Adversarial vulnerability of powerful near out-of-distribution detection

强大的近域外检测的对抗脆弱性

Stanislav Fort 1

Stanislav Fort 1

Abstract

摘要

There has been a significant progress in detecting out-of-distribution (OOD) inputs in neural networks recently, primarily due to the use of large models pretrained on large datasets, and an emerging use of multi-modality. We show a severe adversarial vulnerability of even the strongest current OOD detection techniques. With a small, targeted perturbation to the input pixels, we can change the image assignment from an in-distribution to an out-distribution, and vice versa, easily. In particular, we demonstrate severe adversarial vulnerability on the challenging near OOD CIFAR-100 vs CIFAR-10 task, as well as on the far OOD CIFAR-100 vs SVHN. We study the adversarial robustness of several post-processing techniques, including the simple baseline of Maximum of Softmax Probabilities (MSP), the Mahalanobis distance, and the newly proposed Relative Mahalanobis distance. By comparing the loss of OOD detection performance at various perturbation strengths, we demonstrate the beneficial effect of using ensembles of OOD detectors, and the use of the Relative Mahalanobis distance over other postprocessing methods. In addition, we show that even strong zero-shot OOD detection using CLIP and multi-modality suffers from a severe lack of adversarial robustness as well. Our code is available on GitHub.

最近，在神经网络中检测域外 (OOD) 输入取得了显著进展，这主要得益于在大型数据集上预训练的大型模型的使用，以及多模态的出现。我们展示了即使是当前最强的 OOD 检测技术也存在严重的对抗脆弱性。通过对输入像素进行小规模的有针对性的扰动，我们可以轻松地将图像分配从域内更改为域外，反之亦然。特别是，我们在具有挑战性的近域外 CIFAR-100 与 CIFAR-10 任务上，以及在远域外 CIFAR-100 与 SVHN 上展示了严重的对抗脆弱性。我们研究了几种后处理技术的对抗鲁棒性，包括最大软最大概率 (MSP) 的简单基线、马哈拉诺比斯距离以及新提出的相对马哈拉诺比斯距离。通过比较在不同扰动强度下的 OOD 检测性能损失，我们展示了使用 OOD 检测器集成的有益效果，以及相对马哈拉诺比斯距离相较于其他后处理方法的优势。此外，我们还展示了即使是使用 CLIP 和多模态的强大零样本 OOD 检测也存在严重的对抗鲁棒性缺失。我们的代码可在 GitHub 上获得。

1 Introduction

1 引言

The recent success of deep neural networks has led to their increasing deployment in high-stakes, safety critical applications such as health care [1; 2] , where models are required to be not only accurate but also robust to distribution shift. [3] Neural networks often assign high confidence to inputs that are misclassified, or even do not come from the distribution they were trained on at all [4; 5] . Reliable out-of-distribution (OOD) detection remains a significant challenge.

深度神经网络的近期成功导致其在高风险、安全关键应用中的日益广泛部署，例如医疗保健 [1; 2] ，在这些应用中，模型不仅需要准确，还需要对分布变化具有鲁棒性。[3] 神经网络通常对被错误分类的输入赋予高置信度，甚至对那些根本不来自于其训练分布的输入也是如此 [4; 5] 。可靠的域外 (OOD) 检测仍然是一个重大挑战。

Improving OOD detection has seen progress by training generative models [6; 7; 2; 8] , and modifying objective and loss functions [9]. Exposure to a number of OOD samples during training has also lead to improvements [10].

改进 OOD 检测已经通过训练生成模型 [6; 7; 2; 8] 和修改目标及损失函数 [9] 取得了进展。在训练过程中接触到大量的 OOD 样本也带来了改进 [10]。

Recently, large models (such as the Vision Transformer [11]) pre-trained on large datasets (such as ImageNet21k [12]) produced sufficiently high-quality image embeddings that allowed us to close the gap to human performance in many challenging near-OOD tasks in vision (such as distinguishing CIFAR-100 from CIFAR-10)[1] , as well as to make significant progress in genomics. [13]

最近，大型模型 (如视觉变换器 [11]) 在大型数据集 (如 ImageNet21k [12]) 上进行预训练，生成了足够高质量的图像嵌入，使我们在许多具有挑战性的近 OOD 视觉任务中 (如区分 CIFAR-100 与 CIFAR-10)[1]) 缩小了与人类表现之间的差距，并在基因组学方面取得了显著进展 [13]。

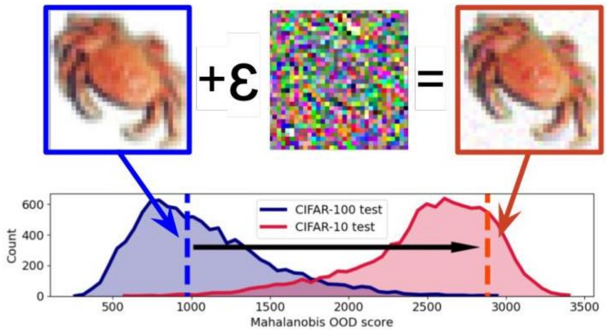


Figure 1: A small adversarial perturbation to the pixels of an in-distribution image (CIFAR-100) changes its out-of-distribution (OOD) score from $\approx 1,000$ (around the mode of the in-distribution) to a confident out-distribution (CIFAR-10) region at $\approx 2,800$ even for a state-of-the-art near-OOD detection method using a large ViT. $\varepsilon = 10^{-4}$ and the attack used is the Fast Gradient Sign Method applied to the Mahalanobis distance score for a ViT-L $_{16}$ as used in [13]. The unperturbed CIFAR-100 → CIFAR-10 AUROC for this model is 97.98% .

图 1: 对一个分布内图像 (CIFAR-100) 像素的小对抗扰动将其 OOD 分数从 $\approx 1,000$ (接近分布内的模式) 改变为一个自信的分布外 (CIFAR-10) 区域，在 $\approx 2,800$ ，即使对于使用大型 ViT 的最先进近 OOD 检测方法也是如此。$\varepsilon = 10^{-4}$ 使用的攻击是应用于 ViT-L 的马哈拉诺比斯距离分数的快速梯度符号方法 $_{16}$ ，如 [13] 中所用。未扰动的 CIFAR-100 → 该模型的 CIFAR-10 AUROC 为 97.98% 。

A Mahalanobis distance (MD) based method [14] is a simple approach for post-processing embedding vectors coming from a neural network for OOD detection. Some of its common failure modes have been improved upon by the introduction of the Relative Mahalanobis Distance (RMD) in [15], generally improving performance while being more agnostic to hyperparameter choice.

基于马哈拉诺比斯距离 (MD) 的方法 [14] 是一种简单的后处理神经网络嵌入向量以进行 OOD 检测的方法。通过在 [15] 中引入相对马哈拉诺比斯距离 (RMD)，对其一些常见的失效模式进行了改进，通常提高了性能，同时对超参数选择更加不敏感。

Mahalanobis distance based methods are good at detecting far OOD samples - for example CIFAR-10 vs SVHN, where the samples are distinct both in their surface-level style as well as in semantics. Near OOD samples - for example CIFAR-100 vs CIFAR-10, where samples are superficially similar and differ only in their semantic content - have remained a challenge until the widespread use of large models and pre-training [13], and multi-modality (for example the use of CLIP [16] in [13] for zero-shot class-name-only exposure OOD detection).

基于马哈拉诺比斯距离的方法在检测远离分布 (OOD) 样本方面表现良好，例如 CIFAR-10 与 SVHN，其中样本在表面风格和语义上都明显不同。近 OOD 样本，例如 CIFAR-100 与 CIFAR-10，样本在表面上相似，仅在语义内容上有所不同，这一直是一个挑战，直到大型模型和预训练的广泛使用 [13]，以及多模态 (例如在 [13] 中使用 CLIP [16] 进行零-shot 类别名称曝光的 OOD 检测)。

The question of adversarial examples is usually framed in the classification setup, where an adversarial perturbation leads to a confident class change [17]. [18] show that OOD detection systems are also vulnerable to such attacks, and propose a robust training algorithm for counteracting it. [19] propose a training algorithm leading to more robust OOD detection as well.

[1] Stanford University. Correspondence to: Stanislav Fort <sfort1@stanford.edu>.

