Boosting the Transferability of Adversarial Samples via Attention

Weibin Wu[1], Yuxin Su[1]*, Xixian Chen[2], Shenglin Zhao[2], Irwin King[1], Michael R. Lyu[1], Yu-Wing Tai[2]
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, [2]Tencent
{wbwu, yxsu, king, lyu}@cse.cuhk.edu.hk, {xixianchen, henryslzhao, yuwingtai}@tencent.com

Abstract

The widespread deployment of deep models necessitates the assessment of model vulnerability in practice, especially for safety- and security-sensitive domains such as autonomous driving and medical diagnosis. Transfer-based attacks against image classifiers thus elicit mounting interest, where attackers are required to craft adversarial images based on local proxy models without the feedback information from remote target ones. However, under such a challenging but practical setup, the synthesized adversarial samples often achieve limited success due to overfitting to the local model employed. In this work, we propose a novel mechanism to alleviate the overfitting issue. It computes model attention over extracted features to regularize the search of adversarial examples, which prioritizes the corruption of critical features that are likely to be adopted by diverse architectures. Consequently, it can promote the transferability of resultant adversarial instances. Extensive experiments on ImageNet classifiers confirm the effectiveness of our strategy and its superiority to state-of-the-art benchmarks in both white-box and black-box settings.

1. Introduction

Deep neural networks (DNNs) have emerged as a cutting-edge solution to a broad spectrum of real-world applications, such as object detection, speech recognition, and machine translation [27]. Despite the impressive performance of these deep learning techniques, they are surprisingly vulnerable to the so-called adversarial samples [36]. For example, by imposing human-imperceptible noises on legitimate images purposefully, the resultant adversarial input can incur erroneous predictions from state-of-the-art image classifiers. It raises growing concerns over the reliability of these high-performance black boxes and hinders the deployment of these models in practice, especially in safety- and security-sensitive domains such as autonomous driving and medical diagnosis [3].

Attacks thus play an important part in evaluating a model and revealing its blind spots before deployment, and one of the most fundamental and recognized tasks is to generate adversarial images against DNN image classifiers [3]. To simulate the threat a DNN image classifier may face, there are generally two kinds of threat models considered in the literature [20]. One is white-box settings, where attackers have full access to the victim model, such as the model architectures and parameters. The other one is black-box settings, where attackers only possess query access to the target model, namely, offering input images and obtaining output predictions.

Corresponding to the threat models that they are tailored for, there exist two sorts of attacks: white-box attacks and black-box ones [20]. White-box attacks can exploit the exact gradient information of the victim model to craft malicious instances [36, 9, 5], while black-box attacks can be further divided into two categories according to the mechanism attackers adopt [8]. One is query-based, and the other one is transfer-based. Query-based black-box attacks usually require excessive queries before a successful trial [16]. On the contrary, without the feedback information from the target model, transfer-based black-box attacks devise adversarial samples with off-the-shelf local models (*i.e.,* source models) and directly harness the resultant example to fool the remote target model (*i.e.,* victim models) [41, 8].

Among these two sorts of black-box attacks, the transfer-based one has attracted ever-increasing attention recently [8]. In general, only costly query access to deployed models is available in practice. Therefore, white-box attacks hardly reflect the possible threat to a model, while query-based attacks have less practical applicability than the transfer-based counterparts due to the prohibitive query cost they may incur [8].

Thanks to the observed cross-model transferability of adversarial samples, a popular practice is to freely employ any white-box attack strategy as transfer-based black-box attacks [21]. Unfortunately, the malicious images synthesized by such a scheme are prone to overfit to the exclusive blind spots of the source model [39, 8, 41, 7]. Specifically, although the crafted adversarial samples can attack

*Corresponding author.

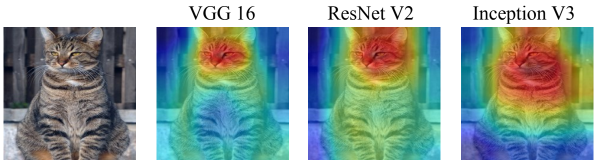

Figure 1: The attention heatmaps of three representative models (VGG 16 [33], ResNet V2 [12, 13], and Inception V3 [35]) for a cat prediction. The visualization is generated with the technique of [30] as detailed in Section 4.2. Redder regions possess higher importance to the model decision.

the source model with near 100% success rates, they suffer from limited success against the target model.

In this work, we aim to promote such transfer-based attacks, which requires improving the transferability of adversarial samples crafted with white-box attack strategies. We expect that the crux is to guide the search of adversarial images towards the common vulnerable directions of both the source and the target models. Therefore, it inspires us to seek for the common characteristics of diverse models and exploit such information to ameliorate the overfitting issue.

We discover that before different models arrive at a correct decision, they should first extract various features and then weigh these features appropriately, namely, allocating suitable **attention** over extracted features[1]. Although some models may adopt exclusive feature extractors, the most critical features that diverse architectures employ tend to overlap largely in our numerous observations. For instance, as demonstrated in Figure 1, when different models recognize a cat image, albeit one of the models (Inception V3) also looks for features extracted from the cat neck, all of them tend to pay attention to the face-related features spontaneously.

