

# What it Thinks is Important is Important: Robustness Transfers through Input Gradients

Alvin Chan<sup>1</sup>, Yi Tay<sup>1</sup>, Yew-Soon Ong<sup>1,2</sup>

<sup>1</sup>Nanyang Technological University, <sup>2</sup> AI3, A STAR, Singapore

## Abstract

*Adversarial perturbations are imperceptible changes to input pixels that can change the prediction of deep learning models. Learned weights of models robust to such perturbations are previously found to be transferable across different tasks but this applies only if the model architecture for the source and target tasks is the same. Input gradients characterize how small changes at each input pixel affect the model output. Using only natural images, we show here that training a student model's input gradients to match those of a robust teacher model can gain robustness close to a strong baseline that is robustly trained from scratch. Through experiments in MNIST, CIFAR-10, CIFAR-100 and Tiny-ImageNet, we show that our proposed method, input gradient adversarial matching, can transfer robustness across different tasks and even across different model architectures. This demonstrates that directly targeting the semantics of input gradients is a feasible way towards adversarial robustness.*

## 1. Introduction

Deep learning models have shown remarkable performances in a wide range of computer vision tasks [15, 28, 17] but can be easily fooled by adversarial examples [27]. These examples are crafted by imperceptible perturbations and can manipulate a model's prediction during test time. Due to its potential security risk in deployment of deep neural networks, adversarial examples have received much research attention with many new attacks [2, 19, 4] and defenses [24, 20, 16, 9, 33, 1] proposed recently.

While there is still a wide gap between accuracy on clean and adversarial samples, the strongest defenses rely mostly on adversarial training (AT) [8, 18, 25]. Adversarial training's main idea, simple yet effective, involves training the model with adversarial samples generated in each training loop. However, crafting strong adversarial training samples is computationally expensive as it entails iterative gradient

steps with respect to the loss function [13, 31].

To circumvent the cost of AT, a recent line of work explores transferring adversarial robustness from robust models to new tasks [11, 26]. To transfer to a target task, current such techniques involve finetuning new layers on top of robust feature extractors that were pre-trained on other domains (source task). While this approach is effective in transferring robustness across different tasks, it assumes that the source task and target task models have similar architecture as pre-trained weights are the medium of transfer.

Here, we propose a robustness transfer method that is both task- and architecture-agnostic with input gradient as the medium of transfer. Our approach, input gradient adversarial matching (IGAM), is inspired by observations [29, 6] that robust AT-trained models display visibly salient input gradients while their non-robust standard trained models have noisy input gradients (Figure 1). The value of input gradient at each pixel defines how a small change there can affect the model's output and can be loosely thought as to how important each pixel is for prediction. Here, we show that learning to emulate how robust models view 'importance' on images through input gradients can result in robust models even without adversarial training examples.

The core idea behind our approach is to train a student model with an adversarial objective to fool a discriminator into perceiving the student's input gradients as those from a robust teacher model. To transfer across different tasks, the teacher model's logit layer is first briefly finetuned on the target task's data, like in [26]. Subsequently, the teacher model's weights are frozen while a student model is adversarially trained with a separate discriminator network in a min-max game so that the input gradients from the student and teacher models are semantically similar, i.e., indistinguishable for the discriminator model [7].

Through experiments in MNIST, CIFAR-10, CIFAR-100 and Tiny-ImageNet, we show that input gradients are a feasible medium to transfer robustness, outperforming finetuning on transferred weights. Surprisingly, student models even outperform their teacher models in both clean accuracy and adversarial robustness. In some cases, the student model's adversarial robustness is close to that of a strong

---

Corresponding author: guowei.al001@ntu.edu.sg

baseline that is adversarially trained from scratch. Though our method does not beat the state of the art robustness, it shows that addressing the semantics of input gradients is a new promising way towards robustness.

In summary, the key contributions of this paper are as follows:

- For the first time, we show that robustness can transfer across different model architectures.
- We achieve this by training the student model's input gradients to semantically match those of a robust teacher model through our proposed method.
- Through extensive experiments, we show that input gradients are a more effective and versatile medium to transfer robustness than pre-trained weights.

## 2. Background

We review the concept of adversarial robustness for image classification and its relationship with input gradients.

**Adversarial Robustness** We express an image classifier as  $f(x; \theta) : \mathcal{X} \rightarrow \mathcal{C}$  that maps an input image  $x$  to output probabilities for  $k$  classes in set  $\mathcal{C}$ , where classifier's parameters is defined as  $\theta$ . Denoting training dataset as  $\mathcal{D}$ , empirical risk minimization is the standard way to train a classifier  $f$ , through  $\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} L(x, y)$ , where  $y \in \mathcal{C}$  is the one-hot label for the image and  $L(x, y)$  is the standard cross-entropy loss:

$$L(x, y) = \mathbb{E}_{(x,y) \sim \mathcal{D}} -y \log f(x) \quad (1)$$

With this training method, deep learning models typically show good performance on clean test samples but fail in the classification of adversarial test samples. With an adversarial perturbation of magnitude  $\epsilon$  at input  $x$ , a model is considered robust against this attack if

$$\arg\max_{i \in \mathcal{C}} f_i(x; \theta) = \arg\max_{i \in \mathcal{C}} f_i(x + \delta; \theta) \quad (2)$$

where  $\|\delta\|_p = \epsilon$ . With small  $\epsilon$ , adversarial perturbation with  $p = \infty$  is often imperceptible and is the focus in this paper.

