Enhancing Targeted Attack Transferability via Diversified Weight Pruning

Hung-Jui Wang, Yu-Yu Wu, Shang-Tse Chen
National Taiwan University
{r10922061, r10922018, stchen}@csie.ntu.edu.tw

Abstract

Malicious attackers generate adversarial instances by introducing imperceptible perturbations into data. Even in the black-box setting where model details are concealed, attackers still exploit networks with cross-model transferability. Despite the notable success of untargeted attacks, achieving targeted attack transferability remains a challenging endeavor. Recent investigations have demonstrated the efficacy of ensemble-based techniques. However, utilizing additional models to carry out ensemble attacks brings extra costs. To reduce the number of white-box models required, model augmentation methods augment the given network to produce a variant of diverse models, contributing useful gradients for attack. In this work, we propose Diversified Weight Pruning (DWP) as an innovative model augmentation technique specifically designed to facilitate the generation of transferable targeted attacks. In contrast to prior techniques, DWP preserves essential connections while simultaneously ensuring diversity among the pruned models, both of which are identified as pivotal factors for targeted transferability. DWP is shown effective with experiments on ImageNet under challenging conditions, with enhancements of up to 10.1%, 6.6%, and 7.0% across adversarially trained models, Non-CNN architectures, and Google Cloud Vision respectively.

1. Introduction

Deep neural networks (DNNs) have achieved noteworthy advancements across domains of applications. However, recent investigations have uncovered vulnerabilities within DNNs. Adversaries can launch adversarial attacks, which introduce imperceptible alterations into benign images, deceiving classification models. Consequently, numerous studies on adversarial attacks have been developed to assess the robustness of DNNs. [2, 22, 45]. Adversarial perturbations can be effectively crafted through gradient-based algorithms in the white-box setting. Even within the black-box scenarios where details of the target models' implementations and parameters are concealed, malicious actors can
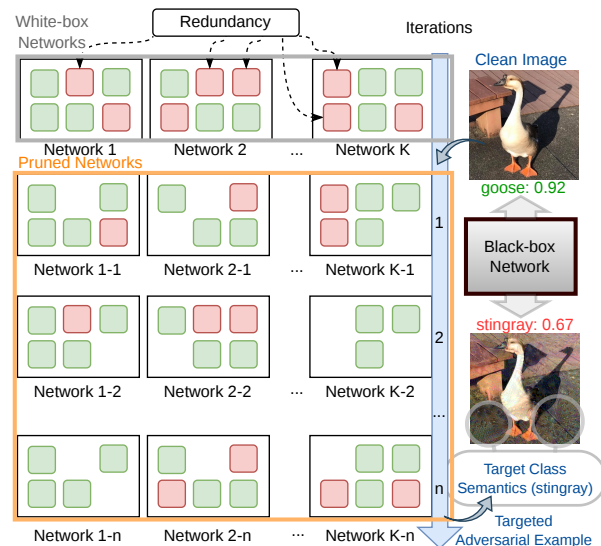

Figure 1. **Big picture of DWP.** We leverage the weight pruning technique to produce additional diversified models from existing white-box networks per attack iteration. By protecting necessary weight connections in each network, the quality of models is well-preserved. These additional pruned models can better impose semantics of the target class, yielding higher targeted transferability.

still exploit the victim by employing cross-model transfer attacks with substitute networks. The ability to transfer adversarial attacks between distinct models poses a significant threat to the reliability of deep learning applications and has drawn substantial attention.

Previous works have introduced diverse methods enhancing the transferability of untargeted attacks. Despite the achievement that untargeted transfer attacks have made, where the attack success rate can be over 90%, obtaining targeted attack transferability remains challenging [5, 27]. Nevertheless, the targeted attack could be more practical in a real-world scenario. For example, transferable targeted attacks can be employed as "honey pots" within CAPTCHA systems designed to distinguish between human users and automated bots. By launching transferable attacks against robotic agents, it is possible to induce these agents to pro-

vide erroneous responses to a predefined class. Given that the probability of a human user providing such a response is low, we can infer that a robot is attempting to subvert the system. Therefore, an effective targeted transferable attack is vital since the bots' implementations are unknown.

Ensemble-based approaches have been shown effective in generating transferable targeted attack examples with multiple networks as substitute models [27, 55]. The gradients provided by the substitute models are accumulated and recalculated to provide a more general update direction for the adversarial examples. Collecting a sufficient amount of models to participate in the ensemble attack is crucial to ensure the diversity of gradients to escape the local minimum of the network. Nonetheless, the necessity of extra white-box networks brings overhead for the attack pipeline. To reduce the resources needed while maintaining the power of the ensemble attack, model augmentation techniques create additional networks by altering the existing ones and developing adversarial examples with these generated networks altogether. Ghost Networks (GN) [24] inserts extra dropout layers and random skip connections into the original networks to produce additional models. Duel-Stage Random Erosion (DSNE) [8] improves GN by introducing uniform erosion on the remaining parameters after applying GN, further increasing the diversity of the generated models.

However, those methods randomly drop neurons away without considering their significance to the prediction and lack of protection on necessary parameters. To avoid destroying network performance, prior works require heavy tuning on the hyper-parameters like dropout, skip connection, and the second erosion rates. The inserted dropout layers' location should also be examined. Those hyper-parameters vary in architecture and require sophisticated investigation to obtain a satisfactory result. In the case of the targeted attack, the quality of white-box substitute models plays a critical role. Rather than merely moving away from the original class, the semantics of targeted adversarial examples need to be close to the target class to acquire higher transferability [23, 31]. Without properly tuning for the previous model augmentation methods, the network performance may be severely affected since important parameters are altered. Dropping or disturbing the significant components in substitute networks can mislead targeted adversarial examples and yield worse transferability.

