Visual Informatics 8 (2024) 71–84

Contents lists available at ScienceDirect

Visual Informatics

journal homepage: www.elsevier.com/locate/visinf

Research article

ATVis: Understanding and diagnosing adversarial training processes through visual analytics

Fang Zhu [a], Xufei Zhu [a], Xumeng Wang [b,*], Yuxin Ma [a,*], Jieqiong Zhao [c]

[a] Southern University of Science and Technology, Shenzhen, China
[b] DISSec, Nankai University, Tianjin, China
[c] Augusta University, Augusta, Georgia, USA

ARTICLE INFO

Article history:
Received 18 August 2024
Received in revised form 7 October 2024
Accepted 19 October 2024
Available online 24 October 2024

Keywords:
Visual analytics
Explainable AI
Adversarial training

ABSTRACT

Adversarial training has emerged as a major strategy against adversarial perturbations in deep neural networks, which mitigates the issue of exploiting model vulnerabilities to generate incorrect predictions. Despite enhancing robustness, adversarial training often results in a trade-off with standard accuracy on normal data, a phenomenon that remains a contentious issue. In addition, the opaque nature of deep neural network models renders it more difficult to inspect and diagnose how adversarial training processes evolve. This paper introduces ATVis, a visual analytics framework for examining and diagnosing adversarial training processes. Through multi-level visualization design, ATVis enables the examination of model robustness from various granularity, facilitating a detailed understanding of the dynamics in the training epochs. The framework reveals the complex relationship between adversarial robustness and standard accuracy, which further offers insights into the mechanisms that drive the trade-offs observed in adversarial training. The effectiveness of the framework is demonstrated through case studies.

1. Introduction

Deep Neural Networks (DNNs) have achieved great success across various application areas, including image classification (He et al., 2016), object detection (Szegedy et al., 2013a), face recognition (Hahn and Marcel, 2022), and autonomous driving (Li et al., 2021). However, extensive research has demonstrated that DNNs are highly vulnerable to attacks (Singh et al., 2024; Li et al., 2024). By introducing carefully crafted and imperceptible perturbations to the original images (Goodfellow et al., 2015), adversarial examples can be generated to trigger incorrect predictions from models. Such vulnerability poses severe challenges for the deployment of DNNs in safety-critical domains (Shen et al., 2021). To defend against adversarial perturbations and enhance model robustness, adversarial training is widely considered the most efficient and essential defense strategy (Bai et al., 2021). It involves augmenting training data with perturbations to generate adversarial examples that are incorporated into every training iteration, and numerous studies have further optimized and enhanced adversarial training from various perspectives (Zhang et al., 2019; Wang et al., 2019; Balaji et al., 2019; Yang et al., 2020b).

Although adversarial training can effectively enhance model robustness against adversarial attacks, it has been shown to significantly compromise standard accuracies of the models on normal data (Madry et al., 2018), thereby potentially posing a trade-off between adversarial robustness and standard performances. However, such a trade-off still remains a subject of debate and has not reached a consensus. Various works (Tsipras et al., 2019; Rade and Moosavi-Dezfooli, 2022) have studied this trade-off phenomenon and proposed different observations towards the balance between robustness and standard accuracy. In addition, some theoretical and technical works (Zhang et al., 2019; Yang et al., 2020a) also demonstrated the possibility of satisfying both criteria in the same model simultaneously. Consequently, there is an urgent need for more comprehensive methods to analyze and explore the impacts of adversarial training on models in a fine-grained manner.

Besides the tradeoff issue, the complex architecture and opaque nature of DNNs, coupled with high-dimensional data and extensive datasets (Singh et al., 2024), pose significant challenges in understanding and analyzing the underlying mechanisms and impact of adversarial training on models. Specifically, evaluating and dynamically tracking changes in model robustness and standard accuracy during adversarial training is essential for domain

* Corresponding authors.
E-mail addresses: zhuf2022@mail.sustech.edu.cn (F. Zhu), zhuxf2020@mail.sustech.edu.cn (X. Zhu), wangxumeng@nankai.edu.cn (X. Wang), mayx@sustech.edu.cn (Y. Ma), jiezhao@augusta.edu (J. Zhao).

https://doi.org/10.1016/j.visinf.2024.10.003

experts to gain valuable insights. Existing work primarily evaluates model adversarial robustness and predictive performance using robust accuracy (Bai et al., 2021; Croce et al., 2020) on adversarially perturbed test sets and standard accuracy on natural test sets, respectively. These individual statistical metrics only provide a basic evaluation of model robustness, which fails to explain the mechanisms behind enhanced robustness and reduced accuracy due to adversarial training. Furthermore, they do not identify the specific samples or clusters with altered predicted labels, nor do they illustrate the variations in the decision boundary and sample margins during training.

Visualization has been utilized to understand the properties and impacts of adversarial examples through various studies (Cao et al., 2020; Das et al., 2020) from a static perspective, i.e., focusing on the model checkpoint from an individual training epoch. However, it is essential to enable a dynamic analysis of how the entire adversarial training process strengthens model robustness and how adversarial examples are predicted in different stages along such a process. To address the aforementioned challenges and requirements, we propose ATVis, an interactive visual analytics framework to support the comprehension and exploration of adversarial training processes. The framework adopts a multi-level visualization scheme to inspect the training processes from multiple granularities, including model, cluster, and instance levels. By integrating various measuring algorithms on instances' resistance to perturbation attacks, model robustness on different instances and neighboring areas can be revealed. Furthermore, the essential part for explaining the properties of adversarial training is to illustrate changes in decision boundaries and the instances in their vicinity during the training process. With the help of high-dimensional visualization techniques, easily attacked and misclassified instances can be identified, which enables quick pinpointing of critical local areas in the feature space. Additional comparative analysis between model checkpoints from varying epochs is designed to examine the development trends of model performance and robustness.

In summary, our contributions include:

- An interactive visual analytics framework, ATVis, that supports understanding and examination of adversarial training processes;
- A suite of visualization and interaction designs that enables multi-level analysis from model, cluster, and instance levels;
- Case studies on real-world datasets that demonstrate the effectiveness and usability of the framework.

2. Related work

Our work aims to explain and analyze the dynamic processes involved in adversarial training. In this section, we review the related work on adversarial training and visual analytics for model explainability.

2.1. Adversarial training

Adversarial training (Goodfellow et al., 2015; Madry et al., 2018) has proven to be an effective method for enhancing the robustness of deep neural network models through training on adversarial examples (Bai et al., 2021). These examples are generated by applying meticulously crafted and imperceptible perturbations to the original examples, deliberately causing misclassification. Numerous studies have explored adversarial training from various perspectives. One approach includes generating data (Gowal et al., 2021) through data augmentation to supplement the training set, thereby improving model robustness. Wang et al. (2023b) proposed using the latest diffusion models for data generation to enhance adversarial training, achieving state-of-the-art performance. Although incorporating a large amount of generated data into adversarial training can significantly improve model robustness, the surge in computational costs can also lead to inefficiencies in the training process.

In addition to utilizing synthetic data, another major category of methods focuses on improving the adversarial training schemes, including loss function optimization (Zhang et al., 2019; Rade and Moosavi-Dezfooli, 2022; Yan et al., 2024), sample reweighting (Xu et al., 2023; Zhang et al., 2021), and adaptively adjusting the attack strength (Ding et al., 2020). Many of these studies pay special attention to changes in the decision boundary during the adversarial training process. TRADES (Zhang et al., 2019) introduces a penalty regularization term for instances close to the decision boundary, pushing data points away from it to enhance robustness against attacks. Rade and Moosavi-Dezfooli (2022) proposed the HAT algorithm to mitigate the issue of adversarial training causing excessively large decision boundary margins, which in turn harms clean accuracy. GAIRAT (Zhang et al., 2021) assigned greater weights to data points closer to the decision boundary, prioritizing the enhancement of robustness for these vulnerable instances. Xu et al. (2023) observed that the decision boundary might move away from some vulnerable points while approaching others during adversarial training. Hence, they proposed the DyART algorithm to prioritize enhancing smaller decision boundary margins. However, these works analyze from a macro-theoretical perspective and do not clearly demonstrate the specific changes in the decision boundary throughout the adversarial training process. Our paper aims to conduct a more detailed analysis and exploration using a visual analytics methodology.

2.2. Visual analysis approaches for machine learning

In the process of model training, machine learning models learn the data facts in training data by iteratively updating parameters. To interact with the training process, visual analysis approaches can provide humans with three-fold assistance, which consists of model understanding, model diagnosis, and model steering (Yuan et al., 2021).

Model Understanding. Understanding how a model works is necessary for model assessment and model improvement. Visual analysis approaches facilitate such understanding by summarizing model behaviors. To demonstrate the effects of neurons in deep convolutional neural networks, CNNVis (Liu et al., 2017) clusters neurons according to both the position in the network and the activation situations. Taking advantage of neuron clusters, humans can quickly inspect a CNN model that includes hundreds of layers. TensorFlow Graph Visualizer (Wongsuphasawat et al., 2018) provides humans with an overview of convolutional networks by describing data flows between group operations. Seeking an intuitive explanation, the What-If tool (Wexler et al., 2020) allows users to figure out model behaviors based on specified instances.

Model Diagnosis. To identify performance issues, there is a need to monitor the training process and diagnose the model performance. The evolution of performance indicators, like accuracy and loss, demonstrates a high-level summary of the training process, which could be visualized by line charts (Wang et al., 2023a, 2018; Liu et al., 2019). The fluctuation reflected by line charts can be used to detect the training stage when training issues have occurred. To further diagnose the issue, humans can review the training stage and examine model updates. For example, ConceptExplorer (Wang et al., 2020) correlates performance issues with the data records that participated in certain rounds of online learning. AEVis (Cao et al., 2020) compares the data-paths of a normal record for model training and an adversarial

one to explain how a neural network is fooled by adversarial examples. Ma et al. (2020) also correlate the performance issue with information, including data features and model structures, to explain how a poisoning attack affects the model.

