Neural Processing Letters (2025) 57:47
https://doi.org/10.1007/s11063-025-11751-z

Dynamic Trade-Offs in Adversarial Training: Exploring Efficiency, Robustness, Forgetting, and Interpretability

Efi Kafali[1] · **Theodoros Semertzidis**[1] · **Petros Daras**[1]

Accepted: 6 March 2025 / Published online: 5 May 2025
© The Author(s) 2025

Abstract
Adversarial attacks pose a threat to neural networks, requiring robust methods to mitigate them. Adversarial Training has emerged as a promising approach; however, its practical application in real-world deep learning systems is hindered by the trade-offs between efficiency and robustness, as optimizing for one aspect may come at cost of the other. This paper presents a comprehensive investigation into the impact of different Adversarial Training approaches and model types on the robustness of adversarially trained models, while considering the dynamic trade-offs involved. Leveraging our previously published method, Delayed Adversarial Training with Non-Sequential Adversarial Epochs – DATNS, we conduct extended empirical analyses through new experiments to effectively balance these trade-offs and navigate the interplay between efficiency and robustness, as well as catastrophic forgetting and interpretability. By providing our insights on the discussed trade-offs this research aims to enable the development of more efficient, robust, and interpretable models against adversarial attacks.

Keywords Deep learning · Adversarial robustness · Catastrophic forgetting · Interpretability

1 Introduction

Deep Learning (DL) models have achieved impressive performance in computer vision tasks. However, they struggle with adversarial samples, which are inputs that undergo subtle perturbations [1–4]. These adversarial samples can cause incorrect predictions even with high confidence, indicating that the models fail to understand essential characteristics of the data [1, 3]. This discovery shifted the research focus from solely improving model performance on clean samples to developing methods that mitigate the models' sensitivity to adversarial attacks. It also highlighted the importance of integrating safety against adversarial attacks, particularly in safety-critical domains such as autonomous driving, healthcare, and critical infrastructures.

Adversarial Training (AT) is acknowledged as one of the most effective strategies for robust computer vision models [5–8]. AT involves training the model on both clean and adversarial inputs, introducing adversarial samples to the training set. It can be formulated as

✉ Efi Kafali
e.kafali92@gmail.com

[1] The Visual Computing Lab, Centre for Research and Technology Hellas, Information Technologies Institute, 6thkm Charilaou-Thermi Road, 57001 Thessaloniki, Greece

a min-max optimization problem, where the inner maximization problem generates the best adversarial version of the input clean sample at each iteration. Simultaneously, stochastic gradient descent updates are used to minimize the loss of adversarial samples [8].

Despite its impressive performance, AT has a significant drawback in terms of computational cost. The need to compute multiple adversarial samples during each update adds to the complexity of the training process, making it impractical for large, high-resolution image datasets and complex architectures. This computational demand is often beyond the reach of many organizations and institutes. Furthermore, the challenge of generalizing well to standard test data has also been extensively discussed in the literature as a trade-off associated with AT [9]. To make AT more feasible on a larger scale, it is crucial to improve its training time efficiency while ensuring that adversarially trained models remain robust and capable of efficiently classifying both natural and adversarial inputs.

This work aims to shed light on the interplay among the dynamic trade-offs involved in AT approaches. Leveraging our latest method, Delayed Adversarial Training with Non-Sequential Adversarial Epochs (DATNS) [10] which managed to reduce the training time of AT without significantly harming accuracy, in this work we aim to thoroughly investigate how less complex AT methods can effectively achieve a well-balanced combination of efficiency, robustness, generalization, and interpretability. Building upon our previous findings that DATNS has the potential to enhance the robustness of trained models while reducing training time, we delve deeper into the trade-offs and challenges inherent in AT. These aspects distinguish our work from the previously published paper, which primarily focuses on the complexity of AT and proposes methods to reduce its training time.

The contributions introduced in this paper can enable the advancement of more efficient, robust, generalizable, and interpretable models to counter adversarial attacks:

- Through extensive experiments, we show that robust AT is achieved by alternating between adversarial and clean training epochs. Unlike traditional AT methods that mix clean and adversarial samples in each epoch, our approach strategically uses adversarial samples only in select epochs, enhancing both robustness and generalization.
- We find that AT works exceptionally well with wide DL architectures when adversarial data are presented in a non-sequential manner. The robustness of wide architectures is proportional to their number of parameters when trained with DATNS.
- Our experimental results demonstrate that DATNS contributes to minimizing catastrophic forgetting.
- We confirm that adversarially trained models have more interpretable loss gradients compared to those trained with standard methods. This holds true even when AT starts from a posterior epoch. In particular, our empirical study for DATNS shows that the model focuses on the overall object rather than texture or color details, resulting in easily interpretable loss gradient visualizations.

The paper is organized as follows: Section 2 provides a detailed analysis on the complexities of training deep neural networks, as well as the elevated complexities of AT, supported by early works and state-of-the-art methods. This section also covers research on catastrophic forgetting and the interpretability of loss gradients in adversarially trained models. Section 3 describes the fundamentals of AT and discusses our previous work [10] with some minor variations to improve clarity. In Sect. 4, we present and analyze the experiments conducted in this study, including the experiments we conducted to tune the hyperparamenters of DATNS, as well as a comparative analysis of DATNS with AT baselines. This section concludes with a discussion on the strengths and limitations of DATNS. Finally, in Sect. 5, we draw con-

clusions based on our interpretation of the experimental results and discuss future research directions.

2 Related Work

Our work centers on AT methods designed to reduce the overhead associated with adversarial defenses. In this section, we explore key areas of research related to AT and its complexity, including its impact on model robustness, catastrophic forgetting, and the interpretability of loss gradients in adversarially trained models.

2.1 Adversarial Training

Recent research has focused on various methods to reduce the complexity of training deep neural networks, which can be particularly relevant in the context of AT. For instance, the MoRR-CNN model introduces a multi-objective optimization framework designed to reduce redundancy in convolutional neural networks (CNNs). This model aims to eliminate redundant parameters and improve computational efficiency, demonstrating effectiveness across several benchmark datasets and CNN architectures [11].

In the domain of time series classification, the NCR-CNNO framework represents another optimization approach. It involves converting raw time series data into matrix representations and optimizing CNNs to enhance classification performance. This method addresses challenges associated with selecting and training deep learning models for time series tasks [12]. Similarly, DropConnect, a stochastic regularization method, has been adapted to dynamically adjusts the dropout rates based on generalization metrics. This technique helps in managing overfitting and improving model regularization by adapting dropout rates during training [13].

Adversarial examples are defined as imperceptible perturbations added to the inputs of a model, which are capable of disrupting their robustness by causing wrong predictions [1]. The introduction of adversarial examples triggered the research community, which soon started to incorporate adversarial examples into the training process, laying the groundwork of AT as a robust defense mechanism against such attacks.

Algorithms such as the Fast Gradient Sign Method (FGSM) started to be used to generate adversarial examples based on the assumption that the vulnerability to adversarial attacks arises from the linear nature of neural networks [2]. However, using single-step attacks, such as FGSM during AT, causes the network to converge to a degenerate global minimum, causing vulnerability to other types of attacks [14].

In order to address the vulnerability of neural networks trained with single-step attacks, researchers have proposed to additionally use perturbations transferred from other similar models. Such approaches can result to increased diversity of the perturbations the model sees during training [15]. Through a similar perspective, the use of multiple random noise layers has been explored as a defense against strong attacks [16], while ensembles of quantized models and full precision networks have also been introduced as a novel approach to improve robustness against adversarial attacks [17].

Recent research findings have shed light on the limitations of the single-step FGSM method. Referred to as "catastrophic overfitting," it has been observed that FGSM harms the model's generalization ability. Attempts to mitigate this issue through random initialization of FGSM have been introduced [18]. While FGSM offers faster training times compared to other

methods [5], the addition of randomness does not effectively address catastrophic overfitting, implying that randomness alone cannot significantly improve robustness. Notably, studies indicate that catastrophic overfitting is not solely influenced by model depth or parameter count. Thus, despite initial improvements in robustness during early training stages, FGSM's overall poor performance is primarily attributed to non-linearity [19].

According to [20], the PGD attack is considered the most reliable method in terms of robustness for learned models. Further enhancement of the PGD attack has been introduced by an algorithm that efficiently recycles gradient information during model parameter updates. This approach enables simultaneous backward passes for both the model's parameters and the crafted perturbations, eliminating the need for separate gradient computations, significantly reducing computational costs, while approaching the efficiency of natural training [5].

AT is a robust optimization framework that incorporates adversarial samples into the training process, improving model robustness [8, 21]. The original work by [8] achieved significant robustness on MNIST and CIFAR10 datasets against diverse attacks, formulating AT as a min-max optimization problem and utilizing the iterative PGD attack. Unlike the single-step FGSM attack, PGD demonstrated higher accuracy for adversarial samples, reducing label leaking and overfitting [22]. The importance of wide architectures in enhancing robustness against adversarial inputs is furthermore supported by several recent studies [8, 10, 23, 24].

Moreover, recent advancements in AT have led to the development of new defense methods, such as TRADES [25]. This method is inspired by theoretical analysis and trades off adversarial robustness against accuracy by providing a differentiable upper bound on the prediction error for adversarial examples. This approach has shown promising results in enhancing model robustness while maintaining high accuracy levels, as demonstrated in real-world datasets. However, TRADES appears to be less efficient in terms of training time than [8], as evidenced by many works, such as [26].