To deal with these problems, we propose an improved model augmentation approach **D**iversified **W**eight **P**runing (DWP) using the idea of model compression. Model compression reduces the storage and computation overhead without substantially affecting performances [9, 12, 21, 28]. With the over-parameterized property [4] of neural networks, weight pruning [12] can compress the model while maintaining accuracy by removing redundant connections only. To generate transferable targeted adversarial exam-

ples, we apply random weight pruning to each accessible network to form additional ones. The attack success rate is improved by ensemble attacks with generated diverse models. Fig. 1 summarizes our proposed pipeline.

In summary, our contributions are as follows:
- We propose a simple yet effective transferable targeted attacks methodology, **D**iversified **W**eight **P**runing (DWP) that leverages the idea of weight pruning to preserve necessary parameters within networks, reducing the time needed for searching optimal hyper-parameters because important connections are protected.
- Comprehensive experiments are conducted on the ImageNet-compatible dataset used in the NeurIPS 2017 adversarial attack competition [20]. The average targeted success rate of DWP reaches 81.30% across CNNs.
- DWP remains competitive in challenging scenarios, improving the targeted success rate with up to 10.1% and 6.6% on average when transferring to adversarially trained models and non-CNN architectures.
- DWP exhibits its efficacy by generating targeted attacks on the real-world Google Cloud Vision service, yielding a notable improvement of 7.0%.

2. Related work

2.1. Transferable attack

We focus on simple transferable attacks [55], which require neither additional data nor further training on networks compared to the resource-intensive ones [11, 17, 18, 47, 52]. Existing attack methods can be categorized into 4 groups: input transformation, gradient optimization, ensemble and model augmentation, and advanced loss function.

Input transformation Motivated by the success that data augmentation has achieved in standard training [37], several works advocate attacking the transformed input to prevent overfitting on white-box models to increase the transferability on black-box ones. DI [49] uses random resizing and padding throughout the iterative attack. TI [6] enumerates several translated inputs and fuses the gradients from all augmented data. SI [25] leverages the scale-invariant property of CNNs and employs multiple scale copies from each input image. Admix [47] extends the concept of mixup [54], attacking the mixup version of the data.

Gradient optimization Optimization-based methods are widely adopted [2, 10, 19, 40] in generating adversarial examples. With iterative optimization-based methods [2, 19], better solutions to the objective can be obtained by iteratively attacking models and updating adversarial examples. Dong *et al*. [5] combine momentum techniques with iterative attacks, accumulating gradients at each iteration to escape local optimum and stabilize the updating direction. Lin

et al. [25] apply Nesterov accelerated gradient for optimization, giving adversarial examples an anticipatory updating to yield faster convergence. Wang and He [46] introduce variance tuning-based momentum to reduce the variance of gradients at each iteration. Huang and Kong [15] leverage integrated gradients to include smoothing, attention modification, and optimization during attacking.

Ensemble and model augmentation   Adversarial examples generated by ensembling multiple white-box networks are more likely to transfer to black-box ones [27]. Instead of simply fusing the output confidence from each model, Xiong *et al.* [50] suggest reducing the gradient variance of collected networks. To further improve ensemble-based approaches, model augmentation produces additional models from the existing one. Li *et al.* [24] acquire ghost networks (GN) by employing dropout and skip connections on the existing model and ensemble all generated models' predictions. DSNE [8] further improves the diversified ensemble via dual-stage erosion. Yuan *et al.* [53] use reinforcement learning to automatically find transformations suitable with white-box networks to yield more diversity.

Advanced loss function   While cross-entropy is a widely used loss function in standard training, it also serves as the objective for many adversarial attack algorithms. However, cross-entropy is found to have a saturation problem in targeted attack scenarios [23], as the output confidence of the target class approaches one. To this end, alternative loss functions attempt to provide more suitable gradients for optimization. Li *et al.* [23] leverage Poincaré distance as the loss function, which amplifies the gradient magnitude as the confidence of the target class grows. Zhao *et al.* [55] propose a simple logit loss, which has constant gradient magnitude regardless of the output probability.

2.2. Network pruning

The intensive cost of computation and storage hinders applications of neural networks, especially on embedding systems. Network Compression aims to reduce the scale of networks while maintaining their performance, making them more feasible for deployment. With the over-parameterized property [4], network pruning is a compression technique that aims at removing redundancy within the model. LeCun *et al.* [21] use the second-derivative information to find redundant weights in networks. Han *et al.* [12] show that neural networks can highly preserve performance even if trimming more than half of their connections. It is also investigated that retraining the pruned model after compression can achieve higher accuracy [9, 28].

3. Methodology

Unlike simply decreasing the accuracy in untargeted attack, the adversarial examples semantics require proximity to the intended class to maximize the targeted transferability [23, 31]. The quality of white-box substitute models plays an important role in assuring attacks' efficacy.

Model augmentation techniques provide an efficient way to generate a group of auxiliary models from the existing one to participate in the ensemble attack. Since the generated models are different from the original network, they can produce diverse gradients given input, which is valuable in enhancing the attack performance. However, extant methodologies employ random neuron dropout without considering their relevance to predictive outcomes. The network's performance may deteriorate substantially as critical parameters are perturbed or dropped. It requires meticulous tuning for hyper-parameters such as dropout and skip connection rates to secure the quality of generated models. As the architecture of models varies, these hyper-parameters exhibit structural variation and demand intricate examination to yield satisfactory results.

