, but the circles in ResNet*'s projector only move minimally (Figure 7.1c). At a perturbation size of 0.03, Zoey is amazed to see that ResNet's accuracy has dropped to merely 57%,

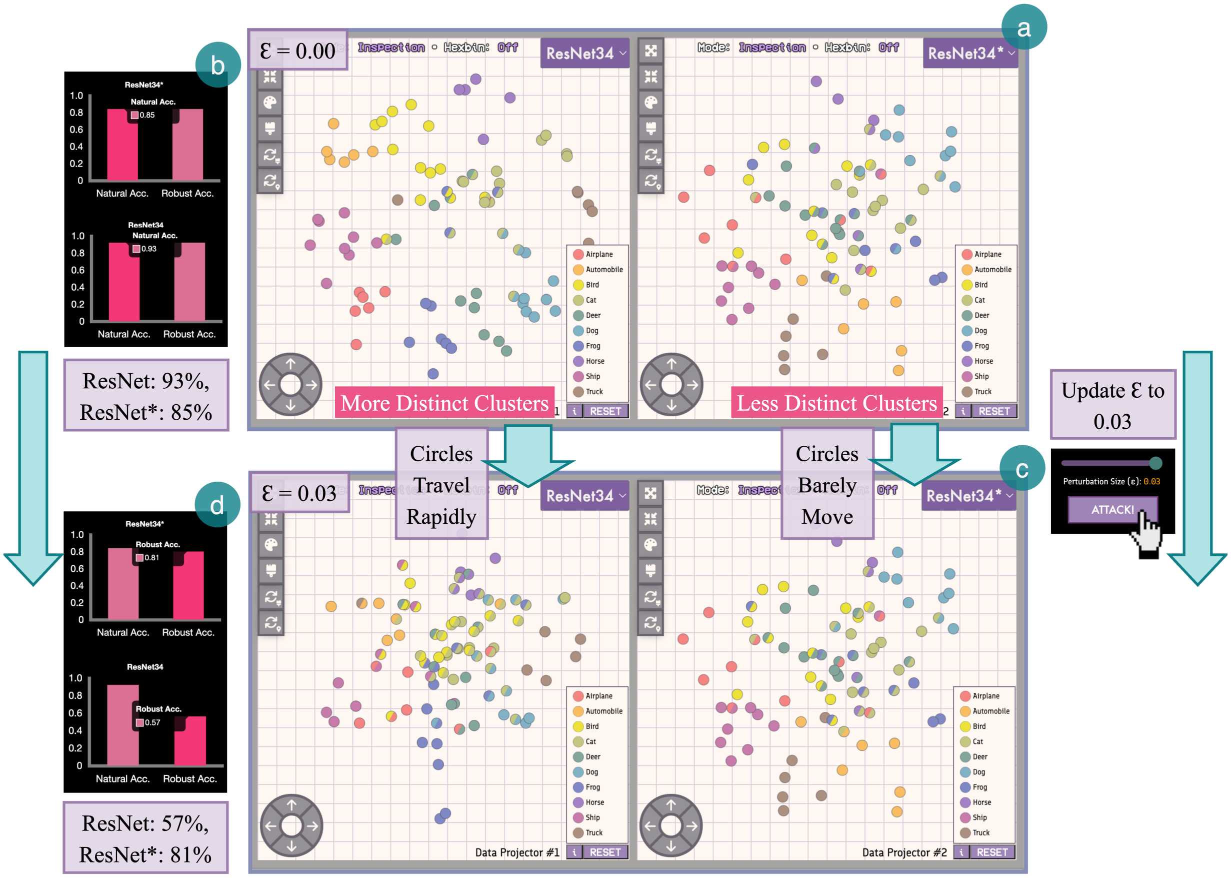
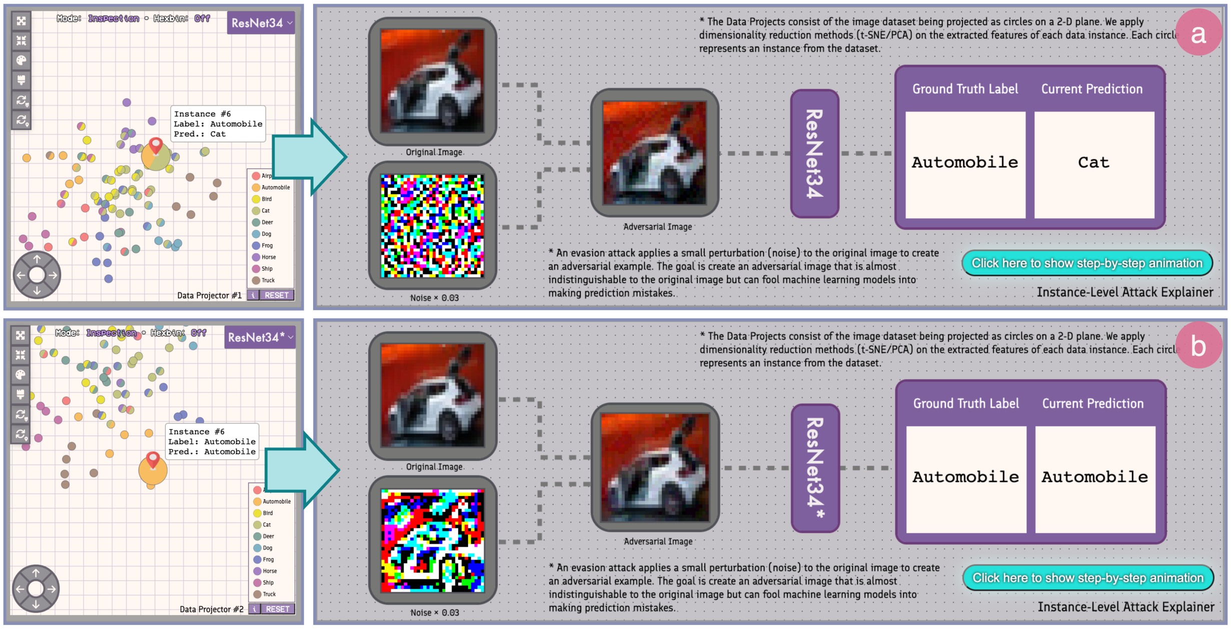