In conclusion, the visual analytics community has proposed various approaches to enhance comprehension, exploration, and diagnosis of complex learning models, especially those that tackle the problems of diagnosing the impact of adversarial attacks (Cao et al., 2020; Ma et al., 2020). Yet, there is an urgent need for taking adversarial training processes into consideration, where the dynamics inside the training epochs reflect complex behaviors of how certain instances can be vulnerable to crafted perturbations as well as the general characteristics of model robustness.

3. Design overview

In this section, we outline the foundational framework of adversarial training, which serves as the primary focus of our exploration and analysis. Through a comprehensive literature review and expert interviews, we have identified key user requirements and analytical tasks to guide the development of our visual analysis framework as shown in the middle of Fig. 1.

3.1. Preliminaries

Given a training dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$ where x_i and y_i correspond to the samples and their respective true labels, to enhance the robustness of a classifier f_θ parameterized by θ, adversarial training can be mathematically formulated as a min–max optimization problem (Madry et al., 2018). The objective function can be written as follows:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{\delta \in \Delta} \ell(f_\theta(x_i + \delta), y_i) \tag{1}$$

where n is the number of training samples, δ is the perturbation added to the original image x_i, and ℓ is the loss function, specifically cross-entropy loss in the context of multi-class classification. Note that Δ is the ϵ-ball defined as $\Delta = \{\|\delta\|_p \leq \epsilon\}$, where p is the norm, which can take values of 1, 2, or ∞, and ϵ indicates the maximum allowable perturbation for the adversarial attack. The inner maximization aims to find the adversarial example within the ϵ-ball that maximizes the loss, representing the most challenging aspect of the process. Based on these adversarial examples, the outer minimization then optimizes the model parameters. The most widely used method for generating adversarial examples to address the inner maximization problem is the Projected Gradient Descent (PGD) (Madry et al., 2018) attack:

$$x_i^{t+1} = \text{Proj}_{x_i+\Delta}\left(x_i^t + \alpha \,\text{sign}\left(\nabla_{x_i^t} \ell(f_\theta(x_i^t), y_i)\right)\right) \tag{2}$$

where t is the current iteration of the attack, α is the step size, and the Proj function ensures that the generated adversarial examples stay within the maximum perturbation range. Subsequent extensive research has demonstrated that adversarial training based on the PGD attack (PGD-AT) can significantly enhance the adversarial robustness of models. As a result, PGD-AT has become the standard method for adversarial training and serves as an important benchmark in the field. However, research has shown that adversarial training can degrade a model's standard accuracy on natural datasets. To balance this trade-off, the most common approach is to introduce a regularization term that penalizes the misclassification of natural samples during training:

$$\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \max_{\delta \in \Delta} \ell(f_\theta(x_i + \delta), y_i) + \lambda R(\ell(f_\theta(x_i), y_i)) \right\} \tag{3}$$

In this formulation, λ is the regularization coefficient that controls the strength of the regularization term, and $R(\ell(f_\theta(x_i), y_i))$ is the regularization function that penalizes the model's tendency to misclassify natural samples. Building on this foundational framework, researchers have developed surrogate loss functions from multiple perspectives, resulting in various variants (Zhang et al., 2019, 2021; Xu et al., 2023). These efforts have substantially advanced the understanding of the impact of adversarial training on model characteristics.

3.2. Requirement analysis

Given the fundamental framework and theoretical basis of adversarial training, we aim to design a visual analysis framework that comprehensively evaluates model robustness and explains adversarial training from the perspective of decision boundaries. To refine our research focus, we conducted a literature review (Bai et al., 2021; Yang et al., 2020a; Singh et al., 2024; Zhang et al., 2019) to preliminarily identify potential research gaps. Subsequently, we carried out semi-structured interviews with two experts, one specializing in the theoretical aspects of adversarial examples and the other focusing on adversarial training methodologies.

During the interviews, we engaged in detailed conversations about the current challenges and gaps in this research domain, which highlighted several critical issues. The experts emphasized that existing evaluation methods, which primarily rely on robust accuracy as a single metric, are insufficient for a comprehensive assessment of model robustness. The experts advocated for a multi-faceted evaluation approach to better understand model robustness against adversarial attacks and the perturbation resistance of data samples, as well as to observe the impact of adversarial training on standard accuracy with natural datasets. Additionally, the experts pointed out that current research on decision boundary changes and instance behavior during adversarial training involves limited samples and does not prioritize critical ones, lacking intuitive visual representation. Furthermore, the experts expressed the need for a multi-level visualization framework to compare original and adversarial training models, providing clearer insights into the effects and effectiveness of adversarial training. These discussions have guided the development of our visual analysis framework to address these identified analysis requirements, summarized as follows:

- **R1: Multi-faceted analysis of adversarial training.** A thorough understanding of adversarial training necessitates analyzing multiple facets (Carlini et al., 2019): overall model performance, sample distribution, and individual instance behavior. Quantitative metrics are frequently used by experts to assess and compare models. Beyond these metrics, examining sample distribution helps identify significant clusters, while analyzing changes in individual instances provides insights into how adversarial training enhances robustness. Additionally, experts emphasize the importance of decision boundaries in explaining the mechanisms and effectiveness of adversarial training (Xu et al., 2023; Zhang et al., 2019; Rade and Moosavi-Dezfooli, 2022).
- **R2: Comprehensive measures for adversarial robustness.** Existing research on evaluating model adversarial robustness relies heavily on robust accuracy (Croce et al., 2020), a single metric that indicates robustness strength with a simple number while fails to explain why or how it improves. To address this, we need more comprehensive metrics (Carlini et al., 2019) that assess overall perturbation resistance and focus on the vulnerability of individual instances, such as their distance to the decision boundary (Zhang et al., 2021).

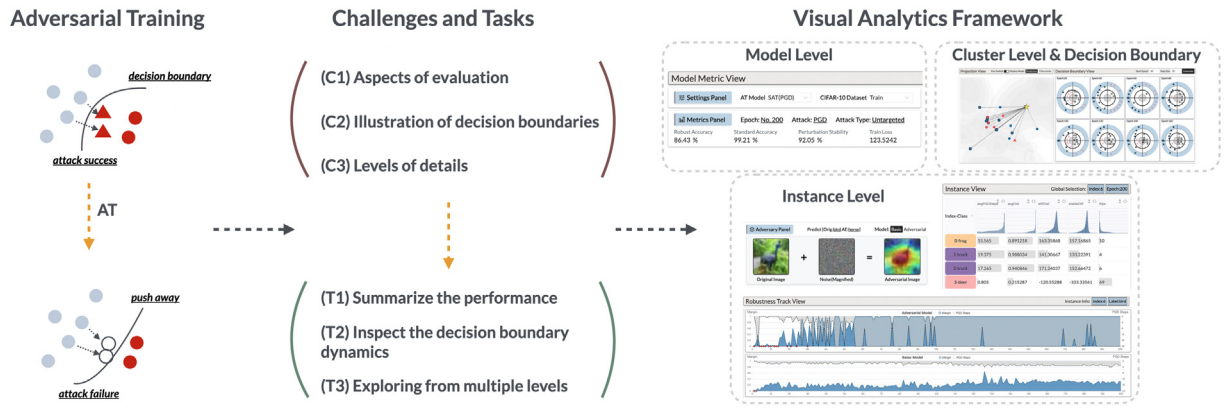

Fig. 1. An overview of the design and implementation workflow for our visual analytics framework. To explore the impact of adversarial training on models and instances, three key challenges (C1–C3) are identified. Corresponding analytical tasks (T1–T3) are then developed, followed by the design of a visual analytics framework to support multi-level analysis across the model, cluster, and instance stages.

Additionally, understanding the trade-off phenomenon and identifying instances with label flips during adversarial training are crucial, as these important instances ultimately impact the model's overall performance.

- **R3: Comparative analysis of adversarial training.** A critical aspect of adversarial research is the comparative analysis between normally trained models and adversarially trained models (Bai et al., 2021), as well as the comparison of models across different training loops within the adversarial training process. Visualization methods such as juxtaposition and superimposition (Gleicher, 2017) effectively highlight these differences, providing experts with clear visual insights. These visualizations facilitate the identification of key factors and patterns that enhance understanding and inspire further research in adversarial robustness.

3.3. Analytical tasks

Based on the previously mentioned requirements, we have outlined the following analytical tasks:

T1: Systematically summarize model robustness evaluation measures. Quantitative metrics are the most common method for experts to assess model performance. By incorporating multiple evaluation perspectives, experts can gain a more comprehensive and in-depth understanding:

- *How do significant clusters in the sample distribution and the robustness of individual instances against adversarial attacks affect overall model robustness?*(**R1, R2**)
- *How consistent is the model's robustness across different evaluation metrics?* (**R2**)

T2: Dynamically investigate changes in models and instances during adversarial training. Tracking the dynamic effects of adversarial training on models and instances is crucial for understanding the nature of adversarial training and the impact of instance behavior changes on model robustness:

- *How does the decision boundary of models evolve, and how do the labels and distances of instances near the decision boundary change over time?* (**R1, R3**)
- *How do different model metrics change during adversarial training, and is there evidence of a trade-off phenomenon at specific stages?* (**R1**)

T3: Inspect differences between models. Multi-level comparative analysis between models before and after adversarial training is important for experts to explore:

- *How do the performance and robustness metrics differ between the naturally trained model and the adversarially trained model?* (**R2, R3**)
- *How do models from temporally adjacent epochs during the adversarial training process change?* (**R3**)

4. Visual analytics framework

Built upon the requirements and analytical tasks, we propose a visual analytics framework, ATVis, designed to explore the dynamic process in adversarial training, as illustrated in Fig. 1. The framework comprises three essential components:

Overview of Model Performance and Instance Metrics. This component serves as the entry point for visual exploration, which combines familiar model performance metrics for users with comprehensive instance metrics. This setup allows domain experts to quickly gain an overview of the adversarial training process. By observing trend changes and sorting instance metrics, experts can rapidly identify critical nodes and key instances in the training process, facilitating more detailed analysis.

