GAME THEORETIC MIXED EXPERTS FOR COMBINATIONAL ADVERSARIAL MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in adversarial machine learning have shown that defenses considered to be robust are actually susceptible to adversarial attacks which are specifically tailored to target their weaknesses. These defenses include Barrage of Random Transforms (BaRT), Friendly Adversarial Training (FAT), Trash is Treasure (TiT) and ensemble models made up of Vision Transformers (ViTs), Big Transfer models and Spiking Neural Networks (SNNs). It remains an open question, however, as to whether the adversarial examples designed to target one defense will be similarly misclassified by another defense. In this paper, we provide the first adversarial defense transferability study, as well as a game theoretic framework for ensemble adversarial attacks and defenses. Our framework is called Game theoretic Mixed Experts (GaME) and is designed to find the Mixed-Nash strategy for an attacker that can employ compositional adversarial attacks. We show that this framework creates an ensemble of defenses with greater robustness than a combinational defense with a uniform or random probability distribution. Overall, our framework and analyses advance the field of adversarial machine learning by yielding new insights into compositional attack and defense formulations.

1 INTRODUCTION

Machine learning models have been shown to be vulnerable to adversarial examples Goodfellow et al. (2014); Papernot et al. (2016). Adversarial examples are inputs with small perturbations added, such that machine learning models misclassify the example with high confidence. Addressing the security risks posed by adversarial examples are critical for the safe deployment of machine learning in areas like health care Finlayson et al. (2019) and self driving vehicles Qayyum et al. (2020). However, current defenses and attacks in adversarial machine learning have trended towards a cat and mouse dynamic where in new defenses are continually being proposed and then broken Carlini & Wagner (2017); Tramer et al. (2020); Mahmood et al. (2021a); Sitawarin et al. (2022) by improved attacks.

In parallel to attack and defense development, studies have also been conducted on the transferability of adversarial examples Liu et al. (2016); Mahmood et al. (2021b); Xu et al. (2022). Transferabiltiy refers to the phenomena where adversarial examples generated for one model are also misclassified by a different machine learning model. However, to the best of our knowledge no analyses have been done on the transferability of adversarial examples designed to attack specific defenses. From these observations several pertinent questions arise:

1. *Do adversarial examples generated for one specific defense transfer to other defenses?*
2. *Based on adversarial transferability, can a game theoretic framework be developed to determine the optimal choices for both attacker and defender?*
3. *Can randomized defense selection yield higher robustness than a single state-of-the-art defense?*

These are precisely the questions our paper seeks to answer. We break from the traditional dynamic of adversarial machine learning which focuses on the single best attack and defense. We instead take a multi-faceted approach and develop a game theoretic framework to answer the above questions. Specifically, we provide the following contributions: On the attack side we develop two new

white-box attacks called the Momentum Iterative Method over Expectation (MIME) and the Auto Expectation Self-Attention Gradient Attack (AE-SAGA). These attacks are necessary for targeting certain randomized defenses and for adapting to multi-defense strategies. We analyze the adversarial transferability of current defenses like Trash is Treasure Xiao & Zheng (2020), Barrage of Random Transforms Raff et al. (2019), Friendly Adversarial Training Zhang et al. (2020) and other new architectures like SNNs Rathi & Roy (2021b); Fang et al. (2021) and ViTs Dosovitskiy et al. (2020). Lastly, and most importantly, we formulate a practical, game-theoretic framework for finding the optimal strategies for an attacker and defender who each employ a set of state-of-the-art adversarial attacks and defenses.

2 Adversarial Machine Learning Defenses

Here we summarize the state-of-the-art defenses we analyze in this paper. In the following subsections, we give an overview of each defense and our reasons for choosing said defense. It is important to note our analyses encompass a broad range of different defenses, including ones based on randomization, ones based on adversarial training and ones based on exploiting model transferability. In addition, we also consider diverse architectures including Big Transfer models (BiTs), Vision Transformers (ViTs) and Spiking Neural Networks (SNNs). Despite our broad range, we do not attempt to test every novel adversarial defense. It is simply infeasible to test every proposed adversarial machine learning defense, as new defenses are constantly being produced. However, based on our game theoretic design and open source code (which will be provided upon publication), any new defense can easily be tested and integrated into our proposed framework.

2.1 Barrage of Random Transforms

Barrage of Random Transforms (BaRT) Raff et al. (2019) utilize a set of image transformations in a random order and with randomized transformation parameters to thwart adversarial attacks. Let $t^i_j(x)$ represent the ith transformation used in the jth order in the sequence. A BaRT defense using n image transformations randomly alters the input x:

$$t(x) = t^{\omega_n}_{\mu_n} \circ t^{\omega_{n-1}}_{\mu_{n-1}} \circ ... \circ t^{\omega_1}_{\mu_1}(x) \tag{1}$$

where ω represents the subset of n transformations randomly selected from a set of N total possible transformations and μ represents the randomized order in which the n transformations are applied. In Equation 1 the parameters of each image transformation $t^i_j(x)$ are also randomized at run time, further adding to the stochastic nature of the defense. In this paper, we work with the original BaRT implementation which includes both differentiable and non-differentiable image transformations.