To reduce the efforts for tuning hyper-parameters in the existing model augmentation methodologies, we design a simple yet effective algorithm **D**iversified **W**eight **P**runing (DWP) that leverages the idea from model compression to generate networks in a performance-aware way. Given that DWP preserves the essential parameters and only alters the redundant neurons, it acquires high-quality auxiliary networks without heavy tunning on hyper-parameters. Additionally, as the vital parameters are protected, ensuring good semantic representation in the auxiliary models, the targeted attack transferability can be further boosted.

In this section, we establish current state-of-the-art techniques for iterative attacks and demonstrate how DWP creates auxiliary models from the given white-box network and combines them with other techniques. Due to its simplicity of design, DWP enables a seamless plug-and-play in combination with relevant methodologies.

3.1. Preliminary

Given a network θ and a benign example x, we generate a targeted adversarial example x^{adv} for the target class y^{target} by solving the following constrained optimization problem:

$$\arg\min_{x^{\text{adv}}} J(x^{\text{adv}}, y^{\text{target}}; \theta) \quad \text{s.t.} \quad \left\| x^{\text{adv}} - x \right\|_{\infty} \leq \epsilon, \quad (1)$$

where J is the loss function for multiclass classification and ϵ is the perturbation budget under l_{∞} norm aligning with previous works. We use logit loss as our objective function J following Zhao *et al.* [55] to circumvent the gradient saturation problem of cross-entropy. To obtain a strong baseline, we choose methods from gradient optimization (NI)

and input transformation (SI, TI, DI) abbreviated as NI-SI-TI-DI in combination with the proposed DWP. Additional baseline details are provided in Appendix Sec. 6.1.

3.2. Diversified Weight Pruning

Our proposed DWP increases the diversity of white-box models for the ensemble via weight pruning techniques. First, we sort the connections of the white-box network by the L1 norm of their weight values since it is better than L2 at preserving accuracy [12]. With a predefined rate r, we only consider the lowest $(100 \cdot r)\%$ "prunable" since weights with small absolute values are shown unnecessary [12]. Networks can preserve accuracy after these connections are pruned away even without retraining [12].

For our pruning operation, we first identify the set of prunable weights. Let γ be the $(100 \cdot (1 - r))$-th percentile of weights in θ. We formulate the prunable set as:

$$\Gamma(\theta, r) = \{w \in \theta | w < \gamma\} \subseteq \theta. \tag{2}$$

With $\Gamma(\theta, r)$ collecting all the prunable weights of θ, we introduce an indicator vector for it:

$$\Pi_{\Gamma(\theta, r)} = (\lambda_1, \lambda_2, ..., \lambda_\kappa), \tag{3}$$

where κ is the total number of weights in $\theta = \{w_1, w_2, ..., w_\kappa\}$ including non-prunable ones. λ_i is determined by whether its corresponding $w_i \in \theta$ is in the prunable subset $\Gamma(\theta, r)$:

$$\lambda_i = \begin{cases} 1, & \text{if } w_i \in \Gamma(\theta, r) \\ 0, & \text{otherwise} \end{cases}. \tag{4}$$

Supported by the indicator vector $\Pi_{\Gamma(\theta, r)}$, pruning operation $P(\cdot)$ can protect the non-prunable weights by masking:

$$P(\theta, r) = (\mathbf{1}_\kappa - \Pi_{\Gamma(\theta, r)} \odot \mathbf{b}) \odot \theta, \tag{5}$$

where $\odot$ denotes the element-wise multiplication, $\mathbf{1}_\kappa = (1, 1, ..., 1) \in R^\kappa$ denotes an all-one vector and $\mathbf{b} = (b_1, b_2, ..., b_\kappa)$ is a vector with $b_i \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(p_{\text{bern}})$. p_{bern} represents the probability of pruning each connection independently. $\Pi_{\Gamma(\theta, r)}$ and $\mathbf{b}$ both are binary masks with identical layout as θ. $\Pi_{\Gamma(\theta, r)}$ is responsible for protecting non-prunable weights, while $\mathbf{b}$ is for random pruning. Each binary element in $\Pi_{\Gamma(\theta, r)} \odot \mathbf{b}$ indicates whether to prune the corresponding weight value in θ. The main difference from Dropout [39] used in previous model augmentations [8, 24], is that DWP only considers dropping prunable weights.

Instead of producing all the pruned models beforehand, we acquire pruned models at each iteration right before gradient computing. With this longitudinal ensemble strategy [8, 24], the storage and computation overhead is almost identical to the original attack procedure. We provide the new attack objective that employs DWP as the following:

$$\arg\min_{x^{\text{adv}}} J(x^{\text{adv}}, y^{\text{target}}; P(\theta, r)) \quad \text{s.t.} \quad \left\| x^{\text{adv}} - x \right\|_\infty \leq \epsilon, \tag{6}$$

Without the need for network retraining or extra data, our proposed DWP demonstrates a notable simplicity and lightweight nature. Owing to its straightforward design, DWP exhibits compatibility with a broad spectrum of gradient-based, input transformation attacks, as well as advanced loss functions, making it an adaptable and versatile solution. Additional integration details of DWP with related works are provided in Appendix Sec. 6.2.