Hierarchical Visual Exploration of Adversarial Training. The **model metric view** mentioned in the previous component serves as the highest level in the multi-layered framework, offering a comprehensive overview, as seen in Fig. 2 (A). Subsequently, the **projection view** elucidates the distribution of samples at the cluster level, aiding analysts in identifying significant areas for more in-depth exploration, as illustrated in Fig. 2 (D). The analysis then narrows to the instance level, depicting changes in individual instance robustness against perturbations depicted in Fig. 2 (B).

Investigation of Decision Boundary Dynamics. After pinpointing areas or instances of interest, the **decision boundary view** dynamically illustrates the decision boundary and the behavioral changes of instances within its vicinity, as shown in Fig. 2 (E). This fine-grained perspective is essential for exploring the impact of adversarial training on the decision boundary, thereby enhancing the understanding of the underlying mechanisms of adversarial training.

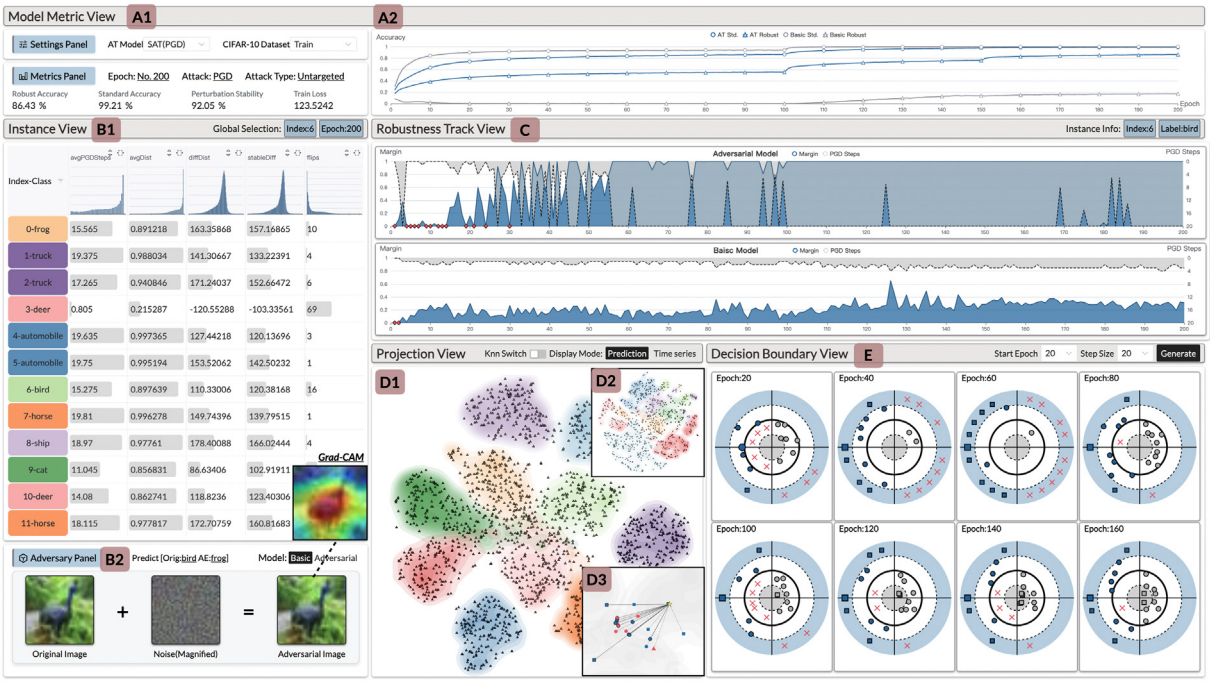

Fig. 2. The ATVis interface comprises five key components. (A1) Quantitative metrics and (A2) the accuracy line chart provide an overview of model performance. (B1) The instance table view and (B2) adversarial attack demonstration offer detailed instance-level insights. (C) Area charts track robustness during training. (D1, D2, D3) t-SNE projections in the embedding space reveal distribution patterns, and (E) the decision boundary view allows dynamic inspection of boundary changes.

4.1. Model performance and instance robustness

Quantitative metrics of the model offer an overview of the model's overall performance. Additionally, evaluating the robustness of individual instances is essential for a finer-grained perspective. This approach allows for the identification of specific instance changes during adversarial training that contribute to overall model performance variations, thereby uncovering the underlying mechanisms of adversarial training. The integration of both metrics provides a comprehensive and nuanced understanding of adversarial training (T1).

Model Performance Metrics. According to Robustbench Croce et al. (2020), a benchmarking platform for evaluating the robustness of machine learning models against adversarial attacks and adhering to the principles proposed by Carlini et al. (2019), four key metrics have been identified for assessing model performance. These metrics are categorized into two types: those estimating adversarial robustness and predictive accuracy via accuracy on specific test sets and those focusing on the robustness of the model against adversarial attacks.

Standard Accuracy and Robust Accuracy: Standard accuracy refers to the predictive accuracy of a model on a natural test set, while robust accuracy denotes the accuracy on a test set subjected to adversarial attacks. In this study, the adversarial attacks were conducted on white-box models with an untargeted PGD attack, using parameters set to 10 PGD steps and an attack size of $\frac{8}{255}$ under ℓ_∞. These settings, chosen based on common practices in adversarial robustness research (Bai et al., 2021; Rade and Moosavi-Dezfooli, 2022; Croce et al., 2020; Xu et al., 2023), balance the trade-off between attack strength and computational feasibility. Furthermore, prior studies have demonstrated their effectiveness in evaluating model robustness under ℓ_∞ constraints, enabling meaningful comparisons across different model architectures. Consequently, both standard and robust accuracy are crucial for evaluating model performance, providing comprehensive insights into robustness and predictive accuracy. Changes in these metrics over time clearly illustrate the impact of adversarial training on model performance.

Perturbation Stability and Flip Probability: Unlike the previous metrics, these two metrics focus solely on the model's robustness against attacks, serving as a supplementary assessment of robustness. Perturbation stability measures the ratio of stable instances, where the predictions remain unchanged, to the total sample set. It evaluates the model's ability to maintain consistent prediction labels in the presence of perturbations. Similarly, Flip probability is the complement of perturbation stability, representing the proportion of samples where the prediction changes due to perturbations.

Instance Robustness Measures. Understanding how adversarial training enhances model robustness and affects standard accuracy on natural datasets requires a detailed examination of individual instances' robustness to adversarial attacks and the specific changes occurring during the training process. Domain experts emphasize the importance of the distance between data points and the decision boundary, as instances susceptible to attacks are closer to the decision boundary. Given the complexity of DNN decision boundaries, accurately calculating these distances is challenging and computationally expensive. To address this, researchers have proposed two methods to approximate the distance of data points to the decision boundary, as illustrated in Fig. 3.

Identifying the closest boundary point. Given a data point x_i, the closest boundary point $\hat{x}_i$ is defined as the point on the decision boundary nearest to x_i:

$$\hat{x}_i = \arg\min_{\hat{x}} \|\hat{x} - x_i\|_p \quad \text{s.t.} \quad g(\hat{x}) = 0. \tag{4}$$

where the constraint $g(\hat{x}) = 0$ ensures that the point $\hat{x}_i$ lies on the decision boundary. To efficiently calculate this closest boundary point, Xu et al. (2023) employed the Fast Adaptive Boundary Attack (FAB) (Croce and Hein, 2020). The core principle of the FAB algorithm is to iteratively move in the direction of the normal to

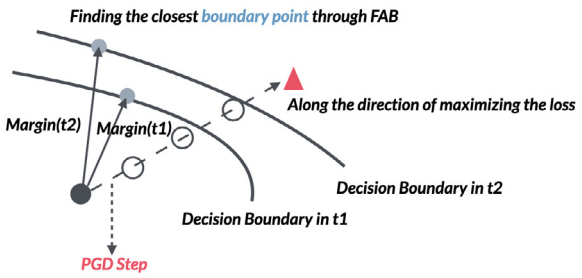

Fig. 3. The illustration of the number of PGD steps and margin for evaluating instance robustness. The figure shows how the number of PGD steps and margin measure instance robustness by illustrating movement toward decision boundaries over time.

the local linear decision boundary until it crosses the boundary shown in Fig. 3. Through experiments, they demonstrated that FAB can reliably find the closest boundary points with sufficient iterations. Consequently, the distance between the closest boundary point $\hat{x}_i$ and the data point x_i is used to approximate the distance of data points to the decision boundary.

Approximating geometric distance by the number of PGD steps. Studies (Zhang et al., 2021; Xu et al., 2023; Rade and Moosavi-Dezfooli, 2022; Ding et al., 2020) indicate that data points vulnerable to attacks tend to have smaller geometric distances from the decision boundary. To approximate this geometric distance, researchers have proposed using the number of PGD steps, which is defined as the minimum number of PGD iterations required to successfully cause the model to misclassify a data point, shown in Fig. 3. This approach provides a direct measure of an instance's susceptibility to attacks. Consequently, a higher number of PGD steps suggests that an instance is more stable and resistant to adversarial perturbations, making it a valuable metric for assessing instance robustness.