Researchers have further investigated the framework proposed by [8] to address the high complexity caused by repeated gradient updates during training. Several works have introduced modifications to reduce this complexity by altering the attack strength employed during training. For instance, a curriculum approach has been proposed, incorporating adversarial examples generated with varying attack strengths, including both weak and strong attacks. This curriculum strategy has demonstrated notable performance improvements [27]. These findings are at contrast with earlier conclusions by [8], which suggested that stronger attacks were necessary for achieving higher robustness. Additionally, researchers have explored the accumulation of attack strength over epochs. Instead of generating adversarial inputs at the beginning of each epoch, a method has been proposed where the adversary from the previous epoch is reused in subsequent epochs, resulting in more efficient AT [26].

Other attempts to confront the complexity of AT involve the reduction in the amount of training data. An informed data selection strategy has been explored for AT, where only selected samples based on the loss are used to update the model parameters. As a result, the training time is reduced and the trade-off between accuracy and robustness is balanced [28]. Random sampling has also been investigated in this context, where only a small subset of the entire dataset is receiving the attack during some of the training epochs, resulting in reduced training time [10].

Furthermore, important observations regarding the impact of adversarial examples at different stages of training are demonstrated. A dynamic training strategy has been introduced, which is progressively improving the convergence quality of crafted adversarial examples. Training on adversarial samples with better convergence quality in the later phase enhances robustness, while in the initial phase, adversarial inputs are unnecessary [6]. Additionally, it has been shown that the initial phase of AT has minimal impact on robustness and can even

negatively affect accuracy. As a result, AT can be initiated from a posterior epoch, omitting the initial phase [29].

Building on the research by [29], our latest work, DATNS, introduces a variation of AT that initiates AT from a posterior epoch and alternates between natural and adversarial training. Through extensive experiments, we discovered that deep learning classification models can maintain their robustness when adversarial data is sporadically injected during training. Specifically, we found that periodically injecting adversarial data yields greater benefits to robustness compared to random epoch injection. With DATNS, we achieved a significant reduction in AT training time while preserving robustness at attainable levels [10].

2.2 Catastrophic Forgetting

When machine learning algorithms are used to solve a sequence of tasks, it is crucial to specify the notation used to describe these tasks. In this context, $J_1 : j$ denotes a sequence of tasks from J_1 to J_j, where J_1 is the first task and J_j is the current or most recent task in the sequence. Catastrophic forgetting occurs when learning a new task J_j leads to the overwriting or degradation of previously learned knowledge from tasks J_1 through J_{j-1} [30].

Catastrophic forgetting, the phenomenon according to which neural networks lose previously acquired knowledge when exposed to new, divergent information, has been a subject of extensive discussion and research [31, 32]. Unlike humans who accumulate knowledge, neural networks tend to fully adapt to new tasks, resulting in a significant decrease in performance on previously learned tasks, a phenomenon also known as catastrophic interference [32]. The connection between defensive methods against adversarial attacks and catastrophic forgetting is evident, as deep neural networks excel in classification tasks but struggle with adversarial inputs. Recent research has thus focused on exploring the relationship between catastrophic forgetting and adversaries.

In a notable work by [33], a sparse coding technique is introduced for adaptive allocation of model capacity to different tasks. The authors employ group sparse regularization to assign parameter groups for each task, freezing them while the rest of the network learns new tasks. They also propose a meta learning technique that facilitates knowledge transfer between tasks. By utilizing episodic training-based optimization, the authors promote the learning of weights expected to be beneficial for future task solving. These approaches effectively reduce task interference.

The work by [34] highlights the strong link between single-step attacks like FGSM and catastrophic forgetting. They demonstrate that fixed perturbation magnitudes, rather than attack directions, lead to decision boundary distortion and highly curved loss surfaces. To address this, they propose an algorithm that dynamically adjusts the perturbation magnitude for each image, effectively mitigating catastrophic forgetting. In a related study by [35], the authors tackle catastrophic forgetting in single-step AT by considering the rapid gradient growth of each sample. They present a variation of AT that restricts the training process to a carefully extracted subspace, controlling gradient growth. This approach achieves state-of-the-art performance in single-step AT. Furthermore, the work by [36] investigates the impact of multiple exit networks on reducing adversarial perturbations. A multi-exit network architecture is proposed, that can produce easily identifiable samples at early exits, preventing catastrophic forgetting in single-step AT.

PGD attack has been shown to be prone to catastrophic forgetting and the use of weak attacks during training has been suggested as an effective mitigation strategy [27]. Moreover, a

novel sequential learning method has been introduced that incorporates adversarial examples. The approach utilizes adversarial subspaces from previous tasks to facilitate the learning of new tasks, preventing catastrophic forgetting [37]. On a related work, it is argued that AT models cannot be conventionally fine-tuned due to severe catastrophic forgetting. The inability of AT models to retain previously learned features has been highlighted. To address this, a novel adversarial fine-tuning method has been inferred [38].

An adversarial feature alignment method has been presented by [39], with the goal to avoid catastrophic forgetting. In their work, the authors proposed that both the low-level visual features and high-level semantic features are used as soft targets. Both features guide the training process in multiple stages and provide adequate supervised information of the old tasks, contributing to forgetting reduction. The proposed method achieves state-of-the-art performance by means of accuracy on new tasks, however, it also preserves the performance to old tasks.

Furthering this discussion, the work by [40] introduces an approach that emphasizes regularizing network parameters to enhance robustness against adversarial attacks while mitigating catastrophic forgetting. This method, termed DIversity via Orthogonality (DIO), incorporates multiple paths within the network and imposes orthogonality constraints to ensure diversity among these paths. By augmenting the model in this manner, the learned features become more adaptable to diverse inputs, including adversarial examples, thus reducing the likelihood of catastrophic forgetting. The DIO method demonstrates its effectiveness across various datasets, structures, and attack scenarios, and it can also be combined with existing data augmentation techniques for further robustness gains.

2.3 Visualizing and Interpreting Gradients of Adversarially Trained Models

AT has been recognized for its ability to enhance the robustness of deep learning (DL) models. However, recent studies have uncovered an additional intriguing aspect of AT: it tends to generate loss gradients that are visually more interpretable compared to DL models trained with standard methods. This unexpected benefit has significant implications as interpretable gradients can offer insights into the model's prediction process and provide a better understanding of both the model's outputs and the inner workings of neural networks. Several works have confirmed and further elucidated this connection.

The interpretability of gradients has been described as an unexpected benefit of AT by [41], highlighting that interpretable gradients can shed light on the decision-making process of classification models. It has also been demonstrated that networks robust to adversarial perturbations exhibit interpretability in saliency maps [42]. Furthermore, AT has been found to improve the alignment between gradients and the human visual system, as the loss gradients tend to lie closer to the image manifold [43]. Moreover, a recent study has visually compared attribution maps produced by different methods and highlighted that adversarially trained models generate attribution maps that are quantitatively more meaningful and visually aligned with human perception [44].

Furthermore, adversarially trained models have been found to prioritize the overall structure of objects over specific details, a characteristic that aligns with human reasoning. In a study by [45], adversarially trained CNNs were shown to be more sensitive to global structures such as shapes and edges, compared to normal CNNs that focused more on texture information. Adversarially trained models were also less affected by texture distortion and

exhibited a stronger emphasis on shape information, leading to improved generalization performance. This preference for object structure over details suggests that adversarially trained models may possess better generalization abilities compared to models trained with standard methods. Furthermore, research by [46] further supports these findings, demonstrating that AT encourages a shift towards shape representation, which is a primary factor in human object recognition.

3 Methodology

3.1 Regular Adversarial Training

We begin by introducing the AT framework proposed by [8] (we refer to it as regular AT). The objective of regular AT is to find adversarial samples that maximize the loss while minimizing it with respect to the model parameters. The framework uses the PGD attack to maximize the loss and its formulation can be described by (1).

$$\begin{aligned} &\min_{\theta} \rho(\theta); \\ &\rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\|\tilde{x}-x\|_\infty \leq \epsilon} \mathcal{L}(f_\theta(\tilde{x}), y)\right] \end{aligned} \tag{1}$$

Here, θ represents the classifier's parameters, x denotes a natural sample, $\tilde{x}$ denotes its adversarial counterpart, and $\mathcal{L}$ is the loss function. The adversarial sample $\tilde{x}$ is generated within the ϵ-ball (ℓ_∞ norm) around the natural sample x with ground truth label y. The objective is to find the adversarial sample $\tilde{x}$ that maximizes the loss $\mathcal{L}$ while searching for the model parameters θ that minimize the loss $\mathcal{L}$ with respect to $\tilde{x}$.

3.2 Delayed Adversarial Training with Non-Sequential Adversarial Epochs (DATNS)

In our previous work [10] we have explored a more efficient variation of AT, namely DATNS. In DATNS, we begin AT from a posterior epoch and alternate between natural and adversarial training until the end of the training process. This approach significantly reduces the training time and computational overhead compared to regular AT.

We introduce the concept of an *initial adversarial epoch*, denoted as s_0, which represents the first epoch after which AT begins. We also define *natural epochs* as epochs during which the model is trained on natural samples and *adversarial epochs* as epochs during which the model is trained adversarially. The total number of training epochs is denoted by N.