**Input Gradients of Robust Models** Input gradients characterize how an infinitesimally small change to the input affects the output of the model. Given a pair of input and label  $(x, y)$ , its corresponding input gradient  $\nabla_x L(x, y)$  can be computed through gradient backpropagation in a neural network to its input layer. For classification tasks, the input gradient can be loosely interpreted as a pixel map of what the model thinks is important for its class prediction.

It was observed [29] that robust models that are adversarially trained display an interesting phenomenon: they produce salient input gradients that loosely resemble input images while less robust standard models display noisier input gradients (Figure 1). [6] shows in linear models that distance from samples to decision boundary increases as the alignment between the input gradient and input image grows but this weakens for non-linear neural networks. While these previous studies show that robustly trained models result in salient input gradients, our paper studies input gradients as a medium to transfer robustness across different models.

Figure 1: Input gradients of (middle) a non-robust model and (right) robust model on CIFAR-10 images. The non-robust model undergoes standard SGD training with natural images while the robust model is trained with 7-step PGD adversarial examples.

## 3. Related Work

We review prior art on defense against adversarial examples and highlight those that are most similar to our work.

**Adversarial Training** With the aim of gaining robustness against adversarial examples, the core idea of adversarial training (AT) is to train models with adversarial training examples. Formally, AT minimizes the loss function:

$$L(x, y) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \max_{B(\cdot)} L(x + \delta, y) \quad (3)$$

where  $\max_{B(\cdot)} L(x + \delta, y)$  is computed via gradient-based optimization methods. One of the strongest defenses employ projected gradient descent (PGD) which carries out the following gradient step iteratively:

$$\text{Proj}[\cdot - \text{sign}(\nabla_x L(x + \delta, y))] \quad (4)$$

where  $\text{Proj}(x) = \arg\min_{B(\cdot)} \|x - \cdot\|$ .

AT has seen many adaptations since its introduction. A recent work [32] seeks to generate more effective adversarial training examples through maximizing feature matching distance between those examples and clean samples. To

smoothen the loss landscape so that model prediction is not drastically affected by small perturbations, [21] proposed minimizing the difference between the linearly estimated and real loss value of adversarial examples. Another work, TRADES [33], reduces the difference between the prediction of natural and adversarial examples through a regularization term to smoothen the model's decision boundary.

**Non-Adversarial Training Defense** Closely linked to our method, there is a line of work that regularizes the input gradients to boost robustness. Those prior art [23, 12] focus on using double backpropagation [5] to minimize the input gradients' Frobenius norm. Those approaches aim to constrain the effect that changes at individual pixels have on the classifier's output but not the overall semantics of the input gradients like our method. [3] show that models can be more robust when regularized to produce input gradients that resemble input images.

Several recent methods fall under the category of provable defenses that seeks to bound minimum adversarial perturbation for a subset of neural networks [10, 22, 30]. These defenses typically first find a theoretical lower bound for the adversarial perturbation and optimize this bound during training to boost adversarial robustness.

**Robustness Transfer** There is a line of work that shows robustness can transfer from one model to another. [11] shows that robustness from adversarial training can be improved if the models are pre-trained from tasks from other domains. Another work shows that adversarially trained learn robust feature extractors that can be directly transferred to a new task by finetuning a new logit layer on top of these extractors [26]. Circumventing adversarial training, these transferred models can still retain a high degree of robustness across tasks. Unlike our method, these two work require that the source and target models both have the same model architecture since pre-trained weights are directly transferred.

## 4 Input Gradient Adversarial Matching

Our proposed training method consists of two phases: 1) finetuning robust teacher model on target task and 2) adversarial regularization of input gradients during the student models' training.

### 4.1. Finetuning Teacher Classifier

The first stage involves finetuning the weights of the teacher model  $f_t$  on the target task. Parameterizing the model weights as  $\theta$ , the finetuning stage minimizes the cross-entropy loss over the target task training data  $(x, y) \in D_{\text{target}}$ :

$$L_{\text{,xent}}(x, y, \theta) = E_{(x,y)} -y \log f_t(x) \quad (5)$$

where  $x \in \mathbb{R}^{hwc}$  for  $h \times w$ -size images with  $c$  channels,  $y \in \mathbb{R}^k$  is one-hot label vector of  $k$  classes.