4.2. Investigating the process of adversarial training

Understanding the adversarial training process is essential for grasping its underlying mechanisms and their effects on overall model performance and instance-level robustness. In this module, we introduce a multi-level, fine-grained visualization approach to explore adversarial training across the model, cluster, and instance Levels, providing a comprehensive view of the process (T1, T2).

Model Metric View. The model metric view, Fig. 2 (A), is designed to provide a concise overview of model parameters and key quantitative metrics, directly highlighting the indicators most relevant to domain experts and enabling a quick assessment of the overall performance. It contains three main components: the setting panel, the metrics panel in Fig. 2 (A1), and the model accuracy line chart, Fig. 2 (A2). The setting panel allows users to select the adversarial training method, as well as the choice to conduct experiments on either the training or test set. Once the parameters are selected, the metrics panel displays the specific experimental settings and the performance metrics of the selected model. The accuracy metric, particularly important to experts, is visualized in the model accuracy line chart in Fig. 2 (A2), which plots accuracy over the adversarial training epochs. The chart features four lines representing the standard accuracy and robust accuracy for both naturally trained and adversarially trained models, differentiated by colors and symbols. This visualization allows users to quickly identify significant changes across epochs. Additionally, the line chart supports filter and click operations, enabling users to select epochs of interest, with real-time updates reflected in the metrics panel.

Projection View. Focusing only on metric changes misses how adversarial training impacts the internal mechanisms of models, while examining instances and clusters reveals deeper insights into how it handles different sample types. The projection view, illustrated in Fig. 2 (D), is intended to visualize the distribution of all dataset samples within the embedding space, enabling users to identify key clusters and specific instances. It supports two modes: prediction mode and time series clustering mode.

Prediction Mode: In prediction mode, as shown in Fig. 2 (D1), the input to the final layer of the model, which represents the feature embeddings extracted by the model, is used for t-SNE (Van der Maaten and Hinton, 2008) projection. This projection preserves the relative positions of instances in the embedding space when mapped to a two-dimensional plane, helping to reveal clustering structures and distribution patterns in a lower-dimensional space.

To avoid visual clutter, a density map combined with a contour map is first employed to reduce information overload, which visualizes the data distribution density through color gradients and contour lines. Important instances are then highlighted on the constructed contour map. For visual encoding, the contour map's categorical color scheme is chosen from the ColorBrewer (Harrower and Brewer, 2003) tool, tailored to match the classes of the dataset. Triangles are used to encode instances that are particularly vulnerable to attacks, with the color of the triangles determined by the predicted labels.

Time Series Clustering Mode: Similar to prediction mode, this mode also employs t-SNE for projecting all data points, Fig. 2 (D2). However, the projection is based on time series data representing the margin, which is the minimum distance of each instance to the decision boundary across training epochs. This time series data is processed to extract relevant features, followed by clustering to organize the data effectively. The parameters for these processes are adjustable, allowing users to refine and discover the most suitable clustering projection outcomes. This approach enables efficient reduction of information complexity, helping users quickly identify instances and clusters that warrant further investigation.

Both modes of the projection view support interactive features. When hovering over a data point, the point is highlighted, and a tooltip displays relevant information and metrics. Clicking on a data point triggers the construction of a k-nearest neighbors graph, showing the distribution of neighboring data points. Additionally, the selected point's index is synchronized across other views, prompting updates in those views and enabling coordinated interaction throughout the framework. Users can also control the data shown in the projection view by selecting an epoch in the accuracy line chart, allowing them to dynamically track changes in the projection distribution throughout the training process.

Instance View. As shown in Fig. 2 (B1), the instance view is designed to offer a fine-grained perspective for exploring the detailed changes in the robustness of individual instances during adversarial training, focusing on the impact of adversarial training from the input sample's standpoint. This view consists of two main sections.

Top Panel: A scrolling table presents the robustness metrics for all instances in the dataset, Fig. 2 (B1). At the top of the table, comprehensive metrics assess the robustness of each instance, including commonly used indicators such as average PGD steps, the average distance to the decision boundary, the difference in distance to the decision boundary before and after adversarial

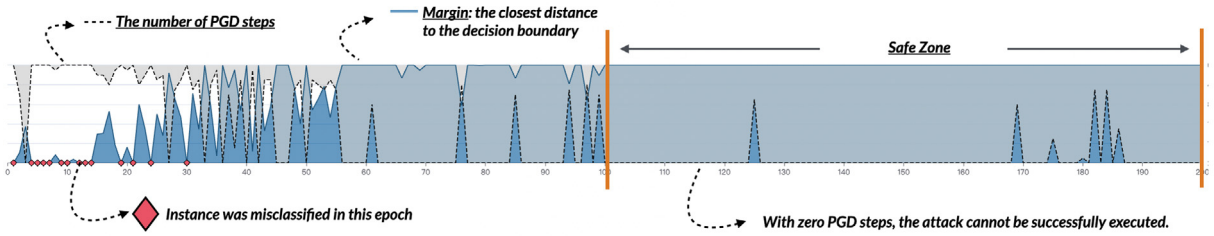

Fig. 4. The visual design of the robustness track view. This figure shows the evolution of PGD steps and margin over time, with the peaks representing PGD steps and the blue area indicating the margin to the decision boundary. Red diamonds highlight misclassification events, while the "Safe Zone" marks periods where attacks fail due to zero PGD steps.

training, and the number of label flips to evaluate stability under adversarial attacks. The table headers feature histograms that display the distribution of all data samples across these metrics. Each row in the table represents an instance, with the leftmost column showing its index and true label. The background color of the table cells corresponds to the label color, and the values of the attributes are visualized using horizontal bar charts. This encoding makes it easy to compare the attributes of different instances, reducing cognitive load.

Bottom Panel: A demonstration panel illustrates how adversarial attacks generate adversarial samples by adding imperceptible noise to the original image, leading to misclassification by the model, shown in Fig. 2 (B2). At the top, the predicted labels for both the original and adversarial samples are displayed to confirm the success of the attack. The leftmost part shows the original image of the selected instance, the middle displays a magnified noise image processed according to the method by Szegedy et al. (2013b), where RGB values are shifted by 128, clamped, and then magnified, and the rightmost part shows the adversarial image generated by adding the noise. Users can switch the image display to Grad-CAM (Selvaraju et al., 2017) by clicking, which uses the gradients from the last convolutional layer to generate feature maps, combined with heatmap visualization to show the areas the model focuses on during prediction, thereby enhancing interpretability.

Interactions: As shown in Fig. 2 (B1), the histograms in the table header support zooming, brush filtering, and sorting, allowing users to quickly find notable instances based on the attributes they are interested in. Each row in the table supports hover highlighting and instance selection, with the corresponding original and adversarial images immediately displayed in the lower section. This selection also triggers linked updates in other views. The demonstration panel below the instance view allows switching between the naturally trained model and the adversarially trained model, enabling users to observe the improvements in an instance's resistance to adversarial attacks due to adversarial training.

4.3. Examining model discrepancies

The comparison module is designed to rigorously analyze the impact of adversarial training by directly contrasting models from both before and after the training process (T3). Specifically, the module utilizes juxtaposed and overlapping visualizations, including the robustness track view, to underscore the distinctions between models prior to and following adversarial training, as well as to examine the variations between models across successive epochs during the adversarial training process. Through this approach, a comprehensive analysis is facilitated, allowing for a deeper understanding of how adversarial training influences model performance over time.

Visualization. The accuracy line chart mentioned earlier compares model performance by simultaneously displaying the accuracy curves of both naturally trained and adversarially trained models. Additionally, the projection view illustrates differences in sample distribution across adversarial training epochs through t-SNE projections for each epoch.

Moreover, the robustness track view, as shown in Fig. 2 (C), offers a nuanced tool for in-depth analysis of how different models influence the robustness of individual instances. This view juxtaposes robustness metrics for selected instances under both adversarial and baseline models, tracking these metrics across training epochs. The horizontal axis is aligned with the accuracy line chart to ensure consistency, facilitating simultaneous observation of model and instance characteristics at specific epochs. Within the timeline module of each model, an area chart superimposes two key robustness metrics: the number of PGD steps (gray, dashed line) and the margin (blue, solid line). This combined visualization enhances the analysis of the relationship between these robustness indicators. In terms of visual encoding, as depicted in Fig. 4, a blue area with a value of 1 denotes a safe zone, indicating that the instance is impervious to adversarial attacks within this range. Additionally, epochs where the model misclassifies the instance are marked with a red diamond, drawing particular attention to these critical points for domain experts.

Interaction. Users can zoom in to focus on specific epochs, allowing for a detailed examination of finer changes during those periods. Hovering over the timeline reveals relevant information about the instance for the current epoch, providing immediate insights. Additionally, by clicking on the corresponding symbol in the legend, users can hide the area chart for specific metrics, enabling them to focus exclusively on the trend of a single indicator. These interactions collectively facilitate a more targeted and in-depth exploration of the data.

Alternative Design. The initial design separated the two metrics, the number of PGD steps and the margin, but this approach required a substantial amount of space, leading to lower information density. Additionally, the early design used vertical solid lines to represent robust instances, while misclassified instances were marked with the color corresponding to their misclassification label. However, this visual encoding resulted in an overly complex and visually overwhelming design, which was eventually discarded. To simplify the design and improve information density, a more concise approach was adopted by integrating both metrics into a single area plot. This approach enhanced the efficiency of information presentation while ensuring clarity and simplicity in the visualization.

