基于模型多样性增强的集成攻击

Abstract

深度神经网络在机器学习任务中取得了巨大进展，然而，对抗样本攻击对其构成严重威胁。这种攻击的可转移性使得模型容易受到影响，尤其对于黑盒应用更为现实和危险。针对提高对抗样本的转移性，研究者们提出了一系列方法，包括基于模型集成的攻击。相较于传统的集成方法，这些新方法更注重模型之间的差异性，并试图提高攻击的泛化能力。然而，在模型架构差异较大的情况下，现有方法的效果仍有限。因此，本文提出了一种基于模型多样性增强的集成攻击方法。我们根据模型输出的多样性评估对不同模型的增强输出进行加权，以生成更具攻击性和迁移性的对抗样本。实验结果显示，该方法在不同数据集和攻击策略下均取得了比当前最优方法更高的攻击成功率，尤其在对抗训练后的模型上表现出更为显著的优势，为集成对抗攻击领域的深入研究提供了新的思路和方法

关键词：模型多样性；集成攻击；迁移攻击

1 引言

深度神经网络（DNNs）包括卷积神经网络（CNNs）[8,15,22] 和视觉变换器（ViTs）[5,12,16]，在各种机器学习任务中带来了巨大的进展。然而，目前却发现它们容易受到对抗样本的攻击 [15]，即对原始输入添加微不可见的精心设计的扰动可能导致 DNNs 出现错误的预测行为。这一发现给 DNNs 的部署带来了严重的安全隐患。更重要的是，一些设计精良的对抗样本可以在模型之间转移。也就是说，从一个替代模型中制作的对抗样本也可以干扰其他模型。对抗样本具有的这种特性，被称为可转移性，允许攻击者攻击目标模型而无需了解其内部情况，因此对于黑盒应用（即用户无法访问架构和参数）构成了更现实的威胁。为了为提高模型的鲁棒性、防止潜在的黑盒攻击威胁，近年来学术界更加关注对提高对抗样本的转移性的研究。攻击的转移成功率取决于替代模型和目标模型之间的差异，替代模型和目标模型越相似，转移成功率就越高。因此，已经提出了一系列方法来通过最大化在 DNNs 之间共享的关键部分的扰动来提高对抗样本的转移性。主流策略包括最大化重要神经元的信息 [19,23]，提高输入的多样性 [1,20]，以及将动量 [3,18] 引入迭代攻击中。尽管这些方法有效，但它们在模型架构差异较大的情况下（即 CNNs 和 ViTs）之间的转移上通常效果较差。与传统的集成方法类似，其依靠多个预测能力不同的弱模型的输出来提高整体准确性，一系列研究提出利用一组替代模型的集成来生成能够成功攻击所有目标模型的对抗样本。更直观地来说，这种方法可以提高对抗样本的转移性，因为它能够捕捉到潜在的固有可转移对抗信息，所以攻击者可以同时欺骗多个有着广泛差异的模型。此外，这样的集成还可以轻松地与现有的基于转移的对抗攻击方法相结合而不会发生冲突。以往的工作探索了几种基于模型集成的方法 [9,11]，

然而，大多数方法只是简单地融合了所有模型的输出以获取用于应用基于梯度的攻击的集成损失，这可能限制了模型集成攻击的潜力。尽管最近的一些研究 [2, 21] 注意到了替代模型之间的梯度差异，但由于忽略了每个模型的个体特征，以及对模型多样性的评估，该集成仍然没有达到最好的效果。

2 相关工作

2.1 梯度下降对抗攻击

为了优化攻击目标，通常会使用梯度信息来最大化模型损失。Goodfellow 等人 [7] 根据对 CNN 线性特性的研究设计了快速梯度符号法（FGSM），以生成更强的对抗样本。Wang 等人 [17] 和 Madry 等人 [13] 进一步将 FGSM 中的单步扰动生成拆分为迭代生成，并提出了 I-FGSM 和投影梯度下降（PGD）攻击。虽然这些攻击可以在白盒模型上表现出很高的攻击成功率，但它们通常在黑盒模型上的转移率较低，因为梯度信息很难近似。