The similarity among different models in the employed features inspires us to exploit the model's attention to guide the search of adversarial perturbations. Figure 2 illustrates the proposed strategy. In short, we first adopt back-propagated gradients to approximate the importance of different features to model decisions (*i.e.,* **attention extraction**). Then we require the adversarial manipulation to contaminate attention-weighed feature outputs. As a result, the synthesized malicious noise can focus on undermining the most vital image features that the local source model employs (*i.e.,* **critical feature destruction**). Since different models strongly rely on such features, we can alleviate overfitting to a specific source model and boost the transferability of resultant adversarial samples.

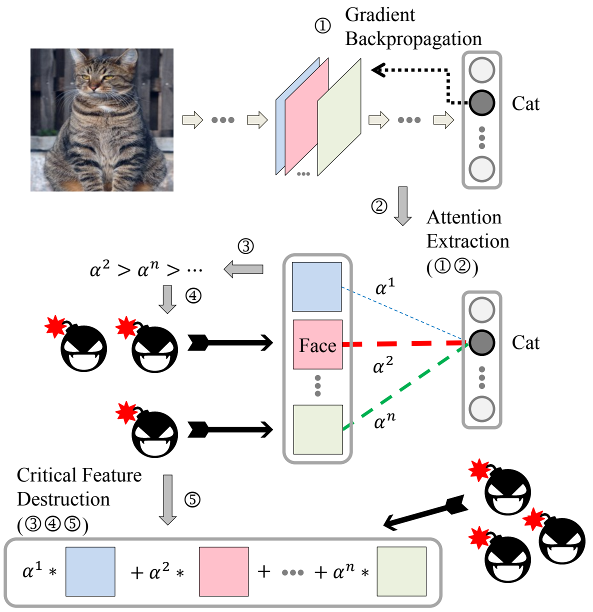
① Gradient Backpropagation
Cat
② Attention Extraction (①②)
③
$\alpha^2 > \alpha^n > \cdots$
④
α^1
Face
α^2
α^n
Cat
Critical Feature Destruction (③④⑤)
⑤
$\alpha^1 *$
$+ \alpha^2 *$
$+ \cdots + \alpha^n *$

Figure 2: The proposed procedure of model **attention extraction** and its application to guide the search of deceptive samples towards **critical feature destruction**. See Section 4 for details.

In summary, we would like to highlight the following contributions of this work:

- We propose a novel strategy to boost the transferability of adversarial images. It features an introduction of model attention to regularize the search of deceptive noises, which mitigates overfitting to specific blind spots of the source model.
- Extensive experiments show that our attention-guided transfer attack (ATA) can severely compromise diverse top-performance image classifiers and defended ones. Empirical results also witness the superior performance of our proposal to state-of-the-art benchmarks in both white-box and black-box scenarios.
- We show that our strategy is generally compatible with other transfer-based attacks and can be conveniently integrated into several state-of-the-art approaches to improve their performance.

2. Related Work

According to the knowledge of attackers, there are generally two categories of threat models in the literature [3]. One is white-box settings where attackers acquire full access to the victim model, for example, the model architecture and parameters. The other one is black-box settings where adversaries only obtain query access, namely, image input uploading and prediction output downloading. Under both scenarios, attackers aim to synthesize adversarial samples to mislead learning algorithms by perturbing legitimate

[1] In this work, we consistently employ the term "feature" to refer to the hidden representations of images extracted by middle layers of DNNs, rather than the raw image pixels.

images in a human-unnoticeable manner. Corresponding to the setting that they are tailored for, attacks are coined white-box attacks and black-box ones [3].

The white-box attack enjoys great popularity among early work on attacking DNNs [36, 9, 18, 5]. Different from the process of model training, they feature an optimization in input space to elevate training loss. Fast gradient sign method (FGSM) alters clean seed images by taking one step along with the sign of the gradient of the model loss function [9]. Its successor, basic iterative method (BIM), iteratively applies FGSM perturbations of smaller magnitude to improve attack success rates [18]. Projected gradient descent (PGD) extends BIM with random start to diversify the synthesized adversarial instances [22]. Carlini and Wagner attacks (C&W) devise a novel attack object to absorb the perturbation budget constraint [5], which also admits the employment of sophisticated optimizers like Adam [17] during the search for deceptive noises. Jacobian-based Saliency Map Attack (JSMA) [25] is tailored for seeking the adversarial noise with the minimal l_0 norm. Therefore, it proposes to prioritize the modification of the most important image pixels to model decisions.

However, white-box attacks hardly reflect the threat to models in practice since only query access is allowed in most realistic cases. Therefore, black-box attacks have attracted increasing attention recently. There are roughly two sorts of black-box attacks according to the mechanism they adopt. One is query-based [24, 2, 10], and the other one is transfer-based [41, 39, 8, 21, 23].