In DATNS, adversarial epochs occur periodically after s_0 and under a time interval t. Therefore, AT takes place on non-sequential epochs inbetween natural training epochs. The hyperparameter t needs to satisfy the condition $t \geq 2$, given that for $t = 1$ adversarial epochs take place sequentially. As soon as the training reaches an epoch $\geq s_0$ satisfying (2), the model is trained adversarially for the current epoch. In any other case the model is trained only on natural samples.

$$epoch \mod t = t \tag{2}$$

The algorithmic description of DATNS is provided in Algorithm 1.

Algorithm 1 Delayed Adversarial Training with Non-Sequential Adversarial Epochs (DATNS)

Input: J training examples $\{(x_j; y_j)\}_{j=1}^{J}$; Number of epochs N; Optimizer; PGD attack $\mathcal{A}_{T,\epsilon,\alpha}$ where T is the number of steps, ϵ is the ball size, and α is the step size; time interval $t \in \{2, 3, ..., M\}$
Output: Model parameters θ;

Initialize θ randomly;
$s_0 \leftarrow n$; {Initialize the first switching point}
$t \leftarrow m$; {Initialize the time interval under which adversarial epochs occur}
for $epoch = 0$ **to** $N - 1$ **do**
  for each batch (x_b, y_b) **do**
    if $epoch \geq s_0$ **then**
      if $(epoch \mod t) = 0$ **then**
        Replace with $x_b \leftarrow \mathcal{A}_{T,\epsilon,\alpha}(\theta, x_b, y_b)$; {If switching point reached, replace natural with adversarial samples}
      else
        Proceed with x_b without modification; {If the switching point is not reached, keep the batch unchanged}
      end if
    end if
    Train the model on the batch of input examples (x_b, y_b) using the optimizer;
  end for
end for

4 Experiments

4.1 Overview of Experimental Setup

In this section, we provide a detailed overview of the experimental setup used to further evaluate our previously proposed DATNS method. Our experiments are designed to rigorously assess the effectiveness of DATNS through two primary avenues: an ablation study and a comparative analysis. The ablation study focuses on systematically examining the impact of various hyperparameters on the performance of DATNS, including different patterns of adversarial data appearance and perturbation levels. This study aims to identify optimal settings for DATNS that balance robustness and efficiency.

Following this, the comparative analysis evaluates DATNS against established baselines to determine its performance in terms of accuracy, training efficiency, robustness, catastrophic forgetting and interpretability of loss gradients. This analysis involves comparing DATNS with both standard AT methods and a recent approach that initiates AT from a posterior epoch. Together, these subsections provide a comprehensive assessment of DATNS and its potential as an effective defense mechanism in adversarial settings.

Previous Work and Revisited Experiments The vast majority of the reported experiments are new, aiming to provide new analyses that extend the evaluation of DATNS beyond our previous work. However, to provide continuity and context, we reference specific experiments from our earlier research. The revisited experiment instances are listed below:

1. **Hyperparameter Tuning and Ablation Study**:
   - l_∞ perturbation of $\epsilon = \frac{8}{255}$ DATNS training with ResNet-18: Experiments on CIFAR-10 and CIFAR-100 datasets, specifically for time intervals $t = 2$, $t = 3$, and $t = 5$ with initial adversarial epochs $s_0 = 55$, $s_0 = 75$, and $s_0 = 100$. The results of these experiments are included in Figure 1 and Figure 3, along with our new experiments on different hyperparameter settings.

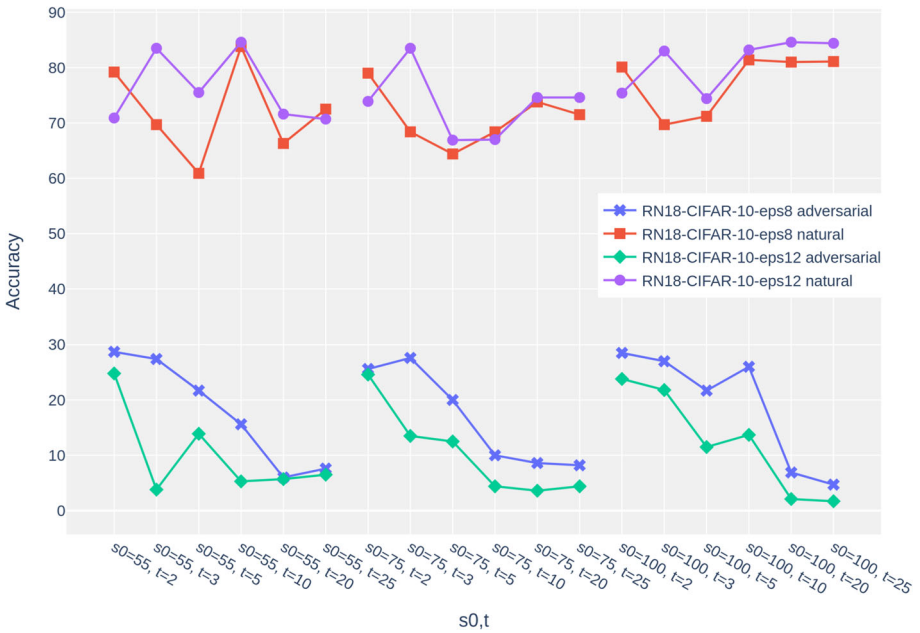

Fig. 1 Natural and adversarial accuracy for ResNet-18 trained with DATNS on CIFAR-10, with a maximum l_∞ perturbation of $\epsilon = \frac{8}{255}$ and $\epsilon = \frac{12}{255}$, versus different values of s_0 and t and grouped by s_0

2. **Comparative Analysis**:
   - *WRN 28-10 on CIFAR-10*: Comparison of DATNS and [29] involving WRN 28-10, which were previously reported. The results of this experiment are included in Table 1, along with new experiments on a variety of Wide ResNet architectures.

These reused experiments serve as a reference point for validating and comparing new results, while all the experiments not listed in this list are new.

4.2 Hyperparameter Tuning and Ablation Study

This subsection explores the impact of different hyperparameters on the performance of DATNS through an ablation study. We aim to investigate how the pattern of adversarial data appearance influence the overall performance of DATNS, which will inform the setup for subsequent comparative experiments.

To evaluate the trade-offs between robustness and efficiency, we consider natural accuracy which measures the model's performance on clean, unperturbed test data, reflecting its generalization ability. Additionally, we consider adversarial accuracy to assess the model's robustness by evaluating its performance on adversarial examples. Finally, we consider training efficiency to examine the impact of adversarial epochs on training resources; specifically, fewer adversarial epochs lead to more efficient training.

Our experiments for this ablation study utilize a ResNet18 model trained on CIFAR-10 and CIFAR-100 datasets. All models are trained for a total of 200 epochs with a batch size of 128. We use SGD as the optimizer with a momentum of 0.9, weight decay of 2×10^{-4}, and an initial learning rate of 0.1, which is decreased by a factor of 10 after 100 and 150

Table 1 Wide ResNets with varying width, depth and total number of parameters trained for 200 epochs on CIFAR-10 with [8], [29] and DATNS with $t = 2$, beginning AT from $s_0 = 100$.

Depth	Width	#Params	nat./adv. accuracy [8]	nat./adv. accuracy [29]	nat./adv. accuracy DATNS
40	1	0.6M	77.0% / **41.5%**	74.6% / 38.0%	**83.9%** / 37.7%
40	2	2.2M	79.4% / 45.3%	79.7% / 41.9%	**86.1%** / **45.6%**
40	4	8.9M	82.1% / 51.1%	81.8% / 48.4%	**86.4%** / **58.2%**
40	8	35.7M	83.6% / 52.1%	84.0% / 53.8%	**86.6%** / **63.0%**
28	10	36.5M	83.8% / 52.1%	84.6% / 52.3%	**86.5%** / **60.8%**
28	12	52.5M	83.6% / 50.7%	84.5% / 52.2%	**86.8%** / **60.1%**
22	8	17.2M	82.2% / 50.6%	83.1% / 50.0%	**86.5%** / **51.0%**
22	10	26.8M	84.1% / 53.3%	84.0% / 51.3%	**85.9%** / **53.5%**
16	8	11.0M	81.1% / **48.1%**	82.9% / 46.4%	**87.0%** / 47.6%
16	10	17.1M	82.9% / **52.0%**	83.5% / 48.8%	**86.8%** / 50.4%
Adv. epochs ratio: #adv./#total			200/200	100/200	**50/200**

Bold values indicate the best-achieved results for each reported experiment

epochs. During adversarial training epochs, the dataset is subjected to PGD attacks with $T = 10$ steps. We investigate both l_∞ perturbations with $\epsilon \in \{\frac{8}{255}, \frac{12}{255}\}$ and l_2 perturbations to examine their effects on model performance.

4.2.1 Pattern of Adversarial Data Appearance

Our previous research [10] has highlighted the importance of the periodic occurrence of adversarial epochs, as opposed to a predefined number of adversarial epochs at random intervals, for achieving better overall performance. To further explore this, we conduct a series of experiments with different time interval values t, such that $t \in \{2, 3, 5, 10, 20, 25\}$ for adversarial epochs, focusing on a ResNet18 model trained on CIFAR-10 and CIFAR-100 datasets. We consider initial adversarial epoch values s_0, such that $s_0 \in \{55, 75, 100\}$, as explored by [29].