Figure 7.1: An AI expert using AdvEx to investigate ResNet models trained naturally & adversarially under the FGSM attack. Here, she increases ϵ from 0.00 to 0.03 and observes the changes in the two models' embedding distributions and prediction accuracy.

but ResNet* maintains a high robust accuracy of 81% (Figure 7.1d). Since the embeddings represent the features each model perceives from the image data, Zoey realizes that ResNet and ResNet* rely on different sets of features from the dataset for classification. While the FGSM attack is drastically changing the way ResNet perceives the CIFAR-10 dataset, the attack is struggling to change how ResNet* perceives the same dataset. From this, Zoey deduces that the features used by ResNet* are well-generalized in both natural and adversarial datasets, and therefore ResNet* is able to maintain a higher accuracy than ResNet under the attack.

When the perturbation size is 0.03, Zoey notices instance #6 in ResNet's projector, which is an image of an automobile being incorrectly classified as a cat. Zoey clicks on



Instance #6
Label: Automobile
Pred.: Cat

Ground Truth Label: Automobile
Current Prediction: Cat

* The Data Projects consist of the image dataset being projected as circles on a 2-D plane. We apply dimensionality reduction methods (t-SNE/PCA) on the extracted features of each data instance. Each circl represents an instance from the dataset.

* An evasion attack applies a small perturbation (noise) to the original image to create an adversarial example. The goal is create an adversarial image that is almost indistinguishable to the original image but can fool machine learning models into making prediction mistakes.

Instance #6
Label: Automobile
Pred.: Automobile

Ground Truth Label: Automobile
Current Prediction: Automobile

* The Data Projects consist of the image dataset being projected as circles on a 2-D plane. We apply dimensionality reduction methods (t-SNE/PCA) on the extracted features of each data instance. Each circl represents an instance from the dataset.