To preserve the robust learned representations in the teacher model [26], we freeze all the weights and replace the final logits layer to finetune. Denoting the frozen weights as  $\theta^\dagger$  and the new logits layer as  $\text{logit}$ , the teacher model finetuning objective is

$$\text{logit} = \underset{\text{logit}}{\text{argmin}} L_{\text{,xent}}(z(x, \theta^\dagger), y, \text{logit}) \quad (6)$$

where  $z(x, \theta^\dagger)$  represents the hidden features before the logit layer. After finetuning the logits layer on the target task, all the teacher model's parameters  $(\theta)$  are fixed, including  $\text{logit}$ .

### 4.2. Input Gradient Matching

The aim of the input gradient matching is to train the student model to generate input gradients that semantically resemble those from the teacher model. The input gradient characterizes how the loss value is affected by small changes to each input pixel.

We express the classification cross entropy loss of the student model  $f_s$  on the target task dataset  $D_{\text{target}}$  as:

$$L_{\text{,xent}}(x, y, \theta) = E_{(x,y)} -y \log f_s(x) \quad (7)$$

Through gradient backpropagation, the input gradient of the student model  $f_s$  is

$$J_s(x) := \nabla_x L_{\text{,xent}} = \frac{\partial L_{\text{,xent}}}{\partial x_1} \parallel \parallel \parallel \frac{\partial L_{\text{,xent}}}{\partial x_d} \quad (8)$$

where  $d = hwc$ .

Correspondingly, the input gradient of the teacher model  $f_t$  is

$$J_t(x) := \nabla_x L_{\text{,xent}} = \frac{\partial L_{\text{,xent}}}{\partial x_1} \parallel \parallel \parallel \frac{\partial L_{\text{,xent}}}{\partial x_d} \quad (9)$$

### 4.2.1 Adversarial Regularization

To achieve the objective of training the student model's input gradient  $J_s$  to resemble those from the teacher model  $J_t$ , we draw inspiration from GANs, a framework comprising a generator and discriminator model. In our case, we train the  $f_s$  to make it hard for the discriminator  $f_{\text{disc}}$  to distinguish between  $J_t$  and  $J_s$ . The discriminator output value  $f_{\text{disc}}(J)$  represents the probability that  $J$  came from the teacher model  $f_t$  rather than  $f_s$ . To train  $f_s$  to produce  $J_s$  that  $f_{\text{disc}}$  perceive as  $J_t$ , we employ the following adversarial loss:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{J_t}[\log f_{\text{disc}}(J_t)] + \mathbb{E}_{J_s}[\log(1 - f_{\text{disc}}(J_s))] \quad (10)$$

Combining this regularization loss with the classification loss function  $\mathcal{L}_{\text{xent}}$  in Equation (7), we can optimize through stochastic gradient descent (SGD) to approximate the optimal parameters for  $f_s$  as follows,

$$= \text{argmin}(\mathcal{L}_{\text{xent}} + \text{adv}\mathcal{L}_{\text{adv}}) \quad (11)$$

where  $\text{adv}$  control how much input gradient adversarial regularization term dominates the training.

In contrast, the discriminator ( $f_{\text{disc}}$ ) learns to correctly distinguish the input gradients by maximizing the adversarial loss term. Parameterizing  $f_{\text{disc}}$  with  $\theta$ , the discriminator is also trained with SGD as such

$$= \text{argmax} \mathcal{L}_{\text{adv}} \quad (12)$$

#### 4.2.2 Reconstruction Regularization

Apart from the adversarial loss term, we also employ a term to penalize the  $l_2$  difference between the  $J_s$  and  $J_t$  generated from the same input image.

$$\mathcal{L}_{\text{diff}} = \|J_s - J_t\|_2^2 \quad (13)$$

The  $\mathcal{L}_{\text{diff}}$  term is analogous to the additional reconstruction loss in a VAE-GAN setup [14] where it has shown to improve performance. For each given input image ( $x$ ) in IGAM, there is a corresponding target input gradient  $J_t$  for the student model's  $J_s$  to match, allowing us to exploit this instance matching loss ( $\mathcal{L}_{\text{diff}}$ ). Adding this term with Equation 11, the final training objective of the student model is

$$= \text{argmin}(\mathcal{L}_{\text{xent}} + \text{adv}\mathcal{L}_{\text{adv}} + \text{diff}\mathcal{L}_{\text{diff}}) \quad (14)$$

where  $\text{diff}$  determines the weight of the  $l_2$  penalty term in the training.

Figure 2 shows a summary of IGAM training phase while Algorithm 1 details the corresponding pseudo-codes.

#### 4.3. Transfer With Different Input Dimensions

In the earlier sections, we assume that the input dimensions of the teacher and student models are the same. Recall that before finetuning, the teacher model  $f_t$  was originally trained on source task samples  $(x_{\text{src}}, y_{\text{src}}) \in \mathcal{D}_{\text{src}}, x_{\text{src}} \in \mathbb{R}^{d_{\text{src}}}$  where each  $x_{\text{src}}$  is a  $h_{\text{src}} \times w_{\text{src}}$ -size image with  $c_{\text{src}}$  channels. In practice, the image dimensions may differ from those from the task target, i.e.,  $d_{\text{src}} \neq d_{\text{tar}}$ . To allow the gradient backpropagation of the losses through the input gradients, we use affine functions to adapt the target task images to match the dimension of the teacher model's input layer:

Figure 2: Training phase of input gradient adversarial matching (IGAM).