Why we selected it: Many defenses are broken soon after being proposed Tramer et al. (2020). BaRT is one of the few defenses that has continued to show robustness even when attacks are specifically tailored to work against it. For example, most recently BaRT achieves 29% robustness on CIFAR-10 against a customized white-box attack Sitawarin et al. (2022). It remains an open question whether using using BaRT with other randomized approaches (i.e. selecting between different defenses) can yield even greater robustness.

2.2 Friendly Adversarial Training

Training classifiers to correctly recognize adversarial examples was originally proposed in Goodfellow et al. (2014) using FGSM. This concept was later expanded to including training on adversarial examples generated by PGD in Madry et al. (2018). In Zhang et al. (2020) it was shown that Friendly Adversarial Training (FAT) could achieve high clean accuracy while maintaining robustness to adversarial examples. This training was accomplished by using a modified version of PGD called PGD-K-τ. In PGD-K-τ, K refers to the number of iterations used for PGD. The τ variable is a hyperparamter used in training which stops the PGD generation of adversarial examples earlier than the normal K number of steps, if the sample is already misclassified.

Why we selected it: There are many different defenses that rely on adversarial training Madry et al. (2018); Zhang et al. (2019); Wang et al. (2019); Maini et al. (2020) and training and testing them all is not computationally feasible. We selected FAT for its good trade off between clean accuracy and robustness, and because we wanted to test adversarial training on both Vision Transformer and

CNN models. In this regard, FAT is one of the adversarial training methods that has already been demonstrated to work across both types of architectures Mahmood et al. (2021b).

2.3 TRASH IS TREASURE

One early direction in adversarial defense design was model ensembles Pang et al. (2019). However, due to the high transferability of adversarial examples between models, such defenses were shown to not be robust Tramer et al. (2020). Trash is Treasure (TiT) Xiao & Zheng (2020) is a two model defense that seeks to overcome the transferability issue by training one model $C_a(\cdot)$ on the adversarial examples from another model $C_b(\cdot)$. At run time both models are used:

$$y = C_a(\psi(x, C_b)) \tag{2}$$

where ψ is an adversarial attack done on model C_b with input x and C_a is the classifier that makes the final class label prediction on the adversarial example generated by ψ with C_b.

Why we selected it: TiT is one of the newest defenses that tries to achieve robustness in a way that is fundamentally different than pure randomization strategies or direct adversarial training. In our paper, we further develop two versions of TiT. One version is based on the original proposed CNN-CNN implementation. We also test a second mixed architecture version using Big Transfer model and Vision Transformers to try and leverage the low transferability phenomena described in Mahmood et al. (2021b).

2.4 NOVEL ARCHITECTURES

In addition to adversarial machine learning defenses, we also include several novel architectures that have recently achieved state-of-the-art or near state-of-the-art performance in image recognition tasks. These include the Vision Transformer (ViT) Dosovitskiy et al. (2020) and Big Transfer models (BiT) Kolesnikov et al. (2020). Both of these types of models utilize pre-training on larger datasets and fine tuning on smaller datasets to achieve high fidelity results. We also test Spiking Neural Network (SNNs) architectures. SNNs are a competitor to artificial neural networks that can be described as a linear time invariant system with a network structure that employs non-differentiable activation functions Xu et al. (2022). A major challenge in SNNs has been matching the depth and model complexity of traditional deep learning models. Two approaches have been used to overcome this challenge, the Spiking Element Wise (SEW) ResNet Fang et al. (2021) and transferring weights from existing CNN architectures to SNNs Rathi & Roy (2021a). We experiment with both approaches in our paper.

Why we selected it: The set of adversarial examples used to attack one type of architecture (e.g. a ViT) have shown to not be misclassified by other architecture types (e.g. a BiT or SNN) Mahmood et al. (2021b); Xu et al. (2022). While certain white-box attacks have been used to break multiple undefended models, it remains an open question if different architectures combined with different defenses can yield better performance.

3 ADVERSARIAL ATTACKS

In our paper we assume a white-box adversarial threat model. This means the attacker is aware of the set of all defenses D that the defender may use for prediction. In addition, $\forall d \in D$ the attacker also knows the classifier weights θ_d, architecture and any input image transformations the defense may apply. To generate adversarial examples the attacker solves the following optimization problem:

$$\max_{\delta} L(C(x+\delta), y) \quad \text{subject to: } ||\delta||_p \leq \epsilon \tag{3}$$

where L is the loss function to be maximized (e.g. cross entropy), (x, y) represents the original input with corresponding class label and δ is the adversarial perturbation the attacker adds to the input. The magnitude of this perturbation δ is typically limited by a certain l_p norm. In this paper we analyze the attacks and defenses using the l_∞ norm.