2.2 基于转移的对抗攻击

为了提高转移性，现有的研究尝试在输入的关键部分最大化失真。Wang 等人 [19] 和 Zhang 等人 [23] 基于 DNN 中神经元的重要性，研究了特征上的失真。Xie 等人 [20] 和 Dong 等人 [4] 将 FGSM 与输入多样性或平移不变策略相结合，以产生多样化的输入模式来生成对抗样本。Gao 等人 [6] 提出了 PI-FGSM，它生成基于补丁而不是像素的扰动，有利于黑盒攻击。虽然这些攻击相对于最初的基于梯度的攻击可以提高转移性，但它们几乎无法转移到新的 DNN 架构，即 ViT 系列。

2.3 模型集成攻击

集成攻击方法通常通过对多个白盒攻击进行加权线性求和来制造对抗样本。Liu 等人 [11] 直接对多个模型的预测进行平均，得到一个用于应用基于梯度攻击的集成损失。Dong 等人 [6] 进一步融合了集成模型的 logits 和损失。Xiong 等人 [21] 注意到集成模型之间的差异性，并提出了随机方差减少集成（SVRE）攻击，以提高攻击的泛化能力。虽然取得了改进，但由于对每个模型的个别优势进行的研究较少，集成仍然不够优化。

3 本文方法

3.1 本文方法概述

提出了一种自适应集成攻击，命名为 AdaEA，通过监测其对对抗目标的贡献的差异比率，自适应地控制每个模型输出的融合。此外，引入了一种额外的减小差异的滤波器，进一步同步更新方向。结果表明，在各种数据集上，我们相对于现有的集成攻击取得了显著的改进，而 AdaEA 还可以提升现有的基于传递性的攻击，进一步证明了其高效性和多功能性，其总体框架入如图 1所示：

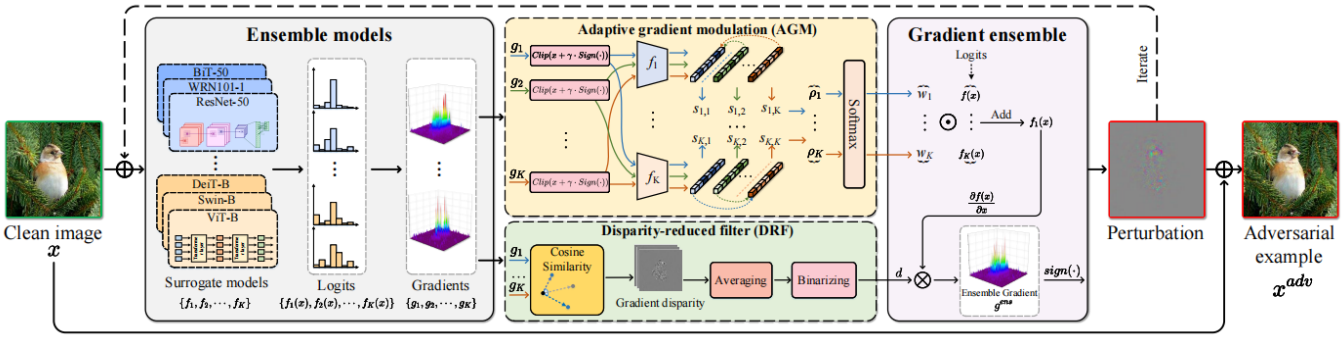

Figure 1. AdaEA 的概述。从卷积神经网络（CNNs）和视觉注意力模型（ViTs）获得的梯度被馈送到自适应梯度融合器（AGM）和差异降低滤波器（DRF），以获取用于生成基于梯度攻击的对抗样本的集成梯度

3.2 自适应梯度模块

在从每个替代模型 f_i 中通过输入图像获得输出 $f_i(x)$ 和梯度信息 g_i 之后，即 $g_i = \nabla_{x_t^{adv}} \mathcal{L}\left(f_i\left(x_t^{adv}\right), y\right)$，我们提出通过监测它们对对抗攻击目标的贡献差异来自适应调制模型集成。具体而言，对于第 i 个集成模型 f_i，我们通过测试从 g_i 生成的对抗样本在其他模型上的攻击性能来评估 g_i 中的潜在对抗传递性，我们将其定义为对抗比率，并根据每个模型的对抗比率调整集成权重。我们首先计算：

$$s_{k,i} = -\mathbf{1}_y \cdot \log\left(\text{softmax}\left(\mathbf{p}_k\left[x_t^{adv} + \alpha\,\text{sign}\left(g_i\right)\right]\right)\right) \tag{1}$$