4. Experiments

In this section, we introduce experiment settings and demonstrate results of transferable attacks under various scenarios such as single-model, multiple-model ensemble, and real-world black box Google Cloud Vision service. A variant of architectures and adversarially trained models are evaluated. We report results for the targeted attack success rate as it is known to be more challenging and realistic in practice. Additional untargeted results are provided in the Appendix. Additionally, we provide time cost analysis of DWP in Appendix Sec. 11. To explore whether auxiliary networks produced by DWP exhibits gradient diversity, an additional ablation study is provided in Appendix Sec. 12.

4.1. Experimental Setup

Dataset Following previous studies [17, 18, 31, 55], we focus on the targeted attack transferability of the ILSVRC 2012 [35] since it is more difficult than other datasets (e.g, MNIST and CIFAR-10) that has fewer classes and smaller images. The ImageNet-compatible dataset [33] which contains 1000 samples provided by the NeurIPS 2017 adversarial attack competition [20] is applied in the following experiments. The dataset contains 1000 class, and each image is officially assigned a target class for a fair comparison.

Models We apply 7 naturally trained CNNs: Inception-v3 (Inc-v3), Inception-v4 (Inc-v4) [41], inception-resnet-v2 (IncRes-v2) [42], ResNet-50 (Res-50), ResNet-101 (Res-101) [13], VGGNet-16 (VGG-16) [38] and DenseNet-121 (Den-121) [14], 4 naturally trained Vision Transformers (ViTs): ViT-Small-Patch16-224 (ViT-S-16-224), ViT-Base-Patch16-224 (ViT-B-16-224)[7], Swin-Small-Patch4-Window7-224 (Swin-S-224), Swin-Base-Patch4-Window7-224 (Swin-B-224)[29], 3 naturally trained Multi-Layer Perceptrons (MLPs): Mixer-Base-Patch16-224 (MLP-Mixer) [43], ResMLP-Layer24-224 (ResMLP) [44], gMLP-Small-Patch16-224 (gMLP) [26], and 2 adversarially trained CNNs: ens3-adv-Inception-v3 (Inc-v3ens3) and ens-adv-inception-resnet-v2 (IncRes-v2ens) [45]. All the networks are publicly accessible in [48].

	Source Model: Res-50			Source Model: VGG-16		
	→VGG-16	→Den-121	→Inc-v3	→Res-50	→Den-121	→Inc-v3
NI-SI-TI-DI	52.0	75.3	31.5	22.9	27.6	14.4
+GN	55.6	76.9	37.1	29.9	32.9	19.5
+DWP	**65.0**	**82.0**	**42.1**	**30.2**	**33.4**	**19.9**
	Source Model: Den-121			Source Model: Inc-v3		
	→Res-50	→VGG-16	→Inc-v3	→Res-50	→VGG-16	→Den-121
NI-SI-TI-DI	37.8	25.5	13.9	1.39	3.70	5.90
+GN	53.3	36.4	28.0	1.40	2.50	1.50
+DWP	**59.2**	**44.8**	**32.5**	**10.9**	**12.7**	**16.0**

Table 1. Targeted success rates of transferring to naturally trained CNNs without the ensemble strategy. The "→" prefix stands for the black-box network. Results with targeted / untargeted attack success rates are reported.

Baselines We compare the transferability of DWP with the related model augmentation method, Ghost Networks (GN) [24], in combination with the state-of-the-art techniques NI-SI-TI-DI. GN drops activation outputs with a dropout rate Λ_{GN} and multiplies the skip connection by a factor sampled from the uniform distribution $U[1-\zeta_{GN}, 1+\zeta_{GN}]$. For non-residual networks like VGG-16 and Inc-v3, we insert dropout layers after each activation function. As for residual networks such as Res-50 and Den-121, skip connection erosion on the blocks of each network is applied. Throughout the experiment, we set $\Lambda_{GN} = 0.012, \zeta_{GN} = 0.22$ following the settings in [24].

Hyper-parameters Following the settings in [50, 55], we use 100 iterations with step size $\alpha = 2/255$ for I-FGSM and set the maximum perturbation budget $\epsilon = 16$ under L_∞ norm in all iterative attacks. Comply with Li *et al*. [23], we set the probability p_{DI} of DI to 0.7 and select a Gaussian kernel with a kernel length of 5 for W in TI. For SI, due to the limited computing resources, we set the number of scale copies $M = 3$. The momentum decay factor μ is set to 1 same as [5, 23, 25, 55]. For our proposed DWP, the probability p_{bern} is 0.5 and the prunable rate r is 0.7. In other words, we prune 35% of the connections of each network in expectation in each iteration.

4.2. Single model attack transferability

In our initial experiment, we conducted a comparative assessment of the single-model transfer targeted attack, specifically focusing on the baseline NI-SI-TI-DI, combing with GN and DWP in Tab. 1. We also provide the untargeted attack results in Appendix Tab. 8. The generation of adversarial examples was executed within a white-box model, subsequently transferring these adversarial instances to previously unseen black-box networks. Notably, it is evident that the DWP model consistently exhibits superior performance across various experimental configurations.