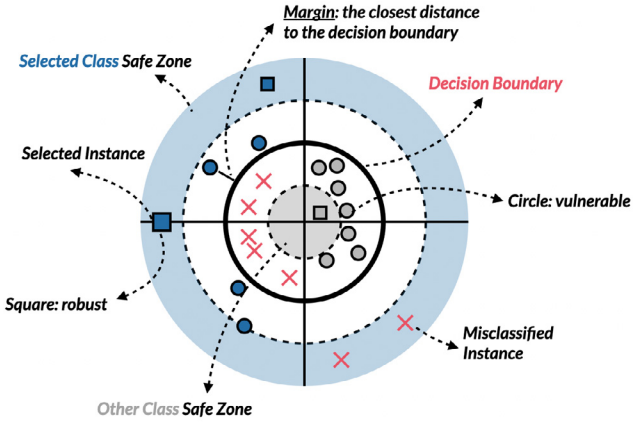

Fig. 5. The example of the glyph for the decision boundary view. The glyph illustrates instance positions relative to the decision boundary, with squares representing robust instances and circles indicating vulnerable ones. Blue and gray areas denote the safe zones for the selected and other classes, while red crosses mark misclassified instances and the margin shows the closest distance to the boundary.

4.4. Inspecting decision boundary dynamics

After conducting a comparative analysis of the model and instance metrics and tracking their robustness changes, it is equally important to examine how predictions of instances near the decision boundaries change. To address this, the decision boundary view is specifically designed to visualize the dynamic changes of the decision boundary and its neighboring regions within the original input space, Fig. 2 (E). This view provides a focused examination of how the decision boundary evolves over time and how points in its vicinity interact, offering deeper insights into the decision-making process of models.

Visualizing KNN Graph in Embedding Space. As illustrated in Fig. 2 (D3), when a user selects a specific instance, the projection view adjusts by changing the contour map's colors to gray, thereby highlighting the k-nearest-neighbor (*k*NN) graph centered on the selected instance. This *k*NN graph reveals the relationships among points within the embedding space. The graph is constructed by identifying the 10 nearest neighbors of the same class and 10 of different classes based on Euclidean distance in the original input space. The *k*NN graph is then superimposed on the t-SNE projection, providing a clear visualization of the relative spatial positioning of instances after feature extraction by the model.

Visual Encoding: In the *k*NN graph, the shape and color of the points are determined by the instance's proximity to the decision boundary at the current epoch and the robustness of the instance. The color scheme employs a context-sensitive encoding: blue represents a safe state, while red indicates a potentially vulnerable or dangerous state. The shape of the points further conveys robustness and classification status—squares denote robust and stable instances, circles indicate correctly classified instances that may be susceptible to attacks, and triangles represent misclassified instances. The combination of color and shape encodes four possible states: robust and distant from the decision boundary, near the decision boundary, close to or crossing the decision boundary leading to misclassification, and safely distant from the boundary. Additionally, the edge and line styles differentiate between points of the same class (no edge, solid line) and those of different classes (with edge, dashed line), as shown in Fig. 2 (D3).

Visualizing the Decision Boundary in Input Space. Given that most adversarial attacks involve adding noise perturbations directly to images, visualizing the decision boundary and the variations in the distance of instances to this boundary within the original input space becomes crucial. The decision boundary view is meticulously designed to capture these dynamic changes, focusing on how the proximity of instances to the decision boundary evolves, thereby revealing the interplay among neighboring data points. This visualization aims to uncover the underlying mechanisms by which these changes impact both model robustness and prediction accuracy (T2, T3).

As illustrated in Fig. 2 (E), the view employs a small multiples visualization approach, with each module modeled after a target-like diagram in Fig. 5. The outermost light blue ring represents the safe zone for the selected instance's class, indicating that instances within this region are sufficiently robust to resist adversarial attacks. Moving inward, the innermost gray circle signifies the safe zones for all other classes, establishing a comparative baseline. The thick middle ring marks the decision boundary, where instances on either side are predicted with differing labels by the model. To enhance clarity, a dashed line delineates the boundary between the safe and danger zones, visually separating these critical regions.

Within this structure, instances that lie in the safe zone are depicted by square symbols, colored according to their class, underscoring their robustness. Misclassified samples are prominently highlighted with a red cross, drawing attention to their critical status. Meanwhile, instances located near the decision boundary are represented by circles, with their distance to the boundary encoded by their proximity to the thick ring. This encoding not only facilitates a clearer understanding of instance distribution but also enhances the identification of points at risk of misclassification. Further enhancing the visualization, instances belonging to the same class as the selected instance are uniformly distributed on the left side, while those from other classes are positioned on the right. The selected instance itself is centrally placed on the left, emphasized by an increased size, ensuring that it remains the focal point of analysis. This structured approach enables analysts to closely monitor the dynamic evolution of the selected instance's *k*NN graph throughout adversarial training, thereby revealing critical interactions between instances.

Interactions: The interface allows users to select the starting epoch and the interval between epochs, providing the flexibility to customize keyframes for investigating the decision boundary. When hovering over a point, the interface dynamically displays detailed information and specific metrics related to that instance. Additionally, clicking on a point not only highlights it within the decision boundary view but also synchronizes updates across all other views, ensuring that the selected instance is consistently represented throughout the entire analysis platform.

Alternative Design: In our exploration of visualization techniques, we initially considered representing the decision boundary as a straight line, shown in Fig. 6, following approaches similar to those proposed by Ma et al. (2020) and Delaforge et al. (2022). However, we ultimately chose to use a circular representation. This decision was informed by the understanding that decision boundaries in deep learning models are typically complex and nonlinear. A straight line, while simple, inadequately captures these complexities and is more suited to linear relationships. Conversely, a circular boundary provides a more nuanced and accurate representation of the intricate decision boundaries characteristic of deep learning models. Additionally, in a two-dimensional space, circular boundaries intuitively illustrate

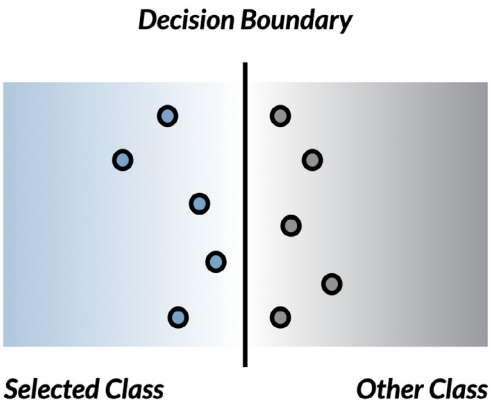

Fig. 6. Linear decision boundary visualization. This diagram presents an alternative design for visualizing a linear decision boundary, separating instances from the selected class (left) and the other class (right).

which sample points are encompassed within the decision region. This method enhances the interpretability of the model's decision-making process, making the underlying mechanisms more transparent and accessible to users.

5. Evaluation

In this section, we demonstrate the effectiveness of our proposed visual analytics framework through case studies conducted with the same two experts (E1 and E2) involved in the requirement analysis. These case studies illustrate how the framework assists domain experts in exploring and analyzing the impact of adversarial training on both model and instance robustness. We also conducted expert interviews to gather feedback on the usability and practical impact of the framework in their respective workflows.

5.1. Case 1: Trade-off between robustness and accuracy

The trade-off between accuracy and robustness is a phenomenon that has sparked intense debate among domain experts. Numerous studies have attempted to explain this phenomenon from various perspectives, yet no definitive conclusion has been reached. In this study, we leveraged our proposed framework to delve deeper into this issue.

To begin, the expert E1 selected the widely recognized and foundational SAT(PGD) as the adversarial training method to train the model. Upon examining the line chart, shown in Fig. 7 (A), E1 observed that, for a model not subjected to adversarial training, the robust accuracy was nearly zero, indicating high susceptibility to attacks. However, after applying adversarial training, robustness significantly improved. Yet, by comparing the curves and hovering over specific epochs to display detailed information, E1 noticed a decrease in accuracy on the natural dataset, as depicted in Fig. 7 (A). This observation empirically demonstrated that while adversarial training enhances robustness, it also compromises standard accuracy. This trade-off was particularly evident when E1 switched to the test set.

E1 then noted a substantial spike in the accuracy curve at epoch 100 on the line chart, as illustrated in Fig. 7 (A). Being curious about this anomaly, E1 selected epoch 100 and examined the relevant metrics through the metrics panel. The expert discovered that, at this stage, the robust accuracy and perturbation stability were relatively low, indicating room for improvement. Concurrently, E1 explored the t-SNE projection for this epoch, revealing that the separation between different classes was still unclear. By zooming into a disordered region in Fig. 7 (B.1) and focusing on a specific area near the decision boundary in Fig. 7 (B.2), E1 identified a zone where samples from various classes were intermingled. The expert then selected an instance with an index of 1836 as the focus of further investigation, Fig. 7 (B.3).

Upon selecting the instance, illustrated in Fig. 7 (C), displayed the corresponding robustness-related metrics. Analyzing the metrics such as label flipping frequency, average distance to the decision boundary, and the number of PGD steps, E1 determined that the selected point was relatively robust and stable throughout the adversarial training process. However, the adversary panel revealed that, at this stage, both the naturally trained model and the adversarially trained model failed to withstand the attack. The true label of the sample was "frog", but after being attacked, the basic model and adversarial model predicted the labels as "horse" and "cat", respectively. This indicated that the sample was still highly vulnerable to attacks at the current epoch.

E1 then examined the robustness track view, which showed the instance's distance from the decision boundary over time as an area chart, Fig. 7 (D). It was observed that, at epoch 100, there was a fluctuation indicating the instance was moving away from the decision boundary, and after epoch 119, the instance maintained sufficient robustness. By vertically comparing the basic model and the adversarial model, it was evident that adversarial training enhanced robustness by pushing the instance away from the decision boundary, thereby improving its ability to resist PGD attacks. Moreover, upon reviewing more instances in the robustness track view, E1 found that, as adversarial training progressed, the distance of most instances from the decision boundary steadily increased until they eventually exceeded the maximum range of perturbation attacks, thus achieving absolute robustness under the current attack strength. This was the fundamental reason for the continuous improvement in robust accuracy.