Static white-box attacks such as the Projected Gradient Descent (PGD) Madry et al. (2018) attack often perform poorly against randomized defenses such as BaRT or TiT. In Xiao & Zheng (2020)

they tested the TiT defense against an attack designed to compensate for randomness, the Expectation over Transformation attack (EOT) attack Athalye et al. (2018). However, it was shown that the EOT attack performs poorly against TiT (e.g. 20% or worse attack success rate). For attacking BaRT, in Sitawarin et al. (2022) they proposed a new white-box attack to break BaRT. However, this new attack requires that the image transformations used in the BaRT defense be differentiable, which is a deviation from the original BaRT implementation.

Attack Contributions: We propose two new white-box attacks for handling randomized defenses. The first attack is designed to work on individual randomized defenses and is called the Momentum Iterative Method over Expectation (MIME). To the best of our knowledge, MIME is the first white-box attack to achieve a high attack success rate ($> 70\%$) against TiT. MIME is also capable of achieving a high attack success rate against BaRT, even when non-differentiable transformations are implemented as part of the defense. Our second attack, is designed to generate adversarial examples that work against multiple type of defenses (both randomized and non-randomized defenses) simultaneously. This compositional attack is called, the Auto Expectation Self-Attention Gradient Attack (AE-SAGA).

3.1 Momentum Iterative Method over Expectation

We develop a new attack white-box attack specifically designed to work on defenses that inherently rely on randomization, like Barrage of Random Transforms (BaRT) Raff et al. (2019) and Trash is Treasure (TiT) Xiao & Zheng (2020). Our new attack is called the Momentum Iterative Method over Expectation (MIME). The attack "mimes" the transformations of the defender in order to more precisely model the gradient of the loss function with respect to the input after the transformations are applied. To this end, MIME utilizes two effective aspects from earlier white-box attacks, momentum from the Momentum Iterative Method (MIM) Dong et al. (2018) attack and repeated sampling Athalye et al. (2018) from the Expectation Over Transformation (EOT) attack:

$$x_{adv}^{(i)} = x_{adv}^{(i-1)} + \epsilon_{\text{step}} g^{(i)} \tag{4}$$

where the attack is computed iteratively with $x_{adv}^{(0)} = x$. In Equation 4 $g^{(i)}$ is the momentum based gradient of the loss function with respect to the input at iteration i and is defined as:

$$g^{(i)} := \gamma g^{(i-1)} + \mathbb{E}_{t \sim T}[\frac{\partial L}{\partial t(x_{adv}^{(i)})}] \tag{5}$$

where γ is the momentum decay factor and t is a random transformation function drawn from the defense's transformation distribution T. In Table 1 we show experimental results for the MIME attack on CIFAR-10 randomized defenses (TiT and BaRT). It can clearly be seen that MIME has a better attack success rate than both APGD Croce & Hein (2020) and MIM Dong et al. (2018).

Attack	BaRT-1	BaRT-5	BaRT-10	TIT (BiT/ViT)	TiT (VGG/RN)
MIME-10	**3.18%**	15.5%	43.2%	10.1%	24.9%
MIME-50	4.3%	**8.22%**	**23.2%**	**8.3%**	**23.3%**
MIM	6.7%	39.5%	59.5%	52%	58.9%
APGD	8.9%	47.7%	70.8%	68.2%	40.7%
Clean	98.4%	95.3%	92.5%	90.1%	76.6%

Table 1: Performance of the MIME attack against CIFAR-10 randomized defenses: Trash is Treasure (TiT) and Barrage of Random Transforms (BaRT). In our experiments $\epsilon_{max} = 0.031$ for all attacks. It can clearly be seen that MIME outperforms both APGD and MIM on these two randomized defenses. Further defense and attack implementation details are given in our supplemental material.

3.2 Auto Expectation Self-Attention Gradient Attack

The use of multi-model attacks are necessary to achieve a high attack success rate when dealing with ensembles that contain both CNN and non-CNN model architectures like the Vision Transformer (ViT) Dosovitskiy et al. (2020) and Spiking Neural Network (SNN) Fang et al. (2021). This is because adversarial examples generated by single model white-box attacks generally do not transfer

well between CNNs, ViTs and SNNs Mahmood et al. (2021b); Xu et al. (2022). In addition it is an open question if multi-model attacks can be effective against the current state-of-the-art defenses. In this paper, we expand the idea of a multi-model attack to include not only different architecture types, but also different defenses. The generalized form of the multi-model attack is:

$$\max_{\delta} \sum_{d \in D} L_d(C_d(x+\delta), y) \quad \text{subject to: } ||\delta||_p \leq \epsilon \tag{6}$$

