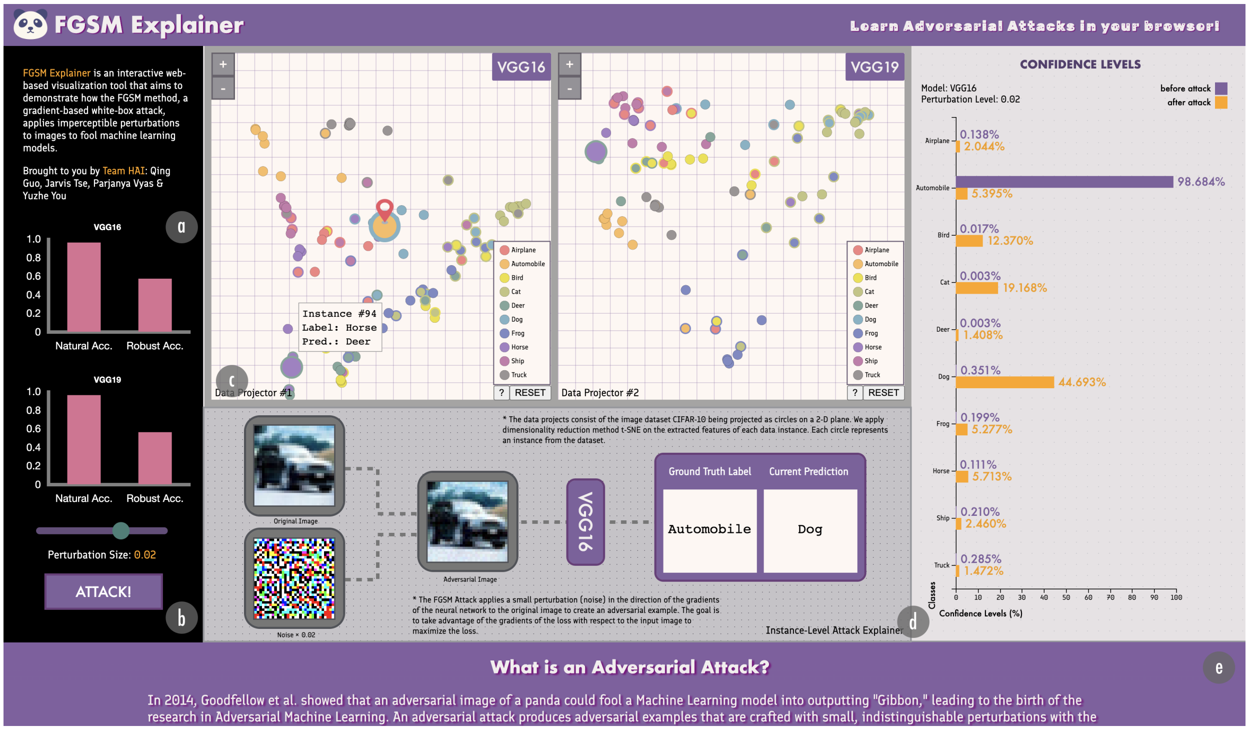
FGSM Explainer: An Interactive Visualization for Understanding Adversarial Attack

Qing Guo, Jarvis Tse, Parjanya Vyas, and Yuzhe You



FGSM Explainer is an interactive web-based visualization tool that aims to demonstrate how the FGSM method, a gradient-based white-box attack, applies imperceptible perturbations to images to fool machine learning models.

Brought to you by Team HAI: Qing Guo, Jarvis Tse, Parjanya Vyas & Yuzhe You

* The data projects consist of the image dataset CIFAR-10 being projected as circles on a 2-D plane. We apply dimensionality reduction method t-SNE on the extracted features of each data instance. Each circle represents an instance from the dataset.

* The FGSM Attack applies a small perturbation (noise) in the direction of the gradients of the neural network to the original image to create an adversarial example. The goal is to take advantage of the gradients of the loss with respect to the input image to maximize the loss.

Instance #94
Label: Horse
Pred.: Deer

Ground Truth Label	Current Prediction
Automobile	Dog

Model: VGG16
Perturbation Level: 0.02

Perturbation Size: 0.02

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," 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

Fig. 1. The FGSM Explainer user interface: (a) Robustness analyzer that displays the models' prediction accuracy pre- and post-attack; (b) Perturbation adjuster that initiates the attack sequence with specified magnitude; (c) Data projectors that project the extracted embeddings of data instances onto a 2-D latent space; (d) Instance-level attack explainer that displays in-depth information regarding the highlighted instance; (e) General information provider that provides some background to the user on adversarial machine learning.

Abstract—Adversarial machine learning is an important emerging technique utilized to exploit various models. Specifically for image classification models, it is difficult to understand the underlying logic behind adversarial examples even for ML practitioners because the differences in such images are indistinguishable to human eyes. Therefore, we create a web-based application that can help visualize the FGSM attack along with data and model performance analytics. We design the web app to be interactive and animated, which can help ML experts visualize the internal logic and properties of an adversarial attack. Additionally, we intend the visualization to be helpful in terms of comparing the robustness of various models with respect to the attack method chosen and illustrating the shift in confidence level, through which ML experts can understand how prediction results change with varying perturbation levels applied to the data instances.

Index Terms—FGSM attack, adversarial machine learning, information visualization

1 Introduction

In 2014, Goodfellow *et al.* [8] showed that an adversarial image of a panda can fool a machine learning (ML) model into labeling it as a gibbon with high confidence, leading to the birth of the research in adversarial ML. An adversarial image is intentionally crafted by an attacker with often imperceptible perturbations to result in prediction mistakes of a model. Some of the most well-known adversarial attacks include the Fast Gradient Sign Method (FGSM) Attack [8], the Basic Iterative Method (BIM) Attack, the Projected [12] Gradient Descent (PGD) Attack [17], etc. For this work, we focus on adversarial attacks in image classifications.

The primary objective of our study is to explore how visualizations can help ML practitioners understand the underlying logic of adversarial attacks. Especially, we aim to explore for practitioners with limited background in adversarial ML, how effectively can visualizations highlight and explain the adversarial images, as they tend to be imperceptible to human eyes but can fool state-of-the-art classifiers [4]. To achieve our goal, we develop a web-based interactive visualization tool that demonstrates how the FGSM attack, a gradient-based white-box attack [14], alters images' projected positions on the two-dimensional feature space. We also display plots representing two ML models' corresponding two-dimensional feature space side-by-side such that the users of our visualization tool can intuitively perceive

- *The authors are from the University of Waterloo*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

how some ML models are more robust against the FGSM attack than the others. Note that we have chosen to illustrate adversarial attack by demonstrating the FGSM attack since existing work has shown that the FGSM attack is capable of effectively fooling various ML models used for classifying images from multiple datasets [2, 23]. In addition, studies have shown that visualizations could be used to effectively help their users understand the underlying logic and architecture of various ML models, including convolutional neural network [30] and generative adversarial network [10]; these visualizations have provided inspiration to the design of our visualization tool. The details of our proposed visualization tool will be discussed in the proposed design section.