Query-based black-box attacks can settle the susceptible direction of the victim model as per the response of the target model to given inputs [10]. Alternatively, attackers can approximate the loss gradient of the target model through training a local replica [24] or finite difference techniques [2]. However, such attacks usually require excessive queries before a successful trial and thus have limited applicability in practice [8].

Transfer-based black-box attacks are motivated by the transferability of adversarial samples across different models. Concretely, attackers first launch attacks on off-the-shelf local models to which they have white-box access. Then the deceptive samples are directly transferred to fool the remote victim model. Therefore, attackers can apply any white-box attack algorithm in this task, such as FGSM and BIM. Unfortunately, such a straightforward strategy frequently suffers from overfitting to specific weaknesses of local source models and manifesting limited success. We show that by introducing regularizers into the optimization process of adversarial samples, we can significantly improve the performance of such transfer-based black-box attacks.

There also exist two sorts of methods to promote adversarial transferability. Ensemble-based mechanisms often require the deduced distortion to remain harmful for an ensemble of models [21, 32] or images [39, 8, 23]. More related to our work is the regularization-based approach: transferable adversarial perturbation (TAP) introduced by [41]. TAP injects two regularization terms into the vanilla training loss function of the model to guide the search of adversarial manipulations, which alleviates the issue of vanishing gradient and reduces the variations of resultant adversarial samples. We reveal that different models share similar attention when making correct predictions. Therefore, we can exploit this property to boost the transferability of malicious images.

There is a huge body of parallel proposals to enhance the robustness of deep models. Unfortunately, defenders appear to lag far behind in the arms race against adversaries due to the prevailing reactive defense methodology [3]. Failed attempts include pre-processing the input images to diminish malicious noises [11, 1], defensive distillation to mask gradients [26, 6], and feature squeezing to detect adversarial samples [40, 14, 38]. Adversarial training arguably remains the most effective and promising defense to date, where defenders proactively craft deceptive images for their model and augment the clean training data with such instances to train the model [9, 22, 20]. Moreover, exploiting the malicious examples tailored for diverse hold-out models can further strengthen defense and confer robustness to transfer-based black-box attacks [37]. Therefore, we also employ state-of-the-art adversarial trained models to investigate the performance of our strategy against defended models.

3. Preliminaries

We represent a DNN image classifier as a function $f(\mathbf{x})$, which is usually a hierarchical composition of layers of neurons. It outputs the probability vector for a given image $\mathbf{x}$, where $f(\mathbf{x})[i]$ denotes the probability of the image $\mathbf{x}$ belonging to class i. We signify the hidden representation of $\mathbf{x}$ in layer k as $A_k(\mathbf{x}) = f_k(\mathbf{x})$, which consists of multiple feature maps. We will omit the input $\mathbf{x}$ when it is clear from the context. Therefore, A_k^c is the c^{th} feature map in layer k, and $A_k^c[m, n]$ is the output of the neuron with the spatial position $[m, n]$ therein.

Given a model f, an adversarial counterpart $\mathbf{x}'$ of the clean image $\mathbf{x}$ with ground truth label t should satisfy the following two conditions:

$$\text{argmax } f(\mathbf{x}') \neq t, \tag{1}$$

and

$$||\mathbf{x}' - \mathbf{x}||_p \leq \epsilon. \tag{2}$$

The first condition formulates the attack object, namely, misleading the target model to a wrong prediction with the malicious instance $\mathbf{x}'$. The second condition ensures that the induced distortion to the original image $\mathbf{x}$ is imperceptible, since ϵ is usually a fairly small number. We adopt l_∞ norm in this work, as it is the most widely advocated in the community [9]. We also note that our method is generally applicable to other norm choices.

Let $l(f(\mathbf{x}), t)$ signify the loss function to guide the training of model f. Attackers can harness the training loss function as a surrogate for the attack object in Eq. (1) and formulate the generation of an adversarial image $\mathbf{x}'$ as the optimization problem below:

$$\begin{aligned} \text{maximize} \quad & l(f(\mathbf{x}'), t), \\ \text{subject to} \quad & ||\mathbf{x}' - \mathbf{x}||_p \leq \epsilon. \end{aligned} \tag{3}$$

4. Method

Under the setup of transfer-based black-box attacks, attackers can only exploit off-the-shelf local models to manufacture deceptive samples. However, the solution to the above optimization problem of Eq. (3) usually exhibits limited transferability due to overfitting to the source model.

To overcome the pitfall, we propose to augment the vanilla training loss function with an attention-based regularization term in Eq. (3). It encourages the search toward harmful directions common to different deep architectures when updating the deceptive perturbations.

As illustrated in Figure 2, we will first approximate model attention over extracted features with corresponding back-propagated gradients (Section 4.1). Then we formulate the destruction of attention-weighed combinations of feature maps as a regularization term to Eq. (3) (Section 4.3). Finally, we explain the algorithm we employ to solve the reformulated optimization problem for adversarial sample generation (Section 4.4).