其中 $\mathbf{p}_k(\cdot)$ 表示从 f_k 输出的 logits，$\mathbf{1}_y$ 是真实标签的 logits。$s_{k,i}$ 可以被视为从第 i 个模型获取梯度的对抗样本上的第 k 个模型损失。然后，我们定义对抗比率 ρ_i 为：

$$\rho_i = \frac{\beta}{K-1} \sum_{k=1,k\neq i}^{K} \frac{s_{k,i}}{s_{k,k}} \tag{2}$$

其中 β 是控制集成权重效果的超参数。较高的 ρ_i 值表示从 g_i 生成的对抗样本具有更好的传递攻击效果，暗示着 g_i 包含更多可传递的对抗信息。

通过这样的方式，我们可以找出哪个模型可以提供更通用的对抗信息，并自适应地分配更高的集成权重。因此，根据每个模型的对抗比率，我们使用 softmax 函数来归一化每个模型的集成权重：

$$w_1^*, w_2^*, \ldots, w_K^* = \text{softmax}(\rho_1, \rho_2, \ldots, \rho_K) \tag{3}$$

通过 Eq.3获得的 w_i^*，具有更多潜在对抗传递性信息的每个替代模型的输出在的集成梯度中得到放大，从而导致在保持隐身的黑盒模型上实现更高的传递攻击成功率。

3.3 差异性过滤模块

替代模型的梯度优化方向在一个大范围内变化巨大，有时梯度朝着彼此相反的方向走，导致过度拟合到集成模型。为了解决这个问题并同步更新方向，我们引入了一个额外的差异降

低滤波器，以减小替代模型之间的梯度变化。我们首先应用余弦相似性来评估替代模型中梯度的偏差，通过平均相似性分数与其他模型的梯度计算得到差异图 d_i，描述如下：

$$d_i^{(p,q)} = \frac{1}{K-1} \sum_{k=1,k\neq i}^{K} \cos\left(\vec{g}_i^{(p,q)}, \vec{g}_k^{(p,q)}\right) \tag{4}$$

其中 $\cos(\cdot)$ 表示余弦相似性函数，$\vec{g}_i^{(p,q)}$ 和 $\vec{g}_k^{(p,q)}$ 分别表示从梯度 g_i 和 g_k 的位置 (p,q) 通过通道提取的向量。集成梯度的最终差异图 d 是通过对所有 d_i 取平均得到的。然后，我们使用滤波器 $\mathbf{B}$ 清除集成梯度中的差异部分，定义如下：

$$\mathbf{B}(p,q) = \begin{cases} 0, & \text{if } d_i^{(p,q)} \leq \eta \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

其中 η 是差异过滤的容差阈值。通过过滤掉集成梯度的差异部分，梯度优化方向可以得到同步。因此，集成梯度可以通过如下方式得到：

$$g_{t+1} = \nabla_{x_t^{adv}} \mathcal{L}\left(\sum_{k=1}^{K} w_k^* f_k\left(x_t^{adv}\right), y\right) \otimes \mathbf{B} \tag{6}$$

其中 $\otimes$ 表示逐元素乘法。因此，替代模型之间的差异可以被抑制。

4 复现细节

4.1 与已有开源代码对比

本章之前的内容皆为主要参考文献 AdaEA [2] 的内容介绍，同时仅使用了其开源代码，并在其基础上进行改进，取得了极大的提升。改动内容为：其中将两个模块皆去除，并新增了模型多样性增强模块。本方向内的所有工作主要以数据指标为评价标准，在不涉及不可感知性指标的提升时无图片可视化。

4.2 创新点

传统的攻击算法通过迭代来更新对抗样本，然而，这可能导致在数据分布相似的区域陷入局部最优解，使得对抗样本更容易过度拟合于源模型。这种过度拟合会导致在攻击其他模型时的不稳定性，从而导致攻击失败。如果每个模型都有相似的结构、训练数据，它们可能会产生相似的对抗样本。在这种情况下，集合攻击的结果可能缺乏多样性，限制了对抗样本的攻击能力。

我们把原始图像输入到代理模型 1 中，可以得到输出分类向量 output。我们将随机噪声加入图像后，模型的输出为 output1，当随机噪声加入到图片中时，输出分量的变化程度体现了模型对于输入变化的敏感程度。如果输出分量的变化大，意味着模型对输入的微小变化具有较高的敏感性，即模型的多样性较好（模型的多样性是指模型在输入空间中的输出分布的多样性）。如果模型对输入变化敏感且输出分布变化较大，说明模型能够更好地捕捉输入中的细微变化，具有更高的多样性。这种多样性对于生成的对抗样本的迁移性很重要。