Additionally, for the previous combinations of s_0 and t values, we extend our exploration by investigating two values of ϵ to represent the maximum l_∞ perturbation, where $\epsilon \in \{8/255, 12/255\}$. This additional dimension enables us to examine the impact of varying levels of perturbation on model performance. Although our focus for the rest of this paper remains on the l_∞ threat model, in this subsection we broaden our scope to include l_2 perturbations. Figures 1 and 2 depict the natural and adversarial accuracy results for a ResNet18 trained on CIFAR-10 dataset. These results are presented across various combinations of s_0, t and ϵ values, showcasing the model's performance under l_∞ (Fig. 1) and l_2 (Fig. 2) threat models.

In Fig. 1 we observe that adversarial accuracy tends to decrease as the time interval t increases, reflecting the susceptibility to adversarial attacks when adversarial examples are presented less frequently to the model. Natural accuracy exhibits significant fluctuation as t values increase across most experiment instances. Despite variations, it's notable that across the range of t values explored in this experimental setup, natural accuracy exhibits closely comparable values for the smallest ($t = 2$) and largest ($t = 25$) t values for most of the experiment instances. In Fig. 2, where the natural and adversarial accuracy results

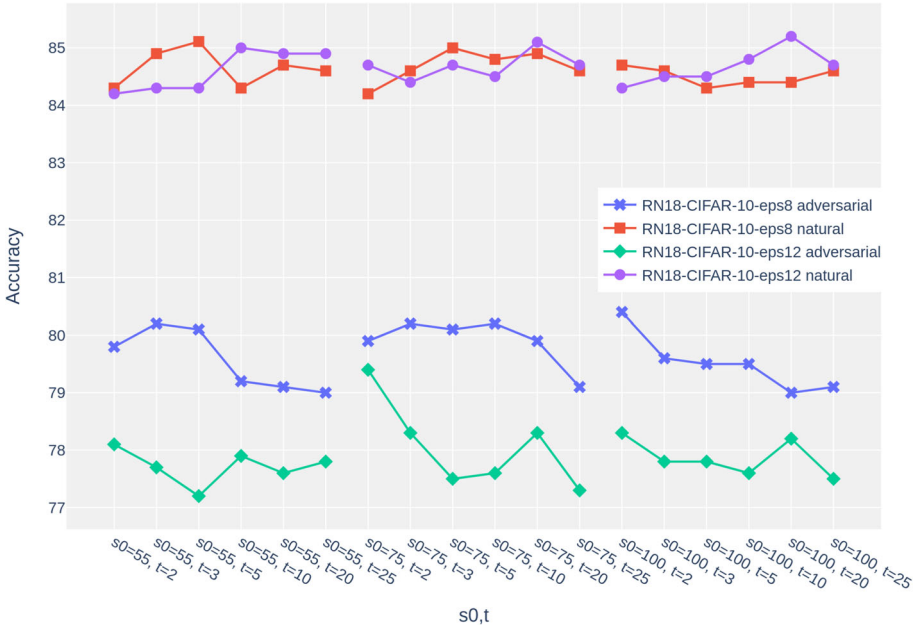

Fig. 2 Natural and adversarial accuracy for ResNet-18 trained with DATNS on CIFAR-10, with a maximum l_2 perturbation of $\epsilon = \frac{8}{255}$ and $\epsilon = \frac{12}{255}$, versus different values of s_0 and t and grouped by s_0

for the same ResNet18 model trained on the CIFAR-10 dataset are presented under the l_2 perturbation threat model, we observe that adversarial accuracy tends to decrease for larger values of t. Additionally, in Fig. 2, the overall natural accuracy levels remain relatively more stable across different combinations of s_0, t, and ϵ values. For both l_∞ and l_2 threat models, adversarial accuracy consistently remains lower for $\epsilon = \frac{12}{255}$ compared to $\epsilon = \frac{8}{255}$, while the s_0 value appears to exert minimal influence on the levels of both natural and adversarial accuracy.

In Fig. 3, we present the natural and adversarial accuracy results for a ResNet18 model trained on the CIFAR-100 dataset, specifically focusing on the l_∞ perturbation threat model. Similarly to the observations from the CIFAR-10 experiments, we note that adversarial accuracy tends to decrease as the time interval t increases. Additionally, the natural accuracy in CIFAR-100 experiments displays a steady decrease with increasing t in most of the experiment instances. This downward trend suggests that the model's robustness diminishes as the frequency of adversarial examples decreases. Notably, natural and adversarial accuracy remain consistently lower for $\epsilon = \frac{12}{255}$ compared to $\epsilon = \frac{8}{255}$. The s_0 value continues to have minimal discernible impact on accuracy levels, underscoring that the epoch at which the adversarial samples first appear may not significantly impact the overall robustness of the model.

4.2.2 Insights Gained

The ablation study reveals several insights:

- **Optimal Configuration**: Values of $s_0 = 100$, $t = 2$, and $\epsilon = 8/255$ consistently deliver reliable results in DATNS.

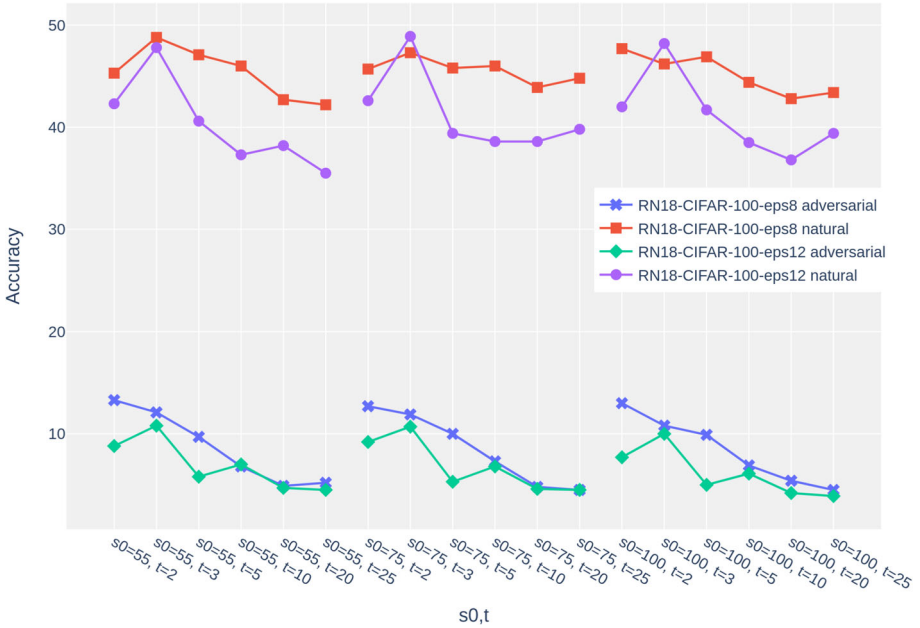

Fig. 3 Natural and adversarial accuracy for ResNet-18 trained with DATNS on CIFAR-100, with a maximum l_∞ perturbation of $\epsilon = \frac{8}{255}$ and $\epsilon = \frac{12}{255}$, versus different values of s_0 and t and grouped by s_0

- **Trade-off Between Robustness and Efficiency**: Choosing a late s_0 minimizes adversarial epochs, reducing training time without significant performance loss. Larger t values occasionally improve natural accuracy but do not significantly enhance adversarial accuracy.
- **Perturbation Levels**: An ϵ value of $8/255$ is preferable for comparisons with existing methods, aligning with more common practices in the literature.

These findings provide a solid foundation for the comparative experiments in the subsequent sections, allowing us to benchmark DATNS effectively against other methods in the literature.

4.3 Comparative Analysis

In this section, we provide a comprehensive analysis of our previously proposed method, DATNS, by evaluating its performance across various aspects and comparing it with established baselines. Our analysis includes multiple dimensions:

1. **Adversarial Training Performance**: We compare DATNS with established AT methods, specifically the framework by Madry et al. [8] and the approach by Gupta et al. [29]. The comparison is conducted on CIFAR-10 and GTSRB datasets, focusing on natural and adversarial accuracy as well as training efficiency. Additionally, we compare our method with [25] to provide a deeper understanding of the tradeoffs between robustness and efficiency. Although exact training times are not reported due to the extensive number of experiments conducted across various GPUs, we use the fraction of adversarial to total number of epochs as a key indicator of computational efficiency. This metric helps to

assess the relative efficiency of each method, given that adversarial epochs are resource intensive compared to natural epochs.
2. **Catastrophic Forgetting**: We investigate how DATNS behaves in relevance to the issue of catastrophic forgetting, a challenge in AT where models may forget previously learned information when trained on adversarial examples. This section explores how DATNS's alternating training approach impacts the model's ability to retain knowledge of both clean and adversarial data.
3. **Interpretability of Loss Gradients**: We qualitatively evaluate the interpretability of loss gradients for models trained using DATNS compared to natural training and other AT methods. This includes examining how different training approaches affect the clarity and usefulness of the gradients in understanding model behavior.