Existing studies have proposed several visualization tools that aim to illustrate adversarial attacks in image classifications; however, these tools all possess certain disadvantages. For instance, Bluff [7] visualizes how adversarial attacks confuse deep neural networks by displaying what features have been induced by pixel perturbations to contribute to the misclassifications; nonetheless, it focuses too much on how neural networks interpret features that exist in the inputted images, as well as how adversarial attacks induce incorrect features, and does not intuitively illustrate the impact of an adversarial attack on different images from the same dataset, or how ML models with different robustness levels react to the attack. Adversarial-Playground [20] is another existing visualization tool that aims to visualize the FGSM attack; this tool includes an interface that displays the original image and the adversarial image side-by-side, with a model's corresponding prediction distribution presented under as a histogram. However, similar to Bluff [7], Adversarial-Playground [20] does not intuitively demonstrate an adversarial attacks' impacts on multiple images or models. Moreover, Steinberg and Munro [26] have successfully visualized different adversarial attacking methods' impacts on multiple dimensionality reduction algorithms and displayed those results as 2-D plots where each marker represents one image entry. Although the study of Steinberg and Munro's [26] visualizes the impacts of adversarial attacks on a dataset-level, it does not provide an interface that can be used to intuitively explore the resulting two-dimensional plots; instead, the resulting plots are static and require user expertise in adversarial ML to understand.

For our study, we parameterize the adversarial perturbation applied to data instances and visualize the process of adversarial attack in the form of interactive visualizations and animated sequences. The target users of our proposed visualization tool are ML practitioners with adequate knowledge of ML models, but unfamiliar with the area of adversarial ML. Therefore, our visualization tool focuses on the visual exploration of the attack process instead of placing too much emphasis on the reasoning logic of ML models. In addition, our tool contains a visual analysis of effects of the adversarial attack on the entire dataset as a group, so that ML practitioners can have a high-level understanding of the changes in prediction results. Our visualization also illustrates the movement of perturbed data in a projected space with dimensionality reduction technique t-SNE, so that practitioners can have an instance-level grasp of adversarial attacks.

In summary, following are the primary contributions of our work:

- We propose a novel interactive visualization tool that helps ML practitioners with limited adversarial knowledge understand the underlying logic and consequences of an adversarial attack using instance-level and high-level visualizations.
- Through this approach, we enable ML developers to evaluate the performance of different ML models with respect to the adversarial attack method being studied.
- We provide an actualization of our application and its evaluation with respect to usability and effectiveness using CIFAR-10 dataset to visualize the adversarial examples generated by FGSM attack.

2 Background

Despite deep neural networks' widespread success in various applications, the lack of robustness of ML models raises essential concerns. Numerous models have recently been discovered to be vulnerable to adversarial examples [8]. Adversarial examples are data instances crafted with small, indistinguishable perturbations to result in model prediction mistakes. The attacker determines the characteristics required in adversarial examples and then adopts particular adversarial attack strategies.

The attack methods can be categorized into two groups according to the attacker's intent: targeted attacks and non-targeted attacks. Targeted attacks misinform machine learning models towards a specific class. Typically, targeted attacks will have the effect of increasing the probability of the targeted adversarial class. In contrast, non-targeted attacks are generic, in the sense that except for the original class, the class of adversarial examples can be arbitrary. Several adversarial attacking methods can be used for both targeted and non-targeted attacks [31].

Adversarial attack methods can also be categorized into two groups according to their targeted threat model: white-box attacks and black-box attacks [31]. White-box attacks presume the attacker has complete knowledge of the machine learning model, such as training data, model parameters, and model architecture. Whereas black-box attacks presume the attacker has no knowledge of the model that they target to attack.

For our study, we choose to visualize the FGSM attack [8], one of the first and most well-known adversarial attacks to date. The FGSM attack is a gradient-based white-box attack that is simple in logic but has proven to be highly effective. The attack adjusts the input image by taking a step toward the sign of the back-propagated gradients for each pixel to maximize $J(X,y)$, where J is the corresponding loss function and ε controls the scale of the perturbation [13]:

$$X' = X + \varepsilon \text{sign}(\bigtriangledown_X J(X,y)).$$

3 Related Work

We provide a summary of the related studies present in the literature in this section. We divide these research works into three parts: (i) adversarial machine learning, (ii) visualizations of adversarial attacks, and (iii) visualizations for learning neural networks. Finally, we compare our work with these existing studies and highlight our contributions.

3.1 Adversarial Machine Learning

In the past few years, a wide range of adversarial attacks have been proposed to work under different threat models, namely white-box and black-box attacks. L-BFGS Attack [28], FGSM [8], BIM [12] and DeepFool [18] 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) [5] and One Pixel Attack [27] have also been explored extensively in the literature.

Several prior studies have tried to understand the characteristics of the FGSM attack. Zhang *et al.* [32] discovered that FGSM attack may create not only two-dimensional adversarial images but also three-dimensional adversarial examples when applied to three-dimensional real-world data. They attempted to deceive pointNet [24], a commonly used network for three-dimensional point cloud data. Their experiments showed that PointNet might be deceived into making erroneous predictions with high confidence even if the points are shifted slightly. They improved the robustness of PointNet by training the model with a combination of clean and adversarial examples. Additionally, t-distributed Stochastic Neighbor Embedding (t-SNE) is applied to map three-dimensional data to two-dimensional data for visualization.

In 2019, Crecchi *et al.* [6] investigated the separability of clean and adversarial samples in the projected space. They experimentally proved that adversarial examples could be distinguished from manifold samples using a nonlinear dimensionality reduction technique such as t-SNE. In the same year, Pan *et al.* [21] conducted experiments on FGSM attacks on deep neural networks (DNN) based image classifiers: they proposed a technique to identify classes susceptible to being attacked by FGSM, explored the DNN model to compute the distance between classes in feature space, generated an adversarial map that takes the distance in feature space between classes as input and generates the probability of the source classes of adversarial examples as output.

They evaluated their approach with benchmark datasets such as MNIST, Fashion MNIST, and CIFAR-10.

In addition, FGSM has been applied in various domains such as image classification and object detection as it is efficient and easy to implement. For instance, Musa *et al.* [19] applied FGSM to deceive a convolutional neural network (CNN) trained with human images and their experimental results showed that FGSM has the desired impact in most instances.

To build FGSM Explainer, we took inspiration from these previous works. Precisely, similar to the works [6] and [21], FGSM Explainer visualizes the distance information in the feature space with dimensionality reduction techniques such as t-SNE. Inspired by the methodology of the works [19] and [21], FGSM Explainer aims to deceive machine learning models trained with CIFAR-10 as the benchmark dataset.

3.2 Visualizations of Adversarial Attacks

A range of visualizations that aim to make adversarial attacks more interpretable have been proposed in past studies. Nonetheless, the problems these proposed visualizations target differ from the ones that FGSM Explainer aims to address.