4.1. Attention Extraction

We suppose that transfer-based attackers can benefit from explicitly attacking hidden feature detectors within DNN image classifiers. Different from traditional image classification approaches that count on hand-designed features, deep learning-based image classifiers are renowned for their competence to automatically extract discriminative features from images [15]. We can thus separate a DNN image classifier into two parts: a hierarchical feature extraction module and a softmax classifier. The learned feature extractors of a DNN image classifier are often so generic that they can adapt to different domains and tasks [31]. Inspired by the fact, we expect that lots of feature descriptors are shared among different architectures for the same task, for example, the edge detector for face recognition. Therefore, if the synthesized adversarial noise can not only fool the final prediction of a target model, but also severely contaminate the extracted intermediate features, it is more likely to transfer across different models.

However, polluting the intermediate features under a restricted perturbation budget may still suffer from overfitting to a specific model, since there are some feature filters exclusive to the source model. To address the issue, we ask the deceptive noise to focus on undermining critical features for the model prediction. We assume that although different models may seek for distinct feature evidence to arrive at the final decision, the most crucial features one model pays attention to are frequently shared among various architectures. For example, for a cat image, it is very likely that different models all need to exploit the face-related features when making a correct prediction (Figure 1).

Consequently, we need to derive the importance of diverse features to model decisions, namely, the model's attention. We regard one whole feature map as basis feature detectors. Therefore, we approximate the importance of feature map A_k^c (*i.e.,* the c-th feature map in layer k) to class t with spatially pooled gradients:

$$\alpha_k^c[t] = \frac{1}{Z} \sum_m \sum_n \frac{\partial f(\mathbf{x})[t]}{\partial A_k^c[m, n]}. \tag{4}$$

Here Z is a normalizing constant such that $\alpha_k^c[i] \in [-1, 1]$. We name $\alpha_k[t]$ the **attention weight** of the model to various features extracted in layer k regarding class t.

4.2. Attention Visualization

Built upon the deduced attention weights, we propose to visualize the attention maps of various models with the technique of [30]. Such visualization aims to explore what the model attention looks like and examine whether distinct models showcase similar attentions for the same correctly classified image. Therefore, it serves as a proof of concept for our idea.

Specifically, we first scale different feature maps with corresponding model attention weights $\alpha_k^c[t]$. Then we perform channel-wise summation of all feature maps in the same layer. After that, we proceed with a ReLU operation to derive the attention map for the label prediction t:

$$H_k^t = \text{ReLU}(\sum_c \alpha_k^c[t] \cdot A_k^c). \tag{5}$$

We apply the ReLU operation here to remove negative pixels in the attention map so that we can focus on supportive features, which have a positive influence on the class of interest. Negative pixels probably stands for features from other classes. We note that H_k^t is of the same spatial resolution as the feature maps in layer k. Since the size of the feature maps varies across different layers and models, we finally bilinearly interpolate the attention map to the same resolution as the input image for better comparison.

For the same cat image, Figure 1 displays the attention heatmaps of various ImageNet classifiers regarding the cat prediction. We note that all these models correctly classify the cat image. It corroborates our assumption that diverse models exhibit similar attention when making a correct prediction.

4.3. Critical Feature Destruction

After obtaining the model attention, we can now ask the adversarial samples to not only mislead the final decision of the target model, but also destroy the vital intermediate features. We combine both goals as a novel surrogate attack object function for Eq. (1):

$$\begin{aligned} \text{maximize} \quad & J(\mathbf{x}, \mathbf{x}', t, f), \\ \text{where} \quad & J(\mathbf{x}, \mathbf{x}', t, f) = l(f(\mathbf{x}'), t) + \\ & \lambda \sum_k ||H_k^t(\mathbf{x}') - H_k^t(\mathbf{x})||^2. \end{aligned} \tag{6}$$

Here the first term in J is the vanilla training loss (*i.e.,* the cross-entropy loss), and we maximize it to achieve the first goal. The second term measures the distance between attention-weighed combinations of original feature outputs and the corrupted counterparts. It corresponds to preferring great alterations to features with large attention weight and thus accounts for the second goal. λ is a tunable weight to control the regularization effect of the second term.

4.4. Optimization Algorithm

After substituting the proposed attack object function (Eq. (6)) for that in Eq. (3), we can now reformulate the manufacture of transferable adversarial samples as the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & J(\mathbf{x}, \mathbf{x}', t, f), \\ \text{where} \quad & J(\mathbf{x}, \mathbf{x}', t, f) = l(f(\mathbf{x}'), t) + \\ & \lambda \sum_k ||H_k^t(\mathbf{x}') - H_k^t(\mathbf{x})||_2, \\ \text{subject to} \quad & ||\mathbf{x}' - \mathbf{x}||_p \leq \epsilon. \end{aligned} \tag{7}$$

Therefore, we can freely apply different backbone optimization algorithms to acquire a solution. For fair comparisons, the optimization strategy we apply in this paper is the same as the white-box benchmark (BIM), which is an iterative refinement of FGSM.