* An evasion attack applies a small perturbation (noise) to the original image to create an adversarial example. The goal is create an adversarial image that is almost indistinguishable to the original image but can fool machine learning models into making prediction mistakes.

Figure 7.2: (a) In ResNet's case, the visualized perturbation appears very noisy and has no human-interpretable shapes. (b) In ResNet*'s case, the noise image is more human-aligned and has perceptually relevant features that resemble the original image of an automobile.

the instance, and the Instance-level Attack Explainer is immediately updated to reflect ResNet's performance on this image. From the confidence score view, Zoey sees ResNet's confidence for each class before and after the attack. She observes that the FGSM attack has drastically decreased ResNet's confidence score for automobile from 98.4% to 1.8%, and increased its confidence score for cat from 0.2% to 95.3%. By hovering over the confidence bars, Zoey can view the precise percentage difference for each class before & after. Interested in inspecting the images more closely, Zoey clicks on one of the image thumbnails shown in the attack explainer and proceeds to enter the comparison mode. Here, the original and adversarial images of instance #6 are displayed side by side, allowing Zoey to observe the exact pixel differences and identify the subtle perturbations used to create the adversarial image. From this, Zoey understands what it means for adversarial attacks to be "human-imperceptible," that even though the adversarial image still looks very similar to the original image, the subtle pixel modifications are meaningful enough to ResNet to change its prediction. To better understand the process behind the FGSM attack, Zoey activates the step-by-step execution view. She substitutes the current image as an example for the view's demonstration of the attack process. By going through the animated step-by-step explanations, Zoey fully grasps FGSM's underlying attack logic. She learns that the FGSM attack leverages gradient information w.r.t the input pixels to generate subtle perturbations to apply to the image. By modifying the pixels in the direction of the sign of the backpropagated gradients, the attack maximizes the loss and fools ResNet into changing its prediction from automobile to cat.

Interested in investigating ResNet*'s performance on the same instance, Zoey clicks on instance #6 in ResNet*'s projector, which is classified correctly as an automobile. The attack explainer is updated once again, but this time reflecting ResNet*'s performance. From the general view, Zoey immediately notices that in contrast to ResNet, ResNet*'s visualized perturbation has more defined shapes that resemble human-interpretable features from the original image (Figure 7.2). Zoey realizes that since the noise is generated based on the model's gradient information, which highlights the input features that affect the loss most strongly [43], this means that ResNet* relies on more human-interpretable features from the images for classification, and is therefore more robust.

Zoey continues her investigation, and whenever she wants to know more about AML research or needs reminders on AdvEx's functionalities, she scrolls down to the bottom of the interface to view the reading materials included in the General Information Provider.

Chapter 8

Discussion

In this chapter, we discuss the limitations of our current implementation and the future directions to enhance our work. In addition, we present the possible avenues to extend and generalize our proposed design.

8.1 Limitations and Future Work

While our study results show that AdvEx is highly effective for helping users understand adversarial attacks and evaluate model robustness, it still has several limitations. First, the current Data Projectors (Figure 1.1c) enable comparisons of two different models under the same perturbation level, but do not support comparing the same models side by side under different levels of perturbation, as commented by our participants. Future extensions should enable this type of comparison without adjusting the slider back and forth. A simple solution is to add additional toggles to the Data Projectors for switching between the different comparison modes.

Second, when the perturbation size is above 0, the Data Projectors do not distinguish instances that are misclassified prior to the attack from those that are misclassified as a result of the attack. This can be easily addressed by implementing additional visual encodings, for example, displaying the two types of misclassifications in different shapes (triangles and crosses); however, this may increase the cognitive load of users. Alternatively, an optional filtering feature can be added to allow users to focus only on either type of misclassification.