[1] 斯坦福大学。通讯联系:Stanislav Fort <sfort1@stanford.edu>。

[1] https://paperswithcode.com/sota/

[1] https://paperswithcode.com/sota/
out-of-distribution-detection-on-cifar-100-vs
out-of-distribution-detection-on-cifar-100-vs

对抗样本的问题通常在分类设置中提出，其中对抗扰动导致自信的类别变化 [17]。[18] 显示 OOD 检测系统也容易受到此类攻击，并提出了一种稳健的训练算法以对抗它。[19] 还提出了一种训练算法，旨在实现更稳健的 OOD 检测。

Key contributions: We show empirically that currently even the most powerful and robust OOD detection systems based on large models and massive data are severely vulnerable to targeted adversarial attacks. We demonstrate that this is the case for different post-processing techniques, including the baseline Max of Softmax Probabilities (MSP), as well as the more advanced Mahalanobis distance. The zero-shot multi-modal approach using CLIP suffers from an even more acute vulnerability to such attacks. We show that working with lower resolution images increases OOD adversarial robustness. The largest positive effect we see comes from the use of ensembles of several OOD detectors, and the use of the Relative Mahalanobis distance. We demonstrate that these two interventions can be successfully combined as well, making the detection system more adversarially robust as well as improving its OOD detection performance in general.

关键贡献: 我们通过实验证明，目前即使是基于大型模型和海量数据的最强大和稳健的 OOD 检测系统也严重容易受到针对性对抗攻击的影响。我们展示了这一点适用于不同的后处理技术，包括基线的最大软最大概率 (MSP) 以及更先进的马哈拉诺比斯距离。使用 CLIP 的零-shot 多模态方法对这种攻击的脆弱性更为严重。我们表明，使用较低分辨率的图像可以提高 OOD 对抗鲁棒性。我们观察到的最大积极效果来自于使用多个 OOD 检测器的集成，以及使用相对马哈拉诺比斯距离。我们证明这两种干预措施也可以成功结合，使检测系统在对抗性上更为鲁棒，同时改善其 OOD 检测性能。

2 Methods

2 方法

In this section, we describe how to get adversarial examples to OOD detection algorithms and briefly review the Maha-lanobis distance and Relative Mahalanobis distance methods. We also discuss the baseline Maximum of Softmax Probabilities, and the use of the multi-modal CLIP model for zero-shot OOD detection. We present a method for attacking ensembles of detectors we well.

在本节中，我们描述如何获取对抗样本以用于 OOD 检测算法，并简要回顾 Maha-lanobis 距离和相对 Mahalnobis 距离方法。我们还讨论了基线的最大软最大概率，以及使用多模态 CLIP 模型进行零样本 OOD 检测的方法。我们还提出了一种攻击检测器集成的方法。

2.1 Generating adversarial examples to OOD score

2.1 生成对抗样本以获得 OOD 评分

Given an out-of-distribution scoring function score(x)that maps an image $\mathbf{x}$ into a floating point value characterizing its distance from the in-distribution, we can use its gradient with respect to the input,

给定一个将图像 $\mathbf{x}$ 映射为浮点值的分布外评分函数 score(x)，该值表征其与内部分布的距离，我们可以使用其相对于输入的梯度，

$$g\left(\mathbf{x}\right)=\left.\frac{\partial\,\mathrm{score}\left(\mathbf{x}'\right)}{\partial\mathbf{x}'}\right|_{\mathbf{x}'=\mathbf{x}}, \tag{1}$$

to gradually change the input $\mathbf{x}$ to have either a higher or lower OOD score. This is exactly the same way adversarial examples, first described in [17], are typically generated. Modifications exist that change the form of the perturbation, for example the Fast Gradient Sign Method in [20] that uses $\mathrm{sign}\left(g\left(\mathbf{x}\right)\right)$ instead of $g\left(\mathbf{x}\right)$ as the step direction. We will primarily be using that in this paper, as it is easy to use and works well out of the box.

逐渐改变输入 $\mathbf{x}$ 以获得更高或更低的 OOD 评分。这正是对抗样本的生成方式，最早在 [17] 中描述。存在一些修改，改变了扰动的形式，例如在 [20] 中使用的快速梯度符号方法，它使用 $\mathrm{sign}\left(g\left(\mathbf{x}\right)\right)$ 而不是 $g\left(\mathbf{x}\right)$ 作为步长方向。我们将在本文中主要使用这种方法，因为它易于使用且开箱即用效果良好。

Starting from an in-distribution image of a low score (confidently in-distribution), taking iterative steps

从一个低评分的内部分布图像 (自信地属于内部分布) 开始，进行迭代步骤

$$\mathbf{x}_{t+1}=\mathbf{x}_t+\varepsilon g\left(\mathbf{x}\right), \tag{2}$$

where ε is the learning rate, we move in the local direction of increasing OOD score. As shown in Figure 1, a very small perturbation to an image of a crab leads to a shift from the center of the in-distribution scores to the higher end of the out-distribution scores. This turns the image from a confidently and correctly in-distribution to a confidently out-distribution, as judged by a well-performing detection method from [13].

其中 ε 是学习率，我们沿着增加 OOD 评分的局部方向移动。如图 1 所示，对一只螃蟹图像进行非常小的扰动会导致其从内部分布评分的中心移动到外部分布评分的高端。这使得图像从自信且正确的内部分布变为自信的外部分布，依据 [13] 中的高效检测方法进行判断。

2.2 Mahalanobis distance based OOD detection

2.2 基于 Mahalnobis 距离的 OOD 检测

The Mahalanobis distance (MD) [14] method and Relative Mahalanobis distance (RMD) [15] method use intermediate features of a trained deep neural network. A frequent choice of the features are the pre-logits - the output of the second to last layer of a network, just before the classification layer. Let us indicate these features as $\mathbf{z}_i = f\left(\mathbf{x}_i\right)$ for an input $\mathbf{x}_i$.

马哈拉诺比斯距离 (MD) [14] 方法和相对马哈拉诺比斯距离 (RMD) [15] 方法使用经过训练的深度神经网络的中间特征。特征的一个常见选择是预日志值——即网络倒数第二层的输出，紧接着分类层。我们将这些特征表示为 $\mathbf{z}_i = f\left(\mathbf{x}_i\right)$ ，用于输入 $\mathbf{x}_i$ 。

For a K -class in-distribution dataset, both methods fit K class-specific Gaussian distributions $\mathcal{N}\left(\mu_k, \sum\right), k = 1, 2, \ldots, K$ to each of the K in-distribution classes using their feature vectors $\mathbf{z}_i$

对于一个 K 类的内部分布数据集，这两种方法为每个 K 内部分布类拟合 K 类特定的高斯分布 $\mathcal{N}\left(\mu_k, \sum\right), k = 1, 2, \ldots, K$ ，使用它们的特征向量 $\mathbf{z}_i$ 。

We compute the class centroids (means) and covariance matrices as: $\mu_k = \frac{1}{N_k} \sum\limits_{i:y_i=k} \mathbf{z}_i$, for $k = 1, \ldots, K$, and $\sum = \frac{1}{N} \sum\limits_{k=1}^{K} \sum\limits_{i:y_i=k} \left(\mathbf{z}_i - \mu_k\right)\left(\mathbf{z}_i - \mu_k\right)^T$. Notice that the class means μ_k are independent for each class, while we use the same covariance matrix $\sum$ for all classes to avoid numerical issues due to under-fitting to the typically smaller than needed numbers of examples.

我们计算类中心 (均值) 和协方差矩阵，如下所示: $\mu_k = \frac{1}{N_k} \sum\limits_{i:y_i=k} \mathbf{z}_i$ ，对于 $k = 1, \ldots, K$ 和 $\sum = \frac{1}{N} \sum\limits_{k=1}^{K} \sum\limits_{i:y_i=k} \left(\mathbf{z}_i - \mu_k\right)\left(\mathbf{z}_i - \mu_k\right)^T$ 。注意，类均值 μ_k 对于每个类是独立的，而我们对所有类使用相同的协方差矩阵 $\sum$ ，以避免由于样本数量通常小于所需数量而导致的数值问题。

For a test input $\mathbf{x}'$ whose in- or out-distribution assignment is to be determined, we compute the Mahalanobis distances from the embedding vector of the test input $\mathbf{z}' = f\left(\mathbf{x}'\right)$ to each of the K in-distribution Gaussian distributions $\mathcal{N}\left(\mu_k, \sum\right), k \in \{1, \ldots, K\}$ given by $\mathrm{MD}_k\left(\mathbf{z}'\right)$ we just computed. We take the minimum of the distances over all classes to be the uncertainty score $\mathcal{U}\left(\mathbf{x}'\right)$ characterizing how far from the in-distribution the input x' is deemed to be. There the score can be seen as the extent to which the sample is OOD. The Mahalanobis distances are computed as

对于一个测试输入 $\mathbf{x}'$ ，其内部或外部分配需要确定，我们计算测试输入的嵌入向量 $\mathbf{z}' = f\left(\mathbf{x}'\right)$ 到每个 K 内部分布高斯分布 $\mathcal{N}\left(\mu_k, \sum\right), k \in \{1, \ldots, K\}$ 的马哈拉诺比斯距离，这些距离由我们刚刚计算的 $\mathrm{MD}_k\left(\mathbf{z}'\right)$ 给出。我们取所有类的距离中的最小值作为不确定性得分 $\mathcal{U}\left(\mathbf{x}'\right)$ ，以表征输入 x' 被认为距离内部分布的远近。因此，该得分可以视为样本是 OOD 的程度。马哈拉诺比斯距离的计算为

$$\mathrm{MD}_k\left(\mathbf{z}'\right) = \left(\mathbf{z}' - \mu_k\right)^T {\sum}^{-1} \left(\mathbf{z}' - \mu_k\right), \tag{3}$$

$$\operatorname{score}\left(\mathbf{x}'\right) = \mathcal{U}\left(\mathbf{x}'\right) = -\min_k \left\{\mathrm{MD}_k\left(\mathbf{z}'\right)\right\}. \tag{4}$$

This confidence score is used to distinguish the in-distribution and out-distribution samples from each other.