To ensure a fair comparison, all methods were trained under consistent conditions, unless otherwise specified (e.g., TRADES), using a total of 200 epochs and a batch size of 128. The SGD optimizer was used with a momentum of 0.9, weight decay of 2×10^{-4}, and an initial learning rate of 0.1, decreasing by a factor of 10 after 100 and 150 epochs. During AT, we employed PGD attacks with $T = 10$ steps, a maximum l_∞ perturbation of $\epsilon = \frac{8}{255}$, and a gradient ascent step size of $\alpha = \frac{2}{255}$.

By addressing these aspects, we aim to provide a well-rounded view of DATNS's strengths, limitations, and practical implications, thus offering insights into its overall efficacy and suitability for different AT scenarios.

4.3.1 Training Wide ResNets with DATNS

In our previous work [10], we found that DATNS increased the robustness of the employed Wide ResNet baseline, confirming the findings of [23] that wide architectures are more robust against adversarial attacks. In this section, we verify that this property is conserved, even when the total number of adversarial epochs is profoundly reduced. Through a new set of experiments we further explore how the robustness of Wide ResNets of different depths, widths and total number of parameters is affected by DATNS.

To evaluate the performance of DATNS compared to [8] and [29] under different hyperparameter settings, we train 10 Wide ResNet baselines on CIFAR-10 and GTSRB datasets for 200 epochs. For [8] we train the models adversarially for the total number of epochs, while for [29] and DATNS, we begin AT from $s_0 = 100$. Additionally, for DATNS, we alternate between adversarial and natural training with a time interval of $t = 2$. For reference, we mention here the results for Wide ResNet 28x10 traned on CIFAR-10 from [29], where it is reported that regular AT [8] achieved an adversarial accuracy of 48.5% and a natural accuracy of 86.8%, while [29] achieved an adversarial accuracy of 49.7% and a natural accuracy of 87.9% without early stop.

Furthermore, to broaden the scope of our comparative analysis, we include an evaluation of DATNS against TRADES [25]. TRADES, known for its efficacy in enhancing model robustness, often comes at the cost of increased computational complexity. Therefore, although we can assume that DATNS is more efficient than TRADES, we include this comparison to explore the balance between robustness and computational efficiency compared to TRADES. The results for TRADES are sourced from the study referenced in [24], focusing on a Wide ResNet 34-10 architecture trained on CIFAR-10. In order to uphold consistency, we train the same model architecture on CIFAR-10 with DATNS for 100 epochs with a batch size of 128, starting AT from $s_0 = 20$, $s_0 = 40$, $s_0 = 50$ and employing the same time interval of $t = 2$. Additionally, for this set of experiments we perform the standard PGD attack using

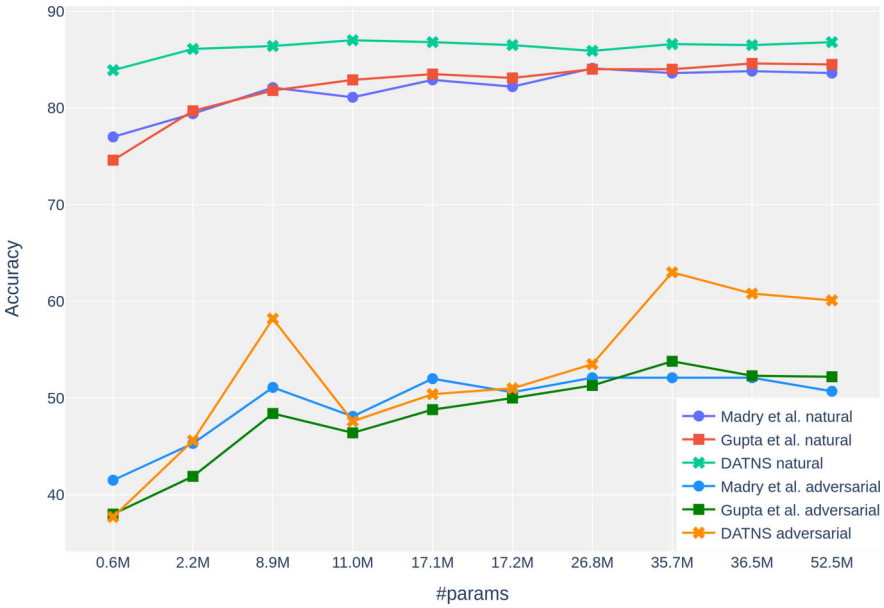

Fig. 4 Natural and adversarial accuracy of DATNS, [29] and [8] for the Wide ResNets of Table 1 (CIFAR-10), sorted by total number of parameters

20 steps with step size 0.007, and epsilon 8/255, following [24]. The initial lr is set to 0.1 and is halved after s_0, following a similar approach as [24].

In summary, our comparative analysis in this subsection sheds light on the complex nature of AT methodologies, revealing trade-offs between adversarial robustness, accuracy, and computational efficiency. We hope that these insights contribute to advancing the understanding of robust machine learning techniques but also provide valuable guidance for the development of practical and effective defense mechanisms against adversarial attacks.

CIFAR-10. Table 1 summarizes the results of [8], [29] and DATNS on CIFAR-10 for various Wide ResNet configurations. It is observed that DATNS consistently leads to increased natural and adversarial accuracy across most experiment instances. The natural accuracy improvement over [29] aligns with expectations, given the increased number of natural epochs. Surprisingly, even with fewer adversarial epochs compared to [29] and [8], DATNS achieves enhanced adversarial accuracy for a wide range of experiments. Although training times are not directly compared due to different GPUs used, all three methods can be assessed based on the fraction of adversarial epochs over the total number of epochs ($\frac{1}{1}$ for [8], $\frac{1}{2}$ for [29] versus $\frac{1}{4}$ for DATNS). Notably, cases where DATNS performs slightly worse than [29] or [8] involve Wide ResNet architectures of smaller capacity.

Furthermore, Fig. 4 illustrates that as the total number of parameters increases, both natural and adversarial accuracy improve for the set of Wide ResNets considered in Table 1. The best performance appears to be achieved by the most complex models, indicating that while DATNS may not significantly improve the performance of shallower models, it is particularly effective when combined with complex wide architectures, providing faster training and competent performance.

In Table 2, we compare the results of DATNS with TRADES [25], which has significant presence within the field of AT. DATNS achieves lower robustness in terms of adversarial

Table 2 Comparison of DATNS with TRADES [25]. The reported results for TRADES are sourced from [24] and involve a Wide ResNet 34-10 trained on CIFAR-10 for 100 epochs, with varying robust regularization parameter values λ. For DATNS, we experiment with $s_0 = 20$, $s_0 = 40$ and $s_0 = 50$ and $t = 2$ for the total of 100 epochs.

	Adversarial accuracy	Natural accuracy	Adv. epochs ratio
λ	TRADES [25]		
6	54.1%	84.9%	100/100
9	55.2%	84.1%	
12	55.9%	83.5%	
15	55.9%	82.8%	
18	**56.4%**	82.2%	
24	56.0%	81.7%	
s_0, t	DATNS		
20, 2	51.2%	**86.6%**	**40/100**
40, 2	52.1%	**87.9%**	**30/100**
50, 2	50.9%	80.8%	**25/100**

Bold values indicate the best-achieved results for each reported experiment

Table 3 Wide ResNets with varying width, depth and total number of parameters trained for 200 epochs on GTSRB dataset with [8], [29] and DATNS with $t = 2$, beginning AT from $s_0 = 100$.

Depth	Width	#Params	nat./adv. accuracy [8]	nat./adv. accuracy [29]	nat./adv. accuracy DATNS
40	1	0.6M	84.4% / 57.2%	78.7% / **59.0%**	**96.0%** / 57.8%
40	2	2.2M	89.0% / 58.7%	85.3% / **59.7%**	**95.7%** / **59.7%**
40	4	8.9M	90.5% / 59.6%	85.3% / 60.7%	**96.2%** / **61.3%**
40	8	35.7M	89.8% / 59.9%	87.4% / 61.3%	**96.2%** / **61.5%**
28	10	36.5M	89.0% / 59.1%	86.2% / 60.1%	**95.9%** / **60.4%**
28	12	52.5M	89.1% / 59.7%	84.8% / 60.5%	**96.5%** / **61.4%**
22	8	17.2M	88.7% / 59.1%	86.9% / 59.8%	**96.2%** / **60.1%**
22	10	26.8M	88.5% / 58.5%	84.1% / **60.9%**	**95.8%** / 60.5%
16	8	11.0M	89.6% / 59.2%	87.4% / 59.3%	**96.3%** / **59.5%**
16	10	17.1M	89.3% / 59.2%	88.3% / **59.8%**	**95.7%** / 59.3%
Adv. epochs ratio: #adv./#total			200/200	100/200	**50/200**

Bold values indicate the best-achieved results for each reported experiment

accuracy compared to TRADES. However, it achieves better natural accuracy even for a smaller total number of epochs (100, as compared to 200 for our previous experiments), indicating that DATNS maintains strong performance on clean data while providing a certain level of robustness against adversarial attacks. These improvements are achieved within a reduced training time compared to TRADES. While there is a trade-off in robustness compared to TRADES, our results note potential advantages of DATNS in practical deployment scenarios.