Concretely speaking, BIM extends FGSM into an iterative procedure with a smaller step size ϵ' in each run:

$$\mathbf{x}'_{k+1} = \text{clip}_{\mathbf{x},\epsilon}\{\mathbf{x}'_k + \epsilon' \ \text{sign}(\frac{\partial l(f(\mathbf{x}'_k), t)}{\partial \mathbf{x}})\}, \tag{8}$$

where $\mathbf{x}'_0 = \mathbf{x}$, and $\text{clip}_{\mathbf{x},\epsilon}\{\mathbf{x}'\}$ conducts pixel-wise clipping for the resultant image $\mathbf{x}'$. Accordingly, it guarantees that $\mathbf{x}'$ stays within the l_∞ ϵ-neighborhood of the seed image $\mathbf{x}$.

Algorithm 1 Attention-guided Transfer Attack (ATA)

```
Require: A classifier $f$, attack object function $J$ (Eq. (6)),
    a clean image $\mathbf{x}$, and its ground-truth label $t$
Require: The perturbation budget $\epsilon$, iteration number $K$
Ensure: $||\mathbf{x}' - \mathbf{x}||_\infty \leq \epsilon$
$\epsilon' = \frac{\epsilon}{K}$
$\mathbf{x}'_0 = \mathbf{x}$
for $k = 0$ to $K - 1$ do
    $\mathbf{x}'_{k+1} = \text{clip}_{\mathbf{x},\epsilon}\{\mathbf{x}'_k + \epsilon' \ \text{sign}(\frac{\partial J(\mathbf{x}, \mathbf{x}'_k, t, f)}{\partial \mathbf{x}})\}$
end for
return $\mathbf{x}' = \mathbf{x}'_K$
```

Algorithm 1 summarizes our algorithm to craft transferable adversarial samples. In short, it features an introduction of attention-based regularization term to the optimization procedure of BIM.

5. Experiments

In this section, we first elucidate the experimental setup in Section 5.1. Then we report the results of our attack against diverse top-performance models and make comparisons with numerous state-of-the-art benchmark approaches in Section 5.2. Subsequently, we investigate the effect of hyper-parameters on our attack success rates in Section 5.3. Finally, we verify the complementing effect of our strategy on compatible algorithms in Section 5.4.

5.1. Experimental Setup

We focus on attacking image classifiers trained on ImageNet [29], which is the most broadly recognized benchmark task for transfer-based black-box attacks [20, 4]. We follow the protocol of the baseline method [41] to curate experimental datasets and target models for fair comparisons.

Dataset. We need two sorts of datasets to develop and assess our attacks, respectively. The development dataset is the ILSVRC 2012 validation dataset [29], where we fine-tune our hyper-parameters. The test data adopted to assess our technique is the ImageNet-compatible dataset released by the NeurIPS 2017 adversarial competition [20]. This test set contains 1000 images that are not included in the original ImageNet dataset. Therefore, it satisfies the requirement of evaluating the generalization capability of attack algorithms in practice.

Target model. We examine our technique with both undefended and defended models. As for undefended models, we employ numerous top-performance models with diverse architectures, including ResNet V2 [12, 13], Inception V3 [35], Inception V4 [34], and Inception-ResNet V2 [34][2].

[2] These pre-trained models are all publicly available at https://github.com/Cadene/pretrained-models.pytorch.

	Attack	ResNet V2	Inception V3	Inception V4	Inception-ResNet V2	Ensemble
	No Perturbation	89.6%	96.4%	97.6%	100%	99.8%
	Random Noise	84.5%	91.7%	94.6%	97.8%	98.1%
ResNet V2	FGSM	14.6%	56.3%	64.8%	66.8%	63.1%
	BIM	**4.4%**	53.2%	62.0%	63.8%	54.3%
	C&W	37.7%	94.5%	96.4%	98.5%	98.5%
	JSMA	27.2%	59.3%	65.2%	62.1%	64.4%
	TAP	9.5%	**51.2%**	60.1%	55.5%	50.3%
	ATA	8.7%	52.9%	**58.3%**	**55.1%**	**49.4%**
Inception V3	FGSM	65.7%	27.2%	70.2%	72.9%	76.2%
	BIM	76.8%	**0.01%**	67.7%	70.2%	73.6%
	C&W	86.9%	24.5%	93.5%	96.2%	96.0%
	JSMA	66.4%	22.4%	57.2%	60.3%	68.9%
	TAP	48.2%	0.1%	24.5%	26.3%	34.2%
	ATA	**47.2%**	0.1%	**22.1%**	**25.7%**	**31.9%**
Inception V4	FGSM	68.3%	67.1%	50.3%	72.8%	76.4%
	BIM	62.1%	40.9%	**0.9%**	69.1%	55.5%
	C&W	86.7%	91.7%	49.5%	93.2%	92.9%
	JSMA	70.7%	68.9%	30.0%	65.2%	68.9%
	TAP	**58.4%**	27.3%	1.8%	24.2%	51.7%
	ATA	59.9%	**24.8%**	**0.9%**	**22.1%**	**50.3%**
Inception-ResNet V2	FGSM	71.7%	69.0%	76.5%	57.2%	78.7%
	BIM	60.4%	41.5%	51.5%	**1.2%**	54.5%
	C&W	85.6%	91.7%	92.4%	49.0%	93.5%
	JSMA	55.4%	62.7%	66.8%	50.3%	64.9%
	TAP	53.3%	25.9%	33.2%	4.8%	48.2%
	ATA	**49.8%**	**22.1%**	**30.1%**	**1.2%**	**45.3%**

Table 1: Accuracy of undefended models under attacks. The first column shows the source model employed, while the first row states the remote target models.