该置信得分用于区分内部分布样本和外部分布样本。

2.3 Relative Mahalanobis Distance

2.3 相对马哈拉诺比斯距离

In [15] the Relative Mahalanobis Distance is proposed which modifies Eq. 4 by subtracting a term to make it more robust to hyperparameter choice as well as generally better at OOD detection for near-OOD tasks in vision and genomics. The approach attempts to model the shape of the in-distribution and subtract its effects from the class-conditional distances. The RMD is defined as

在 [15] 中，提出了相对马哈拉诺比斯距离 (Relative Mahalanobis Distance)，该方法通过减去一个项来修改公式 4，使其对超参数选择更加稳健，并且在视觉和基因组学的近 OOD 任务中通常更适合进行 OOD 检测。该方法试图对在分布内的形状建模，并从类别条件距离中减去其影响。RMD 定义为

$$\mathrm{RMD}_k\left(\mathbf{z}'\right) = \mathrm{MD}_k\left(\mathbf{z}'\right) - \mathrm{MD}_0\left(\mathbf{z}'\right),$$

where $\mathrm{MD}_0\left(\mathbf{z}'\right)$ indicates the Mahalanobis distance to a Gaussian distribution fitted to the whole in-distribution dataset without regard to its label structure, as $\mathcal{N}\left(\mu_0, \sum_0\right)$, where $\mu_0 = \frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i$ and $\sum_0 = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{z}_i - \mu_0\right)\left(\mathbf{z}_i - \mu_0\right)^T$. The goal is to model the background distribution. The resulting uncertainty score using RMD is then

其中 $\mathrm{MD}_0\left(\mathbf{z}'\right)$ 表示与整个在分布数据集拟合的高斯分布的马哈拉诺比斯距离，而不考虑其标签结构，如 $\mathcal{N}\left(\mu_0, \sum_0\right)$ ，其中 $\mu_0 = \frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i$ 和 $\sum_0 = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{z}_i - \mu_0\right)\left(\mathbf{z}_i - \mu_0\right)^T$ 。目标是对背景分布进行建模。使用 RMD 得到的结果不确定性评分为

$$\mathcal{C}^{\mathrm{RMD}}\left(\mathbf{x}'\right) = -\min_k\left\{\mathrm{RMD}_k\left(\mathbf{z}'\right)\right\} \tag{5}$$

This can be extended to more powerful generative models fit ([21; 22]) to the class-specific and full-dataset approximations. [15]

这可以扩展到更强大的生成模型，拟合 ([21; 22]) 到类别特定和完整数据集的近似值。[15]

2.4 Maximum of Softmax Probabilities

2.4 Softmax 概率的最大值

A solid baseline for OOD detection is provided by the simple approach of using the Maximum of Softmax Probabilities as the in-distribution score. For a classification model $f\left(\mathbf{x}\right) = \mathbf{p}$ that maps in input image $\mathbf{x}$ to a vector of probabilities $\mathbf{p}$, the OOD score is $\mathrm{score}\left(\mathbf{x}\right) = \max\left(f\left(\mathbf{x}\right)\right)$. For in-distribution images, for a well trained model the image will belong to one of the output classes that will likely be close to 1 in the probabilities vector. For an OOD sample, the model will likely be confused and will not assign as high a probability to any of the classes. This provides the rational for using this method, which proved to be a good baseline given how simple its implementation is.

使用 Softmax 概率的最大值作为在分布内评分，提供了一个稳固的 OOD 检测基线。对于一个分类模型 $f\left(\mathbf{x}\right) = \mathbf{p}$ ，它将输入图像 $\mathbf{x}$ 映射到概率向量 $\mathbf{p}$ ，OOD 评分为 $\mathrm{score}\left(\mathbf{x}\right) = \max\left(f\left(\mathbf{x}\right)\right)$ 。对于在分布内的图像，对于一个训练良好的模型，该图像将属于输出类之一，该类的概率向量中的值可能接近 1。对于一个 OOD 样本，模型可能会感到困惑，并且不会给任何类别分配如此高的概率。这为使用此方法提供了合理性，考虑到其实现的简单性，这被证明是一个良好的基线。

2.5 Zero-shot multi-modal OOD detection using words to specify distributions

2.5 使用词语指定分布的零样本多模态 OOD 检测

[13] introduce a new kind of OOD detection scenario, where they use a multi-modal CLIP model [16]. CLIP produces a similarity score comparing the semantic content of an image and a text, as logit(I, T). By choosing two sets of words: in-words characterizing the semantic content of the in-distribution,

and out-words, characterizing the semantic content of the out-distribution, for each image I we can compute the in-logits for the in-words as $z_i^{\text{in}} = \text{CLIP}(I,\ \text{inword}_i)$, and the out-logits for the out-words $z_i^{\text{out}} = \text{CLIP}(I,\ \text{outword}_i)$. We construct the score the same way as in [13] as

[13] 引入了一种新的 OOD 检测场景，在该场景中，他们使用多模态 CLIP 模型 [16]。CLIP 生成一个相似度分数，用于比较图像和文本的语义内容，记作 logit(I, T)。通过选择两组词: 描述内分布语义内容的内词和描述外分布语义内容的外词，对于每个图像 I ，我们可以计算内词的内 logits 为 $z_i^{\text{in}} = \text{CLIP}(I,\ \text{inword}_i)$ ，外词的外 logits 为 $z_i^{\text{out}} = \text{CLIP}(I,\ \text{outword}_i)$ 。我们以与 [13] 相同的方式构建分数。

$$\text{score}(\mathbf{x}) = \max\left(\{\text{CLIP}(\mathbf{x}, \text{inword}_i\}_i\right) \tag{6}$$

$$\text{- }\max\left(\{\text{CLIP}(\mathbf{x},\ \text{outword}_i)_i\right)\ . \tag{7}$$

We can modify this score the same way we do for the Ma-halanobis distance or Relative Mahalanobis distance using a gradient step with respect to the image.

我们可以通过对图像进行梯度步长调整，以与 Mahalanobis 距离或相对 Mahalanobis 距离相同的方式修改该分数。

2.6 Ensembling OOD detectors

2.6 集成 OOD 检测器

A simple way to improve the OOD detection capabilities of several OOD detectors is to ensemble them. For example, this is used in [13] to reach the current state-of-the-art performance on the near OOD CIFAR-100 → CIFAR-10 task. The simplest technique we can use is to generate the OOD score for a particular image $\mathbf{x}$ for each of the models $\text{score}_i(\mathbf{x})$, and compute their average

提高多个 OOD 检测器的 OOD 检测能力的一种简单方法是将它们进行集成。例如，这在 [13] 中被用于在近 OOD CIFAR-100 → CIFAR-10 任务上达到当前的最先进性能。我们可以使用的最简单技术是为每个模型 $\text{score}_i(\mathbf{x})$ 生成特定图像 $\mathbf{x}$ 的 OOD 分数，并计算它们的平均值。

$$\text{score}_{\text{ ensemble }}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\text{score}_i(\mathbf{x})\,. \tag{8}$$

The likely reason for why ensembling of the OOD predicted scores over several models works better than the models individually is similar to the reason for why deep ensembles work in general [5]. A loss landscape approach to that is discussed in [23].