German Traffic Sign Recognition Benchmark (GTSRB). We further evaluate DATNS on the German Traffic Sign Recognition Benchmark (GTSRB) dataset, which is a challeng-

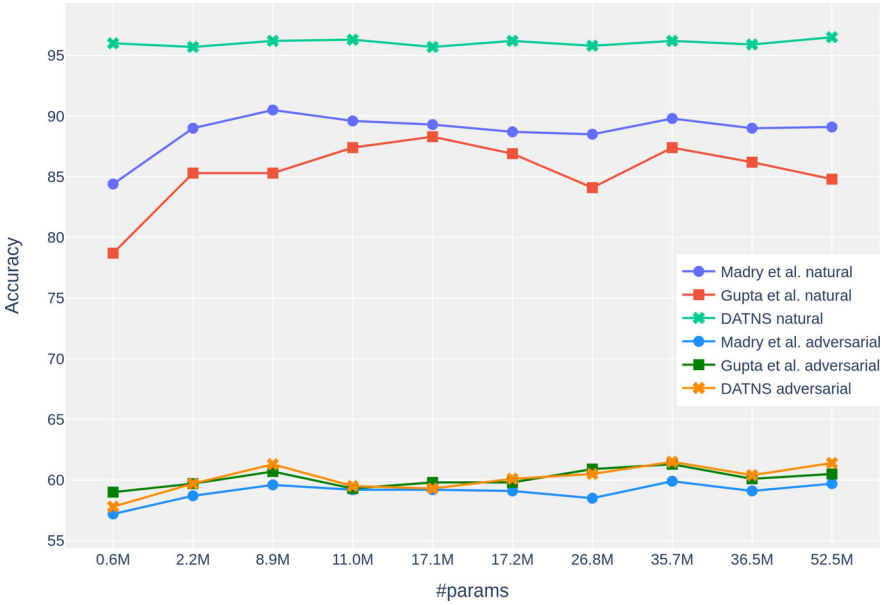

Fig. 5 Natural and adversarial accuracy of DATNS, [29] and [8] for the Wide ResNets of Table 3 (GTSRB), sorted by total number of parameters

ing dataset for traffic sign recognition that includes 43 different classes. GTSRB consists of 39,000 training images and 12,000 test images, preprocessed for our experiments to a resolution of 32x32 pixels. Unlike CIFAR-10, which contains more generic categories, GTSRB involves the classification of traffic signs that often appear in diverse environmental conditions, such as varying lighting and background, making it a suitable benchmark for testing the robustness of AT methods.

Table 3 presents the results of [8], [29] and DATNS on GTSRB for the same set of Wide ResNets. DATNS consistently outperforms [8], as well as [29] in terms of natural accuracy for all experiment instances by an average of approximately 10%. Regarding adversarial accuracy, DATNS either outperforms the compared methods in most cases or achieves similar accuracy with [29]. Notably, these performance gains are obtained with faster training, as DATNS trains the baselines on adversarial data for only a quarter of the total number of epochs. Fig. 5 visualizes the performance of [8], [29] and DATNS with respect to the total number of model parameters. DATNS consistently achieves higher natural accuracy and nearly equal adversarial accuracy, demonstrating its strength in training large capacity models faster while maintaining competitive performance.

4.3.2 DATNS and Catastrophic Forgetting

In this paragraph, we discuss the connection between catastrophic forgetting and the new variations of AT. Regular AT involves training models on a specific task from start to finish. However, in the case of AT starting from a posterior epoch [10, 29], models are trained incrementally on two related but divergent tasks. Initially, models learn to classify clean samples during the classification task. However, when AT begins at the switching epoch (s_0), models start learning to classify both clean and adversarial samples. This transition from

natural to adversarial training raises concerns about catastrophic forgetting, highlighting the need to evaluate these AT variations in terms of their impact on mitigating catastrophic forgetting.

Due to the multiple switches between natural and adversarial training, DATNS bears similarities to multitask and transfer learning. Multitask learning involves training models to solve multiple tasks simultaneously, leveraging knowledge-based inductive bias during training. It has been argued that multitask learning complements AT, enhancing the natural and adversarial performance of single-task models trained using state-of-the-art adversarial techniques [47]. Transfer learning, on the other hand, involves training a classifier on a single task and then using the acquired knowledge as a starting point for learning a related task. AT on the source data generates improved representations, leading to more accurate prediction when fine-tuning on the target data [48]. Exactly like in these studies, where joint training on related tasks enhances robustness, we suggest that DATNS follows a similar approach. Based on our experimental results, DATNS is found to be less sensitive to catastrophic forgetting.

As evidenced in Fig. 6, while DATNS demonstrates lower training adversarial accuracy compared to [29], it achieves better performance in both natural and adversarial test accuracies. This disparity highlights DATNS's robust generalization capabilities to unseen data on the two related tasks. Moreover, Fig. 6 highlights the significant impact of the alternating training process with a time interval of $t = 2$ on the model's generalization ability. Gupta's method [29], which trains models adversarially after s_0 without any switches back to natural training, potentially leads to overfitting on adversarial data post-s0 until the end of training, a challenge that DATNS confronts with frequent switches over the two tasks.

Additionally, when employing smaller time interval values, such as $t = 2$, the model experiences more frequent switches between natural and adversarial training. This frequent alternation enables continual adaptation and refinement of the model's representations to accommodate both clean and adversarially perturbed samples. Consequently, the model is less prone to overfitting specific features of either task and is better equipped to generalize to unseen data, thus reducing catastrophic forgetting. This concept is illustrated in Figures 1-3, showcasing a tendency for adversarial accuracy to decrease as the time interval t increases, emphasizing the significance of balanced exposure to both clean and adversarial samples in reducing catastrophic forgetting.

4.3.3 DATNS and Loss Gradients Interpretability

In this paragraph, we provide a qualitative analysis on the interpretability of loss gradients in adversarially trained models, specifically focusing on the discussed AT variations where models are trained on both natural and adversarial samples. To investigate this, we compare the interpretability of loss gradients in a Wide ResNet 28x10 (WRN) trained on CIFAR-10 using four different methods:

i) **Standard training**: The WRN is trained solely on natural samples, achieving a standard test accuracy of approximately 96%.
ii) **Madry AT**: The WRN is trained using the AT framework proposed in [8]. The model is trained adversarially from $s_0 = 0$.
iii) **Gupta AT**: The WRN is trained using the AT variation proposed by [29], starting with natural samples only until epoch $s_0 = 100$ after which the model is trained adversarially.
iv) **DATNS AT**: The WRN is trained using our method, which involves begining training only on natural samples until $s_0 = 100$. After $s_0 = 100$, the model is trained naturally and adversarially alternately until the end of training.

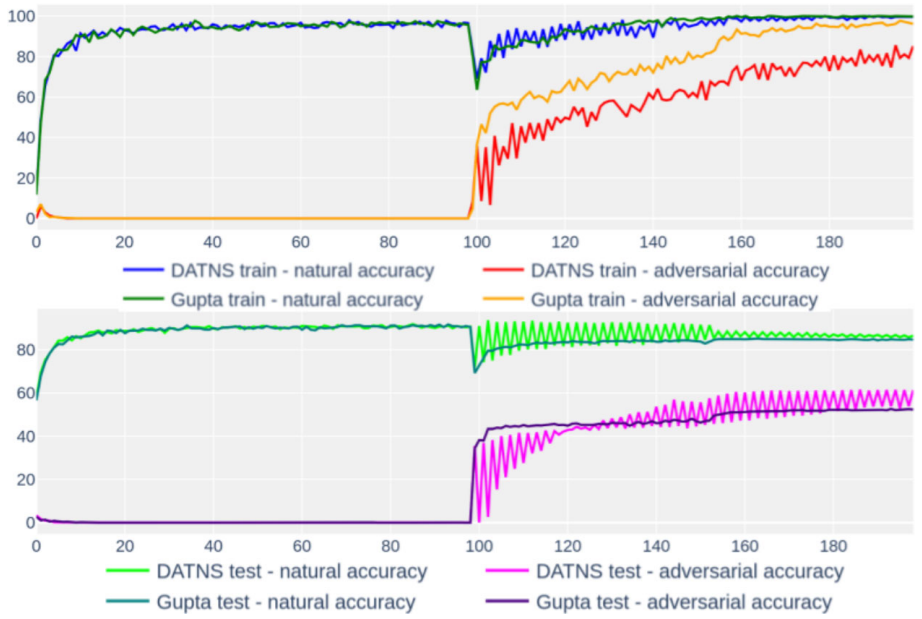

Fig. 6 Comparison of training and test natural and adversarial accuracy of a Wide ResNet 28x10 trained adversarially on CIFAR-10 with [29] and DATNS. Top: Natural and adversarial accuracy during training. While both methods exhibit similar natural accuracy trends, [29] achieves better adversarial accuracy. Bottom: Test results show DATNS surpasses [29] in both natural and adversarial accuracy, implying better generalization. Fluctuations in DATNS accuracy indicate alternating between natural and adversarial epochs

To assess the interpretability of loss gradients, we employ PyTorch and matplotlib. We start by loading a batch of images from the CIFAR-10 test set and subjecting them to the PGD attack. Next, we compute the gradients of the loss function with respect to the input (clean and perturbed) images using the `backward()` function for each model. We visualize the maximum absolute value of the gradients across the three color channels using the `jet` color map.