We also attack the corresponding ensemble model (referred to as **Ensemble**), whose prediction is the average probability output of all the above models.

When it comes to the defended models, we adopt multiple state-of-the-art adversarially trained models as remote targets [37, 19], since adversarial training is arguably the most promising and effective defense to date [22]. These adversarially trained models include adversarially trained Inception V3 (Adv-Inc-v3), adversarially trained Inception-ResNet V2 (Adv-IncRes-v2), adversarially trained Inception V3 with deceptive samples from an ensemble of three models (Ens3-Adv-Inc-v3) and four models (Ens4-Adv-Inc-v3), respectively[3].

Baseline. We compare the performance of our attack with three kinds of benchmark techniques. As a naive baseline attack, we attach Gaussian noise under the same norm constraint to clean images, which is denoted as the **Random Noise** attack. More importantly, we compare our strategies with diverse state-of-the-art white-box attacks, including FGSM [9], BIM [18], C&W [5], and JSMA [25], to showcase the effectiveness of our algorithm in alleviating the overfitting issue and improving the transferability of white-box attacks. Since the original C&W implementation cannot strictly meet the l_∞ budget, we employ the modified l_∞ version of C&W as introduced by [41], which can explicitly satisfy the l_∞ norm constraint. Similar to our strategy, TAP [41] boosts adversarial transferability through two regularization terms and is the state-of-the-art approach under this category. Therefore, we also include TAP in the competing benchmarks.

Metric. We compare different attacks via the top-1 accuracy of target models. Accordingly, lower accuracy of victim models on the synthesized adversarial samples represents better attack performance.

Parameter. We only include the last convolutional layer of the source model in our regularization term based on our preliminary experiments. For fair comparisons, we adopt default parameters as recommended in benchmark approaches and Foolbox [41, 28]. The random noise is sampled from a clipped normal distribution with mean 0 and variance 1.

[3] These models are all publicly available at https://github.com/tensorflow/models/tree/master/research/adv_imagenet_models.

	Attack	Adv-Inc-v3	Adv-IncRes-v2	Ens3-Adv-Inc-v3	Ens4-Adv-Inc-v3
ResNet V2	FGSM	62.1%	85.7%	77.4%	77.8%
	BIM	64.7%	82.6%	72.3%	74.7%
	C&W	94.0%	96.3%	92.8%	90.5%
	JSMA	58.2%	80.3%	75.2%	75.9%
	TAP	**49.2%**	66.5%	59.1%	**56.0%**
	ATA	**49.2%**	**60.3%**	**57.8%**	58.2%
Inception V3	FGSM	72.1%	93.6%	85.1%	86.4%
	BIM	82.4%	93.9%	88.2%	88.5%
	C&W	93.0%	96.4%	92.3%	90.0%
	JSMA	81.4%	93.6%	89.5%	87.4%
	TAP	55.8%	68.8%	61.3%	60.6%
	ATA	**54.1%**	**61.3%**	**60.2%**	**60.2%**
Inception V4	FGSM	74.8%	93.8%	88.1%	86.9%
	BIM	71.9%	92.9%	85.3%	85.3%
	C&W	92.8%	94.8%	91.9%	90.0%
	JSMA	70.6%	91.7%	87.9%	88.4%
	TAP	**65.3%**	90.4%	83.2%	87.3%
	ATA	69.1%	**89.8%**	**80.9%**	**82.9%**
Inception-ResNet V2	FGSM	73.9%	92.7%	86.9%	87.3%
	BIM	70.8%	92.9%	84.8%	86.9%
	C&W	91.8%	94.9%	91.9%	89.3%
	JSMA	72.1%	94.9%	83.3%	84.6%
	TAP	60.5%	87.8%	81.2%	84.3%
	ATA	**58.9%**	**85.9%**	**80.9%**	**81.4%**

Table 2: Accuracy of adversarially trained models under attacks. The first column shows the source model employed, while the first row states the remote target models.

Following [41], we fix the perturbation budget ϵ to 16 for all methods. We conduct grid search on the development dataset to settle the best hyper-parameter for our algorithm. In all our experiments, the attack iteration number K is set to 5. The regularization weight λ roughly balances the contribution of each term in the loss function J (Eq. (6)).