The expert then further explored the mutual influence among data points within the selected instance's neighborhood. Initially, by examining the *k*NN graph mode in Fig. 7 (B1), E1 observed that most same-class points at the current epoch were sufficiently robust or far from the decision boundary. However, the nearest neighbors of different classes, being close to the decision boundary, exhibited non-uniform state distributions, with some moving away from and others moving closer to the decision boundary. By setting parameters to review keyframes from epoch 100 to 103 in Fig. 7 (E), E1 identified that the selected instance first approached the decision boundary, then quickly moved away, finally settling into a safe zone where it was not vulnerable to attacks. The surrounding instances generally exhibited a trend of moving away from the decision boundary.

Notably, there was an instance with an index of 41120 that continually approached the decision boundary during epochs 100 to 103, ultimately crossing it, causing the model to misclassify the unperturbed sample. Further examination of this instance's robustness track view revealed that, during adversarial training, the sample oscillated back and forth across the decision boundary, a phenomenon not observed during normal training. Coupling this observation with previous information, E1 reasonably speculated that the movement of the decision boundary was influenced by other samples in the neighborhood pushing it away, thereby causing it to encroach upon the current instance, leading to such fluctuations. The expert believed that this phenomenon could explain why adversarial training tends to degrade the standard accuracy on the original dataset.

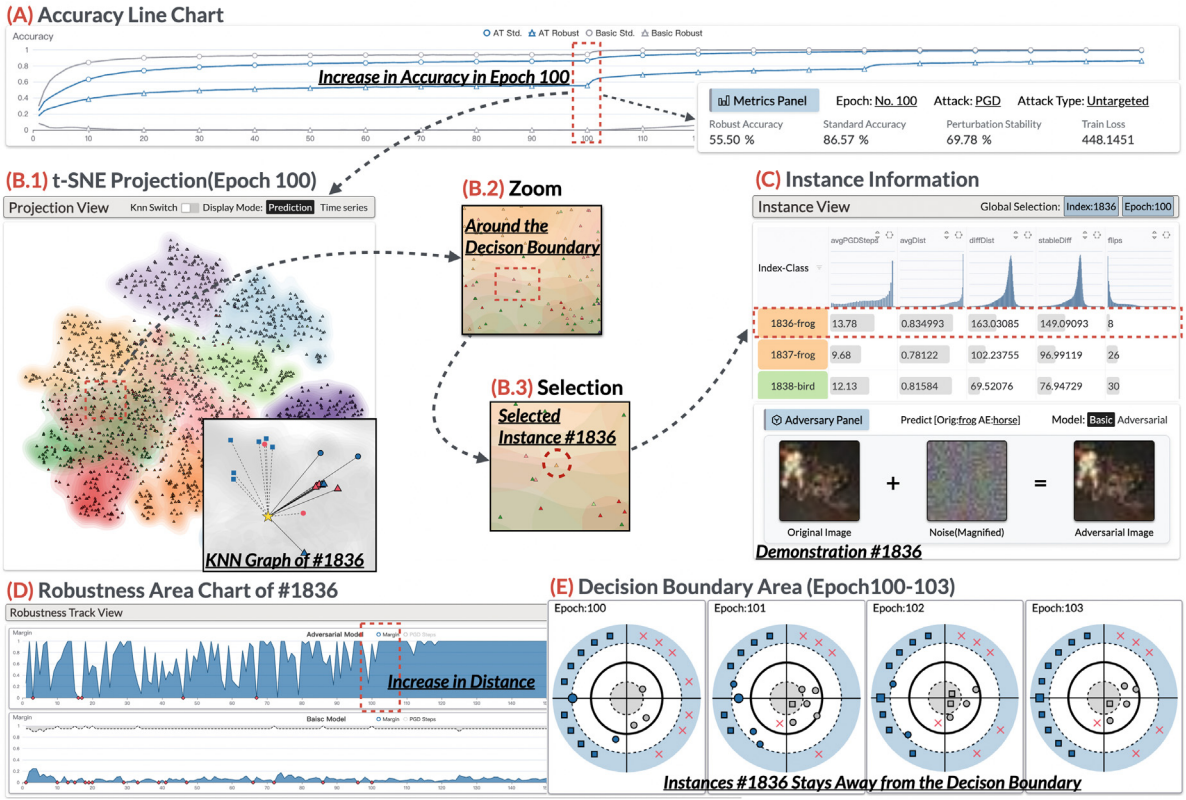

Fig. 7. A case demonstrating how ATVis assists experts in exploring the trade-off between robustness and accuracy caused by adversarial training. The complete analytical workflow spans from the model Level (A) to the cluster level (B.1, B.2, B.3), instance level (C, D), and finally, the decision boundary level (E).

5.2. Case 2: The impact of misclassified instances

Beyond instances that are close to the decision boundary and thus more susceptible to attacks, misclassified instances also play a significant role in influencing adversarial training. Expert E2 sought to utilize ATVis to explore and analyze the impact of these consistently misclassified samples on adversarial training.

Initial Screening and Observation: As shown in Fig. 8 (A), E2 began by sorting instances based on the number of PGD steps required for successful attacks. This sorting helped quickly identify instances that were consistently misclassified throughout the entire adversarial training process. The criterion for this screening was that a PGD step count of zero indicates the prediction on the original image was already incorrect. Upon reviewing the results in Fig. 8 (A), E2 noticed a significant number of instances with the true label "cat" that were consistently misclassified.

To investigate the underlying causes of this phenomenon, E2 selected an instance with the index 49940 in Fig. 8 (A) for further analysis. In the robustness track view, Fig. 8 (B.1), E2 analyzed the area chart of the adversarial model's margin, confirming the consistent misclassification. Meanwhile, for the naturally trained model, although it correctly classified the instance in most cases, it remained very close to the decision boundary, indicating that the separation from other classes was not well defined. Consequently, E2 shifted attention to the adversary panel to examine the original image, revealing that the image itself was of low quality, making it difficult for even the human eye to recognize cat features and make a correct prediction. In response, E2 recommended incorporating higher-quality cat images in future training to help the model learn a richer set of cat features.

Further Analysis of Neighboring Instances: To conduct a more rigorous investigation, E2 sought to identify other potential causes. By examining the decision boundary map of instance 49940, Fig. 8 (C.1), E2 discovered that many of its k-nearest neighbors from the same class were also consistently misclassified. This finding suggested that misclassified instances significantly impaired the model's predictions for surrounding instances, greatly undermining the overall predictive performance. Notably, within the neighborhood of instance 49940, E2 identified another consistently misclassified instance, 20070, as shown in Fig. 8 (C.2).

Upon further examination of the robustness track view for 20070 in Fig. 8 (B.2), E2 was surprised to find a different pattern: the naturally trained model had almost always correctly classified this instance, with a relatively greater and more stable distance from the decision boundary. Intrigued, E2 inspected the corresponding original image and found that, this time, the image clearly depicted a cat. However, the adversarially trained model still failed to classify it correctly. E2 reasonably inferred that the model had not learned sufficiently robust features to distinguish this category effectively. The expert suggested that further optimization of adversarial training, specifically for the cat category, would be necessary to enhance the robustness of the model.

Finally, within the neighborhood of 20070, E2 identified yet another consistently misclassified instance, 8246, as shown in Fig. 8 (C.3). This series of findings highlighted that consistently misclassified instances tend to cluster together, exerting mutual influence and degrading the standard accuracy. For this group of instances, E2 recommended directly reweighting them during adversarial training to mitigate their impact, thereby improving both the robustness and standard accuracy.

5.3. Expert interview

To further evaluate the performance of our framework, we conducted expert interviews with three specialists. Two of them,

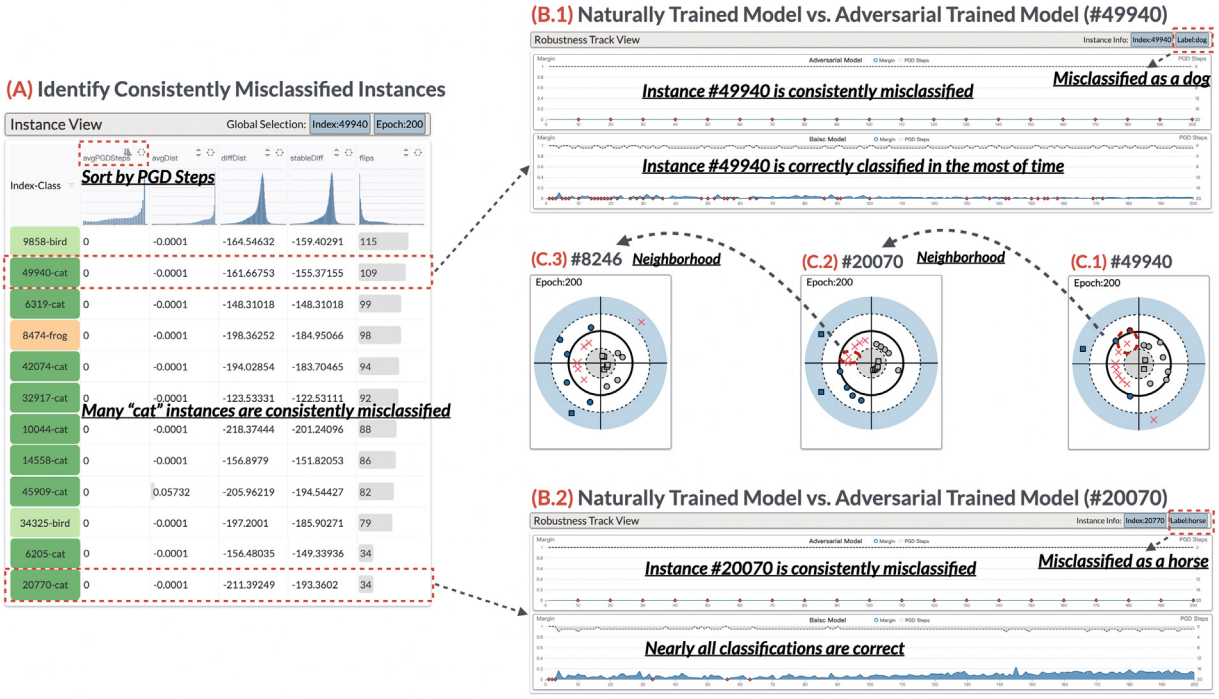

Fig. 8. A case study analyzing the impact of misclassified instances on adversarial training. The analysis involves (A) filtering and identifying misclassified instances, (B) examining the robustness track view of the instances for comparative analysis, and (C) exploring the mutual influence among neighboring instances through the decision boundary view.