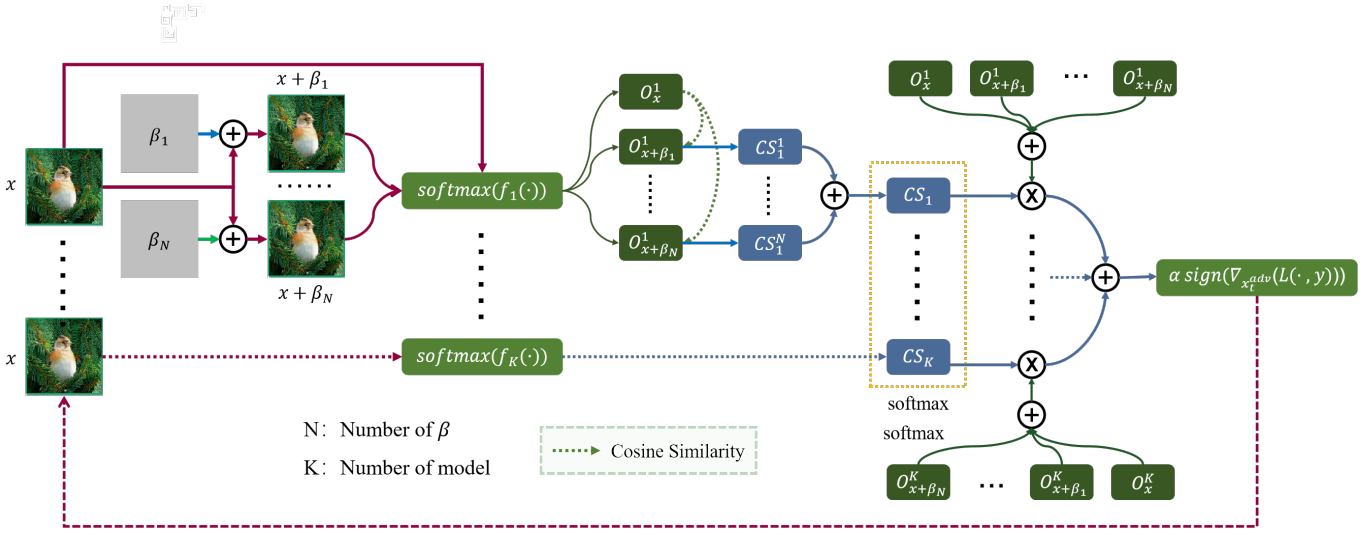

Figure 2. 方法主体架构图

对于我们的方法而言，我们对各个模型进行加权，给予不同模型输出不同的权重来更好的保留迁移性较大的模型（多样性更好）的输出来提高生成的对抗样本的迁移性。我们方法的整体架构图如图2所示。

我们首先根据噪声强度 ϕ 产生 N 个随机噪声 β：

$$\beta_n = \varphi \,\mathrm{rand\_like}(x) \tag{7}$$

然后将噪声加入到原始图像 x 中得到 N 个随机扰动样本并将所有随机扰动后的样本送入模型分别得到他们的输出:

$$O_{x+\beta_n}^k = \mathrm{softmax}\left(f_k\left(x+\beta_n\right)\right) \tag{8}$$

同时将原始样本也送入模型得到输出，此时将原始样本的输出和加噪后的样本的输出计算其余弦相似度，然后将所有的相似度相加后得到本模型对应的多样性的权重值:

$$CS_k = \sum_{n=1}^{N} 1 - \frac{O_x^k \cdot O_{x+\beta_n}^k}{\left\|O_x^k\right\| \cdot \left\|O_{x+\beta_n}^k\right\|} \tag{9}$$

针对每个模型进行相同的操作便可得到每个模型的 CS_k，当观察输出分量的变化程度较大时，可以认为模型具有较好的多样性，生成的对抗样本可能具有更好的迁移性。这是因为模型对输入的微小变化具有较高的敏感性，生成的对抗样本在不同模型上可能会导致类似的误分类结果，从而提高了对抗样本的迁移性。计算每个模型的 CS_k，然后通过 softmax 函数进行归一化处理获得每个模型的权重:

$$w_1^*, w_2^*, \ldots, w_K^* = \mathrm{softmax}\left(CS_1, CS_2, \ldots, CS_K\right) \tag{10}$$