#### Algorithm 1: Input gradient adversarial matching

---

**Input:** Target task training data  $\mathcal{D}_{\text{train}}$ , Learning rates for teacher model  $f_t$ , student model  $f_s$  and discriminator  $f_{\text{disc}}$ : ( $\eta_t, \eta_s, \eta_d$ )

**for each finetuning iteration do**

Sample  $(x, y) \sim \mathcal{D}_{\text{train}}$

$\mathcal{L}_{\text{xent}} = -y \log f_t(x)$       Classification loss

logit  $\leftarrow$  logit  $-$  logit      Update teacher  $f_t$  to minimize  $\mathcal{L}_{\text{xent}}$

**for each training iteration do**

Sample  $(x, y) \sim \mathcal{D}_{\text{train}}$

$\mathcal{L}_{\text{xent}} = -y \log f_t(x)$       Classification loss for teacher

$J_t = x \cdot \mathcal{L}_{\text{xent}}$       Compute teacher input gradient

$\mathcal{L}_{\text{xent}} = -y \log f_s(x)$       Classification loss for student

$J_s = x \cdot \mathcal{L}_{\text{xent}}$       Compute student input gradient

$\mathcal{L}_{\text{adv}} = \log f_{\text{disc}}(J_t) + \log(1 - f_{\text{disc}}(J_s))$       Adversarial loss

$\mathcal{L}_{\text{diff}} = \|J_s - J_t\|_2^2$        $l_2$  penalty loss

$\text{loss} = (\mathcal{L}_{\text{xent}} + \text{adv}\mathcal{L}_{\text{adv}} + \text{diff}\mathcal{L}_{\text{diff}})$       Update the student  $f_s$  to minimize  $\mathcal{L}_{\text{xent}}, \mathcal{L}_{\text{adv}}$  and  $\mathcal{L}_{\text{diff}}$

$\text{loss} = \mathcal{L}_{\text{adv}}$       Update discriminator  $f_{\text{disc}}$  to maximize  $\mathcal{L}_{\text{adv}}$

---

$$x_{\text{tar}} = A \# x_{\text{tar}} + b \quad (15)$$

where  $x_{\text{tar}} \in \mathbb{R}^{d_{\text{src}}}$ ,  $x_{\text{tar}} \in \mathbb{R}^{d_{\text{tar}}}$  and  $A \in \mathbb{R}^{d_{\text{src}} \times d_{\text{tar}}}$ .

Subsequently, cross-entropy loss for the teacher model can be computed:

$$\mathcal{L}_{\text{xent}}(x_{\text{tar}}, y_{\text{tar}}) = \mathbb{E}_{(x_{\text{tar}}, y_{\text{tar}})} -y_{\text{tar}} \log f_t(x_{\text{tar}}) \quad (16)$$

Since affine functions are continuously differentiable, we can backprop to get the input gradient:

$$J_t(x_{\text{tar}}) = x_{\text{tar}} \cdot \mathcal{L}_{\text{xent}} \quad (17)$$

We use a range of such transformations in our experiments to cater for the difference of input dimensions from various source-target dataset pairs.

##### 4.3.1 Input Resizing

Image resizing is one such transformation where the resized image can be expressed as the output of an affine function, i.e.,  $x_{\text{tar}} = A \# x_{\text{tar}}$ . In the case where the teacher

model’s input dimension is smaller than the student model, i.e.,  $d_{\text{tar}} > d_{\text{src}}$ , we can use average pooling to downsize the image. A  $2 \times 2$  average pooling is equivalent to resizing with bilinear interpolation when  $d_{\text{tar}}$  is a multiple of  $d_{\text{src}}$ . Figure 3a shows how we use input resizing to generate the input gradient from the teacher model. For cases of  $d_{\text{tar}} < d_{\text{src}}$ , we use image resizing with bilinear interpolation to upscale the input dimension before feeding into the teacher model. For the source-target pair of MNIST-CIFAR, we can similarly reduce the number of channels by averaging the RGB values of the CIFAR images before feeding to the teacher model (trained on MNIST).

#### 4.3.2 Input Cropping

Cropping is another way to downsize the image to fit a smaller teacher model’s input dimension, i.e.,  $d_{\text{tar}} > d_{\text{src}}$ . The cropped image is output of  $x_{\text{tar}} = A \hat{\wedge} x_{\text{tar}}$  where  $A$  is a row-truncated identity matrix. For input cropping, the initial  $J_t$  would have zero values at the region where the image was cropped out since those pixel values are multiplied by zero. To prevent the discriminator from exploiting this property to distinguish  $J_t$  from  $J_s$ , we feed into the discriminator  $J_t$  and  $J_s$  that are cropped to size  $d_{\text{src}}$ . Figure 3b shows how we use cropping to generate the cropped input gradient from the teacher model.