From our findings, DWP achieves a significant enhance-

	NI-SI-TI-DI	+GN	+DWP
-Inc-v3	65.2	77.5	**83.1**
-Inc-v4	71.3	70.0	**86.1**
-IncRes-v2	73.2	69.0	**85.4**
-Res-101	20.9	26.1	**40.6**
Average	57.65	60.65	**73.80**

Table 2. The targeted success rates of transferring across similar CNNs. "-" stands for the black-box network with the other three serving as the white-box ones for the ensemble.

ment for the baseline in attack success rates when transferring from different source models. Specifically, we observe a 10.1% improvement when transferring from Res-50, a 6.2% improvement from VGG-16, a substantial 19.8% improvement from Den-121, and a notable 9.5% improvement from Inc-v3, on average. In addition, when comparing DWP with GN, which introduces connection dropout to enhance model diversity, DWP consistently outperforms GN across all models. An advancement of attack success rate is achieved for 6.5%, 0.4%, 6.27% and 11.4% on average for Res-50, VGG-16, Den-121 and Inc-v3 respectively.

Our experimental results reveal an intriguing observation: the extent of improvement achieved by DWP over GN is contingent upon the redundancy of the source model. As illustrated in Fig. 2, we investigate how the elimination of network connections, through weight pruning, affects model accuracy. VGG-16 exhibits the highest degree of redundancy, as evidenced by its ability to maintain accuracy even when up to 80% of its connections are pruned. In contrast, other networks experience a near-total accuracy loss under the same pruning conditions. DWP demonstrates substantial improvements over GN for models like Res-50, Den-121, and Inception-v3. However, its performance aligns more closely with GN for VGG-16, primarily due to the latter's abundance of redundant parameters.
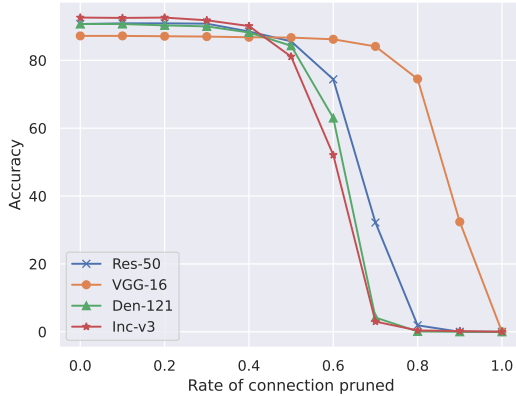
Figure 2. The accuracy (%) of networks when different rates of connections are pruned.

4.3. Ensemble transfer in various scenarios

Our study encompasses the exploration of targeted transferability across 4 distinct scenarios: transferability between CNNs, transferability to adversarially trained models, transferability to non-CNN architectures, and transferability to the real-world Google Cloud Vision service. We craft adversarial examples using an ensemble comprising multiple white-box networks and evaluate targeted success rates on the specified black-box model. Each of the K white-box models in the ensemble is weighted equally $\beta_k = 1/K$.

4.3.1 Transferability between CNNs

CNNs with similar architectures   Tab. 2 summarizes the targeted attack success rates across Inc-v3, Inc-v4, IncRes-v2 and Res-101. The group of CNNs was popular for evaluating attacks [5, 6, 23, 49, 51] with architecture resemblance. From the results, DWP shows 16.15% average improvement over NI-SI-TI-DI and outperforms leading methods GN by 13.15%.

CNNs with distinct architectures   Given the ubiquity of CNNs in contemporary applications, we examine the transferability among distinct architectures suggested by Zhao *et al*. [55]. We selected 4 well-established and canonical CNN models: Res-50, VGG-16, Den-121, and Inc-v3. The results of targeted attack transferability between these CNNs are presented in Tab. 3. Our results demonstrate that DWP substantially enhances the efficacy of attack methodologies, surpassing the performance of competing techniques such as GN by 4.75% on average. Notably, the 4 chosen CNN features distinctive design characteristics, incorporating elements such as Residual, Dense, and Inception blocks. Our findings underscore the advantages of employing a diversi-

	NI-SI-TI-DI	+GN	+DWP
-Res-50	70.1	68.7	**77.7**
-Den-121	86.7	85.0	**89.4**
-VGG-16	77.1	80.1	**87.2**
-Inc-v3	66.9	**72.4**	70.9
Average	75.20	76.55	**81.30**

Table 3. The targeted success rates of transferring across distinct CNNs. The "-" prefix stands for the black-box network with the other three serving as the white-box ones for ensemble.

	NI-SI-TI-DI	+GN	+DWP
Inc-v3ens3	50.0	51.6	**65.3**
IncRes-v2ens	19.4	29.8	**39.0**
Average	34.7	40.7	**52.15**

Table 4. Targeted attack success rates for defensed models.

	NI-SI-TI-DI	+GN	+DWP
Res-18 ($|\epsilon|_\infty = 1$)	33.2	33.6	**37.0**
Res-50 ($|\epsilon|_\infty = 1$)	40.5	39.4	**41.4**
WideRes-50-2 ($|\epsilon|_\infty = 1$)	37.8	35.4	**39.5**
Res-18 ($|\epsilon|_2 = 3$)	12.6	12.6	**15.2**
Den-121 ($|\epsilon|_2 = 3$)	17.4	18.0	**19.2**
VGG16 ($|\epsilon|_2 = 3$)	12.5	13.3	**15.5**
Resnext-50 ($|\epsilon|_2 = 3$)	19.1	19.2	**21.0**
Res-50	21.9	16.8	**22.3**
Den-121	27.6	29.0	**39.0**
VGG-16	8.60	8.80	**18.6**
Inc-v3	17.4	17.9	**26.7**
Inc-v3ens3	22.4	23.3	**30.5**
IncRes-v2ens	22.3	22.6	**30.0**

Table 5. The targeted success rates of transferring to adversarially trained networks from the ones with different architectures and ϵ.

fied ensemble approach when targeting black-box CNNs.