Bluff [7] and Adversarial-Playground [20] are capable of demonstrating adversarial attacks at high-level aspects, but they do not visualize the attack's impacts on an instance level. A study proposed by Steinberg and Munro [26] visualized adversarial attacks through static plots, but these static plots require expertise in adversarial attack to comprehend. Lin *et al.* proposed AdVis [15], an interactive visualization tool that visualizes FGSM attacks by comparing original images and their corresponding adversarial images side-by-side and shows the heat map of the perturbed pixels. Cao *et al.* [3] proposed an interactive visualization tool called AEVis that explains adversarial attacks on DNNs by extracting critical neurons and their connections and showing how those adversarial examples deactivate and activate specific features to fool the ML models through the datapaths that go through those neurons [3].

t-SNE has also been extensively used to visualize the adversarial attacks on the instance level. Ma *et al.* [16] have proposed an interactive visualization tool to visualize adversarial attacks on binary classification ML models. This visualization tool contains a projection view that utilizes the t-SNE plot to display the data on an instance level, which helps the tool depict how the ML models are affected by the adversarial attacks. However, this visualization tool does not show how adversarial attacks modify an instance, support classification tasks that involve more than two classes, and is designed specifically to debug ML models instead of for educational purposes. Lastly, Park *et al.* [22] have proposed VATUN, an interactive visualization tool that utilizes t-SNE to build plots of the instances; although VATUN allows its users to interactively visualize the impacts of FGSM and data augmentation such as blurring, rotation, or brightness adjustment, on the images both through t-SNE plots and through direct comparisons between the original and the modified images, VATUN focuses on displaying the performance of the ML model's prediction accuracy after such image modifications and does not visualize the perturbations that have been implemented on each image nor provide visualization for more than one ML model simultaneously for its users to compare their robustness.

3.3 Visualizations for Learning Neural Networks

Several visualization tools that aim to help people learn about ML have been previously proposed. For instance, GAN (i.e., generative adversarial network) Lab [10] is a visualization tool that focuses on interactively and simultaneously visualizing how different instances are interpreted by GAN; it has been demonstrated that GAN Lab is capable of helping its users learn about the underlying logic of GAN. Also, CNN (i.e., convolutional neural network) Explainer [30], another previously proposed visualization tool, not only displays the structure of a CNN but also allows its users to smoothly transit to make the interface focus on specific mathematical operations (e.g., convolution); consequently, CNN Explainer's users could effectively learn the underlying logic of CNNs. Moreover, a past study has proposed Summit [9], an interactive visualization tool designed to intuitively display what image features are detected by deep neural networks (DNNs), those features' relationships as interpreted by the DNNs, and how those features are utilized for classification tasks in the DNNs.

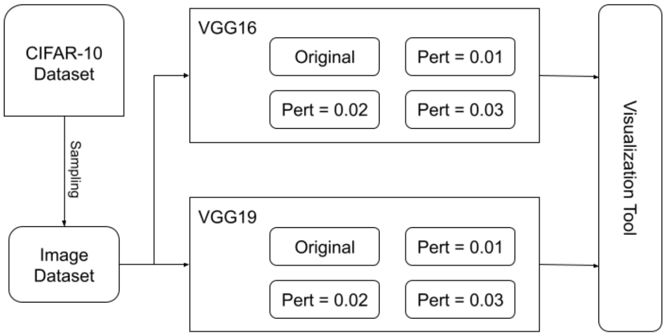


Fig. 2. The schematic diagram depicting the system architecture.

Despite focusing on visualizing neural networks instead of adversarial attacks, all three aforementioned studies have provided us with inspiration for the design of FGSM Explainer. Specifically, similar to GAN Lab [10], FGSM Explainer also visualizes the data on an instance level. Although that is not the case for CNN Explainer [30] and Summit [9], these two visualization tools have inspired us to visualize FGSM on multiple abstraction levels including both instance- and group- levels.

3.4 Our Contributions

As discussed in the previous sections, although many adversarial attack and general ML based visualizations already exist, none of these visualizations has the same design goals as FGSM Explainer. Specifically, most of the existing adversarial attack visualizations are designed for analytical purposes by adversarial ML practitioners instead of for educational purposes. Additionally, most of these visualizations do not demonstrate the impacts of adversarial attacks on a dataset-level. Even for the few existing visualizations that are designed for educational purposes and/or show adversarial attacks' impacts on a dataset level, such as VATUN [22] and AdVis [15], they do not provide an interface that allows their users to easily compare the robustness of multiple ML models. FGSM Explainer aims to allow its users who have little or no knowledge of adversarial ML to:

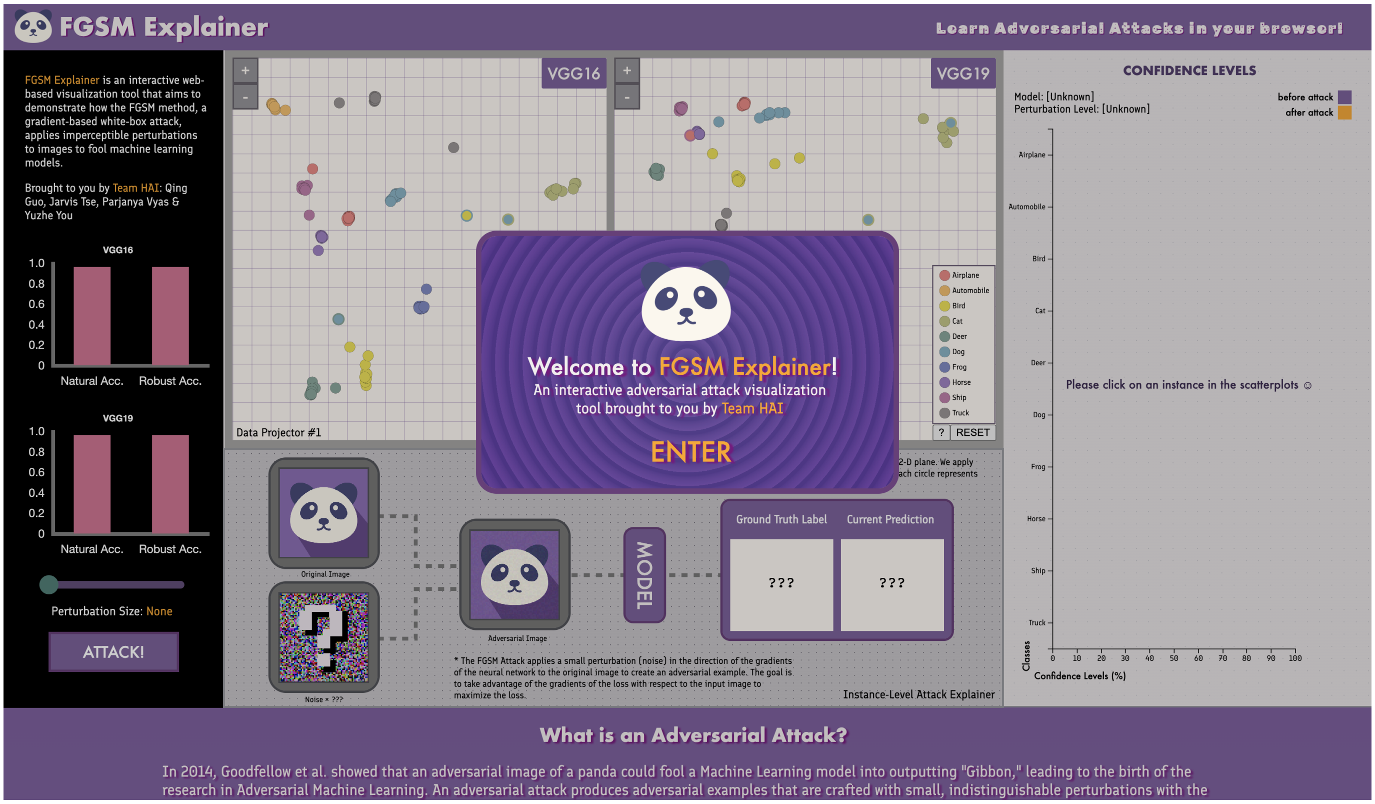
- learn about adversarial attacks through viewing their impacts on images on both a dataset level and an instance level;
- compare the robustness of different ML models by viewing the t-SNE plots of different ML models simultaneously and through a designated robustness analyzer panel;
- adjust perturbation levels and compare different pairs of ML models.