#### 4.3.3 Input Padding

In contrast to cropping, padding can be used for cases where  $d_{\text{tar}} < d_{\text{src}}$ . With the same form of affine function  $x_{\text{tar}} = A \hat{\wedge} x_{\text{tar}}$ ,  $A$  is a identity matrix prepended and appended with zero-valued rows. Figure 3c shows how we generate the input gradient from the teacher model with input padding.

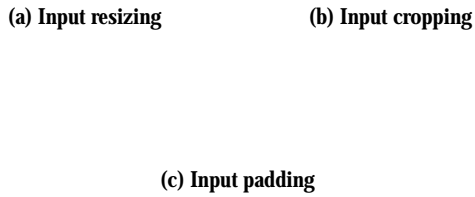


Figure 3: Transformations to fit images to teacher model’s input dimensions.

## 5. Experiments

We conducted experiments with IGAM on source-target data pairs comprising of MNIST, CIFAR-10, CIFAR-100

and Tiny-ImageNet. These datasets allow us to validate the effectiveness of IGAM in transferring across tasks with different image dimensions. Unless otherwise stated, adversarial robustness is evaluated based on  $l_{\infty}$  adversarial examples with  $\epsilon = \frac{8}{255}$ . IGAM’s hyperparameters such as  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$  and  $\eta$  for each experiment are included in the supplementary material.

### 5.1. CIFAR-10 Target Task

In our experiments with CIFAR-10 as the target task, we study two types of robustness transfer. The upwards transfer involves employing IGAM to transfer robustness from a smaller model trained on the simpler MNIST dataset to a larger CIFAR-10 classifier. Conversely, the downwards transfer experiments involve transferring robustness from a 200-class Tiny-ImageNet model to a CIFAR-10 classifier.

#### 5.1.1 Upwards Transfer

**Setup** CIFAR-10 is a 10-class colored image dataset comprising of 50k training and 10k test images, each of size  $32 \times 32 \times 3$ . For the CIFAR-10 student model, we use a Wide-Resnet 32-10 model with similar hyperparameters to [18] and train it for 200 epochs on natural training images with IGAM. The MNIST dataset consists of 60k training and 10k test binary-colored images, each of size  $28 \times 28 \times 1$ . For the robust teacher model trained on MNIST, we also follow the same adversarial training setting and 2-CNN layered architecture as [18]<sup>1</sup>. The teacher model is finetuned on natural CIFAR-10 images for 10 epochs before using it to train the student model with IGAM. Since the input dimensions of CIFAR-10 and MNIST are different, we average pool pixel values across the color channels of CIFAR-10 images to get dimension  $32 \times 32 \times 1$  and subsequently center crop them into  $28 \times 28 \times 1$  input images for the MNIST teacher model. With this same input transformation, we also finetune the final logit layer of a robust MNIST model on CIFAR-10 images similar to [26] for 100 epochs, to compare as a baseline (FT-MNIST). We also train a strong robust classifier, with 7-step PGD adversarial training like in [18], with the same architecture as the IGAM student model to compare.

**Results** In the face of adversarial examples, the IGAM-trained student model outperforms the standard and finetuned baselines by large margins (Table 1). Despite the difference between the dataset domains and model architectures, IGAM can transfer robustness from the teacher to the student model to almost match that from a strong adversarially trained (AT) model. The IGAM student model has higher clean test accuracy than the robust PGD7-trained

<sup>1</sup>Robust MNIST pre-trained model downloaded from [https://github.com/MadryLab/MNIST\\_challenge](https://github.com/MadryLab/MNIST_challenge)

baseline which we believe is a result of using natural (not adversarially perturbed) images as training data in IGAM.

We note that though finetuning was previously showed to have positive results in transferring robustness across relatively similar domains like between CIFAR10 and CIFAR100 [25], it fails to transfer successfully here. This is likely due to the bigger difference between the MNIST and CIFAR-10 dataset, as well as the requirement of a more sophisticated model architecture for the more challenging CIFAR-10 dataset.

Table 1: Accuracy (%) on clean and adversarial CIFAR-10 test samples with upwards transfer.