4.3.2 Transferability to adversarially trained models

Adversarial training [30, 45] is the most effective technique for defending against malicious attacks by being trained with adversarial examples. Successfully attacking such models indicates the ability to break the strongest defense. Under the scenario of transferring to adversarially trained models, we ensemble 4 naturally trained networks (Res-50, Den-121, VGG-16, and Inc-v3) as white-box models to simulate the situation when attackers know little about defense. The 2 one-step adversarially trained networks (Inc-v3ens3 and IncRes-v2ens) will act as our black-box model with defense separately.

Tab. 4 summarizes the targeted transferability results to adversarially trained networks. The result of the untargeted counterpart is provided in Appendix Tab. 9. Under
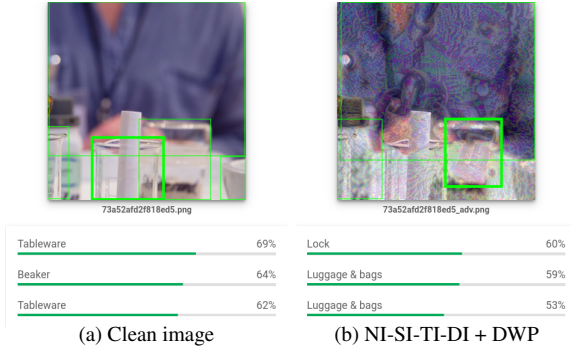
(a) Clean image (b) NI-SI-TI-DI + DWP

Figure 3. A demo of our DWP attack on Google Cloud Vision. The attacked image with the ground truth label of "Beakers" is recognized as the target class "Padlocks" assigned by the NeurIPS 2017 ImageNet-compatible dataset.

Attack Method	NI-SI-TI-DI	+GN	+DWP
ViT-S-16-224	25.9	31.5	**37.3**
ViT-B-16-224	24.8	29.9	**37.4**
Swin-S-224	26.7	29.1	**36.7**
Swin-B-224	23.9	27.1	**32.9**
MLP-Mixer	21.7	24.2	**30.9**
ResMLP	51.3	56.5	**64.1**
gMLP	20.4	25.3	**30.4**
Average	27.81	31.94	**38.53**

Table 6. Targeted attack success rates for Non-CNN architectures.

	NI-SI-TI-DI	+GN	+DWP
Google Cloud Vision	27	43	**50**

Table 7. Targeted success rates (%) on Google Cloud Vision out of 100 randomly selected images.

such challenging cases, DWP can still alleviate the discrepancy between white-box naturally-trained and black-box adversarially-trained networks, bringing about up to 17.45% improvement to the baseline on average. The efficacy of the diversified ensemble safeguarding essential connections is highlighted again when black-box networks exhibit substantial distinctions from white-box models.

Authors in [45] propose "ensemble adversarial training", which trains the network with adversarial examples generated from external models. While the single-step attack in the procedure is less costly, the models fall short of resisting iterative attacks even in black-box scenarios. Therefore, we also explore the black-box targeted attack results on the models with multi-step adversarial training [36]. With the multi-step adversarially trained white-box networks joining the ensemble, the victim network is vulnerable to DWP attack even if it undergoes multi-step adversarial training.

Tab. 5 summarizes the targeted attack results of the ensemble composed of Res-18 ($|\epsilon|_\infty = 2$), Res-50 ($|\epsilon|_\infty = 2$) and WideRes-50-2 ($|\epsilon|_\infty = 2$). The two upper groups in Tab. 5 report the targeted success rates on different CNN architectures and the norm of ϵ used in adversarial training. The attack success rates on naturally-trained CNNs and ensemble adversarial-trained models are reported in the latter groups. We provide additional experiments for transferring from naturally-trained models to multi-step adversarially trained networks in the Appendix Tab. 10.

4.3.3 Transferring to non-CNN architectures

In practice, implementation details of defenders' models remain undisclosed to potential attackers, and various architectures other than CNNs might be utilized. Beyond CNNs, contemporary works have successfully addressed computer vision tasks through Vision Transformers (ViTs) [7, 29] and Multi-Layer Perceptrons (MLPs) [26, 43, 44]. To ensure the effectiveness of DWP stands out in those non-CNN scenar-

ios, we conduct a comprehensive study evaluating the targeted transferability from CNNs to these models. Targeted adversarial images were generated on an ensemble comprising 4 naturally trained CNNs and subsequently transferred to the non-CNN network. From Tab. 6, the efficacy of model augmentations persists, even in instances where black-box networks lack convolution operations beyond input projections. DWP improves the results on both ViTs and MLPs, outperforming NI-SI-TI-DI and GN for 10.72% and 6.59% on average respectively. Appendix Tab. 11 reports an additional untargeted result.

4.3.4 Transferring to Google Cloud Vision

Google Cloud Vision is a publicly accessible service that enables users to create their vision application with pre-trained APIs. As the design behind the tool remains concealed, we use Google Cloud Vision to evaluate our adversarial examples, assuring DWP achieves strong black-box targeted transferability. Google Cloud Vision predicts a list of labels with their corresponding confidence scores. It returns label annotations only when the confidence score is above 50%. Neither gradients nor parameters of the underlying system are accessible. Previous works leverage query-based attacks [1, 3, 16] or black-box transferability [27, 55]. However, query-based methods often require large numbers of queries, and the existing transferable attacks still have substantial room for improvement.