将每个模型的所有原始样本的输出以及加噪后的输出求和后再加权求和便可得到最终的输出，然后将输出对真实标签进行求导便可对原始样本进行一次迭代攻击:

$$x_{t+1}^{adv} = x_{t+1}^{adv} + \alpha \,\mathrm{sign}\left(\nabla_{x_t^{adv}} L\left(\sum_{k=1}^{K} w_k \left[O_x^k + \sum_{n=1}^{N} O_{x+\beta_n}^k\right], y\right)\right) \tag{11}$$

通过以上过程我们可以得到更加具有攻击能力的对抗样本，因为我们不仅根据模型多样性评估来对不同模型进行加权，还使用了多模型进行加噪后的数据输出结果，更多的输出可以模拟更多的模型的输出结果，进一步提升攻击能力。

4.3 数据集选择

我们在 CIFAR-10、CIFAR-100 和 ImageNet 数据集上进行实验，这些数据集被广泛用于分类和对抗攻击任务。其中 CIFAR-10：包含 60,000 张 32x32 像素的彩色图像，分为 10 个类别，每个类别有 6,000 张图像。类别：飞机、汽车、鸟类、猫、鹿、狗、青蛙、马、船和卡车。CIFAR-100 同样包含 60,000 张 32x32 像素的彩色图像，但分为 100 个类别，每个类别有 600 张图像。类别是 CIFAR-10 类别的细分，包括不同种类的动物、植物、交通工具和日常用品等。适用于更复杂的分类任务，要求模型能够处理更多的细粒度类别。ImageNet 是一个大规模的图像数据库，包含数百万张高分辨率图像，分为 1,000 个不同的类别。类别包括各种各样的物体、动物和场景，是一个包罗万象的图像集合。因为代理模型使用了训练集进行训练，所以对于所有数据集我们使用其测试集来对攻击效果进行测试。

4.4 所用模型

我们选择了来自 CNN 和 ViT 两个分支的目标模型，用于黑盒攻击任务，包括 CNN 分支的 ResNet-50（Res-50）[8]、WideResNet-50（WRN-50）[22]、BiT-M-R50×1（BiT-50）[10] 和 BiT-M-R101（BiT-101）[10]；以及 ViT 分支的 ViT-Base（ViT-B）[5]、DeiT-Base（DeiT-B）[16]、Swin-Base（Swin-B）[12][11] 和 Swin-Small（Swin-S）[12]。至于替代模型，我们在后续实验中默认选择了 ResNet-18（Res-18）[8]、Inception v3（Inc-v3）[14]、ViT-Tiny（ViT-T）[5] 和 DeiT-Tiny（DeiT-T）[16]。

4.5 对比方法

我们选择了三种开创性的集成攻击方法，即 Ens [11]、SVRE [21] 和 AdaEA [2] 作为基线与我们的方法进行比较。所有的集成方法在实验中遵循相同的集成设置。其中 Ens 是最基础的将所有输出进行平均的方法

4.6 超参数

对于基线、SVRE、AdaEA 以及我们的方法，我们使用 I-FGSM 进行 20 次迭代，受无穷范数约束，作为基本的攻击方法，并在对抗样本生成时设置 $\epsilon = 8/255$ 和 $\alpha = 2/255$。至于超参数，AdaEA 方法的 DRF 中设置 η = -0.3，AGM 中设置 $\beta = 10$。SVRE 中的内部更新时间设置为 4，遵循其默认设置，我们超参数 N 选择 3，ϕ 选择 0.25。所有实验都是在具有 24GB 图形内存的 NVIDIA RTX3090 GPU 上使用 PyTorch 实现的。