Model	Clean	FGSM	PGD5	PGD10	PGD20
Standard	<b>95.0</b>	13.4	0	0	0
FT-MNIST	33.4	1.51	0.44	0.15	0.12
IGAM-MNIST	93.6	<b>67.8</b>	<b>63.6</b>	<b>56.9</b>	<b>43.5</b>
PGD7-trained	87.3	56.2	55.5	47.3	45.9

### 5.1.2 Downwards Transfer

**Setup** Tiny-ImageNet is a 200-class image dataset where each class contains 500 training and 50 test images. Each Tiny-ImageNet image has dimension of  $64 \times 64 \times 3$ . For the robust teacher model trained on Tiny-ImageNet, we use a similar Wide-Resnet 32-10 model since it is compatible with a larger input dimension due to its global average pooling operation of the feature maps before fully connected layers. We robustly train this teacher model on Tiny-ImageNet, following the same adversarial training hyperparameters in [18] where robust models are trained with 1 adversarial examples generated by 7-step PGD. Before using it to train the student model with IGAM, the teacher model is finetuned on natural CIFAR-10 images for 6 epochs. Since the input dimensions of CIFAR-10 and Tiny-ImageNet are different, we resize the  $32 \times 32 \times 3$  CIFAR-10 images with bilinear interpolation to get dimension  $64 \times 64 \times 3$  for finetuning the teacher model. For the IGAM student model, we use the same Wide-Resnet 32-10 model and hyperparameters as in § 5.1.1. We also finetune the final logit layer of a robust Tiny-ImageNet model on upsized CIFAR-10 images similar to [26] for 100 epochs, to compare as a baseline (FT-TinyImagenet). We also investigate two more types of input transformation for IGAM here. The first is a trained  $3 \times 3$  transpose convolutional filter, with stride 2, to upscale the CIFAR-10 images to size  $64 \times 64 \times 3$ . This single transpose convolutional layer is trained together with the teacher model while finetuning on natural CIFAR-10 images. The second type of input transformation is padding, as detailed in § 4.3.3, of which we explore two variants: center-padding and random-padding.

**Results** With input padding or input resizing, the IGAM-trained student model outperforms the standard and finetuned baselines in adversarial robustness (Table 2). From our experiments, using padding or resizing is more effective for downwards transfer of robustness, with slightly better results for resizing. With the downwards transfer, the student model can match the strong PGD7-trained baseline even more closely than in the upwards transfer case (Table 1). This is expected since the teacher model was robustly trained in a more challenging Tiny-ImageNet task and would likely learn even more robust representations than if it were trained on the simpler datasets like MNIST. Compared to upwards transfer, the finetuning baseline transfers robustness and clean accuracy performance to a larger extent but is still outperformed by IGAM.

### 5.1.3 Input Gradients

When comparing the input gradients of the various baseline and IGAM models (Figure 4), we can observe that there is a diverse degree of saliency. The IGAM models' input gradients appear less noisy than a standard trained model as what we aim to achieve with our proposed method. Interestingly, the IGAM-MNIST model's input gradients have a degree of saliency despite the sparse input gradients from its FT-MNIST teacher model. For IGAM models with a Tiny-ImageNet teacher, the more robust variants like IGAM-Upsize and IGAM-Pad display less noisy input gradients than the less robust IGAM-RandomPad and IGAM-TransposeConv. More input gradient samples are displayed in Figure 6 of the supplementary material.

### 5.2. CIFAR-100 Target Task

We further study IGAM performance in upwards transfer of robustness with CIFAR-100 as the target task, MNIST and CIFAR-10 as the source task.

**Setup** CIFAR-100 is a 100-class colored image dataset comprising of 50k training and 10k test images. Similar to CIFAR-10, each image has a dimension of  $32 \times 32 \times 3$ . For the CIFAR-100 student model, we use a Wide-Resnet 32-10 model with similar hyperparameters as § 5.1.1 except for the final logit layer, which has 100 instead of 10 class outputs. We train the student model for 200 epochs on natural CIFAR-100 training images with IGAM. The robust MNIST teacher model used is similar to the one in § 5.1.1. For the robust CIFAR-10 teacher model, we also follow the same adversarial training setting and architecture as [18]<sup>2</sup>. During IGAM training with MNIST as the source task, the input transformation same as in § 5.1.1 is used to resize CIFAR-100 images into  $28 \times 28 \times 1$  inputs for the

<sup>2</sup>Robust CIFAR-10 pre-trained model downloaded from [https://github.com/MadryLab/cifar10\\_challenge](https://github.com/MadryLab/cifar10_challenge)

Table 2: Accuracy (%) on clean and adversarial CIFAR-10 test samples with downwards transfer.

Model	Clean	FGSM	PGD5	PGD10	PGD20	PGD50	PGD100
Standard	<b>95.0</b>	13.4	0	0	0	0	0
FT-TinyImagenet	77.2	37.7	33.9	28.0	24.9	23.0	22.5
IGAM-TransposeConv	93.2	<b>65.0</b>	<b>58.8</b>	44.5	32.4	22.4	18.7
IGAM-RandomPad	88.3	35.8	43.9	40.1	<b>38.6</b>	37.8	37.6
IGAM-Pad	87.9	51.6	52.2	46.6	44.0	43.0	42.5
IGAM-Upsize	88.7	54.0	52.5	<b>47.6</b>	<b>45.1</b>	<b>43.5</b>	<b>43.0</b>
PGD7-trained	87.25	56.22	55.5	47.3	45.9	45.4	45.3

Figure 4: Input gradients of different models.

teacher model. No input transformation is used when the source task is CIFAR-10 since its images have the same dimensions as CIFAR-100's. The final logit layers of MNIST and CIFAR-10 teacher models are finetuned for 10 and 6 epochs, respectively, on natural CIFAR-100 images before being used to transfer robustness in IGAM. We also finetune the final logit layer of a robust CIFAR-10 model on CIFAR-100 for 100 epochs, to compare as a baseline (FT-CIFAR10). We also train a strong robust classifier, with 7-step PGD adversarial training like in [18], with the same architecture as the IGAM student model to compare.