In this experiment, we randomly select 100 correctly labeled images by Google Cloud Vision from the Imagenet-compatible dataset. Four naturally trained CNNs, Res-50, VGG-16, Den-121, and Inc-v3, are applied to generate adversarial examples. We identified a successful attack if at least one of the returns by the API is semantically close to its corresponding target class given an example. We sum-
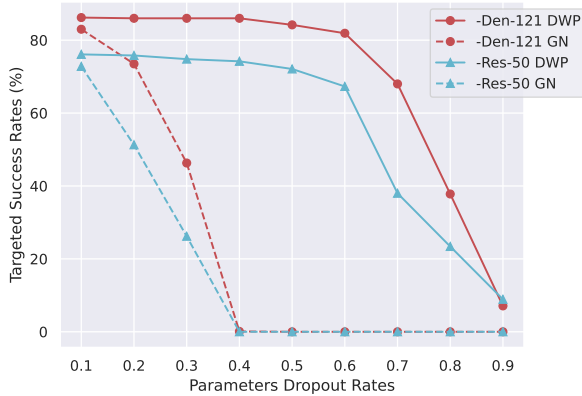
Figure 4. The comparison of DWP and GN for the targeted success rates under different prunable rates.

Figure 5. **The targeted success rates under different prunable rates** r **on each black-box model.** Each curve shows the trade-off between the diversity and stability of pruned models. The curve for mean targeted success rates reaches its maximum at $r = 0.7$.

marize the results in Tab. 7. DWP outperforms the baseline and GN by 23% and 7%, respectively. Fig. 3 demonstrates an example on Google Cloud Vision. More demos can be found in Appendix Sec. 13.

4.4. Ablation analysis on prunable rates

DWP achieves a wider hyper-parameter tolerance Prior model augmentation methods employ random parameter dropout without considering parameter significance, rendering them sensitive to hyperparameter choices. In the absence of meticulous configuration of the dropout rate, these methods experience rapid and unacceptable deterioration in performance. DWP circumvents this challenge by selectively altering only unnecessary parameters while preserving the integrity of crucial ones. Fig. 4 illustrates the variation in targeted attack success rates under different parameter dropout rates. Specifically, in DWP, all parameters with the lowest $100 \cdot r\%$ weight value are pruned away, aligning with the number of connections dropped in GN by setting $p_{berm} = 1$. Observably, DWP exhibits reduced sensitivity to the parameter dropout rate, in contrast to the pronounced decline in attack success rates witnessed in GN as the dropout rate increases. In practical scenarios involving the involvement of a group of white-box models in the ensemble, the manual adjustment of the optimal rate for each model in GN poses significant challenges. Given its broader tolerance for the dropout ratio, DWP emerges as a more user-friendly and effective approach for obtaining high-quality auxiliary models through model augmentation.

Finding the optimal prunable rates This investigation delves into the targeted attack success rates across various prunable rates r. Parameters featuring the lowest $100 \cdot r\%$ weight values are identified as prunable and subsequently pruned with a probability of $p_{berm} = 0.5$. The parameter
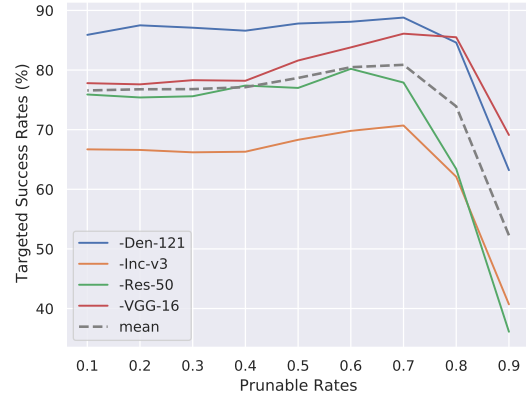
r plays a pivotal role in determining the group size of connections eligible for weight pruning, allowing for the generation of more diverse auxiliary models at higher prunable rates due to increased parameter flexibility. Nevertheless, as a trade-off, excessive pruning of connections results in a decline in the quality of auxiliary networks, leading to performance instability. To strike a balance in this trade-off, we systematically explore different prunable rates and observe the consequential changes in the targeted success rate with four CNNs, as illustrated in Fig. 5. Notably, the mean attack success rates reach their peak when $r = 0.7$. At this optimal prunable rate, DWP prunes approximately 35% of weight connections on average.

5. Conclusion

In this paper, we introduce **D**iversified **W**eight **P**runing (DWP), a novel approach harnessing network compression to enhance the targeted transferability of adversarial attacks. The safeguarding of crucial parameters within the network ensures the preservation of auxiliary model quality generated by DWP. Through comprehensive evaluations on ImageNet, our study demonstrates that DWP surpasses the performance of state-of-the-art model augmentation methods in the realm of transferable targeted attacks. This improvement is particularly pronounced in challenging scenarios, such as the transfer to adversarially trained models and non-CNN architectures. Notably, DWP distinguishes itself through its design simplicity and wide tolerance for hyperparameter selection, facilitating seamless integration with other related techniques. This characteristic renders DWP amenable to plug-and-play implementation without the necessity for extensive parameter tuning. In conclusion, DWP emerges as a potent and versatile solution for enhancing attack transferability.