where D is the set of all possible defenses (models) under consideration in the attack, L_d is the loss funcation associated with defense d, $C_d(\cdot)$ is the class label output of defense d and δ is the adversarial perturbation. For a single input x with corresponding class label y, an untargeted multi-model attack is considered successful if $(\forall d \in D, C_d(x+\delta) \neq y) \wedge (||\delta||_p \leq \epsilon)$. One formulation of the multi-model attack is the Auto Self-Attention Gradient Attack (Auto SAGA) which was proposed in Xu et al. (2022) to iteratively attack combinations of ViTs, SNNs and CNNs:

$$x_{adv}^{(i+1)} = x_{adv}^{(i)} + \epsilon_{step} * \text{sign}(G_{blend}(x_{adv}^{(i)})) \tag{7}$$

where ϵ_{step} was the step size used in the attack. In the original formulation of Auto-SAGA, G_{blend} was a weighted average of the gradients of each model $d \in D$. By combining gradient estimates from different models, Auto-SAGA is able to create adversarial examples that are simultaneously misclassified by multiple models. One limitation of Auto-SAGA attack is that it does not account for defenses that utilize random transformations. Motivated by this, we can integrate the previously proposed MIME attack into the gradient calculations for Auto-SAGA. We denote this new attack as the Auto Expectation Self-Attention Gradient Attack (AE-SAGA). Both SAGA and AE-SAGA uses the same iterative update (Equation 7). However, AE-SAGA uses the following gradient estimator:

$$G_{blend}(x_{adv}^{(i)}) = \gamma G_{blend}(x_{adv}^{(i-1)}) + \sum_{k \in D \backslash R} \alpha_k^{(i)} \phi_k^{(i)} \odot \frac{\partial L_k}{\partial x_{adv}^{(i)}} + \sum_{r \in R} \alpha_r^{(i)} \phi_r^{(i)} \odot (\mathbb{E}_{t \sim T}[\frac{\partial L_r}{\partial t(x_{adv}^{(i)})}]) \tag{8}$$

In Equation 8 the two summations represent the gradient contributions of sets $D \backslash R$ and R, respectively. Here we define R as the set of randomized defenses and D as the set of all defenses being modeled by AE-SAGA. In each summation ϕ is the self-attention map Abnar & Zuidema (2020) which is replaced with a matrix of all ones for any defense that does not use ViT models. α_k and α_r are the associated weighting factors for the gradients for each deterministic defense k and randomized defense r, respectively. Details of how the weighting factors are derived are given in Xu et al. (2022).

4 Transferability Experiments

Adversarial transferability refers to the phenomena in which adversarial examples generated to attack one model are also misclassified by a different model. Adversarial transferability studies have been done on a variety of machine learning models Liu et al. (2016); Mahmood et al. (2021b); Xu et al. (2022). However, to the best of our knowledge, adversarial transferability between different state-of-the-art defenses has not been conducted. This transferability property is of significant interest because a lack of transferability between different defenses may indicate a new way to improve adversarial robustness.

In Table 2 we show the different single defenses we analyze in this paper and the best attack on each of them from the set of attacks (MIM Dong et al. (2018), APGD Croce & Hein (2020) and MIME (proposed in this work). In Figure 1 we visually show the transferability results of these attacks for CIFAR-10. We give detailed discussions of these results in our supplementary material and briefly summarize of the key takeaways from these experiments.

In general adversarial examples generated using the best attack on one defense *do not* transfer well to other defenses. For example, only 1% of the adversarial examples generated by the TiT (BiT/ViT) defense transfer to the FAT ViT defense. The average transferability for the 8 different defenses shown in Figure 1 is only 16.71% and there is no single model attack that achieves strong performance (i.e. $> 50\%$) across all defenses. These results in turn motivate the development of a game

Defense	Best Attack	Robust Acc	Clean Acc
BaRT-1	MIME	5.10%	98.40%
BaRT-10	MIME	45.20%	92.50%
TiT (BiT/ViT)	MIME	13.20%	90.10%
TiT (Vgg/RN)	MIME	29.50%	76.60%
ViT FAT ($\tau = 2, K = 10$)	APGD	41.20%	92.36%
RN FAT ($\tau = 10, K = 10$)	APGD	67.80%	81.89%
SNN-BP	APGD	2.50%	81.16%
SNN-T	APGD	0.00%	91.54%

Table 2: Single defense implementations for CIFAR-10 with the corresponding strongest attack on the defense and the clean accuracy of the defense. The robust accuracy is measured using 1000 adversarial examples. The examples are class wise balanced and also correctly recognized by all the defenses in their original clean form. All attacks are done using $\epsilon_{max} = 0.031$. Complete attack and defense implementation details are given in our supplementary material.