4 Proposed Design

Our proposed design can be divided into three major parts: image data, machine learning models, and the visualization. In this section, we will first provide an overview of system; we then describe in detail the steps involved in data preparation and the selection of the machine learning models; we conclude this section by discussing the functionalities of each component in our visualization.

4.1 System Overview

The image data, the machine learning models, and the visualization web app combined constitute the FGSM Explainer, as demonstrated in Figure. 2. Specifically, the images were first loaded to the FGSM Explainer. Next, we performed adversarial attacks on the loaded images and inputted them into the selected machine learning models (i.e., VGG16 and VGG19). Lastly, we displayed the machine learning models' predictions as well as the inputted images to the users through the FGSM Explainer's visualization.



FGSM Explainer is an interactive web-based visualization tool that aims to demonstrate how the FGSM method, a gradient-based white-box attack, applies imperceptible perturbations to images to fool machine learning models.

Brought to you by Team HAI: Qing Guo, Jarvis Tse, Parjanya Vyas & Yuzhe You

Welcome to FGSM Explainer!

An interactive adversarial attack visualization tool brought to you by Team HAI

2-D plane. We apply
ach circle represents

Model: [Unknown]
Perturbation Level: [Unknown]

Please click on an instance in the scatterplots ☺

* The FGSM Attack applies a small perturbation (noise) in the direction of the gradients of the neural network to the original image to create an adversarial example. The goal is to take advantage of the gradients of the loss with respect to the input image to maximize the loss.

In 2014, Goodfellow et al. showed that an adversarial image of a panda could fool a Machine Learning model into outputting "Gibbon," 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

Fig. 3. The initial interface of the FGSM Explainer upon launching as an adversarial attack visualization software. The user is first greeted by a welcome window that introduces the FGSM Explainer. Upon clicking the enter button, the panda icon transforms into a gibbon and brings the user to the actual interface of the interactive software.

4.2 Data Preparation

For our proposed design, we have sampled a smaller subset from CIFAR-10's testing dataset. The CIFAR-10 dataset consists of 60,000 32x32 coloured images in 10 classes (50,000 training data and 10,000 testing data), with 6,000 images per class. Since the main focus of our design is to visualize how each data entry alters before and after the adversarial attack, we believe that displaying all 10,000 testing data on screen may be too overwhelming for the users to digest instance-level changes. Therefore, we have randomly sampled 10 images from each class (with a total of 100 images) for our visualization.

To prepare the data for our application, we conduct an attack on both selected models separately with perturbation sizes of 0.00, 0.01, 0.02, and 0.03. For this study, we select to visualize the FGSM attack [8], one of the first and most well-known adversarial attacks to date. Though simple in logic, the FGSM attack has been demonstrated to be extremely effective; the attack utilizes a gradient-based approach that adjusts the input data to maximize the corresponding loss. To prepare our data for visualization, we use PyTorch to download the 10,000 raw CIFAR-10 testing data and load the pre-trained models to perform attacks on a subset that includes 100 images. We do so by taking the backpropagated gradients with respect to each input image and perturbing each input data by taking a step in the direction of the sign of the obtained gradients to maximize the calculated loss. The resulting adversarial dataset is saved as a NumPy array in the format of a .npy file. A separate Python script is written to project extracted features of both the original and adversarial datasets onto a 2-D plane by applying t-SNE [29], a nonlinear dimensionality reduction technique and a variation of Stochastic Neighbor Embedding that visualizes high-dimensional data by giving each data point a location in a two-dimensional map. The resulting coordinates of each instance are then combined with both their original label and current prediction (with confidence scores across all classes) and are outputted as CSV files that will later be read by our web application.

4.3 Model Selection

Since FGSM is a gradient-based white-box attack, the set of adversarial examples generated differ depending on the specific model being attacked. To demonstrate the impacts of the FGSM attack on different models, we include visualizations of two specific models side by side. For our design, we use readily available models that were pre-trained on the CIFAR-10 training dataset. Specifically, we select VGG16_BN (VGG-16 with batch normalization) and VGG19_BN (VGG-19 with batch normalization) for our demonstration.

The VGG-Network architecture is a convolution neural network first proposed by Simonyan and Zisserman in [25]; compared to previous derivatives of AlexNet, the VGG-Network investigates the effects of convolutional network depth on its accuracy in a large-scale image recognition setting. The original study [25] has demonstrated that using networks of increasing depth and an architecture with very small (3x3) convolution filters results in a significant improvement over the prior-art configurations by pushing the depth to 16-19 weight layers. For our proposed design, we select the VGG-Networks with 16 and 19 weight layers, respectively, as VGG16 achieved a top-5 accuracy of 91.9% in the ImageNet competition and a validation accuracy of 94.0% on the CIFAR-10 dataset, while VGG19 achieved a top-5 accuracy of 92.0% in the ImageNet competition and a validation accuracy of 93.95% on the CIFAR-10 dataset.

We want to investigate how the performance of each model is affected before and after the FGSM attack, and if increasing the backbone depth of the same model may have effects on each model's adversarial robustness. Additionally, we believe that directly applying t-SNE on the raw pixel values of CIFAR-10 images may be undesirable as high-dimensional image data with multiple channels often contains both useful and useless features for classification, and thus may result in clusters that are less meaningful. Therefore, we first extract the features of each instance by feeding the data through the backbones of the VGG models and then apply dimensionality reduction on the resulting embeddings that are obtained prior to the final linear layer.

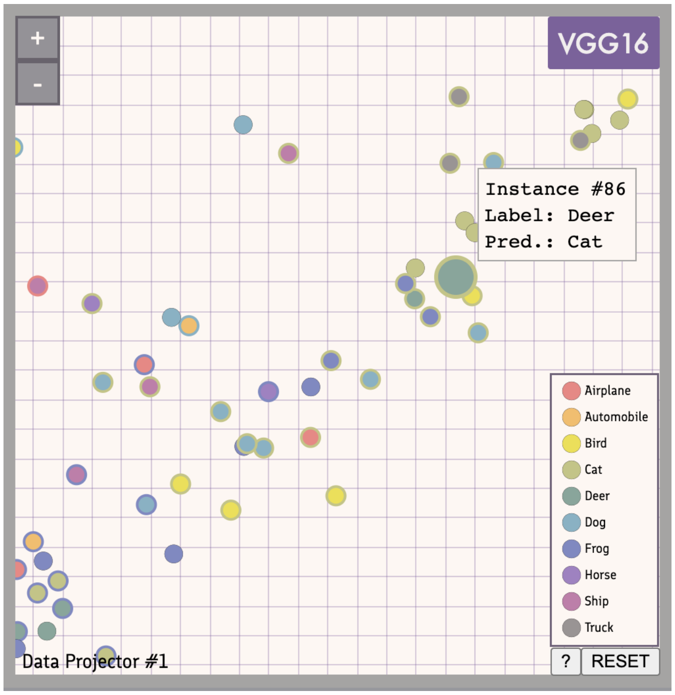


Fig. 4. An example of the user zooming into the data projector and hovering over a data instance that is being mislabelled by the VGG16 model. The reset button in the bottom right corner allows the user to quickly revert the scatterplot back to its original scaling.