集成多个模型的 OOD 预测分数之所以比单个模型效果更好的可能原因与深度集成通常有效的原因相似 [5]。在 [23] 中讨论了这种损失景观的方法。

2.7 Attacks on model ensembles

2.7 对模型集成的攻击

Attacking an ensemble of OOD detectors, as discussed in Section 2.6, is the same as attacking a single model. The only difference is that we replace the single model OOD scoring function score(x)with the ensemble scoring function $\text{score e}_{\text{ ensemble }}(\mathbf{x})$.

攻击一个 OOD 检测器集成，如第 2.6 节所讨论的，实际上与攻击单个模型是相同的。唯一的区别是我们将单个模型的 OOD 评分函数 score(x) 替换为集成评分函数 $\text{score e}_{\text{ ensemble }}(\mathbf{x})$ 。

3 Experiments and Results

3 实验与结果

We studied the adversarial robustness of the currently best performing methods on the near-OOD task of distinguishing CIFAR-100 (in-distribution) from CIFAR-10 (out-distribution). [2] The best performing approach is an ensemble of pre-trained Vision Transformers finetuned on CIFAR-100 with the Mahalanobis distance post-processing method applied on top of their embeddings. This reaches an AUROC of 97.98% [13], as compared to a human benchmark of AUROC ≈ 96.0% . The best approach not using an ensemble of detectors differs in using a single ViT only.

我们研究了当前表现最佳的方法在近 OOD 任务上区分 CIFAR-100(内分布) 和 CIFAR-10(外分布) 的对抗鲁棒性。[2] 表现最佳的方法是一个经过预训练的视觉变换器的集成，这些变换器在 CIFAR-100 上进行了微调，并在其嵌入上应用了马氏距离后处理方法。这达到了 AUROC 97.98% [13]，而与人类基准 AUROC ≈ 96.0% 相比。未使用检测器集成的最佳方法仅使用单个 ViT。

We chose the pre-trained and finetuned ViT-L_ 16[3] to develop OOD adversarial attacks to. Its default resolution is 384×384 and we used the standard tf.image.resize to up-sample the 32×32 CIFAR images to it, as done in the standard ViT preprocessing pipeline.

我们选择了预训练并微调的 ViT-L_ 16[3] 来开发 OOD 对抗攻击。其默认分辨率为 384×384 ，我们使用标准的 tf.image.resize 将 32×32 CIFAR 图像上采样到该分辨率，正如在标准 ViT 预处理管道中所做的那样。

3.1 Attacks on CIFAR-100 vs CIFAR-10 for different post-processing techniques

3.1 针对不同后处理技术的 CIFAR-100 与 CIFAR-10 的攻击

Mahalanobis distance We focused on the challenging near-OOD CIFAR-100 vs CIFAR-10 task. Figure 2 shows an image of an airplane (CIFAR-10, out-distribution) being adversarially modified using the Fast Gradient Sign Method to read as a confident in-distribution image do the ViT based Mahalanobis distance OOD detector. The figure also shows the shift of the OOD score against the histograms of the in- and out-distribution test set images. This is similar to Figure 1, where the direction of change was from the in-distribution to the out-distribution. A small change in the pixel values of the input image resulted in a large change of the OOD score assigned.

马氏距离我们专注于具有挑战性的近 OOD CIFAR-100 与 CIFAR-10 任务。图 2 展示了一张飞机的图像 (CIFAR-10，外分布)，该图像通过快速梯度符号方法被对抗性地修改，以使其看起来像是 ViT 基于马氏距离的 OOD 检测器所认为的可信内分布图像。该图还显示了 OOD 分数相对于内外分布测试集图像直方图的变化。这与图 1 类似，其中变化的方向是从内分布到外分布。输入图像的像素值的小变化导致了分配的 OOD 分数的大变化。

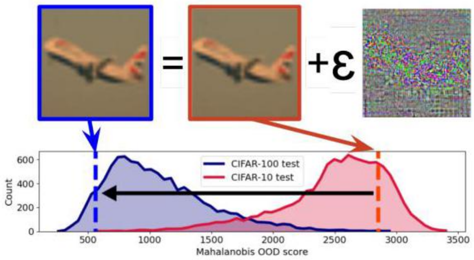


Figure 2: A small adversarial perturbation to the pixels of an out-distribution image (CIFAR-10) changes its out-of-distribution (OOD) score from $\approx 2,800$ (around the mode of the out-distribution) to a confident in-distribution (CIFAR-100) region at ≈ 600 even for a state-of-the-art near-OOD detection method using a large ViT. $\varepsilon = 10^{-4}$ and the attack used is the Fast Gradient Sign Method applied to the Mahalanobis distance score for a ViT-L $_{16}$ as used in [13]. The unperturbed CIFAR-100 → CIFAR-10 AUROC for this model is 97.98% .

图 2: 对一个分布外图像 (CIFAR-10) 像素的小对抗扰动将其分布外 (OOD) 评分从 $\approx 2,800$ (大约在分布外的模式附近) 改变为在 ≈ 600 的一个自信的分布内 (CIFAR-100) 区域，即使对于使用大型 ViT 的

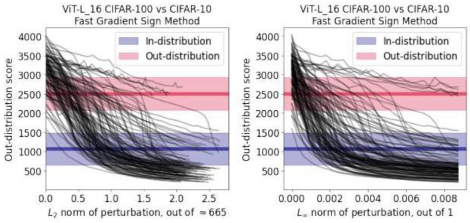
[2] https://paperswithcode.com/sota/

[2] https://paperswithcode.com/sota/
out-of-distribution-detection-on-cifar-100-vs
在 CIFAR-100 与 CIFAR-10 的外分布检测

最先进的近 OOD 检测方法也是如此。$\varepsilon = 10^{-4}$ ，所使用的攻击方法是应用于 Mahalanobis 距离评分的快速梯度符号方法，针对 ViT-L $_{16}$ ，如 [13] 中所用。该模型的未扰动 CIFAR-100 → CIFAR-10 AUROC 为 97.98% 。

Applying the same procedure to 128 test images, we were able to generate a set of perturbed out-distribution images that read as confidently in-distribution to the detector, as shown in Figure 3a as a function of the L_2 norm of the image perturbation and in Figure 3b as a function of the L_∞ norm.

将相同的程序应用于 128 个测试图像，我们能够生成一组扰动的分布外图像，这些图像在检测器看来是自信的分布内，如图 3a 所示，作为图像扰动的 L_2 范数的函数，以及在图 3b 中作为 L_∞ 范数的函数。



(a) (b)

Figure 3: Changing the out-distribution score for a set of 128 CIFAR-10 test images (out-distribution) by applying the Fast Gradient Sign Method to the Mahalanobis OOD score based on a ViT-L_16. (a) shows the score as a function of the L_2 norm of the image perturbation, while (b) shows the L_∞ norm.

图 3: 通过将快速梯度符号方法应用于基于 ViT-L_16 的 Mahalanobis OOD 评分，改变一组 128 个 CIFAR-10 测试图像 (分布外) 的分布外评分。(a) 显示了作为图像扰动的 L_2 范数的函数的评分，而 (b) 显示了 L_∞ 范数。

Relative Mahalanobis distance Using the proposed Relative Mahalanobis distance [15], that we discuss in Section 2.3, as an OOD score, we show an equivalent effect of a small adversarial perturbation on the OOD score in Figure 4. Applying this attack to 128 out-distribution images and their gradual score change with the L_2 and L_∞ norms of the perturbation are shown in Figure 5a and Figure 5b respectively.

使用所提出的相对马哈拉诺比斯距离 [15]，我们在第 2.3 节中讨论，将其作为 OOD 分数，我们展示了小的对抗扰动对 OOD 分数的等效影响，如图 4 所示。将此攻击应用于 128 张分布外图像，其与扰动的 L_2 和 L_∞ 范数的逐渐分数变化分别如图 5a 和图 5b 所示。

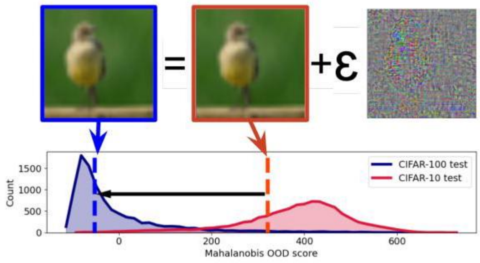
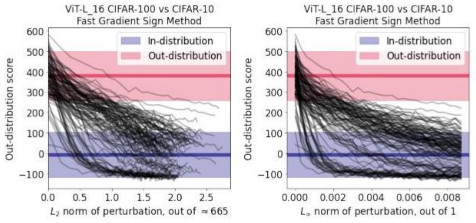


Figure 4: A small adversarial perturbation to the pixels of an out-distribution image (CIFAR-10) changes its out-of-distribution (OOD) score from ≈ 300 (around the mode of the out-distribution) to a confident in-distribution (CIFAR- 100) region at ≈ −50 even for a state-of-the-art near-OOD detection method using a large ViT. $\varepsilon = 10^{-4}$ and the attack used is the Fast Gradient Sign Method applied to the Relative Mahalanobis distance score for a ViT-L $_{16}$ as used in [13]. The unperturbed CIFAR-100 → CIFAR-10 AU-ROC for this model is 97.11% .