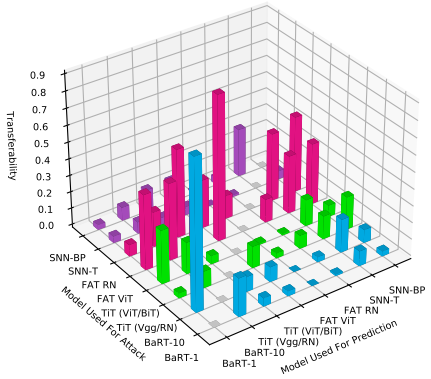

Figure 1: Visual representation of the transferability of adversarial examples between defenses for CIFAR-10. The blue bars represent the transferability of adversarial examples generated using the best attack for BaRT models, the green bars represent adversarial examples generated using the best attack for TiT, the pink bars represent the adversarial examples generated using the best attack for FAT and lastly the purple bars represent the adversarial examples generated using SNNs. Higher bars means more of the adversarial examples are misclassified by other defenses. Full numerical tables used to generate this figure are given in our supplemental material.

theoretic framework from both the attacker and defender. For the attacker, this prompts the need to use multi-model attack like AE-SAGA that was proposed in Section 3, as no single attack (APGD, MIM or MIME) is ubiquitous. For the defender these results highlight the opportunity to increase robustness by randomly selecting between different defenses with different probabilities.

5 Game theoretic Mixed Experts

In this section we derive our framework, Game theoretic Mixed Experts (GaME), for playing the adversarial examples game. In comparison to other works Meunier et al. (2021) Pinot et al. (2020) Pal & Vidal (2020) we take a more discretized approach to playing and solving the adversarial examples game, which ultimately leads to the creation of a finite, tabular, zero-sum game that can be solved using linear programming techniques. In particular, we leverage the state-of-the-art attacks and defenses established in Sections 3 and 4 as "experts" in the adversarial examples game.

5.1 The Adversarial Examples Game

We reformulate the adversarial examples game explored in Meunier et al. (2021). The adversarial examples game is a zero-sum game played between two players: the attacker, p_a, and the defender p_d. Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ the output space of p_d's classifiers, and let Θ represent the

space of classifier parameters. Additionally, let $P_{\epsilon,\mathcal{X}} = \{x \in \mathcal{X} : ||x||_p \leq \epsilon\}$ be the set of valid adversarial perturbations for norm p and $\epsilon \in \mathbb{R}^+$.

Let $A^*_\epsilon = \{(f : \Theta \times \mathcal{X} \times \mathcal{Y} \to P_{\epsilon,\mathcal{X}})\}$ be the set of all valid attack functions. The goal of p_a is to choose $a \in A^*$, which maximizes the expected loss of p_d's classifier, θ, given some pair of input and ground truth label $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The goal of p_d is to minimize this loss through its choice of $\theta \in \Theta$. We can thus formulate the adversarial examples game as a mini-max optimization problem:

$$\inf_{\theta\in\Theta} \sup_{a\in A^*_\epsilon} \mathbb{E}_{(x,y)\sim\mathcal{X}\times\mathcal{Y}}[L(\theta(x + a(\theta, x, y)), y)] \tag{9}$$

Due to the vastness of Θ and A^*_ϵ, solving this optimization problem directly is currently computationally intractable. To this end, in the next subsections we will formulate GaME$_1$ and GaME$_n$ which discretize Θ and A^*_ϵ by enlisting a set of state-of-the-art attacks and defenses.

5.2 GAME$_1$

Let $A' \subset A^*_\epsilon$ be a subset of all valid adversarial attack functions chosen by p_a. Additionally, let $D \subset \Theta$ be a set of defense classifiers chosen by p_d. We further impose that all $a \in A'$ are white-box attacks (see Section 3 for our adversarial threat model) and that A', D are finite, i.e. $|A'| \leq N_a$ and $|D| \leq N_d$ for some $N_a, N_d \in \mathbb{N}$.

It is important to note that each $a \in A'$ is a function of some classifier, $\theta \in \Theta$, in addition to the input and ground truth label $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Due to this it is possible for p_a to choose to attack defense $d \in D$ with attack $a \in A'$, while p_d chooses to evaluate the sample using defense $d' \in D$ where $d \neq d'$. Therefore, for convenience, we will define a new, more general set of attack strategies for p_a:

$$A \subseteq \{(f : \mathcal{X} \times \mathcal{Y} \to P_{\epsilon,\mathcal{X}}) : f(x, y) = a_i(U, x, y),\ a_i \in A',\ U \subseteq D\} \tag{10}$$

where we extend the definition of $A' \subseteq A^*_\epsilon$ to attack functions that can take subset of defense parameters $U \subseteq D$ as input. This comes into play with multi-model attacks like AE-SAGA. Thus we will let D be the strategy set of p_d, and A be the strategy set of p_a. We can then reformulate a discretized version of the adversarial examples game as follows:

$$\min_{d\in D} \max_{a\in A} \mathbb{E}_{(x,y)\sim\mathcal{X}\times\mathcal{Y}}[L(d(x + a(x, y)), y)] \tag{11}$$