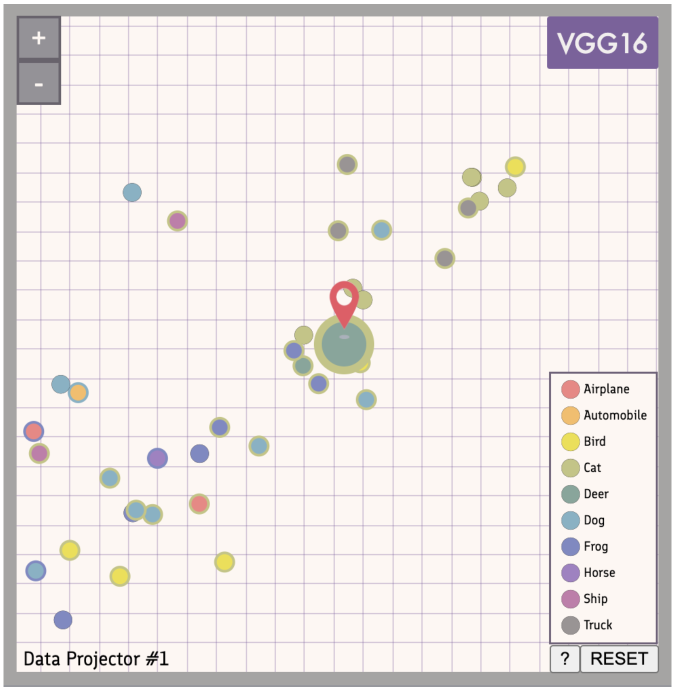


Fig. 5. Instance #86 when it is clicked and highlighted. An animated pin will be placed on top of the selected circle, and the scatterplot will be relocated such that the current instance is emphasized in the center of the projector.

4.4 Interface Components of the Visualization

For our actualization of the web-based visualization application, we implement a NodeJS-based web app and use a combination of D3 and React to realize the interactive visualizations. Our visualization application can be divided into the following components: (1) data projector, (2) instance-level attack explainer, (3) robustness analyzer, (4) perturbation adjuster, and (5) general information provider.

4.4.1 Data Projector

The data projector consists of the image dataset being projected as circles on a 2-D plane. Each circle represents an instance from the dataset and will be highlighted by an assigned colour that represents its ground truth label. When the user clicks on the attack button on the left panel after selecting a perturbation size larger than zero, the dataset will be updated to its corresponding adversarial counterpart and this change will be visualized by the circles' transitions in positions and outline colours. We have achieved this through animated transitions that emphasize the moving trajectory of each circle and the change in its outline colour with an increase in stroke size. For example, suppose we assign the class "airplane" with the colour red and the class "bird" with the colour purple. A red circle with a barely discernible outline will represent an image with the ground truth label "airplane" and that the model is predicting it correctly. On the other hand, a red circle with a thick purple outline on the outside represents an "airplane" image that is being predicted to be a "bird" by the current model. When the user hovers over each circle, a summary that consists of its original label and the current prediction is displayed in the form of a tooltip.

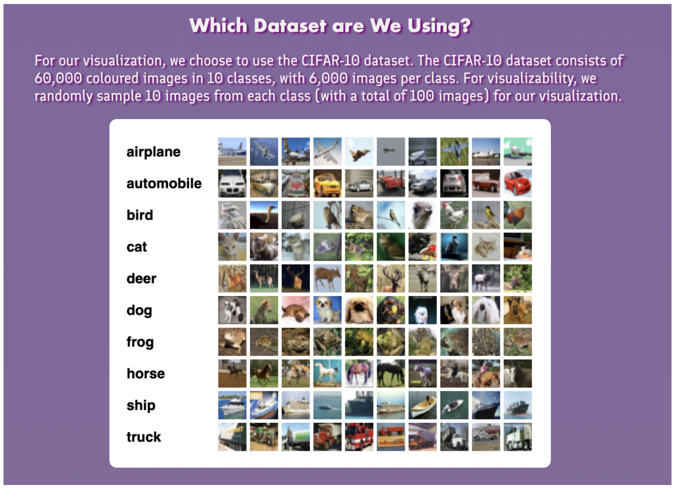
Moreover, we recognize that if multiple images share very similar features, they may end up getting projected on top of each other due to the small differences between their coordinates. Therefore, to prevent this potential inconvenience, we also allow the user to zoom into and drag the scatterplot. Clicking on a specific circle will highlight the instance being inspected by moving the highlighted circle to the center of the canvas via panning the entire scatterplot. We also include a reset button at the bottom right of each canvas to allow the user to easily restore each scatterplot back to its initial scaling and position, and a guidance button that provides quick instructions to the user on how to navigate the scatterplots.

In short, the data projector is designed to illustrate the impact of the adversarial attacks on a dataset-level through the use of t-SNE scatterplots. In addition, through the data projector provides, the users can intuitively see each the corresponding machine learning model's prediction for each instance as well as can easily select specific instances to view their detailed information (e.g., their images before and after the attack).

4.4.2 Instance-level Attack Explainer

When the user wishes to see more detailed instance-level information and how the attack is conducted for a specific image, they may click on the circle that represents the entry and an attack explainer panel will be displayed. Taking inspiration from GAN Lab [11], the goal of the instance-level attack explainer is to provide more details on the underlying logic of the FGSM attack and visualize the perturbation applied to the specific image, as the resulting adversarial example may often be imperceptible to human eyes. For our visualization, we utilize a combination of animations to visualize this explainer: for instance, we have implemented an animated sequence in which the original image and the generated perturbation slowly move towards each other with reduced transparency and overlay on top of each other, and then gradually fades in the final perturbed image to demonstrate the result of the attack. We have also animated the dashed lines involved in the visualization to illustrate the general flow of the attack. If the user wishes to inspect the images more closely, they may hover over the thumbnails of the images displayed and click on them, which will display an enlarged version of the image selected. An comparison mode is also provided if the user wishes to inspect the original image and the adversarial image side by side to investigate the exact difference in pixels. Additionally, an interactive grouped bar chart is available as a separate panel to display the confidence levels across all ten classes, in which the confidence levels of each class pre- and post-attack are grouped together. Hovering over each pair of bars displays the exact difference in confidence value of one class before and after the conducted attack.

The main design goal of the the instance-level attack explainer is to intuitively show that the perturbed image is the combination of the original image and the perturbation. Also, through the confidence level



Which Dataset are We Using?