Finally, the evaluation of AdvEx can be further enhanced. A larger sample size should be obtained to better evaluate the effectiveness of AdvEx. Also, the current study was designed with fixed models, and only the FGSM attack and the CIFAR-10 dataset were used to assess the learning effect and usability of AdvEx. In the future, deployment studies with other types of adversarial attacks and datasets should also be conducted to investigate how AdvEx can be used in various real-world scenarios. This will thoroughly examine the strengths and weaknesses of AdvEx, and help us understand how AdvEx can be effectively incorporated into model developers' existing workflows. In addition, although we did not conduct a comparative study with learners due to existing AML learning tools visualizing either too limited or drastically different information, there could still be value in comparing AdvEx to other learning methods in the future. Such a comparison could provide us valuable knowledge on the users' behaviors and experiences when utilizing different learning approaches.

8.2 Generalization and Extension

We designed AdvEx as a system for visualizing adversarial attacks, but the tool is flexible enough to be adapted to visualize the properties and impacts of other data augmentations. For example, AdvEx can be extended to visualize noise applications (e.g., Gaussian noise, salt-and-pepper noise) and other forms of image degradation (e.g., motion blur, Gaussian blur, JPEG compression). Learners may use AdvEx to understand how visual quality impacts the performance of various models, and experts may use AdvEx to evaluate model robustness in those alternative scenarios. Moreover, AdvEx's Data Projectors provide an intuitive way for practitioners to evaluate the accuracy and embeddings of classification models. Though this study focuses on image classifications, the design of the Data Projectors can be extended to assess other classification models (e.g., audio and text classification).

In addition, AdvEx leverages a balanced combination of active visualizations and passive text-based information to help users understand AML, and this design can be applied to visualization tools for learning other machine learning concepts. In fact, many existing tools (e.g., [28, 5]) only focus on their interactive visualizations and place little emphasis on their text-based information, not providing enough guidance and background knowledge to the users. On the other hand, interactive articles [13] usually involve mainly text and provide insufficient visualizations. AdvEx places more balanced weights on both components, ensuring that the users may gain detailed and accurate AML knowledge from our General Information Provider (Figure 1.1e) in addition to exploration with the visualizations. Our design not only reinforces learning by presenting content in multiple formats, but also allows the learners to quickly grasp complex topics that require visual interpretations, which could shed light on future research on the spectrum of modalities for teaching abstract machine learning concepts.

Chapter 9

Conclusion

We have presented AdvEx, an interactive web-based application for visualizing adversarial attacks. AdvEx is intended to help AML learners understand the properties and impacts of adversarial attacks, and allow experienced practitioners to evaluate the adversarial robustness of trained models. Our tool addresses the limitations of existing tools that visualize adversarial attacks and provides both population-level and instance-level information on the attacks' properties and consequences on different machine learning models. We quantitatively and qualitatively assessed AdvEx in a two-part evaluation, including a user study with 12 AML learners and an interview study with three AML experts. Our results show that AdvEx is highly effective both as an educational tool for learning AML and a visual analytics system for evaluating model robustness. Additionally, we discuss the future directions to enhance our work and present potential avenues to extend and generalize AdvEx to other applications.

References

[1] Kelei Cao, Mengchen Liu, Hang Su, Jing Wu, Jun Zhu, and Shixia Liu. Analyzing the noise robustness of deep neural networks. *IEEE transactions on visualization and computer graphics*, 27(7):3289–3304, 2020.

[2] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.

[3] Francesco Crecchi, Davide Bacciu, and Battista Biggio. Detecting adversarial examples through nonlinear dimensionality reduction. *arXiv preprint arXiv:1904.13094*, 2019.

[4] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[5] Nilaksh Das, Haekyu Park, Zijie J Wang, Fred Hohman, Robert Firstman, Emily Rogers, and Duen Horng Polo Chau. Bluff: Interactively deciphering adversarial attacks on deep neural networks. In *2020 IEEE Visualization Conference (VIS)*, pages 271–275. IEEE, 2020.

[6] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[7] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[9] Guodong Guo and Na Zhang. A survey on deep learning based face recognition. *Computer vision and image understanding*, 189:102805, 2019.

[10] 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*, pages 770–778, 2016.

[11] Dan Hendrycks and Kevin Gimpel. Early methods for detecting adversarial images. *arXiv preprint arXiv:1608.00530*, 2016.