References

[1] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 7

[2] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017. 1, 2

[3] Yuxuan Chen, Xuejing Yuan, Jiangshan Zhang, Yue Zhao, Shengzhi Zhang, Kai Chen, and XiaoFeng Wang. Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices. In *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, pages 2667–2684. USENIX Association, 2020. 7

[4] Misha Denil, Babak Shakibi, Laurent Dinh, Marc'Aurelio Ranzato, and Nando de Freitas. Predicting parameters in deep learning. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2148–2156, 2013. 2, 3

[5] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9185–9193. Computer Vision Foundation / IEEE Computer Society, 2018. 1, 2, 5, 6

[6] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4312–4321. Computer Vision Foundation / IEEE, 2019. 2, 6, 1, 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 4, 7

[8] Yexin Duan, Junhua Zou, Xingyu Zhou, Wu Zhang, Jin Zhang, and Zhisong Pan. Adversarial attack via dual-stage network erosion, 2022. 2, 3, 4

[9] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 2, 3

[10] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 2

[11] Martin Gubri, Maxime Cordy, Mike Papadakis, Yves Le Traon, and Koushik Sen. LGV: boosting adversarial example transferability from large geometric vicinity. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, pages 603–618. Springer, 2022. 2

[12] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1135–1143, 2015. 2, 3, 4

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 4

[14] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017. 4

[15] Yi Huang and Adams Wai-Kin Kong. Transferable adversarial attack based on integrated gradients. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[16] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2142–2151. PMLR, 2018. 7

[17] Nathan Inkawhich, Kevin J. Liang, Lawrence Carin, and Yiran Chen. Transferable perturbations of deep feature distributions. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 2, 4

[18] Nathan Inkawhich, Kevin J. Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 4

[19] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. 2

[20] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, Alan Yuille, Sangxia Huang, Yao Zhao,

Yuzhe Zhao, Zhonglin Han, Junjiajia Long, Yerkebulan Berdibekov, Takuya Akiba, Seiya Tokui, and Motoki Abe. Adversarial attacks and defences competition, 2018. 2, 4

[21] Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems 2, [NIPS Conference, Denver, Colorado, USA, November 27-30, 1989]*, pages 598–605. Morgan Kaufmann, 1989. 2, 3

[22] Chao Li, Shangqian Gao, Cheng Deng, De Xie, and Wei Liu. Cross-modal learning with adversarial samples. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10791–10801, 2019. 1

[23] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 638–646. Computer Vision Foundation / IEEE, 2020. 2, 3, 5, 6

[24] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan L. Yuille. Learning transferable adversarial examples via ghost networks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11458–11465. AAAI Press, 2020. 2, 3, 4, 5

[25] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 2, 3, 5, 1

[26] Hanxiao Liu, Zihang Dai, David R. So, and Quoc V. Le. Pay attention to mlps. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9204–9215, 2021. 4, 7

[27] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and blackbox attacks. In *International Conference on Learning Representations*, 2017. 1, 2, 3, 7

[28] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 2, 3

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. 4, 7

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 6

[31] Muzammal Naseer, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. On generating transferable targeted perturbations. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 7688–7697. IEEE, 2021. 2, 3, 4

[32] Y NESTEROV. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady AN USSR*, pages 543–547, 1983. 1

[33] Nicolas Papernot, Ian Goodfellow, Ryan Sheatsley, Reuben Feinman, and Patrick McDaniel. cleverhans v1.0.0: an adversarial machine learning library. *arXiv preprint arXiv:1610.00768*, 2016. 4

[34] Zeyu Qin, Yanbo Fan, Yi Liu, Li Shen, Yong Zhang, Jue Wang, and Baoyuan Wu. Boosting the transferability of adversarial attacks with reverse adversarial perturbation. In *Advances in Neural Information Processing Systems*, 2022. 2

[35] 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 (IJCV)*, 115 (3):211–252, 2015. 4

[36] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 7, 2

[37] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *J. Big Data*, 6: 60, 2019. 2, 1

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 4

[39] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014. 4

[40] 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*, 2014. 2

[41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016. 4

[42] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the

impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4278–4284. AAAI Press, 2017. 4

[43] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 24261–24272, 2021. 4, 7

[44] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. Resmlp: Feedforward networks for image classification with data-efficient training. *CoRR*, abs/2105.03404, 2021. 4, 7

[45] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 1, 4, 6, 7

[46] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1924–1933. Computer Vision Foundation / IEEE, 2021. 3

[47] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 16138–16147. IEEE, 2021. 2

[48] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 4

[49] 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2730–2739. Computer Vision Foundation / IEEE, 2019. 2, 6, 1

[50] Yifeng Xiong, Jiadong Lin, Min Zhang, John E. Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14983–14992, 2022. 3, 5

[51] Yifeng Xiong, Jiadong Lin, Min Zhang, John E. Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14963–14972. IEEE, 2022. 6

[52] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Boosting transferability of targeted adversarial exam-
ples via hierarchical generative networks. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, pages 725–742. Springer, 2022. 2

[53] Haojie Yuan, Qi Chu, Feng Zhu, Rui Zhao, Bin Liu, and Neng-Hai Yu. Automa: Towards automatic model augmentation for transferable adversarial attacks. *IEEE Transactions on Multimedia*, pages 1–1, 2021. 3

[54] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2

[55] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. On success and simplicity: A second look at transferable targeted attacks. In *Advances in Neural Information Processing Systems*, pages 6115–6128. Curran Associates, Inc., 2021. 2, 3, 4, 5, 6, 7

[56] Junhua Zou, Yexin Duan, Boyu Li, Wu Zhang, Yu Pan, and Zhisong Pan. Making adversarial examples more transferable and indistinguishable. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 3662–3670. AAAI Press, 2022. 1