For our visualization, we choose to use the CIFAR-10 dataset. The CIFAR-10 dataset consists of 60,000 coloured images in 10 classes, with 6,000 images per class. For visualizability, we randomly sample 10 images from each class (with a total of 100 images) for our visualization.

Fig. 6. An example of the general information we provide below the interactive software. We provide background information such as what an adversarial attack is, what the FGSM attack is, what models and dataset we are using for our visualization, etc.

panel, the users can intuitively observe how the corresponding machine learning model's confidence level changes between the original and the perturbed images (i.e., before and after an adversarial attack).

4.4.3 Robustness Analyzer

The robustness analyzer is presented as two small interactive bar charts on the left-most panel that illustrate the natural accuracy and robust accuracy of the VGG-Networks before and after the attack. The natural accuracy stands for the VGG model's prediction accuracy on the unperturbed CIFAR-10 dataset, while the robust accuracy stands for the model's performance on the corresponding adversarial dataset. The height of the bar representing the robust accuracy transitions up and down depending on the perturbation size the user has selected for the visualization.

With the robustness analyzer, the users can easily compare the robustness of the selected machine learning models against the FGSM attack with specified perturbation while seeing how many instances are misclassified in each machine learning model's corresponding scatterplot. By doing so, the users can intuitively learn the concept of robustness against adversarial attacks.

4.4.4 Perturbation Adjuster

The perturbation adjuster is presented in the form of a slider on the left-most panel of the visualizer. The user can adjust the slider horizontally to choose the perturbation of interest (None, 0.01, 0.02, & 0.03). After a perturbation size has been selected, the user may click on the attack button below the perturbation slider to initiate the corresponding animated sequence that simulates the attack. An example sequence would be that after the user has clicked on the attack button, the circles of both scatterplots begin to transition to locations with potential changes in outline colours and stroke sizes, and the bars in the robustness analyzer that represent robust accuracy will either increase or decrease their heights depending on the performance of the VGG models after the conducted attack.

By using the perturbation adjuster, the users can easily change the perturbation size of the adversarial attacks and correspondingly observe the impacts that are proportional to the perturbation sizes. We have also included the perturbation size of zero such that the users can explore the images' representations in the latent space when no perturbation has been applied.

4.4.5 General Information Provider

We have also included a short paragraph above the robustness analyzer that provides general information on the FGSM Explainer. If the user wishes to read more about our work and the research of adversarial machine learning, they may read the information page placed under the interactive visual analysis tool (See Fig. 6), which provides more in-depth explanations regarding those topics.

The goal of our visualization is to explain the underlying logic and consequences of adversarial attacks to ML practitioners who are unfamiliar with the area of adversarial ML. Therefore, by including text explanations of the core concepts and background information of adversarial attacks, the users can gain detailed and accurate knowledge of the adversarial attacks in addition to perceiving adversarial attacks through interactive visualizations.

5 Case Study

In this section, we describe a hypothetical scenario that illustrates how the FGSM Explainer can be utilized by ML experts to learn about the FGSM attack, and its effects on the image classification models.

We first describe the hypothetical usage scenario in detail, followed by the use case for each section of the tool and how it benefits the user in gaining an in-depth understanding of the effects and working of the FGSM attack.

5.1 Usage Scenario

We assume a hypothetical user called Bob, who is an expert in developing state-of-the-art image classification ML models. Bob has recently discovered that one of the models he has deployed has been subjected to the FGSM adversarial attack and as a result its accuracy has been found to suffer drastically for attack images. Therefore, Bob is now eager to learn more about these attacks. Specifically, he intends to see how the attack works in action, what are the visible changes to the attacked images, what is the attack's overall effect on the accuracy of the model, and how are the miss-classified images distributed. Moreover, Bob also wants to compare the effects of the attack on his model with one of the models available from the literature. For the purpose of illustration, we will assume that the model that Bob deployed is "VGG19", whereas the model he wants to compare with is "VGG16".

5.2 Robustness Analyzer

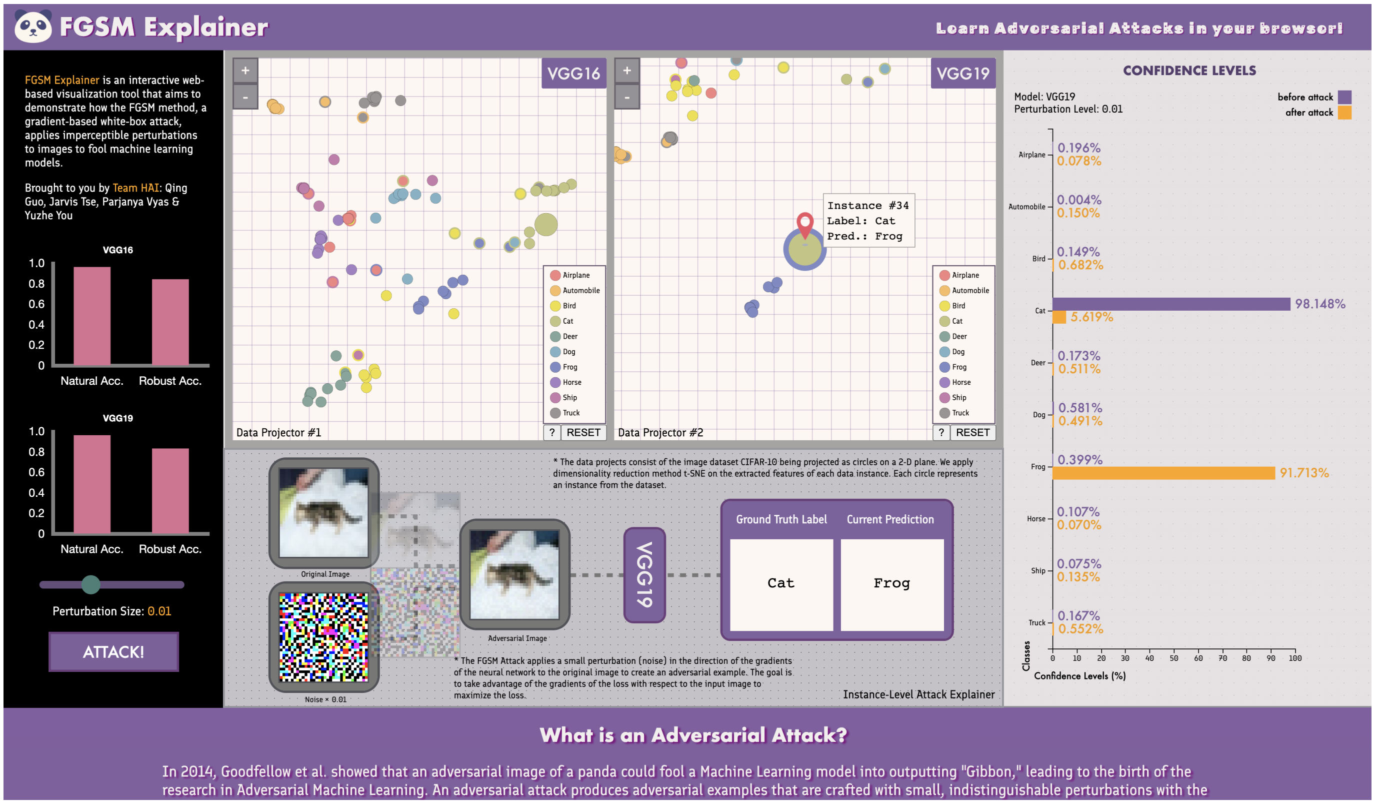
After being greeted by the welcome window and proceeding to the interface of the FGSM Explainer, the first visualization that helps Bob understand the basics of the tool is the *Robustness Analyzer*, which is situated on the left panel. By reading the general information at the top, Bob gets to know how FGSM the attack works in a nutshell.