By comparing the gradients of both natural and adversarial samples for all models, we gain insights into how each model makes predictions and the factors that contribute to its performance. This comparison allows us to understand how AT, but also how the alternating process introduced by DATNS affects the interpretability of the model's behavior.

To provide a comprehensive evaluation, we visualize the loss gradients for the entire CIFAR-10 test set. While this analysis may not represent every test sample, we summarize our conclusions based on a significant portion of the data, referencing the loss gradient visualizations shown in Fig. 7. The `jet` color map is used, where cooler colors (blue or green) indicate small gradient values and warmer colors (yellow or red) represent large gradient values. Cooler colors suggest regions where the loss function changes slowly with respect to the input, indicating lower sensitivity of the model in those regions. Warmer colors indicate regions where the loss function changes rapidly, suggesting areas of greater attention during prediction. The observations are as follows:

Standard training: The model trained solely on natural samples exhibits smoother, less noisy gradients for natural images. However, the gradient patterns for perturbed images appear more unstructured, making their visualizations less interpretable. Additionally, larger

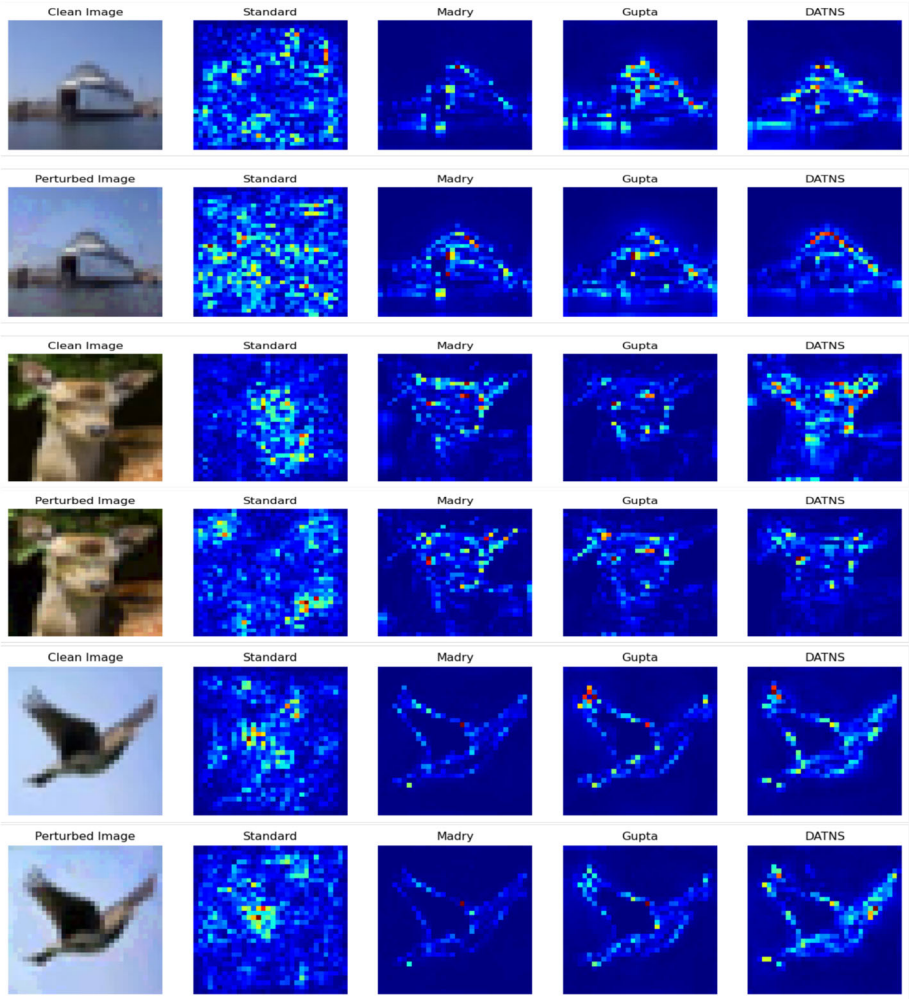

Fig. 7 Selected examples from the CIFAR-10 test set and the loss gradients for a Wide ResNet 28x10 trained with Standard training (natural samples only), [8] (AT beginning from $s_0 = 0$), [29] (both standard and adversarial training, beginning AT from $s_0 = 100$ until the end of the training) and DATNS (both standard and adversarial training, beginning AT from $s_0 = 100$ and alternating between natural and adversarial training until the end) respectfully

gradient values seem to be associated with the texture or color of the pixels rather than the shape or edges of the object. For example, despite achieving 96% standard accuracy, the loss gradients for the "ship" class indicate larger values for pixels related to the background (sky, sea).

Madry AT and **Gupta AT**: Both methods produce more structured and self-explanatory loss gradients for clean and perturbed inputs. The heatmaps clearly differentiate between the background pixels and those related to the foreground object. The gradient magnitude is highest in regions depicting the object's outline. The model trained with Gupta's method tends to have more interpretable gradients compared to the one trained with Madry's method.

While the object's shape and edges receive larger gradient values, there are occasional slightly larger gradient values for pixels related to the outline or details inside the object compared to Madry's method.

DATNS AT: Despite training for fewer adversarial epochs than Madry AT and Gupta AT (reducing the adversarial epochs by 3/4 and 1/2, respectively), DATNS maintains the property of producing interpretable gradients for both clean and perturbed images. The loss gradients are more structured, with higher magnitudes in regions associated with the object's outline and shape. The foreground and background pixels are more distinguishable, resulting in more interpretable visualizations. Moreover, compared to Madry AT and Gupta AT, DATNS tends to exhibit higher gradient magnitudes for details inside the object outline. While we cannot claim that DATNS always produces more interpretable loss gradients than Madry AT or Gupta AT, we can conclude that DATNS tends to focus more on the overall object (shape and details) for many test samples.

These findings suggest that DATNS maintains the property of producing interpretable loss gradients while reducing the number of adversarial epochs, enhancing both robustness and interpretability.

4.4 Discussion and Limitations

The improved performance of DATNS in AT, as evidenced by our comparative analysis, can be attributed to several advantages inherent in its design. DATNS alternates between adversarial and clean training epochs, allowing the model to periodically recalibrate its understanding of clean data while also defending against adversarial perturbations. This approach prevents the model from overfitting to adversarial examples, maintaining a balanced representation of both clean and adversarial data. In contrast, traditional PGD-based AT exposes the model to adversarial examples in every epoch, which can lead to overfitting and a subsequent decrease in clean accuracy.

By alternating between these training modes, DATNS retains strong generalization capabilities while enhancing robustness against adversarial attacks. This can result in higher adversarial accuracy compared to methods that rely solely on continuous adversarial training, as DATNS is found to mitigate adversarial overfitting more effectively.

In addition to its robustness benefits, DATNS offers significant computational advantages. Its design requires fewer epochs dedicated to adversarial examples, thereby reducing the overall training time compared to PGD-based methods. This makes DATNS particularly suitable for scenarios with limited computational resources, where achieving a minimum level of adversarial robustness is crucial without excessive resource expenditure. The flexibility of DATNS allows for adjustments in the training strategy, such as starting AT earlier or tailoring the intervals between adversarial epochs to specific needs, making it adaptable to various resource constraints.

However, some limitations of DATNS should be acknowledged as well. While DATNS has demonstrated strong performance, particularly with Wide ResNet architectures, its effectiveness may not generalize across all model types. This study focused on evaluating DATNS across a diverse range of model architectures commonly used as baselines in adversarial training research. Consequently, our findings can may not be generalized to other computer vision or machine learning architectures that were not assessed within this study's scope. Furthermore, while DATNS addresses some of the computational challenges associated with AT, it may not fully mitigate robustness issues against more sophisticated or novel attack methods beyond those tested in this study.

In practice, the choice of AT method should consider the specific context and priorities. For instance, if natural accuracy is the primary concern, DATNS is a reliable solution. However, in applications where adversarial accuracy is critical-such as in human-in-the-loop systems where natural samples may be more easily recognized-methods such as [8] or [25], which consistently achieve high adversarial accuracy, may be more suitable, provided that computational overhead is not a constraint.

These observations underscore the need for continued research to enhance DATNS's applicability across a broader range of models and attack scenarios. Future work should focus on optimizing DATNS to further improve its robustness, generalization, and adaptability, ensuring it can meet the demands of diverse real-world applications.

5 Conclusions

In this work, we explored conditions that can enable robust AT with a reduced computational overhead. We extended our previous work with new experiments and provided an empirical analysis regarding the pattern of adversarial data appearance during training and discovered that robust AT is facilitated when natural and adversarial training occur alternately. By experimenting with various time intervals for injecting adversarial data in the training process, we observed that DATNS excels when adversarial data is presented non-sequentially and not at the full length of AT. Particularly, we found that the robustness of wide deep learning architectures correlates with their number of parameters, emphasizing the importance of architecture selection in AT.

Moreover, we argued in this work that variations of AT, which introduce adversarial data from a posterior epoch, should be evaluated based on their ability to minimize catastrophic forgetting. Models trained with such AT variations should undergo comprehensive analysis to ensure they maintain their performance on previously learned tasks while adapting to new adversarial challenges.