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。

5.1 超参数敏感度分析

我们首先对加噪的数量 N 进行挑选，加噪次数越多所耗费的时间就越多，所以我们尽可能去平衡攻击成功率以及所消耗的时间，对不同的 N 的测试数据折线图如图3所示。

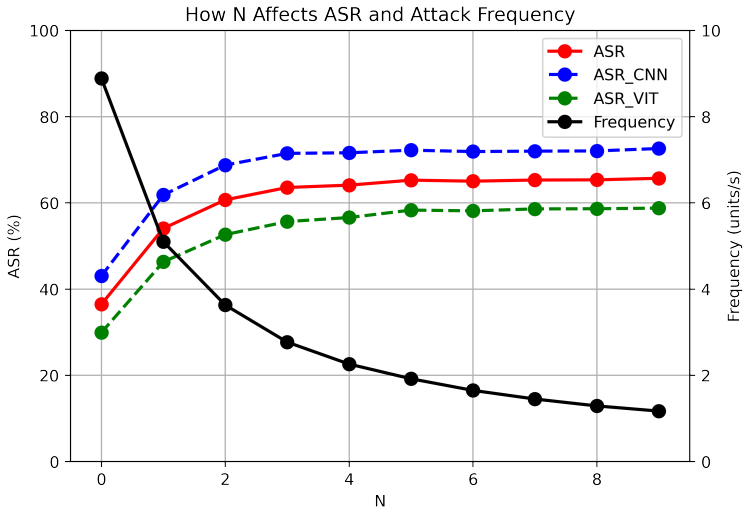

Figure 3. 超参数 N 敏感性分析

其中黑色折线图表示每秒可以生成的对抗样本的数量，即生成频率，红色线是平均攻击成功率，绿色和蓝色的折现分别表示在 VIT 簇模型以及 CNN 簇模型上的效果。ASR 表示攻击成功率（Attack Success Rate）可以看到在 N=3 时在生成速度和攻击成功率之间达到了较为平衡的点，所以我们挑选 N=3。

我们接着对添加的噪声强度 ϕ 进行测试，测试曲线如图4所示。

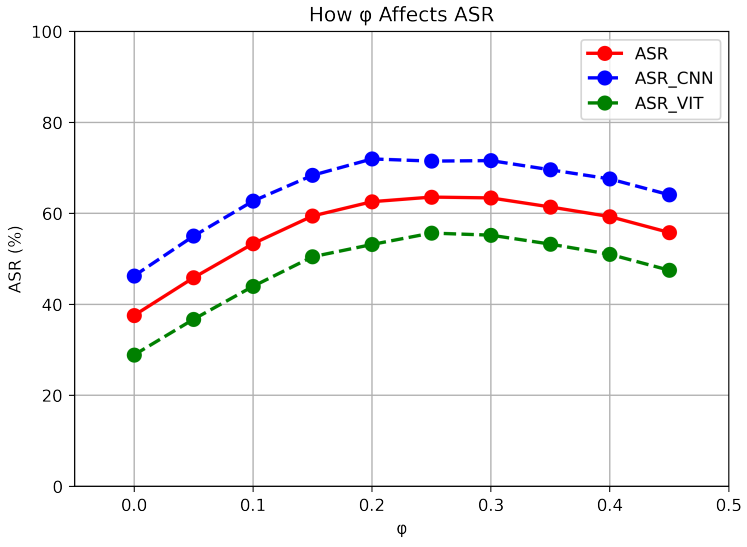

Figure 4. 超参数 ϕ 敏感性分析

ϕ 表示我们加入的随机噪声的范围，即其强度。从图中可以发现当 ϕ=0.25 时达到了攻击成功率最高，所以我们选择 ϕ=0.25 作为我们最终使用的数值。

5.2 不同数据集测试

我们接着对本方法在不同的数据集上进行攻击测试，同时对比方法有 Ens，以及两个已发表的效果较好的方法 SVRE，AdaEA，其中 AdaEA 是目前此类攻击中能达到最高攻击成功率的方法。

dataset	attack	res-50	wrn101-2	bit-50	bit-101	vit-B	deit-b	swin-b	swin-s	avg
cifar10	ens	47.76	24.9	36.79	17.09	10.95	25.66	24.46	31.31	**27.37**
	svre	53.97	26.64	41.97	19.83	13.29	31.26	29.74	36.35	**31.63（+4.26）**
	adaEA	67.21	39.75	55.75	31.43	23.63	45.82	45.8	53.24	**45.33（+17.96）**
	ours	74.85	46.31	63.52	35.76	28.5	59.16	59.69	67.97	**54.47（+27.10）**
cifar100	ens	73.84	61.58	69.7	41.83	41.67	64.78	63.03	66.48	**60.36**
	svre	78.94	63.31	74.21	44.01	45.72	69.1	65.43	69.88	**63.83（+3.47）**
	adaEA	86.46	72.23	81.55	49.46	48.42	70.76	69.57	73.82	**69.03（+8.67）**
	ours	95.37	84.89	90.43	62.4	69.59	89.81	86.26	89.16	**83.49（+23.10）**
ImageNet	ens	48.69	44.24	45.93	35.16	28.12	45.01	13.79	24.01	**35.62**
	svre	53.09	47.65	47.6	36.04	30.94	48.2	13.79	23.28	**37.57(+1.95)**
	adaEA	59.37	52.77	53.07	40.99	33.33	53.5	16.65	27.75	**42.18(+6.56)**
	ours	75.29	73.88	76.59	64.51	53.33	77.07	34.78	46.99	**52.49(+16.87)**