Bob can use the slider to adjust perturbation levels and press the *Attack* button, which results in an updated natural vs robust accuracy bar chart that clearly shows Bob that as the perturbation increases, the accuracy of his model decreases. Moreover, the other bar chart at the top also shows Bob that the robust accuracy of his model ("VGG19") is marginally lesser than the robust accuracy of the "VGG16" model for each attack scenario, proving that the robustness of his model is slightly worse than VGG16.

5.3 Data Projector

To further investigate the overall effect of the attack on the dataset, Bob then moves on to the data projector situated in the middle of the screen. Using this visualization, Bob can see how each image class is initially grouped correctly together as a cluster for the original images, but changes significantly for each attack scenario with increasing perturbation. Bob can also compare the clusters of instances for VGG16 and VGG19. Moreover, he can clearly distinguish the correctly predicted instances from the wrong predictions (along with what the original and prediction labels are) by the filled and border colours of the projected dots. Finally, Bob can map each instance from VGG19 to the same instance in VGG16 to compare the results for the same instance in both models by hovering on the dot representing the instance.

Apart from these, Bob can interact with the scatterplot by zooming in and out, as well as dragging the mouse to move around the scatterplot. By analyzing the scatterplot, Bob finds that there is a particular instance (Instance #34) in the attack scenario with perturbation 0.01, which is correctly classified by the VGG16 model but results in a wrong



FGSM Explainer is an interactive web-based visualization tool that aims to demonstrate how the FGSM method, a gradient-based white-box attack, applies imperceptible perturbations to images to fool machine learning models.

Brought to you by Team HAI: Qing Guo, Jarvis Tse, Parjanya Vyas & Yuzhe You

Perturbation Size: 0.01

* The data projects consist of the image dataset CIFAR-10 being projected as circles on a 2-D plane. We apply dimensionality reduction method t-SNE on the extracted features of each data instance. Each circle represents an instance from the dataset.

Ground Truth Label	Current Prediction
Cat	Frog

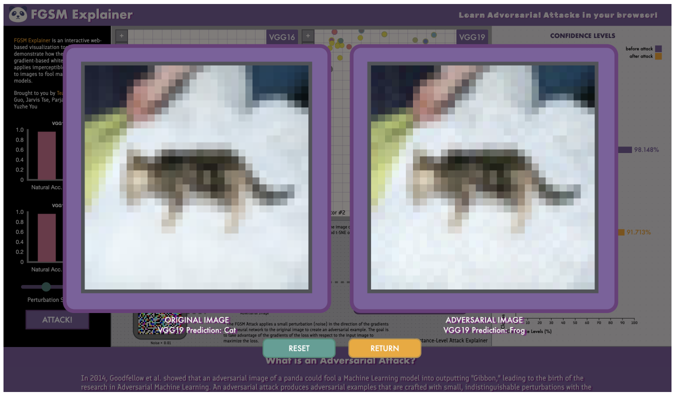
* The FGSM Attack applies a small perturbation (noise) in the direction of the gradients of the neural network to the original image to create an adversarial example. The goal is to take advantage of the gradients of the loss with respect to the input image to maximize the loss.

Model: VGG19
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," 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

Fig. 7. When the perturbation size is 0.01, Bob notices that there is a particular instance (Instance #34) that is being correctly classified by the VGG16 model but not by the VGG19 model. Bob clicks on this instance to analyze it further individually, which results in the updates of the instance-level attack explainer along with its accompanied confidence level bar chart.



ORIGINAL IMAGE
VGG19 Prediction: Cat

ADVERSARIAL IMAGE
VGG19 Prediction: Frog

Fig. 8. The side-by-side comparison of Instance #34 original and adversarial images.

prediction by VGG19. Bob can click on this instance to analyze it further individually. This operation is illustrated in Figure 7.

5.4 Instance-level Attack Explainer

When Bob clicks on Instance #34 he can see the original and the adversarial images in the instance-level attack explainer at the bottom of the screen. The animation shows how the original image and the noise are combined to create the adversarial image by the FGSM attack. Bob can also click on the original or the adversarial image to see a blown up (intentionally pixelated) version of the image and compare the original and the adversarial images side-by-side to clearly visualize the changes in the image pixels as shown in Figure 8. Additionally, Bob can also click on the noise image to visualize the perturbation.

When Bob looks at the right panel, he sees how the confidence levels for each class change in the attack image as compared to the original image. Moreover, Bob can select the same image instance in the VGG16 model from the data projector as described in the previous section to see how the confidence levels still change in the attack image but the final prediction remains the same, making the model more robust than Bob's model.

5.5 General Information Provider

Finally, after analyzing various image instances and comparing the effects of the FGSM attack on each of these using the tool sections described above, Bob now has a clear understanding of how the FGSM attack affects the original images to fool the model by generating and adding noise to create adversarial examples. To understand more about how this noise and the analogous adversarial image are created, Bob then scrolls through the rest of the page, which gives him an overview of various kinds of other adversarial attacks as well as a brief introduction to the FGSM attack. The page also details other information about the tool such as the image dataset used to create the tool.

Therefore, after going through this entire process, Bob has gained the following valuable knowledge and insights:

1. what is the FGSM attack, and how it affect Bob's model in terms of robust accuracy;
2. how the FGSM attack affects individual image instances (the actual changes in the pixels and the changes in confidence levels) as well as the overall distribution of the predictions in the entire dataset;
3. how does Bob's model performs when compared to one of the other well-known models from the literature.

6 Discussion

Our contribution is proposing a novel interactive visualization tool that uses instance-level and high-level visualizations to assist machine learning practitioners with little adversarial experience in understanding the underlying logic of the FGSM attack. To our knowledge, this is the first interactive visualization application to demonstrate FGSM attack.

Our visualization is a good starter for machine learning developers and researchers to evaluate the robustness of machine learning models against the FGSM attack. Our visualization prioritizes demonstrating the shift across various perturbation levels and focuses on expressing the underlying logic. This work emphasizes the ease of understanding, which may be overlooked by other researchers.

Due to the computation cost, our tool has several limitations in terms of interactivity and scalability. Currently, it supports only a few perturbation levels and data instances. In addition, test data is projected with t-SNE, a dimensionality reduction technique. One drawback of dimensionality reduction is information loss; it is not a complete reflection of the dataset. Also, t-SNE is computationally costly with regard to large datasets, which limits scalability. One future direction is to apply a better dimensionality reduction technique like LDA.

7 Future Work

As a tool that aims to visualize adversarial attacks, FGSM Explainer has several possibilities for further algorithmic and visualization improvement.