**Results** Similar to our findings in § 5.1, IGAM-trained models outperform standard and finetuned baselines in adversarial robustness (Table 3). Expectedly, using CIFAR-10 as the source task yields higher transferred robustness than using MNIST for IGAM. Since CIFAR-10 is closer to CIFAR-100 and more challenging than MNIST, the CIFAR-10 teacher model likely has more robust and relevant representations that are reflected as more robust input gradients.

We note that though CIFAR-10 and CIFAR-100 are the most similar datasets in our experiments, the finetuned baseline has lower clean accuracy and adversarial robustness compared to IGAM models. Finetuned models' weights are frozen up until the final logit layer to retain learned robust representations. While weight freezing maintains a degree of robustness to outperform standard training, it may restrict the model from learning new representations relevant to the target task, explaining its lower clean accuracy. We believe this restriction also explains its lower robustness compared to IGAM since IGAM models are free to learn representations important for the target task.

Table 3: Accuracy (%) on clean and adversarial CIFAR-100 test samples.

Model	Clean	FGSM	PGD5	PGD10	PGD20
Standard	<b>78.7</b>	7.95	0.13	0.03	0
FT-CIFAR10	49.3	17.2	15.3	11.7	10.5
IGAM-MNIST	73.16	<b>41.41</b>	<b>33.09</b>	23.35	17.67
IGAM-CIFAR10	62.39	34.31	29.59	<b>24.05</b>	<b>21.74</b>
PGD7-trained	60.4	29.1	29.3	24.3	23.5

**Roles of Loss Terms** Improvements from the two terms are additive to each other, as reflected in Table 4 and 5. From Figure 7 in the supplementary material, we observe that both the  $\mathcal{L}_{adv}$  and  $\mathcal{L}_{diff}$  smoothen the decision boundaries and lower cross-entropy values in the loss landscape compared to the standard trained baseline.

Table 4: IGAM-CIFAR10 accuracy (%) with varying  $\mathcal{L}_{diff}$ .

$\mathcal{L}_{diff}$	0	2.5	5	10
PGD20	16.0	16.3	21.7	21.7
Clean	58.9	61.8	62.9	62.4

Table 5: IGAM-CIFAR10 accuracy (%) with varying  $\mathcal{L}_{adv}$ .

$\mathcal{L}_{adv}$	0	0.5	1	2
PGD20	3.9	4.34	7.37	21.7
Clean	78.4	77.4	74.3	62.4

**Compute Time** Since finetuning is conducted once, we amortize its time taken over each IGAM epoch to arrive at 347s, which is lower than the 815s taken for a 7-step PGD epoch. Even though IGAM involves an additional discriminator update step on top of standard training, the parameter size of the discriminator is much smaller than the classifier model.

### 5.3. Tiny-ImageNet Target Task

We study if robustness can transfer through the input gradients when the target task has significantly larger input dimensions than the source task, with Tiny-ImageNet as the target task and CIFAR-10/100 as the source task.

**Setup** For the robust CIFAR-10/100 teacher model, we follow the same adversarial training setting and Wide-Resnet 32-10 architecture as [18]. We use a similar Wide-Resnet 32-10 model for the Tiny-ImageNet student model due to its compatible with larger input dimension due to its global average pooling layer. The robust CIFAR-10/100 teacher models are finetuned for 5 epochs on natural Tiny-ImageNet images before being used for IGAM. Since the input dimensions of Tiny-ImageNet and CIFAR-10/100 are different, we study two types of input transformation to reshape the image dimension from  $64 \times 64 \times 3$  to  $32 \times 32 \times 3$  for finetuning the teacher model. The first is image resizing with bilinear interpolation (§ 4.3.1), which is equivalent to a  $2 \times 2$  average pooling layer with stride 2. The second transformation is center-cropping as detailed in § 4.3.2. The models' adversarial robustness is evaluated based on 5-step PGD attacks on test Tiny-ImageNet samples.

**Results** Similar to previous target-source task pairs, IGAM can transfer robustness even to much more challenging dataset, to a degree to outperform the standard trained and finetuned baselines (Figure 5). There is no visible difference in robustness transferred when using image resizing or center-cropping as the input transformation.

Figure 5: Accuracy (%) on clean and adversarial Tiny-ImageNet test samples.