5.2. Transferability of Attacks

Here we study the performance of our attack against both undefended and defended models. Specifically, we first fix a source model and run our algorithm on the model to produce adversarial samples. The resultant samples are then directly fed to the source model and other different models to simulate the white-box and black-box setups, respectively.

We first attack undefended models, and Table 1 reports the results. We make the following observations. First, all these models possess impressive clean accuracy and appear resistant to random noise. Models with higher capacity usually exhibit better performance. Second, under white-box setups, BIM is the winning attack. Our algorithm achieves matching results to BIM and significantly outperforms the others. Third, under black-box settings, our attack significantly boosts the transferability of BIM. For example, when employing Inception V3 as the source model, our attack witnesses an average gain of 40.4% on attack success rates compared to BIM. Moreover, we defeat all the other benchmark methods with a significant margin, except for two cases, where we only lag a little behind TAP. We note that TAP employs two regularization terms, one for maximizing internal feature distances and the other for smoothing resultant perturbations. Contrarily, by applying only one regularization term to maximize attention-weighed internal feature distances, our method outperforms TAP in almost all cases.

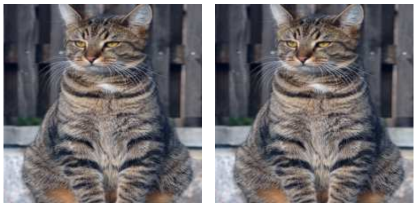
(a) Clean (b) ATA

Figure 3: A clean source image and the corresponding adversarial image crafted with the proposed ATA. The target model is Inception V3. Although the perturbation is imperceptible to humans, it can successfully fool top-performance models.

We next attack models defended by adversarial training. For fair comparisons with the baseline approach [41], we stick to employing undefended models as local source models. Therefore, we explore a more challenging black-box scenario where the source and target models possess more distinct property. We present the results in Table 2. We draw the following conclusions. First, we consistently improve the transferability of BIM to a great extent. For example, we increase the attack success rate of BIM by 29.3% on average, when applying Inception V3 as the source model. Second, our ATA remarkably outperforms all the other benchmarks except for two cases, where we only slightly lag behind TAP.

Figure 3 displays one generated adversarial image against Inception V3 with our attack. We note that the deduced manipulations to the clean image are hardly visible. It confirms that our attack is stealthy.

5.3. Effect of Hyper-parameters on Attack Success Rates

The regularization weight λ is the dominant hyper-parameter in our algorithm, and here we explore its effect on our attack success rates. Specifically, we vary λ while keeping the other parameters fixed to synthesize adversarial samples. Similar to previous experiments, we report the top-1 accuracy of target models on the resultant malicious

Attack	ResNet V2	Inception V3	Inception V4	Inception-ResNet V2	Ensemble	Adv-Inc-v3	Adv-IncRes-v2	Ens3-Adv-Inc-v3	Ens4-Adv-Inc-v3
TAP	58.4%	27.3%	1.8%	24.2%	51.7%	65.3%	90.4%	83.2%	87.3%
TAP+ATA	53.6%	22.7%	0.8%	19.8%	48.1%	57.9%	85.3%	73.2%	72.9%
TI	57.1%	30.9%	2.1%	26.9%	58.3%	62.7%	91.4%	81.9%	83.5%
TI+ATA	56.2%	24.9%	0.7%	24.2%	50.1%	57.9%	88.2%	76.9%	77.6%

Table 3: Accuracy of models under attacks that combine the proposed ATA and compatible algorithms.

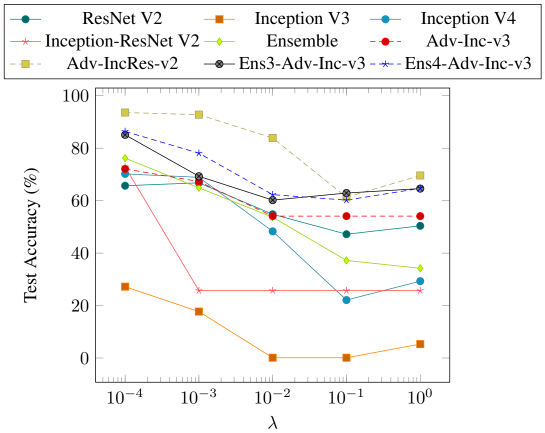

Figure 4: The effect of hyper-parameter λ on our attack success rates.

examples to measure the attack success rates.

Figure 4 illustrates the effect of λ on attack success rates against all undefended and defended models, where the source model is Inception V3. We vary λ from 1×10^{-4} to 1 with a step size of 1 in log scale. We observe similar trends when employing other source models and thus omit such results. We note that there is generally a trade-off between the two terms in J (Eq. (6)). Because under a restricted perturbation budget, it is crucial to balance the contribution from each term to alleviate overfitting.

5.4. Complementing Effect of the Proposed Strategy

In principle, our strategy is compatible with other transfer-based black-box attacks. Therefore, we can conveniently integrate the proposed technique with such algorithms. We select two sorts of cutting-edge transfer-based attacks to corroborate the complementing effect introduced by our strategy. One is the ensemble-based translation-invariant attack (TI) developed by [8], and the other one is the regularization-based transferable adversarial perturbation (TAP) proposed by [41]. With the integrated attacks, we conduct experiments similar to Section 5.2.