7.1 Adding More Perturbation Size & Data Instances

For computational efficiency, FGSM Explainer allows user to select only three perturbation scales: 0.01, 0.02 and 0.03. Although the trend and effect of adjusting perturbation scales is displayed in our visualization tool, supporting users to select more perturbation scales is one future direction for improvement. This adds one more layer of interaction to the visualization.

In addition, we feel that presenting all 10,000 testing data on screen may be too overwhelming for users to comprehend instance-level changes because the major focus of our design is to visualise how each data input changes before and after the adversarial attack. Therefore, we randomly select 10 images from each class. In total, our visualization has 100 data instances. Another future direction is to add more data instances, so it is easier for users to view the movement of data clusters after the adversarial attack.

7.2 Visualizing More Types of Adversarial Attacks

For this study, we choose to visualize the FGSM attack due to it being one of the most commonly used adversarial attack algorithms. Future work should aim to visualize other adversarial attack algorithms. It is beneficial for end users to have a chance to visualize several adversarial attacks simultaneously and see their effects on the same dataset.

7.3 Supporting More Machine Learning Models

Currently, we select VGG16_BN and VGG19_BN for visualization. From our visualization, users are able to see the robustness of VGG models against the FGSM attack. However, the effectiveness of FGSM depends on model architectures. Some machine learning models are much more robust to the FGSM attack, while others are less robust. It is rewarding to include more machine learning models in the visualization since then users would be able to learn which models are more robust to the FGSM attack, which helps them select models in real work.

7.4 Improving the Encoding of Class Labels

In our current implementation of the data projectors, the prediction class is encoded as outline colours of the circles. This approach may be indistinguishable for the users in certain cases. Although we made efforts to adopt another encoding method [1], substantial changes need to be made to the architecture of the data projectors. Therefore, improving the encoding method on prediction labels is another future direction of our work.

8 Conclusion

In this study, we propose FGSM Explainer, an interactive web-based application for visualizing the FGSM attack on the CIFAR-10 dataset. Our application is intended to assist machine learning practitioners in understanding the logic of the FGSM attack, a gradient-based white-box adversarial attack. Our visualization provides interpretability by demonstrating models' performance in classifying original and adversarial images, and it increases interactivity by allowing users to adjust the perturbation scale. By doing so, we help ML practitioners understand how the predictions of ML models change with respect to varying perturbation levels. Nonetheless, more works remain to further enhance FGSM Explainer's interactivity and scalability.

Acknowledgments

Below are the main contributions of our team members:

- **Qing Guo**: Writing part of the Introduction and Related Work section. Writing Background section, Discussion section, Future Work section and Conclusion section. Implementing robustness analyzer, and assisting with the general information provider. Recording video demo. Proofreading and touching up the final paper.
- **Jarvis Tse**: Writing part of the Introduction and Related Work sections and adding content to the Proposed Design section. Implementing the confidence level component of FGSM Explainer and improving FGSM Explainer's user interface. Proofreading and touching up the final paper. Doing part of the final presentation.
- **Parjanya Vyas**: Writing Case Study section. Setting up the initial project skeleton, implementing the conversion from Numpy data output to images for original and attack images, and loading them into the instance-level attack explainer on the bottom panel. Proofreading and touch ups on the final paper. Doing part of the mid-class presentation.
- **Yuzhe You**: Writing the Proposed Design section. Implementing data preprocessing pipeline, scatterplot functionalities & the frontend animations of instance-level attack explainer. Designing and setting up the UI/UX of the web app. Combining and polishing all components of FGSM Explainer, and implementing additional features. Proofreading and touching up the final paper. Doing part of the mid-class presentation and the final presentation.

References

[1] B. Alper, B. Bach, N. Henry Riche, T. Isenberg, and J.-D. Fekete. Weighted graph comparison techniques for brain connectivity analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, p. 483–492. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2470654.2470724

[2] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10231–10241, 2021.

[3] K. Cao, M. Liu, H. Su, J. Wu, J. Zhu, and S. Liu. Analyzing the noise robustness of deep neural networks. *IEEE transactions on visualization and computer graphics*, 27(7):3289–3304, 2020.

[4] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45, 2021.

[5] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. 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*, pp. 15–26, 2017.

[6] F. Crecchi, D. Bacciu, and B. Biggio. Detecting adversarial examples through nonlinear dimensionality reduction. *arXiv preprint arXiv:1904.13094*, 2019.

[7] N. Das, H. Park, Z. J. Wang, F. Hohman, R. Firstman, E. Rogers, and D. H. P. Chau. Bluff: Interactively deciphering adversarial attacks on deep neural networks. In *2020 IEEE Visualization Conference (VIS)*, pp. 271–275. IEEE, 2020.

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

[9] F. Hohman, H. Park, C. Robinson, and D. H. P. Chau. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE transactions on visualization and computer graphics*, 26(1):1096–1106, 2019.

[10] M. Kahng and D. H. 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.

[11] M. Kahng, N. Thorat, D. H. Chau, F. B. Viégas, and M. Wattenberg. Gan lab: Understanding complex deep generative models using interactive visual experimentation. *IEEE transactions on visualization and computer graphics*, 25(1):310–320, 2018.

[12] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

[13] D. Li, W. Wang, H. Fan, and J. Dong. Exploring adversarial fake images on face manifold. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5789–5798, 2021.

[14] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1778–1787, 2018.

[15] J. Lin and D. Soylu. Advis: Visualizing and attributing ml attacks to adversarial examples.

[16] Y. Ma, T. Xie, J. Li, and R. Maciejewski. Explaining vulnerabilities to adversarial machine learning through visual analytics. *IEEE transactions on visualization and computer graphics*, 26(1):1075–1085, 2019.

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

[18] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

[19] A. Musa, K. Vishi, and B. Rexha. Attack analysis of face recognition authentication systems using fast gradient sign method. *Applied Artificial Intelligence*, pp. 1–15, 2021.

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

[21] R. Pan, M. J. Islam, S. Ahmed, and H. Rajan. Identifying classes susceptible to adversarial attacks. *arXiv preprint arXiv:1905.13284*, 2019.

[22] C. Park, S. Yang, I. Na, S. Chung, S. Shin, B. C. Kwon, D. Park, and J. Choo. Vatun: Visual analytics for testing and understanding convolutional neural networks. 2021.

[23] R. Paul, M. Schabath, R. Gillies, L. Hall, and D. Goldgof. Mitigating adversarial attacks on medical image understanding systems. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1517–1521. IEEE, 2020.

[24] C. R. Qi, H. Su, K. Mo, and L. 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.

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

[26] D. Steinberg and P. Munro. Visualizing representations of adversarially perturbed inputs. *arXiv preprint arXiv:2105.14116*, 2021.

[27] J. Su, D. V. Vargas, and K. Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

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

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

[30] Z. J. Wang, R. Turko, O. Shaikh, H. Park, N. Das, F. Hohman, M. Kahng, and D. H. P. Chau. Cnn explainer: Learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1396–1406, 2020.

[31] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.

[32] Y. Zhang, G. Liang, T. Salem, and N. Jacobs. Defense-pointnet: Protecting pointnet against adversarial attacks. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 5654–5660. IEEE, 2019.