In the above two formulations we optimize over the set of pure strategies for the attacker and defender. However, as previously explored in Araujo et al. (2020) Meunier et al. (2021), limiting ourselves to pure strategies severely inhibits the strength of both the attacker and defender. Thus we create the following mixed strategy vectors for p_a, p_d:

$$\lambda^A \in \{r \in [0, 1]^{|A|} : ||r||_1 = 1\}, \quad \lambda^D \in \{r \in [0, 1]^{|D|} : ||r||_1 = 1\} \tag{12}$$

here λ^A and λ^D represent the mixed strategies of p_a and p_d respectively. Let each $a_i \in A$ and $d_i \in D$ correspond to the ith elements of λ^A and λ^D, λ^A_i and λ^D_i, respectively. Formally we can say:

$$\mathbb{P}(\{a_i \in A : a = a_i\}) = \lambda^A_i, \quad \mathbb{P}(\{d_i \in D : d = d_i\}) = \lambda^D_i \tag{13}$$

where $a \in A$ and $d \in D$ are random variables. With these mixed strategy vectors we can then reformulate the adversarial examples game as a mini-max optimization problem over p_d's choice of λ^D and p_a's choice of λ^A:

$$\begin{aligned} &\min_{\lambda^D} \max_{\lambda^A} \mathbb{E}_{(x,y)\sim\mathcal{X}\times\mathcal{Y}}[\mathbb{E}_{(a,d)\sim A\times D}[L(d(x + a(x, y)), y)]] \\ &= \min_{\lambda^D} \max_{\lambda^A} \mathbb{E}_{(x,y)\sim\mathcal{X}\times\mathcal{Y}}[\sum_{a_i\in A} \lambda^A_i \sum_{d_i\in D} \lambda^D_i [L(d_i(x + a_i(x, y)), y)]] \end{aligned} \tag{14}$$

For sufficiently complex $\mathcal{X}$, D, and A solving the above optimization problem is currently computationally intractable. Thus we can instead approximate the mini-max optimization by taking N Monte-Carlo samples with respect to $(x_j, y_j) \in \mathcal{X} \times \mathcal{Y}$:

$$\min_{\lambda^D} \max_{\lambda^A} \frac{1}{N} \sum_{j=0}^{N} \sum_{a_i\in A} \lambda^A_i \sum_{d_i\in D} \lambda^D_i [L(d_i(x_j + a_i(x_j, y_j)), y_j)] \tag{15}$$

For convenience we will denote $r_{d_i,a_i} = \frac{1}{N}\sum_{j=0}^{N}[L(d_i(x_j + a_i(x_j, y_j)), y_j)]$. Colloquially r_{d_i,a_i} represents the expected robustness of defense d_i when evaluating adversarial samples generated by attack a_i. From a game theoretic perspective, $r_{d,a}$ is the payoff for p_d when they play strategy d and p_a plays strategy a. The payoff for p_a given strategies d, a is $-r_{d,a}$. Our mini-max optimization problem can then be simplified to:

$$\min_{\lambda^D} \max_{\lambda^A} \sum_{a_i \in A} \lambda_i^A \sum_{d_i \in D} \lambda_i^D r_{d_i,a_i} \tag{16}$$

Using all of this we can create a finite, tabular, zero-sum game defined by the following game-frame in strategic form:

$$\langle \{p_a, p_d\},\ (A, D),\ O,\ f \rangle \tag{17}$$

where $O = \{r_{d,a}\ \forall a \in A,\ d \in D\}$ and f is a function $f : A \times D \to O$ defined by $f(d, a) = r_{d,a}$. Since this is a finite, tabular, zero-sum game, we know that it must have a Nash-Equilibrium as proven in Nash (1951). Let R be the payoff matrix for p_d where $R_{d,a} = r_{d,a}$. It then becomes the goal of p_d to maximize their guaranteed, expected payoff. Formally, p_d must solve the following optimization problem:

$$\max_{r^*;\lambda^D} r^* \quad \text{subject to } \lambda^D R \geq (r^*, \cdots r^*) \text{ and } ||\lambda^D||_1 \leq 1 \tag{18}$$

This optimization problem is a linear program, the explicit form of which we provide in the supplemental material. All linear programs have a dual problem, in this case the dual problem finds a mixed Nash strategy for p_a. This can be done by changing the problem to a minimization problem and trasposing R. In the interest of space we give the explicit form of the dual problem in the supplemental material. These linear programs can be solved using weakly polynomial time algorithms like the interior point method Karmarkar (1984), or those developed in Vaidya (1989).