Specifically, the combination of TI and ATA will only modify the update rule of Algorithm 1 as:

$$\mathbf{x'}_{k+1} = \text{clip}_{\mathbf{x},\epsilon}\{\mathbf{x'}_k + \epsilon' \,\text{sign}(\mathbf{W} * \frac{\partial J(\mathbf{x}, \mathbf{x'}_k, t, f)}{\partial \mathbf{x}})\}, \quad (9)$$

where $\mathbf{W}$ is a pre-defined kernel, and $*$ signifies convolution operation. The integration of TAP and ATA only adds the following term into the attack object function J (Eq. (6)):

$$\eta||\mathbf{S} * (\mathbf{x'} - \mathbf{x})||_1, \quad (10)$$

where $\mathbf{S}$ is a pre-defined convolution filter. We abandon the other term in TAP for simplicity because we do not have the issue of vanishing gradients.

Table 3 shows the results with Inception V4 as the source model. In black-box settings, our strategy promotes the average attack success rate of TAP and TI by 6.8% and 4.6%, respectively. In white-box settings, our strategy can also further improve their attack success rates. Therefore, it corroborates the complementing effect of our technique to existing efforts.

6. Conclusion

In this work, we introduce an attention-guided transfer attack to synthesize adversarial samples against black-box DNNs without any feedback information from the target model. The proposed strategy exploits the attention of the source model to regularize the search direction for adversarial samples. Consequently, it can focus on undermining critical features that different models count on and manifest remarkable transferability. We conduct extensive experiments to validate the effectiveness of our approach and confirm its superiority to state-of-the-art baselines. Therefore, our attack can more faithfully expose the vulnerability of deep models and serve as a strong benchmark when examining defenses.

Acknowledgment

We thank anonymous reviewers for their valuable comments. The work described in this paper was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK 14210717 of the General Research Fund and CUHK 2300174 of the Collaborative Research Fund, No. C5026-18GF).

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. 3

[2] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *The European Conference on Computer Vision (ECCV)*, pages 158–174. Springer, 2018. 3

[3] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018. 1, 2, 3

[4] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On evaluating adversarial robustness. *arXiv:1902.06705*, 2019. 5

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, 2017. 1, 3, 6

[6] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: Elastic-net attacks to deep neural networks via adversarial examples. In *The Thirty-second AAAI Conference on Artificial Intelligence*, 2018. 3

[7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193, 2018. 1

[8] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3, 8

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 1, 3, 4, 6

[10] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning (ICML)*, 2019. 3

[11] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 770–778, 2016. 2, 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *The European Conference on Computer Vision (ECCV)*, pages 630–645. Springer, 2016. 2, 5

[14] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defense: Ensembles of weak defenses are not strong. In *11th USENIX Workshop on Offensive Technologies (WOOT 17)*, 2017. 3

[15] Geoffrey E Hinton. Learning multiple layers of representation. *Trends in Cognitive Sciences*, 2007. 4

[16] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning (ICML)*, pages 2142–2151, 2018. 1

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 3

[18] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLR Workshop*, 2017. 3, 6

[19] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*, 2017. 6

[20] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*, 2018. 1, 3, 5

[21] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations (ICLR)*, 2017. 1, 3

[22] 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 (ICLR)*, 2018. 3, 6

[23] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1765–1773, 2017. 3

[24] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *The Asia Conference on Computer and Communications Security (ASIA CCS)*. ACM, 2017. 3

[25] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *The IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387, 2016. 3, 6

[26] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 582–597, 2016. 3

[27] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and SS Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 2018. 1

[28] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *ICML Workshop*, 2017. 6

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015. 5

[30] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 4

[31] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *CVPR Workshop*, 2014. 4

[32] Yash Sharma, Tien-Dung Le, and Moustafa Alzantot. CAAD 2018: Generating transferable adversarial examples. *arXiv preprint arXiv:1810.01268*, 2018. 3

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 2

[34] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *The Thirty-first AAAI Conference on Artificial Intelligence*, 2017. 5

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE International Conference on Computer Vision (ICCV)*, 2016. 2, 5

[36] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 1, 3

[37] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018. 3, 6

[38] Weibin Wu, Hui Xu, Sanqiang Zhong, Michael R. Lyu, and Irwin King. Deep Validation: Toward detecting real-world corner cases for deep neural networks. In *International Conference on Dependable Systems and Networks (DSN)*, pages 125–137. IEEE, 2019. 3

[39] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3

[40] Weilin Xu, David Evans, and Yanjun Qi. Feature Squeezing: Detecting adversarial examples in deep neural networks. In *The Network and Distributed System Security Symposium (NDSS)*, 2018. 3

[41] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *The European Conference on Computer Vision (ECCV)*, 2018. 1, 3, 5, 6, 7, 8