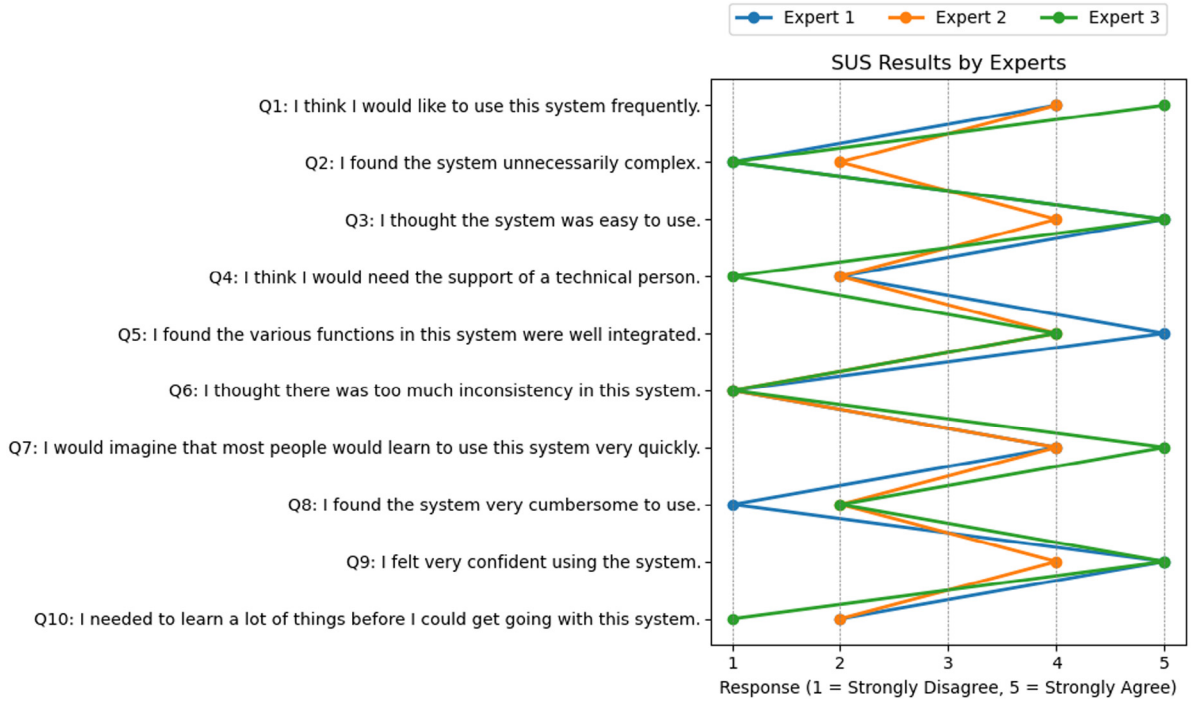

Fig. 9. The System Usability Scale (SUS) evaluation results from three domain experts (Expert 1, Expert 2, Expert 3) on ten key usability questions. The responses, measured on a Likert scale (1 = Strongly Disagree, 5 = Strongly Agree), illustrate the experts' varying perceptions of the system's usability across different dimensions.

E1 and E2, had previously contributed to task development during the requirements analysis phase, while the third expert, E3, specializes in model robustness and adversarial training. The interviews were structured into three stages. First, we presented the context, analysis tasks, and visualization design of the visual analytics framework, followed by a demonstration of the framework's core functionalities using two case studies. Next, the experts were given ample time for exploratory use of the system. Finally, feedback on the framework design and functionality was gathered through interviews and by having the experts complete the System Usability Scale (SUS) questionnaire (Brooke, 1996). The results of the SUS questionnaire are depicted in Fig. 9.

All experts highlighted the effectiveness of the framework in exploring and analyzing the adversarial training process. E1

and E2 highly praised for the multi-level design structure of the framework, noting that it enables comprehensive top-down analysis, from an overview of the model and dataset to detailed instance-level examination. The tight integration between views, along with quick and responsive interactions, significantly enhances the fluidity and efficiency of the analysis process, offering an excellent user experience. Additionally, E2, who participated in case study 2, emphasized that the ATVis framework uncovered a previously unnoticed pattern where misclassified instances tend to cluster, offering key insights for optimizing adversarial training algorithms. These findings collectively demonstrate the usefulness of the framework in supporting effective and detailed adversarial analysis.

In terms of visualization design, the experts also provided positive feedback. E3 noted that the decision boundary view is both innovative and effective, clearly displaying the positions, robustness, and variations of instances near the decision boundary. Additionally, E2 and E3 emphasized that the time series module, which fully presents the detailed information for each epoch during the training process, significantly aids in understanding the impact of adversarial training on the model. Previous research often displayed only a few data points or time periods. Furthermore, the comparative view between naturally trained and adversarially trained models offers a more intuitive representation of the differences, further enhancing the analytical capabilities of the framework.

The experts also provided useful suggestions to enhance the framework. Two of them recommended adding support for customizable PGD attack parameters and incorporating more types of attacks to increase flexibility. Additionally, E3 suggested including adversarial example information in the decision boundary and projection views, allowing for a comparison between adversarial and original examples to more effectively assess the impact of adversarial attacks.

6. Discussion

We propose ATVis, a framework designed to facilitate comparative analysis, enabling experts to conduct a comprehensive evaluation of both model and instance robustness. This framework enables a detailed examination of the differences between naturally trained models and adversarially trained models. Compared to existing studies (Cao et al., 2020; Das et al., 2020), ATVis offers a finer-grained analysis focusing on the changes in individual instances throughout the adversarial training process. Moreover, while other studies primarily focus on how adversarial samples cause misclassification compared to original samples, our approach is aligned with the latest research trends, investigating the impact of adversarial training on both model robustness and predictive performance. Although ATVis provides a powerful tool for exploring and analyzing adversarial training, there are certain issues that we must address to further enhance its effectiveness and applicability.

Comparison with Existing Work. While existing adversarial robustness analysis tools (Cao et al., 2020; Das et al., 2020; Xu et al., 2023) provide valuable insights, they are typically limited to static evaluations or focus on specific training epochs, often relying on sampling a subset of data points. In contrast, ATVis offers an interactive, multi-level analysis framework that tracks the entire adversarial training process, enabling a more comprehensive examination of model behavior across different stages of training. Moreover, ATVis presents a comprehensive adversarial robustness evaluation framework that enhances the single-metric approaches commonly employed by existing tools, allowing for more detailed insights into both robustness and model performance. A key distinguishing feature of ATVis is its in-depth analysis of interactions between instances near decision boundaries, a critical but underexplored area in current methodologies. By focusing on these interactions, ATVis provides a more granular and dynamic understanding of adversarial robustness, which is essential for developing more robust models.

Scalability. Our current framework operates on the CIFAR-10 dataset, which includes 50,000 samples across 10 classes, a considerable data volume. To manage this, we implemented techniques such as sampling, density mapping, and pre-screening to reduce the number of samples requiring analysis. However, when extending to larger and higher-quality datasets like ImageNet (Deng et al., 2009), the framework faces significant challenges in effectively processing and visualizing the data. On one hand, we need to develop better methods for managing visual clutter when displaying large datasets. On the other hand, we must identify more efficient data organization strategies to quickly filter and pinpoint the most critical targets, which may necessitate the integration of recommendation algorithms. Furthermore, to effectively handle real-world industrial-scale datasets, we intend to explore distributed processing and data partitioning techniques to address the associated computational challenges. Additionally, adaptive sampling methods will be considered to dynamically adjust the level of detail, thereby ensuring both computational efficiency and usability when working with large-scale data.

Generalizability. Most adversarial training research focuses on ResNet and its variants, and our framework is designed accordingly. However, it can theoretically generalize to support more advanced models like Vision Transformer (ViT) (Dosovitskiy, 2020), Graph Neural Networks (GNNs) (Kipf and Welling, 2016), and Large Language Models (LLMs) (Brown, 2020), as these models also rely on gradient-based optimization and are vulnerable to gradient-based adversarial attacks, such as PGD. Nonetheless, the main challenge in applying adversarial training to these more complex models is the significant computational cost and time required, as they typically have larger architectures and parameter spaces. As hardware and algorithmic advancements reduce these costs, our framework is designed to adapt, ensuring future support for cutting-edge models as research in adversarial robustness evolves.