5.3 GAME$_n$

GaME$_n$ is a more general family of games of which GaME$_1$ is a special case. In GaME$_n$, for $n > 1$, p_d can calculate their final prediction based upon the output logits of multiple $d \in D$ evaluated on the same input x. In order to do this, p_d must also choose a function to map the output of multiple defenses to a final prediction. Formally, the strategy set of p_d becomes $D = D' \times F$, where F is a set of prediction functions and D' is defined as follows.

$$D' \subseteq \{U : U \subseteq D,\ |U| \leq n\} \tag{19}$$

Multi-model prediction functions can increase the overall robustness of an ensemble by requiring an adversarial sample to be misclassified by multiple models simultaneously Mahmood et al. (2022). In this paper we will focus on two voting functions: the majority vote Raff et al. (2019), and the largest softmax probability vote Sitawarin et al. (2022). We will refer to these as F_h and F_s respectively:

$$F_h(x, U) = \arg\max_{y \in \mathcal{Y}} \sum_{d \in U} \mathbb{1}\{y = \arg\max_{j \in \mathcal{Y}} d_j(x)\} \tag{20}$$

$$F_s(x, U) = \arg\max_{y \in \mathcal{Y}} \frac{1}{|U|} \sum_{d \in U} \sigma(d(x)) \tag{21}$$

where $\mathbb{1}$ is the indicator function, σ is the softmax function, and $d_j(x)$ represents the jth output logit of defense d when evaluated on input x. Solving GaME$_n$ is the same as solving GaME$_1$ in the previous subsection, except now $|D|$ will be larger. Notationally the mini-max optimization problem in terms of $r_{d,a}$ remains the same as we have simply redefined D, however we can redefine $r_{d,a}$ as follows:

$$r_{(U_i,f_i),a_i} = \sum_{j=0}^{N} [L(f_i(U, (x_j + a_i(x_j, y_j))), y_j)] \tag{22}$$

where $(U_i, f_i) \in D = D' \times F$. While $|D|$ is often much larger in GaME$_n$ than in GaME$_1$, solving an instance of GaME$_1$ often easily lends itself to solving additional instances of GaME$_n$. One can first choose a static subset $S \subset \mathcal{X} \times \mathcal{Y}$ and set of attacks A, then calculate the following set of adversarial examples $X^{adv} = \{(a(x, y), y) : a \in A,\ (x, y) \in S\}$ one time. Finally, the payoff table computed by calculating $R_{a,d} = \frac{1}{|S|}\sum_{x,y \in X_a^{adv}} L(d(x, y), y)$ for any set of defenses, D. While completing R does have computational cost that scales with $|D|$, evaluating models is often much quicker than generating X^{adv}.

6 EXPERIMENTAL RESULTS

For our experimental results, we test on two datasets, CIFAR-10 Krizhevsky et al. and Tiny-ImageNet Le & Yang (2015). For CIFAR-10 we solved an instance of GaME_1 using the following defenses: BaRT-1 (B1), BaRT-5 (B5), ResNet-164-FAT (RF), ViT-L_16-FAT (VF), SNN Transfer (ST), Backprop SNN (SB), and TiT using ViT and BiT (BVT). Table 3 shows the Mixed Nash Strategy found for the defender in this instance of GaME_1.

Defender Mixed Nash Strategy (CIFAR-10, GaME_1)							
Defense	B1	B5	RF	VF	ST	SB	BVT
λ^D	0	.16	.75	.02	0	.07	0
Min Robust	3%	16%	50%	29%	0%	1%	3%
Defender Mixed Nash Strategy (Tiny ImageNet, GaME_1)							
λ^D	0.152	0.443	-	0.243	-	-	0.162
Min Robust	4.75%	10.62%	-	3.87%	-	-	1.87%

Table 3: $GaME_1$ results for CIFAR-10 and Tiny-ImageNet. Full numerical results corresponding to the robust accuracies for the individual attacks (APGD, MIM, MIME and AE-SAGA) and the clean accuracies of different defense are given in our supplementary material.

For CIFAR-10 instance $r^* = .573$, meaning the worst expected performance of the ensemble against these attacks is to get a robust accuracy of 57.5%. The clean accuracy for this ensemble is 84.1%. In addition to the mixed Nash strategy, the table above also shows the minimum robust accuracy of every given defense, this shows the expected performance of each defense if the attacker plays a best response strategy. Most notably, if the defender were to make a single model defense, their best choice would be to use the FAT ResNet-164 where they would be expected to have a robust accuracy of 50% (and corresponding clean accuracy of only 81.89%.)

For Tiny-ImageNet, on the other hand, $r^* = .2523$, this is a significant increase of over 14.42% more robustness when compared to playing the best pure strategy of BaRT-5. For this ensemble the clean accuracy is 61.1% as opposed to BaRT-5 which only achieves a clean accuracy on average of 56.92%. Overall, our experimental results demonstrate that our framework yields a compositional defense for both CIFAR-10 and Tiny ImageNet that is more robust and has a higher clean accuracy, than any single state-of-the-art defense we analyzed. Due to space limitations, we provide further results for $GaME_2$ in our supplementary material.