Figure 5. 不同数据集上的测试结果

如图5中数据所示，可以看到我们的方法在所有数据集上都比当前最好的方法的攻击成功率要高，并且不同于 AdaEA，我们在各个数据集上相对于 Ens 的提升都较大，尤其是难度较大的 ImageNet 上，因为 cifar10 与 cifar100 的图像像素点较少，攻击的偶然性更强，并且我们的方法比 AdaEA 平均高出 10 个点以上，足以证明我们的方法的有效性。

5.3 不同攻击策略测试

接着我们测试了我们的方法在不同的攻击策略下的攻击效果，结果如图6：

		target								avg
dataset	attack	res-50	wrn101-2	bit-50	bit-101	vit-B	deit-b	swin-b	swin-s	
FGSM	ens	32.46	31.88	34.45	28.9	22.6	31.32	12.73	18.71	**26.63**
	svre	40.42	38.81	41.25	34.95	31.35	43.21	17.5	25.57	**34.13(+7.50)**
	adaEA	36.44	37.21	41.14	35.16	34.17	45.44	18.66	26.51	**34.34(+7.71)**
	ours	48.27	49.68	52.06	42.86	35.21	48.3	21.74	30.46	**41.07(+14.44)**
I-FGSM	ens	48.69	44.24	45.93	35.16	28.12	45.01	13.79	24.01	**35.62**
	svre	53.09	47.65	47.6	36.04	30.94	48.2	13.79	23.28	**37.57(+1.95)**
	adaEA	59.37	52.77	53.07	40.99	33.33	53.5	16.65	27.75	**42.18(+6.56)**
	ours	75.29	73.88	76.59	64.51	53.33	77.07	34.78	46.99	**52.49(+16.87)**
MI-FGSM	ens	54.66	51.6	50.84	40.66	34.27	52.34	16.65	26.3	**40.92**
	svre	57.49	51.6	51.95	41.43	37.92	56.48	17.6	28.48	**42.87(+1.95)**
	adaEA	68.69	63.11	62.32	51.32	44.69	65.5	24.39	34.51	**51.82(+10.90)**
	ours	81.47	79.74	80.82	70.44	60.52	82.91	38.81	52.6	**68.41(+27.49)**
DI-FGSM	ens	78.53	77.29	78.04	67.25	55.83	74.63	38.07	50.42	**65.01**
	svre	61.78	61.73	63.1	53.08	50.21	65.07	31.81	42.62	**53.68(−11.33)**
	adaEA	82.3	82.62	84.62	73.74	61.56	82.27	42.84	55.41	**70.67(+5.66)**
	ours	90.58	90.62	91.64	86.26	74.79	91.08	54.4	69.85	**81.15(+16.14)**

Figure 6. 不同攻击策略的测试结果

其中 FGSM 为单步攻击，I-FGSM 为多步迭代攻击，MI-FGSM 是带有动量的多步迭代攻击，DI-FGSM 是带有数据增强的带动量多步迭代攻击。可以看到我们的方法仍然在所有方法中都达到了最佳的攻击效果，并且远超当前的最好方法的效果。