[12] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

[13] Fred Hohman, Matthew Conlen, Jeffrey Heer, and Duen Horng Polo Chau. Communicating with interactive articles. *Distill*, 5(9):e28, 2020.

[14] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE transactions on visualization and computer graphics*, 26(1):1096–1106, 2019.

[15] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[16] Minsuk Kahng and Duen Horng Chau. How does visualization help people learn deep learning? evaluation of gan lab. In *IEEE VIS 2019 Workshop on EValuation of Interactive VisuAl Machine Learning Systems*, 2019.

[17] Minsuk Kahng, Nikhil Thorat, Duen Horng Chau, Fernanda B Viégas, and Martin Wattenberg. Gan lab: Understanding complex deep generative models using interactive visual experimentation. *IEEE transactions on visualization and computer graphics*, 25(1):310–320, 2018.

[18] A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tronto*, 2009.

[19] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[20] Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 22(2):712–733, 2020.

[21] Hyeungill Lee, Sungyeob Han, and Jungwoo Lee. Generative adversarial trainer: Defense to adversarial perturbations with gan. *arXiv preprint arXiv:1705.03387*, 2017.

[22] Chang-Sheng Lin, Chia-Yi Hsu, Pin-Yu Chen, and Chia-Mu Yu. Real-world adversarial examples involving makeup application. *arXiv preprint arXiv:2109.03329*, 2021.

[23] Jason Lin and Dilara Soylu. Advis.js. http://jlin.xyz/advis/, 2019.

[24] Lauro Lins, James T Klosowski, and Carlos Scheidegger. Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2456–2465, 2013.

[25] Yuxin Ma, Tiankai Xie, Jundong Li, and Ross Maciejewski. Explaining vulnerabilities to adversarial machine learning through visual analytics. *IEEE transactions on visualization and computer graphics*, 26(1):1075–1085, 2019.

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[27] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 135–147, 2017.

[28] Andrew P Norton and Yanjun Qi. Adversarial-playground: A visualization suite showing how adversarial examples fool deep learning. In *2017 IEEE symposium on visualization for cyber security (VizSec)*, pages 1–4. IEEE, 2017.

[29] Rangeet Pan, Md Johirul Islam, Shibbir Ahmed, and Hridesh Rajan. Identifying classes susceptible to adversarial attacks. *arXiv preprint arXiv:1905.13284*, 2019.

[30] Priyadarshini Panda and Kaushik Roy. Implicit adversarial data augmentation and robustness with noise-based learning. *Neural Networks*, 141:120–132, 2021.

[31] Cheonbok Park, Soyoung Yang, Inyoup Na, Sunghyo Chung, Sungbok Shin, Bum Chul Kwon, Deokgun Park, and Jaegul Choo. Vatun: Visual analytics for testing and understanding convolutional neural networks. 2021.

[32] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[33] Huy Phan. huyvnphan/pytorch_cifar10. Jan 2021.

[34] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[35] Pradeep Rathore, Arghya Basak, Sri Harsha Nistala, and Venkataramana Runkana. Untargeted, targeted and universal adversarial attacks and defenses on time series. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[36] Donghao Ren, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D Williams. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE transactions on visualization and computer graphics*, 23(1):61–70, 2016.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[38] Streamlit. Streamlit • the fastest way to build and share data apps. https://streamlit.io/, 2023.

[39] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[40] Lu Sun, Mingtian Tan, and Zhe Zhou. A survey of practical adversarial example attacks. *Cybersecurity*, 1:1–9, 2018.

[41] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[42] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[43] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[45] Zijie J Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Polo Chau. Cnn explainer: Learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1396–1406, 2020.

[46] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[47] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[48] Jiawei Zhang, Yang Wang, Piero Molino, Lezhi Li, and David S Ebert. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE transactions on visualization and computer graphics*, 25(1):364–373, 2018.

[49] Yu Zhang, Gongbo Liang, Tawfiq Salem, and Nathan Jacobs. Defense-pointnet: Protecting pointnet against adversarial attacks. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 5654–5660. IEEE, 2019.