## 6. Theoretical Discussion

To understand how robustness transfer across input gradients of the student and teacher models, we first look at the link between robustness and saliency of input gradients in a single network. The link is formalized in Theorem 2 of [6] which states that a network's linearized robustness  $\mathcal{R}(\mathbf{x})$  around an input  $\mathbf{x}$  is upper bounded by alignment term  $\mathcal{A}(\mathbf{x})$ :

$$\mathcal{R}(\mathbf{x}) \leq \mathcal{A}(\mathbf{x}) + \frac{C}{g} \quad (18)$$

where  $g$  is the Jacobian of the difference between the top two logits,  $g(\mathbf{x}) = \frac{\|\mathbf{x}, \mathbf{g}\|}{g}$  and  $C$  is a positive constant. An important notion here is that a model with high linearized robustness  $\mathcal{R}(\mathbf{x})$  retains its original prediction in face of large perturbation but may still perform poorly on clean test data with incorrect original outputs, such as finetuned teachers.

Different finetuned teacher models (FT-MINST and FT-TinyImagnet) display visually different input gradients which we speculate to be a result of being 'locked' into their dataset-specific robust features. Different from natural images which have smooth pixel value distributions, MNIST pixels take extreme binary values. From the robustness-alignment link, one can expect the input gradient to also take extreme values, explaining the sparse  $J$  of FT-MINST.

With Theorem 6.1 below, IGAM's  $\mathcal{L}_{adv}$  term encourages the teacher and student models' input gradients and, consequently, their input alignment terms  $\mathcal{A}(\mathbf{x})$  to match well.

**Theorem 6.1.** *The global minimum of  $\mathcal{L}_{adv}$  is achieved when  $J_s = J_t$ .*

Its proof is in the supplementary material (§ B). As a result, the high linearized robustness upper bound of teacher model is transferred to the student model. Though input gradients are approximations of  $g$  and the upper bound is not tight, we observe that such transfer is feasible in our experiments. On top of this transferred robustness bound, all of the student model's weights are free to learn features relevant to the target task in boosting its clean accuracy, hence the improved performance over its teacher models.

## 7. Conclusions

We showed that input gradients are an effective medium to transfer adversarial robustness across different tasks and even across different model architectures. To train a student model's input gradients to semantically match those of a robust teacher model, we proposed input gradient adversarial matching (IGAM) to optimize for the input gradients' source to be indistinguishable for a discriminator network. Through extensive experiments on image classification, IGAM models outperform standard trained models and models finetuned on pre-trained robust feature extractors. This demonstrates that input gradients are a more versatile and effective medium of robustness transfer. We hope that this will encourage new defenses that also target the semantics of input gradients to achieve adversarial robustness.

## Acknowledgments

This work is funded by the National Research Foundation, Singapore [Award No.: AISG-RP-2018-004] and DSAIR at Nanyang Technological University.



## References

- [1] Maksym Andriushchenko and Matthias Hein. Provably robust boosted decision stumps and trees against adversarial attacks. *arXiv preprint arXiv:1906.03526*, 2019.
- [2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.
- [3] Alvin Chan, Yi Tay, Yew Soon Ong, and Jie Fu. Jacobian adversarially regularized networks for robustness. *arXiv preprint arXiv:1912.10185*, 2019.
- [4] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. *arXiv preprint arXiv:1907.02044*, 2019.
- [5] Harris Drucker and Yann Le Cun. Double backpropagation increasing generalization performance. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, volume 2, pages 145–150. IEEE, 1991.
- [6] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172*, 2019.
- [7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [9] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.
- [10] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, pages 2266–2276, 2017.
- [11] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. *arXiv preprint arXiv:1901.09960*, 2019.
- [12] Daniel Jakubovitz and Raja Giryes. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 514–529, 2018.
- [13] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.
- [14] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.
- [15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.
- [16] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018.
- [17] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Cogan: Generation by parts via conditional coordinating. *arXiv preprint arXiv:1904.00284*, 2019.
- [18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [19] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018.
- [20] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8571–8580, 2018.
- [21] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Alhussein Fawzi, Soham De, Robert Stanforth, Pushmeet Kohli, et al. Adversarial robustness through local linearization. *arXiv preprint arXiv:1907.02610*, 2019.
- [22] Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pages 10877–10887, 2018.
- [23] Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Thirty-second AAAI conference on artificial intelligence*, 2018.
- [24] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. *arXiv preprint arXiv:1805.09190*, 2018.
- [25] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *arXiv preprint arXiv:1904.12843*, 2019.
- [26] Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David W. Jacobs, and Tom Goldstein. Adversarially robust transfer learning. *CoRR*, abs/1905.08232, 2019.
- [27] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [28] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *arXiv preprint arXiv:1906.06423*, 2019.
- [29] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may

- be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.
- [30] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, pages 8400–8409, 2018.
  - [31] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 501–509, 2019.
  - [32] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. *arXiv preprint arXiv:1907.10764*, 2019.
  - [33] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.