Building on this foundation, our visual analytics framework currently supports the widely used SAT (PGD) algorithm and several state-of-the-art (SOTA) methods, including HAT and DyART. However, the significant differences among various approaches to enhancing adversarial robustness present a challenge. Most current methods rely on data augmentation to generate large amounts of synthetic data to improve model robustness, but our framework does not yet support the import and analysis of such additional data. Future work will involve incorporating a dedicated module for handling augmented data, making the framework applicable to a broader range of SOTA robustness enhancement methods. Additionally, because adversarial training is time-consuming, we pre-process the training and store the results to ensure real-time responsiveness. In the future, we plan to develop an automated pipeline that will allow users to customize and analyze their own adversarial training models.

Limitations and Future Work. ATVis provides valuable contributions to understanding adversarial training processes but faces limitations in its current scope, particularly due to its focus on untargeted attacks and challenges in scaling to larger datasets. To address these scalability issues, future developments will optimize the framework's ability to efficiently handle and visualize large-scale datasets, minimizing visual clutter and improving performance for real-time, comprehensive analysis. Future work

will also seek to broaden the framework's capabilities by incorporating evaluations of model robustness under more complex adversarial scenarios and targeted attacks. Additionally, further exploration into adversarial transferability is planned, examining how adversarial examples generated for one model transfer to others, with the aim of enhancing defense strategies and identifying cross-model vulnerabilities.

7. Conclusion

In this paper, we present a visual analytics framework designed for the dynamic exploration and analysis of the adversarial training process. The multi-level design of our framework enables experts to conduct a comprehensive and in-depth analysis, encompassing both model-level evaluations and instance-level tracking. This design allows for the summarization and assessment of performance metrics at the model level, providing a broad overview, while also enabling the detailed comparison and monitoring of changes in instance robustness. Notably, to dynamically inspect the subtle changes in the decision boundary and its surrounding neighborhood, we have designed an innovative target chart to illustrate the decision boundary. The effectiveness and usability of our framework have been rigorously validated through two case studies conducted in close collaboration with domain experts.

CRediT authorship contribution statement

Fang Zhu: Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization. **Xufei Zhu:** Visualization, Software, Formal analysis, Data curation. **Xumeng Wang:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Conceptualization. **Yuxin Ma:** Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition, Data curation, Conceptualization. **Jieqiong Zhao:** Writing – review & editing, Validation, Supervision, Project administration, Conceptualization.

Ethical approval

This study does not contain any studies with human or animal subjects performed by any of the authors. All data used in the study are taken from public databases that were published in the past.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Xumeng Wang and Yuxin Ma are Young Advisory Board Members of Visual Informatics.

Acknowledgments

This work was supported in part by the NSFC (62202217, 62202244), Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515012889), and Guangdong Key Program (No. 2021QN02X794).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.visinf.2024.10.003.

References

Bai, T., Luo, J., Zhao, J., Wen, B., Wang, Q., 2021. Recent advances in adversarial training for adversarial robustness. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, pp. 4312–4321, Survey Track.

Balaji, Y., Goldstein, T., Hoffman, J., 2019. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. arXiv preprint arXiv:1910.08051.

Brooke, J., 1996. SUS: A quick and dirty usability scale. Usability Eval. Ind..

Brown, T.B., 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Cao, K., Liu, M., Su, H., Wu, J., Zhu, J., Liu, S., 2020. Analyzing the noise robustness of deep neural networks. IEEE Trans. Vis. Comput. Graphics 27 (7), 3289–3304.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., Kurakin, A., 2019. On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., Hein, M., 2020. RobustBench: a standardized adversarial robustness benchmark. arXiv preprint arXiv:2010.09670.

Croce, F., Hein, M., 2020. Minimally distorted adversarial examples with a fast adaptive boundary attack. In: International Conference on Machine Learning. PMLR, pp. 2196–2205.

Das, N., Park, H., Wang, Z.J., Hohman, F., Firstman, R., Rogers, E., Chau, D.H.P., 2020. Bluff: Interactively deciphering adversarial attacks on deep neural networks. In: Proceedings of the IEEE Visualization Conference. VIS, pp. 271–275.

Delaforge, A., Azé, J., Bringay, S., Mollevi, C., Sallaberry, A., Servajean, M., 2022. EBBE-text: Explaining neural networks by exploring text classification decision boundaries. IEEE Trans. Vis. Comput. Graphics 29 (10), 4154–4171.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Ding, G.W., Sharma, Y., Lui, K.Y.C., Huang, R., 2020. MMA training: Direct input space margin maximization through adversarial training. In: International Conference on Learning Representations.

Dosovitskiy, A., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Gleicher, M., 2017. Considerations for visualizing comparison. IEEE Trans. Vis. Comput. Graphics 24 (1), 413–423.

Goodfellow, I., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. In: International Conference on Learning Representations.

Gowal, S., Rebuffi, S.-A., Wiles, O., Stimberg, F., Calian, D.A., Mann, T.A., 2021. Improving robustness using generated data. Adv. Neural Inf. Process. Syst. 34, 4218–4233.

Hahn, V.K., Marcel, S., 2022. Biometric template protection for neural-network-based face recognition systems: A survey of methods and evaluation techniques. IEEE Trans. Inf. Forensics Secur. 18, 639–666.

Harrower, M., Brewer, C.A., 2003. ColorBrewer. org: an online tool for selecting colour schemes for maps. Cartogr. J. 40 (1), 27–37.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 770–778.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Li, Z., Pan, M., Zhang, T., Li, X., 2021. Testing dnn-based autonomous driving systems under critical environmental conditions. In: International Conference on Machine Learning. PMLR, pp. 6471–6482.

Li, A., Wang, Y., Guo, Y., Wang, Y., 2024. Adversarial examples are not real features. Adv. Neural Inf. Process. Syst. 36.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J., 2019. On the variance of the adaptive learning rate and beyond. In: International Conference on Learning Representations.

Liu, M., Shi, J., Li, Z., Li, C., Zhu, J., Liu, S., 2017. Towards better analysis of deep convolutional neural networks. IEEE Trans. Vis. Comput. Graphics 23 (1), 91–100.

Ma, Y., Xie, T., Li, J., Maciejewski, R., 2020. Explaining vulnerabilities to adversarial machine learning through visual analytics. IEEE Trans. Vis. Comput. Graphics 26 (1), 1075–1085.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2018. Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations.

Rade, R., Moosavi-Dezfooli, S.-M., 2022. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In: International Conference on Learning Representations.

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

Shen, M., Yu, H., Zhu, L., Xu, K., Li, Q., Hu, J., 2021. Effective and robust physical-world attacks on deep learning face recognition systems. IEEE Trans. Inf. Forensics Secur. 16, 4063–4077.

Singh, N.D., Croce, F., Hein, M., 2024. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. Adv. Neural Inf. Process. Syst. 36.

Szegedy, C., Toshev, A., Erhan, D., 2013a. Deep neural networks for object detection. Adv. Neural Inf. Process. Syst. 26.

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

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A., 2019. Robustness may be at odds with accuracy. In: International Conference on Learning Representations.

Wang, X., Chen, W., Xia, J., Chen, Z., Xu, D., Wu, X., Xu, M., Schreck, T., 2020. ConceptExplorer: Visual analysis of concept drifts in multi-source time-series data. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology. pp. 1–11.

Wang, X., Chen, W., Xia, J., Wen, Z., Zhu, R., Schreck, T., 2023a. HetVis: A visual analysis approach for identifying data heterogeneity in horizontal federated learning. IEEE Trans. Vis. Comput. Graphics 29 (1), 310–319.

Wang, J., Gou, L., Yang, H., Shen, H.-W., 2018. GANViz: A visual analytics approach to understand the adversarial game. IEEE Trans. Vis. Comput. Graphics 24 (6), 1905–1917.

Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., Yan, S., 2023b. Better diffusion models further improve adversarial training. In: International Conference on Machine Learning. PMLR, pp. 36246–36263.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q., 2019. Improving adversarial robustness requires revisiting misclassified examples. In: International Conference on Learning Representations.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., Wilson, J., 2020. The What-If Tool: Interactive probing of machine learning models. IEEE Trans. Vis. Comput. Graphics 26 (1), 56–65.

Wongsuphasawat, K., Smilkov, D., Wexler, J., Wilson, J., Mane, D., Fritz, D., Krishnan, D., Viégas, F.B., Wattenberg, M., 2018. Visualizing dataflow graphs of deep learning models in tensorflow. IEEE Trans. Vis. Comput. Graphics 24 (1), 1–12.

Xu, Y., Sun, Y., Goldblum, M., Goldstein, T., Huang, F., 2023. Exploring and exploiting decision boundary dynamics for adversarial robustness. In: International Conference on Learning Representations.

Yan, J., Yin, H., Zhao, Z., Ge, W., Zhang, J., 2024. Enhance adversarial robustness via geodesic distance. IEEE Trans. Artif. Intell..

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R.R., Chaudhuri, K., 2020a. A closer look at accuracy vs. robustness. Adv. Neural Inf. Process. Syst. 33, 8588–8601.

Yang, H., Zhang, J., Dong, H., Inkawhich, N., Gardner, A., Touchet, A., Wilkes, W., Berry, H., Li, H., 2020b. Dverge: diversifying vulnerabilities for enhanced robust generation of ensembles. Adv. Neural Inf. Process. Syst. 33, 5505–5515.

Yuan, J., Chen, C., Yang, W., Liu, M., Xia, J., Liu, S., 2021. A survey of visual analytics techniques for machine learning. Comput. Vis. Media 7, 3–36.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M., 2019. Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning. PMLR, pp. 7472–7482.

Zhang, J., Zhu, J., Niu, G., Han, B., Sugiyama, M., Kankanhalli, M., 2021. Geometry-aware instance-reweighted adversarial training. In: International Conference on Learning Representations.