图 4: 对分布外图像 (CIFAR-10) 像素的小对抗扰动将其分布外 (OOD) 分数从 ≈ 300 (大约在分布外的模式附近) 改变为一个自信的分布内 (CIFAR-100) 区域，在 ≈ −50 ，即使对于使用大型 ViT 的最先进的近 OOD 检测方法也是如此。$\varepsilon = 10^{-4}$ 并且使用的攻击是应用于相对马哈拉诺比斯距离分数的快速梯度符号方法，针对 ViT-L $_{16}$ ，如 [13] 中所使用的。未扰动的 CIFAR-100 → CIFAR-10 AU-ROC 对于该模型为 97.11% 。

Maximum of Softmax Probabilities We used the Maximum of Softmax Probabilities (MSP) as a baseline postprocessing method for OOD detection, as discussed in Section 2.4. Figure 6a and Figure 6b show the change in the score of 128 out-distribution test images as a function of the L_2 and L_∞ norms of the image perturbation.

最大软最大概率我们使用最大软最大概率 (MSP) 作为 OOD 检测的基线后处理方法，如第 2.4 节中讨论的。图 6a 和图 6b 显示了 128 张分布外测试图像的分数变化，作为图像扰动的 L_2 和 L_∞ 范数的函数。



(a) (b)

Figure 5: Changing the out-distribution score for a set of 128 CIFAR-10 test images (out-distribution) by applying the Fast Gradient Sign Method to the Relative Mahalanobis OOD score based on a ViT-L_16. (a) shows the score as a function of the L_2 norm of the image perturbation, while (b) shows the L_∞ norm.

图 5: 通过将快速梯度符号方法应用于基于 ViT-L_16 的相对马哈拉诺比斯 OOD 分数，改变一组 128 个 CIFAR-10 测试图像 (外部分布) 的分数。(a) 显示了图像扰动的 L_2 范数作为分数的函数，而 (b) 显示了 L_∞ 范数。

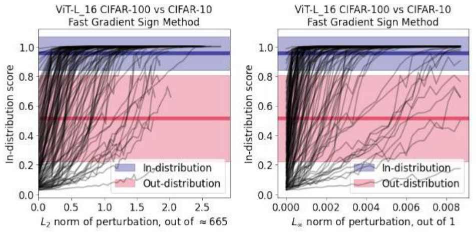


Figure 6: Changing the out-distribution score for a set of 128 CIFAR-10 test images (out-distribution) by applying the Fast Gradient Sign Method to the Maximum of Soft-max Probabilities (MSP) score based on a ViT-L_16. (a) shows the score as a function of the L_2 norm of the image perturbation, while (b) shows the L_∞ norm.

图 6: 通过将快速梯度符号方法应用于基于 ViT-L_16 的最大软最大概率 (MSP) 分数，改变一组 128 个 CIFAR-10 测试图像 (外部分布) 的分数。(a) 显示了图像扰动的 L_2 范数作为分数的函数，而 (b) 显示了 L_∞ 范数。

Robustness comparison The stronger the adversarial attack, the more we can change the out-distribution samples in order for them to be perceived as in-distribution by the detection system. Table 1 summarizes the loss of the AU-ROC on the CIFAR-100 vs CIFAR-10 task for the standard Mahahalanobis distance, the Relative Mahalanobis distance, and the Maximum of Softmax Probabilities (comparison baseline). Figure 7a and Figure 7b show the loss of AUROC as a function of the perturbation strength measured by their L_2 and L_∞ norms. The results in Table 1 can be read off from Figure 7b by looking at $L_\infty = 1/255$.

鲁棒性比较对抗攻击越强，我们就越能改变外部分布样本，使其被检测系统视为内部分布。表 1 总结了标准马哈拉诺比斯距离、相对马哈拉诺比斯距离和最大软最大概率 (比较基线) 在 CIFAR-100 与 CIFAR-10 任务上的 AU-ROC 损失。图 7a 和图 7b 显示了 AUROC 损失与通过其 L_2 和 L_∞ 范数测量的扰动强度的关系。表 1 中的结果可以通过查看图 7b 中的 $L_\infty = 1/255$ 来读取。

The way we turned Figures 3 b, 5 b and 6 b into the summary in Figure 7b was as follows. Each image is adversar-ially modified in T steps. Its OOD score and L_∞ perturbation norm change as a function of T . We used a piecewise linear interpolation to make an image-specific function score (L_∞) . Then, when making Figure 7b, we sampled the L_∞ perturbation norms we wanted to explore, and for each computed

[3] https://github.com/google-research/vision_transformer

[3] https://github.com/google-research/vision_transformer

the interpolated OOD score for each of the 128 images based on their individual linear interpolations. The resulting distribution of scores was then com-

我们将图形 3 b, 5 b 和 6 b 转换为图 7b 中的摘要的方式如下。每个图像在 T 步骤中被对抗性地修改。其 OOD 分数和 L_∞ 扰动范数随着 T 的变化而变化。我们使用分段线性插值来生成特定于图像的函数分数 (L_∞) 。然后，在制作图 7b 时，我们采样了想要探索的 L_∞ 扰动范数，并为每个图像计算了基于其个体线性插值的插值 OOD 分数。最终的分数分布随后被比较

Table 1: The loss of AUROC on the near OOD CIFAR-100 vs CIFAR-10 task for several OOD detection approaches. The strength of the attack is fixed by the L_∞ norm of the adversarial perturbation at 1/255. All approaches use the pretrained and finetuned ViT-L $_{16}$ to generate probability outputs and embeddings. The baseline of using the Max of Softmax Probabilities (MSP) is the least robust, followed by the Standard Mahalanobis Distance. The newly proposed Relative Mahalanobis Distance has the highest adversarial robustness by a significant margin. pared to the scores of the in-distribution test set to obtain the AUROC. For the L_2 norm in Figure 7a the process was analogous, swapping L_∞ for L_2 everywhere.

表 1: 在近 OOD CIFAR-100 与 CIFAR-10 任务中，几种 OOD 检测方法的 AUROC 损失。攻击的强度由对抗扰动的 L_∞ 范数固定为 1/255。所有方法均使用预训练和微调的 ViT-L $_{16}$ 生成概率输出和嵌入。使用最大软最大概率 (MSP) 的基线是最不稳健的，其次是标准马哈拉诺比斯距离。新提出的相对马哈拉诺比斯距离在对抗鲁棒性方面具有显著优势。将其与分布内测试集的分数进行比较以获得 AUROC。对于图 7a 中的 L_2 范数，过程类似，只是在每个地方将 L_∞ 替换为 L_2 。

Post-process method	AUROC before	AUROC l_∞ 1/255	Δ AUROC
Max of Softmax Probs	94.28%	27.48%	-66.8%
Maha	97.98%	41.33%	-56.65%
Relative Maha	97.11%	71.84%	-25.27%

后处理方法	AUROC 之前	AUROC l_∞ 1/255	Δ AUROC
Softmax 概率的最大值	94.28%	27.48%	-66.8%
马哈	97.98%	41.33%	-56.65%
相对马哈	97.11%	71.84%	-25.27%

To compare the robustness of the standard Mahalanobis distance and the Relative Mahalanobis distance to OOD adversarial attacks, we used the Fast Gradient Sign Method of finding the adversary, with a learning rate of 3×10^{-4} (arbitrarily chosen), and ran it for 30 steps on 128 test set images of CIFAR-10 (the out-distribution). We ran the attack against both the Mahalanobis distance score as well as the Relative Mahalanobis distance score. For each of the images, we measured its score, its L_2 distance from the unperturbed image (out of $\sqrt{384 \times 384 \times 3} \approx 665$ for fully saturated pixels in the $[0, 1]$ range), and its L_∞ distance from the unperturbed image (out of 1). For both the L_2 and L_∞ perturbation strength norms, the Relative Mahalanobis distance is significantly more robust to OOD adversarial perturbations, retaining a higher AUROC on the near-OOD CIFAR-100 vs CIFAR-10 task at a given strength of perturbation. This is in line with the observation of higher stability of the relative distance method [15]. The baseline method of Maximum of Softmax Probabilities (in orange) performs the worst, losing AUROC the fastest with perturbation strength.