5.4 防御模型测试

对于进行了对抗训练的模型而言，攻击成功显得更加困难，所以在进行对抗训练后的模型上进行测试更能体现我们方法的有效性，我们分别选择了在 1 个、3 个、4 个对抗攻击方法上进行对抗训练的模型进行测试，测试结果如图7：

		surr models				adv			avg
dataset	method	resnet18	inc_v3	vit_t	deit_t	ens3	ens4	ens	
FGSM	ens	81.04	85.67	77.99	71.79	27.79	29.62	20.92	**26.11**
	svre	79.77	80.3	82.46	77.2	31.84	32.43	23.85	**29.37(+3.26)**
	adaEA	89.95	89.39	96.02	67.88	28.45	31.53	21.34	**27.11(+1.00)**
	ours	93.13	93.53	95.4	89.17	44.42	41.67	32.11	**39.4(+13.29)**
I-FGSM	ens	100	99.86	100	100	26.91	23.99	17.89	**22.93**
	svre	98.73	98.62	99.13	98.87	27.02	25.79	19.35	**24.05(+1.12)**
	adaEA	100	99.96	100	100	31.84	28.27	20.82	**26.98(+4.05)**
	ours	100	100	100	100	50.55	49.89	36.4	**45.6(+22.67)**
MI-FGSM	ens	100	100	100	100	32.06	31.53	23.43	**29.01**
	svre	91.86	91.18	92.16	92.19	34.68	33.9	24.37	**30.98(+1.97)**
	adaEA	100	100	100	100	39.72	36.37	26.26	**34.12(+5.11)**
	ours	100	100	100	100	59.63	57.21	44.04	**53.63(+24.62)**
DI-FGSM	ens	99.75	99.86	99.75	99.62	62.47	61.15	47.8	**57.14**
	svre	93.38	92.7	94.03	94.08	47.37	46.96	37.03	**43.79(−13.35)**
	adaEA	100	99.86	100	100	67.72	63.29	49.79	**60.27(+3.13)**
	ours	100	100	100	99.87	80.74	79.5	68.2	**76.15(+19.01)**

Figure 7. 防御模型测试结果

可以发现我们的方法在对抗训练后的模型上的效果要远好于其他方法，并且提升幅度比普通模型的更大。在 DI-FGSM 上 SVRE 的效果甚至不如 Ens，AdaEA 方法的提升也非常小。

6 总结与展望

以上实验都充分的表明了我们方法的优越性，尤其是我们的方法几乎在所有实验中的效果都远超当前最好的方法的最优结果。充分挖掘了集成对抗攻击的潜力，促进了本方向更深入的研究。而模型多样性增强方面仍有较大发展前景。

References

[1] Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Hee-Seon Kim, and Changick Kim. Improving the transferability of targeted adversarial examples through object-based diverse input. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 15244–15253, June 2022.

[2] Bin Chen, Jiali Yin, Shukai Chen, Bohao Chen, and Ximeng Liu. An adaptive model ensemble adversarial attack for boosting adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4489–4498, 2023.

[3] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, June 2018.

[4] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, June 2019.

[5] 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 *Proc. Int'l Conf. Learn. Repres.*, 2021.

[6] Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *Proc. Euro. Conf. Comput. Vis.*, pages 307–322, 2020.

[7] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proc. Int'l Conf. Learn. Repres.*, 2015.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 770–778, June 2016.

[9] Ziwen He, Wei Wang, Xinsheng Xuan, Jing Dong, and Tieniu Tan. A new ensemble method for concessively targeted multi-model attack. *arXiv preprint arXiv:1912.10833*, 2019.

[10] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Proc. Euro. Conf. Comput. Vis.*, pages 491–507, 2020.

[11] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proc. Int'l Conf. Learn. Repres.*, 2017.

[12] 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 *Proc. IEEE Int'l Conf. Comput. Vis.*, pages 9992–10002, 2021.

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. Int'l Conf. Learn. Repres.*, 2018.

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 1–9, June 2015.

[15] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[16] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *Proc. Int'l Conf. Machine Learn.*, volume 139, pages 10347–10357, 18–24 Jul 2021.

[17] Jiakai Wang. Adversarial examples in physical world. In *Proc. Int'l Joint Conf. Artif. Intell.*, pages 4925–4926, 8 2021.

[18] Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting adversarial transferability through enhanced momentum. In *Proc. British Conf. Machine Vis.*, 2021.

[19] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren. Feature importance-aware transferable adversarial attacks. In *Proc. IEEE Int'l Conf. Comput. Vis.*, pages 7619–7628, 2021.

[20] 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 *Proc. Conf. Comput. Vis. Pattern Recognit.*, June 2019.

[21] Yifeng Xiong, Jiadong Lin, Min Zhang, John E. Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 14983–14992, June 2022.

[22] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proc. British Conf. Machine Vis.*, pages 87.1–87.12, September 2016.

[23] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R. Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 14993–15002, June 2022.