Our experimental results demonstrated that DATNS achieves superior performance in terms of natural and adversarial accuracy within a reduced training time, while preserving the interpretability of adversarially trained models. The loss gradients produced by DATNS exhibited more structure and higher magnitudes in regions relevant to the object's shape and details. This property of DATNS enhances its ability to focus on important features and generalize better on unseen data.

Our work highlights the potential of non-sequential AT variations, combined with principles from widely known DL techniques such as transfer learning, multitask learning, and meta-learning, to develop robust models at a lower computational cost. We hope that our work stimulates further research on AT methods aimed at reducing the computational overhead and advancing the robustness of computer vision models against adversarial attacks in real-world systems.

Acknowledgements This work was partially supported by the EU funded project KINAITICS (Grant Agreement Number 101070176).

Author Contributions Efi Kafali contributed in the conceptualisation and design of the study and experiments, the implementation of experiments, interpretation and evaluation of the results, and the writing of the manuscript. Theodoros Semertzidis contributed in the conceptualisation and design of the study and experiments, interpretation and evaluation of results, proof reading and securing funding for the research. Petros Daras contributed to the evaluation of results, proof reading the article and securing funding for the research. All authors reviewed the manuscript.

Funding Open access funding provided by HEAL-Link Greece. Partial financial support was received from the EU funded project KINAITICS (Grant Agreement Number 101070176).

Data Availability The datasets analysed in the study are publicly available from their corresponding authors at https://www.cs.toronto.edu/~kriz/cifar.html.

Code Availability The code for this study is publicly available at https://github.com/efkaf/DATNS.

Declarations

Conflict of interests Efi Kafali, Theodoros Semertzidis and Petros Daras have no competing interests to declare that are relevant to the content of this article.

Ethical approval Not applicable

Consent to participate Not applicable

Consent for publication All authors consent for the publication of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. arxiv 2013. arXiv preprint arXiv:1312.6199
2. Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572
3. Nguyen A, Yosinski J, Clune J (2015) Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 427–436
4. Moosavi-Dezfooli S-M, Fawzi A, Frossard P (2016) 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
5. Shafahi A, Najibi M, Ghiasi MA, Xu Z, Dickerson J, Studer C, Davis LS, Taylor G, Goldstein T (2019) Adversarial training for free! In: Advances in Neural Information Processing Systems, pp 3358–3369
6. Wang Y, Ma X, Bailey J, Yi J, Zhou B, Gu Q (2019) On the convergence and robustness of adversarial training. In: ICML 1:2
7. Duesterwald E, Murthi A, Venkataraman G, Sinn M, Vijaykeerthy D (2019) Exploring the hyperparameter landscape of adversarial robustness. arXiv preprint arXiv:1905.03837
8. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2017) Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083
9. Zhou X, Tsang IW, Yin J (2022) Latent boundary-guided adversarial training. arXiv preprint arXiv:2206.03717
10. Kafali E, Semertzidis T, Daras P (2021) Delayed adversarial training with non-sequential adversarial epochs. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), pp 1363–1367 . IEEE
11. Boufssasse A, Hssayni Eh, Joudar N-E, Ettaouil M (2023) A multi-objective optimization model for redundancy reduction in convolutional neural networks. Neural Process Lett. 55(7):9721–9741

12. Hssayni EH, Joudar N-E, Ettaouil M (2022) A deep learning framework for time series classification using normal cloud representation and convolutional neural network optimization. Comput Intell 38(6):2056–2074
13. Joudar N-E, Ettaouil M et al (2022) An adaptive drop method for deep neural networks regularization: estimation of dropconnect hyperparameter using generalization gap. Knowl-Based Syst 253:109567
14. Silva SH, Najafirad P (2020) Opportunities and challenges in deep learning adversarial robustness: a survey. arXiv preprint arXiv:2007.00753
15. Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P (2017) Ensemble adversarial training: attacks and defenses. arXiv preprint arXiv:1705.07204
16. Liu X, Cheng M, Zhang H, Hsieh C-J (2018) Towards robust neural networks via random self-ensemble. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 369–385
17. Sen S, Ravindran B, Raghunathan A (2020) Empir: ensembles of mixed precision deep networks for increased robustness against adversarial attacks. arXiv preprint arXiv:2004.10162
18. Wong E, Rice L, Kolter JZ (2020) Fast is better than free: revisiting adversarial training. arXiv preprint arXiv:2001.03994
19. Andriushchenko M, Flammarion N (2020) Understanding and improving fast adversarial training. arXiv preprint arXiv:2007.02617
20. Croce F, Hein M (2020) Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. arXiv preprint arXiv:2003.01690
21. Shaham U, Yamada Y, Negahban S (2015) Understanding adversarial training: increasing local stability of neural nets through robust optimization. arXiv preprint arXiv:1511.05432
22. Akhtar N, Mian A, Kardan N, Shah M (2021) Advances in adversarial attacks and defenses in computer vision: a survey. IEEE Access 9:155161–155196. https://doi.org/10.1109/ACCESS.2021.3127960
23. Gao R, Cai T, Li H, Hsieh C-J, Wang L, Lee JD (2019) Convergence of adversarial training in over-parametrized neural networks. In: Advances in Neural Information Processing Systems, pp 13029–13040
24. Wu B, Chen J, Cai D, He X, Gu Q (2021) Do wider neural networks really help adversarial robustness? Adv Neural Inform Process Syst 34:7054–7067
25. 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, pp 7472–7482 . PMLR
26. Zheng H, Zhang Z, Gu J, Lee H, Prakash A (2020) Efficient adversarial training with transferable adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1181–1190
27. Cai Q-Z, Du M, Liu C, Song D (2018) Curriculum adversarial training. arXiv preprint arXiv:1805.04807
28. Mendonça MO, Maroto J, Frossard P, Diniz PS (2022) Adversarial training with informed data selection. In: 2022 30th European Signal Processing Conference (EUSIPCO), pp 608–612 . IEEE
29. Gupta S, Dube P, Verma A (2020) Improving the affordability of robustness training for dnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 780–781
30. Hihn H, Braun DA (2023) Hierarchically structured task-agnostic continual learning. Mach Learn 112(2):655–686
31. Lee S-W, Kim J-H, Jun J, Ha J-W, Zhang B-T (2017) Overcoming catastrophic forgetting by incremental moment matching. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. https://proceedings.neurips.cc/paper/2017/file/f708f064faaf32a43e4d3c784e6af9ea-Paper.pdf
32. Hasselmo ME (2017) Avoiding catastrophic forgetting. Trends Cognit Sci 21(6):407–408. https://doi.org/10.1016/j.tics.2017.04.001
33. Hurtado J, Lobel H, Soto A (2021) Overcoming catastrophic forgetting using sparse coding and meta learning. IEEE Access 9:88279–88290. https://doi.org/10.1109/ACCESS.2021.3090672
34. Kim H, Lee W, Lee J (2021) Understanding catastrophic overfitting in single-step adversarial training. Proc AAAI Conf Artif Intell 35:8119–8127
35. Li T, Wu Y, Chen S, Fang K, Huang X (2022) Subspace adversarial training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 13409–13418
36. Chen S, Shen H, Wang R, Wang X (2022) Towards improving fast adversarial training in multi-exit network. Neural Netw 150:1–11
37. Wen S, Itti L (2018) Overcoming catastrophic forgetting problem by weight consolidation and long-term memory. arXiv preprint arXiv:1805.07441
38. Dong X, Luu AT, Lin M, Yan S, Zhang H (2021) How should pre-trained language models be fine-tuned towards adversarial robustness? Adv Neural Inform Process Syst 34:4356–4369
39. Yao X, Huang T, Wu C, Zhang R-X, Sun L (2019) Adversarial feature alignment: avoid catastrophic forgetting in incremental task lifelong learning. Neural Comput 31(11):2266–2291

40. Fang K, Tao Q, Wu Y, Li T, Cai J, Cai F, Huang X, Yang J (2024) Towards robust neural networks via orthogonal diversity. Pattern Recognit 149:110281
41. Tsipras D, Santurkar S, Engstrom L, Turner A, Madry A (2019) Robustness may be at odds with accuracy. In: International Conference on Learning Representations
42. Etmann C, Lunz S, Maass P, Schoenlieb C (2019) On the connection between adversarial robustness and saliency map interpretability. In: International Conference on Machine Learning, pp 1823–1832 . PMLR
43. Kim B, Seo J, Jeon T (2019) Bridging adversarial robustness and gradient interpretability. arXiv:1903.11626
44. Nielsen IE, Dera D, Rasool G, Ramachandran RP, Bouaynaya NC (2022) Robust explainability: a tutorial on gradient-based attribution methods for deep neural networks. IEEE Signal Process Mag 39(4):73–84
45. Zhang T, Zhu Z (2019) Interpreting adversarially trained convolutional neural networks. In: International Conference on Machine Learning, pp 7502–7511 . PMLR
46. Gavrikov P, Keuper J, Keuper M (2023) An extended study of human-like behavior under adversarial training. arXiv preprint arXiv:2303.12669
47. Mao C, Gupta A, Nitin V, Ray B, Song S, Yang J, Vondrick C (2020) Multitask learning strengthens adversarial robustness. In: European Conference on Computer Vision, pp 158–174 . Springer
48. Deng Z, Zhang L, Vodrahalli K, Kawaguchi K, Zou JY (2021) Adversarial training helps transfer learning via better representations. Adv Neural Inform Process Syst 34:25179–25191

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.