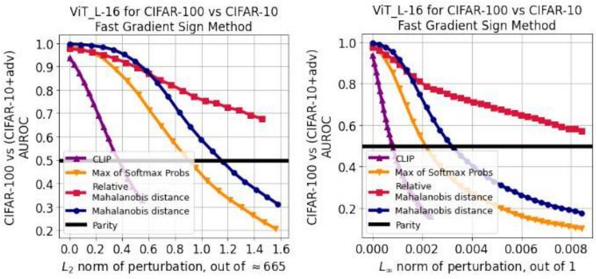
为了比较标准马哈拉诺比斯距离和相对马哈拉诺比斯距离对 OOD 对抗攻击的鲁棒性，我们使用了快速梯度符号法来寻找对手，学习率为 3×10^{-4} (任意选择)，并在 CIFAR-10(分布外)128 个测试集图像上运行了 30 步。我们对马哈拉诺比斯距离分数和相对马哈拉诺比斯距离分数进行了攻击。对于每个图像，我们测量了其分数、与未扰动图像的 L_2 距离 (在 $\sqrt{384 \times 384 \times 3} \approx 665$ 中完全饱和像素的范围内) 以及与未扰动图像的 L_∞ 距离 (范围为 1)。对于 L_2 和 L_∞ 扰动强度范数，相对马哈拉诺比斯距离对 OOD 对抗扰动的鲁棒性显著更强，在给定的扰动强度下，在近 OOD CIFAR-100 与 CIFAR-10 任务中保持了更高的 AUROC。这与相对距离方法的更高稳定性观察结果一致 [15]。基线方法最大软最大概率 (以橙色表示) 表现最差，随着扰动强度的增加，AUROC 损失最快。

3.2 Zero-shot OOD using CLIP

3.2 使用 CLIP 的零-shot OOD

We use the zero-shot OOD detection setup using the multimodal CLIP model described in Section 2.5 and introduced in [13]. In Figures 7a and 7b we show that its adversarial robustness is lower than for other methods, including the baseline Max of Softmax Probabilities (MSP). In Table 2 we show the underlying numbers in detail. Despite its versatility and power, CLIP does not perform very well

我们使用第 2.5 节中描述的多模态 CLIP 模型进行零-shot OOD 检测设置，并在 [13] 中介绍。在图 7a 和 7b 中，我们显示其对抗鲁棒性低于其他方法，包括基线最大软最大概率 (MSP)。在表 2 中，我们详细展示了基础数据。尽管 CLIP 具有多样性和强大功能，但其表现并不理想。



(a) (b)

Figure 7: AUROC of CIFAR-100 vs CIFAR-10 where the out-distribution CIFAR-10 is represented by 128 adversar-ially perturbed images to lower the Mahalanobis distance OOD score (blue) or Relative Mahalanobis distance score (red), and as a baseline the Maximum of Softmax Probabilities (orange). We add the CLIP zero-shot OOD detection as comparison in purple. (a) shows the perturbation strength measured by its L_2 norm, and (b) by its L_∞ norm. Details in Table 1 and Table 2.

图 7:CIFAR-100 与 CIFAR-10 的 AUROC，其中离散分布的 CIFAR-10 由 128 张对抗性扰动图像表示，以降低马哈拉诺比斯距离 OOD 分数 (蓝色) 或相对马哈拉诺比斯距离分数 (红色)，作为基线的最大软最大概率 (橙色)。我们将 CLIP 零样本 OOD 检测作为比较添加为紫色。(a) 显示了通过其 L_2 范数测量的扰动强度，(b) 则通过其 L_∞ 范数测量。详细信息见表 1 和表 2。

Table 2: The loss of AUROC on the near OOD CIFAR- 100 vs CIFAR-10 task for the CLIP zero-shot method using class names at L_∞ of 1/255 perturbation strength. CLIP is by far the least robust technique we studied in this paper.

表 2: 在近 OOD CIFAR-100 与 CIFAR-10 任务中，使用类名在 L_∞ 的 1/255 扰动强度下，CLIP 零样本方法的 AUROC 损失。CLIP 是我们在本文中研究的最不鲁棒的技术。

Post-process method	AUROC before	AUROC l_∞ 1/255	Δ AUROC
CLIP	94.68%	<10%	a lot

后处理方法	AUROC 之前	AUROC l_∞ 1/255	Δ AUROC
CLIP	94.68%	<10%	很多

when under a targeted adversarial attack to its OOD capabilities, underperforming even a simple post-processing baseline (albeit with very strong embeddings from a large, pretrained ViT).

在针对其 OOD 能力的有针对性的对抗攻击下，表现不佳，甚至不如一个简单的后处理基线 (尽管使用来自大型预训练 ViT 的非常强的嵌入)。

The change in the OOD score for 128 test set images from the out-distribution under an adversarial attack against the CLIP-based detector is shown in Figure 8a for the L_2 norm of the perturbation strength and in Figure 8 b for the L_∞ norm.

在对抗攻击下，针对基于 CLIP 的检测器，128 张测试集图像的 OOD 分数变化如图 8a 所示，针对扰动强度的 L_2 范数，以及图 8 b 中的 L_∞ 范数。

3.3 Model ensembles

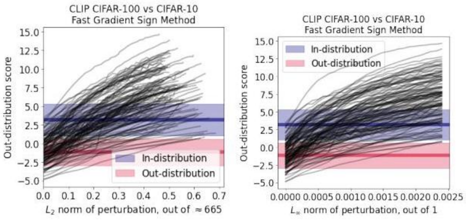
3.3 模型集成

We studied ensembles of OOD detectors, as discussed in Section 2.6. We used the standard setup using the Fast Gradient Sign Method (keeping only the sign of each element of the gradient), learning rate of 3×10^{-4} (arbitrarily chosen) and ran it for 30 steps on 128 test set images of CIFAR-10 (the out-distribution). We identified two well performing models finetuned on CIFAR-100 (training set), the ViT-L_{16} and R50+ViT- L_{32} , both with input resolution of $224 \times 224 \times 3$.

我们研究了 OOD 检测器的集成，如第 2.6 节所讨论的。我们使用了标准设置，采用快速梯度符号方法 (仅保留梯度每个元素的符号)，学习率为 3×10^{-4} (任意选择)，并在 CIFAR-10(外部分布) 的 128 个测试集图像上运行了 30 步。我们确定了两个在 CIFAR-100(训练集) 上微调的表现良好的模型，ViT- L_{16} 和 R50+ViT- L_{32} ，两者的输入分辨率均为 $224 \times 224 \times 3$ 。

We found that OOD model ensembling: 1) improves OOD

我们发现 OOD 模型集成:1) 改善了 OOD



(a) (b)

Figure 8: Changing the out-distribution score for a set of 128 CIFAR-10 test images (out-distribution) by applying the Fast Gradient Sign Method to the CLIP model. (a) shows the score as a function of the L_2 norm of the image perturbation, while (b) shows the L_∞ norm.

图 8: 通过将快速梯度符号法应用于 CLIP 模型，改变 128 个 CIFAR-10 测试图像 (分布外) 的输出分数。(a) 显示了图像扰动的 L_2 范数作为函数的分数，而 (b) 显示了 L_∞ 范数。

detection AUROC, 2) makes it more robust to adversarial attacks, and 3) its benefit combines well with the benefit of using the Relative Mahalanobis distance.

1) 检测 AUROC，2) 使其对对抗攻击更加稳健，以及 3) 其优势与使用相对马哈拉诺比斯距离的优势相结合良好。

We show the detailed results in Table 3 and in Figure 9 and Figure 10.

我们在表 3 以及图 9 和图 10 中展示了详细结果。

We look at the performance of two models individually, and perform adversarial attacks on their OOD score. We record the drop in AUROC for distinguishing the unperturbed CIFAR-100 from the adversarially perturbed CIFAR-10 at the perturbation level $\ell_\infty = 1/255$. We do the same for the ensemble of the two models. Ensembles suffer from a smaller drop in AUROC at a given perturbation level. Its benefit can be combined with the large robustness benefit of the Relative Mahalanobis distance. The AUROC as a func-

我们分别查看两个模型的性能，并对它们的 OOD 评分进行对抗攻击。我们记录了在扰动水平 $\ell_\infty = 1/255$ 下，区分未扰动的 CIFAR-100 与对抗扰动的 CIFAR-10 时 AUROC 的下降情况。我们对这两个模型的集成做了同样的分析。在给定的扰动水平下，集成模型的 AUROC 下降幅度较小。其优势可以与相对马哈拉诺比斯距离的大鲁棒性优势相结合。AUROC 作为一个函数-

Table 3: The benefit of OOD detector ensembling for adversarial robustness. The results are shown for the near OOD CIFAR-100 → CIFAR-10 task. We evaluate two separate models, and their ensemble, each for using the Ma-halanobis distance and the Relative Mahalanobis distance post-processing. Using an ensemble increases adversarial robustness, and can be combined to increase its benefit with the Relative Mahalanobis distance.