7 CONCLUSION

The field of adversarial machine learning has begun to cycle between new defenses, followed by new specific attacks that break those defenses, followed by even more defenses. In this paper, we seek to go beyond this cat and mouse dynamic. We consider adversarial defense transferability, multi-model attacks and a game theoretic framework for compositional adversarial machine learning.

In terms of specific contributions, we develop two new white-box attacks, the Momentum Iterative Method over Expectation (MIME) for attacking single randomized defenses like TiT and Auto Expectation Self-Attention Gradient Attack (AE-SAGA) for dealing with a combination of randomized and non-randomized defenses. We are the first to show the transferability of adversarial examples generated by MIM, APGD, MIME and AE-SAGA on state-of-the-art defenses like FAT, BaRT and TiT.

Lastly, and most importantly, we develop a game theoretic framework for determining the optimal attack and defense strategy. Any newly proposed defense or attack can be easily integrated into our framework. Using a set of state-of-the-art attacks and defenses we demonstrate that our game theoretic framework can create a compositional defense that achieve a 7% increase in robustness over a uniform strategy defense on CIFAR-10 and a 14.42% increase in robustness using a multi-defense, mixed Nash strategy (as opposed to using the best single defense) for Tiny-ImageNet. Both compostional defenses for CIFAR-10 and Tiny-ImageNet come with slightly higher clean accuracies than the best single defense.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

Alexandre Araujo, Laurent Meunier, Rafael Pinot, and Benjamin Negrevergne. Advocating for multiple defense strategies against adversarial examples. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 165–177. Springer, 2020.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pp. 284–293. PMLR, 2018.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 3–14, 2017.

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

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. *Advances in Neural Information Processing Systems*, 34:21056–21069, 2021.

Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

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

Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pp. 302–311, 1984.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pp. 491–507. Springer, 2020.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Kaleel Mahmood, Deniz Gurevin, Marten van Dijk, and Phuoung Ha Nguyen. Beware the black-box: On the robustness of recent defenses to adversarial examples. *Entropy*, 23(10):1359, 2021a.

Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7838–7847, 2021b.

Kaleel Mahmood, Phuong Ha Nguyen, Lam M. Nguyen, Thanh Nguyen, and Marten Van Dijk. Besting the black-box: Barrier zones for adversarial example defense. *IEEE Access*, 10:1451–1474, 2022. doi: 10.1109/ACCESS.2021.3138966.

Pratyush Maini, Eric Wong, and Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*, pp. 6640–6650. PMLR, 2020.

Laurent Meunier, Meyer Scetbon, Rafael B Pinot, Jamal Atif, and Yann Chevaleyre. Mixed nash equilibria in the adversarial examples game. In *International Conference on Machine Learning*, pp. 7677–7687. PMLR, 2021.

John Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951. ISSN 0003486X. URL `http://www.jstor.org/stable/1969529`.

Ambar Pal and René Vidal. A game theoretic analysis of additive adversarial attacks and defenses. *Advances in Neural Information Processing Systems*, 33:1345–1355, 2020.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pp. 4970–4979. PMLR, 2019.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

Rafael Pinot, Raphael Ettedgui, Geovani Rizk, Yann Chevaleyre, and Jamal Atif. Randomization matters how to defend against strong adversarial attacks. In *International Conference on Machine Learning*, pp. 7717–7727. PMLR, 2020.

Adnan Qayyum, Muhammad Usama, Junaid Qadir, and Ala Al-Fuqaha. Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward. *IEEE Communications Surveys & Tutorials*, 22(2):998–1026, 2020. doi: 10.1109/COMST.2020.2975048.

Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean. Barrage of random transforms for adversarially robust defense. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6521–6530, 2019. doi: 10.1109/CVPR.2019.00669.

Nitin Rathi and Kaushik Roy. Diet-snn: A low-latency spiking neural network with direct input encoding and leakage and threshold optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2021a.

Nitin Rathi and Kaushik Roy. Diet-snn: A low-latency spiking neural network with direct input encoding and leakage and threshold optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2021b.

Chawin Sitawarin, Zachary J Golan-Strieb, and David Wagner. Demystifying the adversarial robustness of random transformation defenses. In *International Conference on Machine Learning*, pp. 20232–20252. PMLR, 2022.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, (33), 2020.

P.M. Vaidya. Speeding-up linear programming using fast matrix multiplication. In *30th Annual Symposium on Foundations of Computer Science*, pp. 332–337, 1989. doi: 10.1109/SFCS.1989.63499.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019.

Chang Xiao and Changxi Zheng. One man's trash is another man's treasure: Resisting adversarial examples by adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 412–421, 2020.

Nuo Xu, Kaleel Mahmood, Haowen Fang, Ethan Rathbun, Caiwen Ding, and Wujie Wen. Securing the spike: On the transferabilty and security of spiking neural networks to adversarial examples. *arXiv preprint arXiv:2209.03358*, 2022.

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

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International conference on machine learning*, pp. 11278–11287. PMLR, 2020.