表 3:OOD 检测器集成对对抗鲁棒性的益处。结果显示了近 OOD CIFAR-100 → CIFAR-10 任务的情况。我们评估了两个独立模型及其集成，分别使用马哈拉诺比斯距离和相对马哈拉诺比斯距离后处理。使用集成可以提高对抗鲁棒性，并且可以与相对马哈拉诺比斯距离结合以增加其益处。

Model	Post- process method	AUROC before	AUROC ℓ_∞ 1/255	Δ AUROC
ViT L_{16}	Maha	97.72%	56.14%	-41.58%
R50+L32	Maha	96.95%	54.94%	-42.01%
Ensemble	Maha	97.91%	68.67%	-29.24%
ViT L_{16}	Relative	96.92%	69.82%	-27.10%
R50+L32	Relative	97.09%	68.53%	-28.56%
Ensemble	Relative	97.69%	78.64%	-9.05%

模型	后处理方法	AUROC 之前	AUROC ℓ_∞ 1/255	Δ AUROC
ViT L_{16}	玛哈	97.72%	56.14%	-41.58%
R50+L32	玛哈	96.95%	54.94%	-42.01%
集成	玛哈	97.91%	68.67%	-29.24%
ViT L_{16}	相对	96.92%	69.82%	-27.10%
R50+L32	相对	97.09%	68.53%	-28.56%
集成	相对	97.69%	78.64%	-9.05%

tion of the perturbation strength, both for the L_2 and L_∞ perturbation norms, is shown in Figure 9 and Figure 10. For all perturbation strengths measured by both norms, the en-

对于 L_2 和 L_∞ 扰动范数的扰动强度的影响，如图 9 和图 10 所示。对于通过这两种范数测量的所有扰动强度，集成模型的表现优于单个模型。

semble performs better than the individual models. We see

我们看到

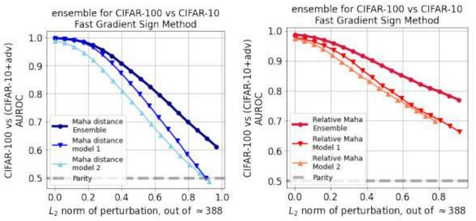


Figure 9: AUROC of CIFAR-100 vs CIFAR-10 where the out-distribution CIFAR-10 is represented by 128 adversar-ially perturbed images to lower the Mahalanobis distance OOD score (left panel, blue) and Relative Mahalanobis distnace score (right panel, red). We show the perturbation strength measured by its L_2 norm. The model ensemble (darker lines) is more robust to adversarial perturbations both for the standard and relative distance post-processing.

图 9:CIFAR-100 与 CIFAR-10 的 AUROC，其中离散分布 CIFAR-10 由 128 张对抗性扰动图像表示，以降低 Mahalanobis 距离 OOD 分数 (左侧面板，蓝色) 和相对 Mahalanobis 距离分数 (右侧面板，红色)。我们展示了通过其 L_2 范数测量的扰动强度。模型集成 (较深的线条) 在标准和相对距离后处理方面对对抗性扰动更具鲁棒性。

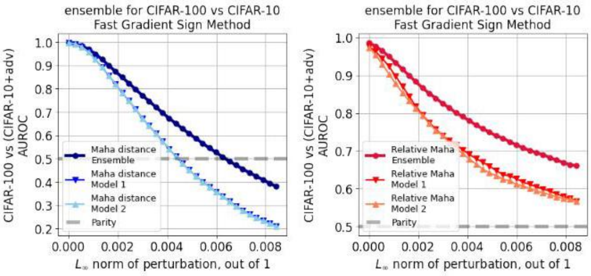


Figure 10: AUROC of CIFAR-100 vs CIFAR-10 where the out-distribution CIFAR-10 is represented by 128 adversar-ially perturbed images to lower the Mahalanobis distance OOD score (left panel, blue) and Relative Mahalanobis distnace score (right panel, red). We show the perturbation strength measured by its L_∞ norm. The model ensemble (darker lines) is more robust to adversarial perturbations both for the standard and relative distance post-processing.

图 10:CIFAR-100 与 CIFAR-10 的 AUROC，其中离散分布 CIFAR-10 由 128 张对抗性扰动图像表示，以降低 Mahalanobis 距离 OOD 分数 (左侧面板，蓝色) 和相对 Mahalanobis 距离分数 (右侧面板，红色)。我们展示了通过其 L_∞ 范数测量的扰动强度。模型集成 (较深的线条) 在标准和相对距离后处理方面对对抗性扰动更具鲁棒性。

a clear benefit of OOD detector ensembling both on the unperturbed AUROC as well as on the adversarial robustness of the resulting detector. This benefit combines well with the benefit of using the Relative Mahalanobis distance, suggesting that using both could be the correct strategy when deploying OOD detection systems.

OOD 检测器集成在未扰动的 AUROC 以及结果检测器的对抗鲁棒性方面的明显好处。这一好处与使用相对马哈拉诺比斯距离的好处相结合，表明在部署 OOD 检测系统时同时使用这两者可能是正确的策略。

3.4 The effect of image resolution

3.4 图像分辨率的影响

The input to the Vision Transformer is either 384×384 (or 224×224) while the resolution of both CIFAR-10 and CIFAR-100 is 32×32 . To resolve that, we upsample images to the correct resolution using the tf.image.resize function prior to feeding them into the network. This means that the image $\mathbf{x}$ coming in has the high resolution required, and that the gradient $\overrightarrow{g}(\mathbf{x}) = \partial\,\mathrm{score}(\mathbf{x})/\partial\mathbf{x}$ will be of the same resolution. This gives the attack many more

视觉变换器的输入为 384×384 (或 224×224)，而 CIFAR-10 和 CIFAR-100 的分辨率为 32×32 。为了解决这个问题，我们在将图像输入网络之前，使用 tf.image.resize 函数将图像上采样到正确的分辨率。这意味着输入的图像 $\mathbf{x}$ 具有所需的高分辨率，并且梯度 $\overrightarrow{g}(\mathbf{x}) = \partial\,\mathrm{score}(\mathbf{x})/\partial\mathbf{x}$ 将具有相同的分辨率。这使得攻击有更多的像素可以改变和潜在利用，从而可能导致更容易找到的对抗样本。

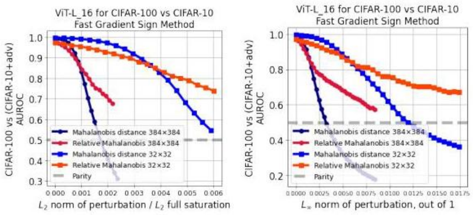


Figure 11: AUROC of CIFAR-100 vs CIFAR-10 where the out-distribution CIFAR-10 is represented by 32 adversar-ially perturbed images to lower the Mahalanobis distance OOD score (blue) and Relative Mahalanobis distance score (red). We show the perturbation strength measured by its L_2 norm (left panel) and L_∞ norm (right panel). The lighter lines show results for images and their gradients at the original CIFAR 32×32 resolution, while the darker lines show the 384×384 resolution. The lower resolution images are harder to adversarially perturb.

图 11:CIFAR-100 与 CIFAR-10 的 AUROC，其中分布外的 CIFAR-10 由 32 张对抗性扰动图像表示，以降低马哈拉诺比斯距离 OOD 分数 (蓝色) 和相对马哈拉诺比斯距离分数 (红色)。我们展示了通过其 L_2 范数 (左面板) 和 L_∞ 范数 (右面板) 测量的扰动强度。较浅的线条显示了在原始 CIFAR 32×32 分辨率下图像及其梯度的结果，而较深的线条显示了 384×384 分辨率。较低分辨率的图像更难以进行对抗扰动。

Table 4: The effect of image and gradient resolution on OOD robustness. Using differential image up-sampling, we show that working with lower resolution images provides adversarial robustness as compared to working with high resolution even for strong near-OOD detectors. pixels to change and potentially exploit, plausibly leading to an easier to find adversarial example.

表 4: 图像和梯度分辨率对 OOD 鲁棒性的影响。通过差分图像上采样，我们展示了与高分辨率图像相比，使用低分辨率图像提供了对抗鲁棒性，即使对于强近 OOD 检测器也是如此。

Resolution	Post- process method	AUROC before	AUROC l_∞ 1/255	Δ AUROC
32×32	Maha	97.98%	93.11%	-4.87%
384×384	Maha	97.98%	41.33%	-56.65%
32×32	Relative	97.11%	90.13%	-6.98%
384×384	Relative	97.11%	71.84%	-25.27%

分辨率	后处理方法	AUROC 前	AUROC l_∞ 1/255	Δ AUROC
32×32	玛哈	97.98%	93.11%	-4.87%
384×384	玛哈	97.98%	41.33%	-56.65%
32×32	相对	97.11%	90.13%	-6.98%
384×384	相对	97.11%	71.84%	-25.27%

To measure the difference between the adversarial robustness of low and high resolution images, we compared the attacks on the images upsampled prior to their use and gradient computation to working with the low res-olutuion images directly. For the latter case, we compute the image score as score(resize(x)) and its derivative as $\partial\,\mathrm{score}(\mathrm{resize}(\mathbf{x}))/\partial\mathbf{x}$, working directly with the small resolution image and modifying it using the small resolution gradient.

为了测量低分辨率和高分辨率图像的对抗鲁棒性之间的差异，我们比较了在使用图像之前进行上采样和梯度计算的攻击与直接处理低分辨率图像的攻击。在后者的情况下，我们将图像得分计算为 score(resize(x))，其导数为 $\partial\,\text{score}\,(\text{resize}\,(\mathbf{x}))\,/\partial\mathbf{x}$ ，直接处理小分辨率图像并使用小分辨率梯度对其进行修改。

The results for both the standard Mahalanobis distance and the Relative Mahalanobis distance, as well as the perturbation strength L_2 and L_∞ norms, are shown in Figure 11 and in Table 4. The lower resolution images are harder to perturb at a given perturbation strength, however, the benefit (or at least comparable performance at low strength) of the Relative Mahalanobis distance persists.

标准马哈拉诺比斯距离和相对马哈拉诺比斯距离的结果，以及扰动强度 L_2 和 L_∞ 的范数，如图 11 和表 4 所示。低分辨率图像在给定扰动强度下更难以扰动，然而，相对马哈拉诺比斯距离的好处 (或至少在低强度下的可比性能) 仍然存在。

3.5 Exploring far OOD CIFAR-100 vs SVHN

3.5 探索远离 OOD 的 CIFAR-100 与 SVHN

We studied the adversarial vulnerability on another, easier, far OOD task. In particular, we looked at the CIFAR-100 (in-distribution) vs SVHN (out-distribution) [24]. We show an example of the adversarial modification in Figure 12. The very large benefit of the Relative Mahalanobis distance for adversarial robustness of the OOD classification seen for near OOD tasks, such as in Figure 7a, Figure 7b and Table 1, is not prominent or does not exist at all for this far OOD task. The results are summarized in Table 5.

我们研究了另一个更简单的远离 OOD 任务的对抗脆弱性。特别是，我们观察了 CIFAR-100(在分布内) 与 SVHN(分布外)[24]。图 12 展示了对抗修改的一个例子。相对马哈拉诺比斯距离在近 OOD 任务 (如图 7a、图 7b 和表 1) 中对对抗鲁棒性的巨大好处，在这个远 OOD 任务中并不明显或根本不存在。结果总结在表 5 中。

Table 5: A comparison of OOD adversarial robustness of the Mahalanobis and Relative Mahalanobis distances for the far OOD CIFAR-100 vs SVHN.

表 5: 马哈拉诺比斯距离和相对马哈拉诺比斯距离在远离 OOD 的 CIFAR-100 与 SVHN 的对抗鲁棒性比较。

Post- process method	AUROC before	AUROC l_∞ 1/255	Δ AUROC
Maha	99.40%	34.47%	-64.93%
Relative	97.19%	43.22%	-53.97%

后处理方法	AUROC 之前	AUROC l_∞ 1/255	Δ AUROC
马哈	99.40%	34.47%	-64.93%
相对	97.19%	43.22%	-53.97%

The loss of AUROC from the unperturbed 99.40% as a function of the L_2 and L_∞ norm of the image perturbation are shown in Figure 13. At the $\ell = 1/255$ level

从未扰动的 99.40% 中损失的 AUROC 作为图像扰动的 L_2 和 L_∞ 范数的函数如图 13 所示。在 $\ell = 1/255$ 水平上

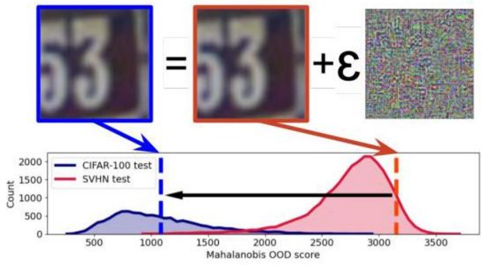


Figure 12: A small adversarial perturbation to the pixels of the out-distribution image (SVHN) changes its out-of-distribution score from $\approx 3,000$ to a confident in-distribution (CIFAR-100) region at $\approx 1,000$ even for a state-of-the-art near OOD detection method. $\varepsilon = 10^{-4}$ and the attack used is the Fast Gradient Sign Method applied to the Mahalanobis distance score for a ViT- L_{16} as used in [13]. The unperturbed CIFAR-100 → SVHN AUROC for this model is 99.40% .

图 12: 对分布外图像 (SVHN) 像素的小对抗扰动将其分布外得分从 ≈ 3,000 改变为在 ≈ 1,000 处的可信分布内 (CIFAR-100) 区域，即使对于一种最先进的近 OOD 检测方法也是如此。$\varepsilon = 10^{-4}$ 使用的攻击方法是快速梯度符号法，应用于 ViT- L_{16} 的马哈拉诺比斯距离得分，如 [13] 中所用。该模型的未扰动 CIFAR-100 → SVHN AUROC 为 99.40% 。

of L_∞ perturbation the AUROC is 34.47% . At the same level with the very same adversary-generation procedure, CIFAR-100 vs CIFAR-10 (near OOD) AUROC drops to 41.33% (see Table 1 for more details). It seems that, based on this example, there is a weak evidence that far OOD tasks might be more susceptible to adversarial attacks on the OOD score.

在 L_∞ 扰动下，AUROC 为 34.47% 。在同一水平上，采用完全相同的对手生成程序，CIFAR-100 与 CIFAR-10(近 OOD)AUROC 降至 41.33% (有关更多详细信息，请参见表 1)。根据这个例子，似乎有微弱的证据表明，远 OOD 任务可能更容易受到对抗攻击对 OOD 得分的影响。

4 Conclusion

4 结论

Even very powerful, near out-of-distribution detection methods based on large, pre-trained models, such as the Vision Transformer [13] and multi-modal text-image models, such as CLIP, suffer from severe adversarial vulnerability to their OOD detection score. Well-targeted, small modifications to the image pixels cause these detection systems to change their classification from confidently in-distribution to confidently out-distribution and vice versa. This might come as a surprise given the recent large improvements on near OOD tasks (such as distinguishing CIFAR-100 vs CIFAR-10) these models brought about. We show that orthogonally to their representational robustness that we can infer from their near-OOD performance, they still suffer from a severe adversarial vulnerability.

即使是基于大型预训练模型的非常强大的近分布外检测方法，例如 Vi-变换器 [13] 和多模态文本-图像模型，如 CLIP，遭受严重的对抗性脆弱性，影响其 OOD 检测得分。对图像像素进行精确的小幅修改会导致这些检测系统将其分类从自信的分布内更改为自信的分布外，反之亦然。这可能会让人感到惊讶，因为这些模型在近 OOD 任务 (例如区分 CIFAR-100 与 CIFAR-10) 上最近取得了显著的进展。我们表明，与我们可以从其近 OOD 性能推断出的表征鲁棒性正交，它们仍然遭受严重的对抗性脆弱性。

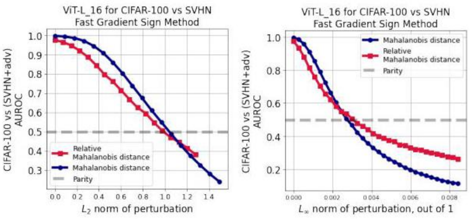


Figure 13: AUROC of CIFAR-100 vs SVHN where the out-distribution SVHN is represented by 128 adversarially perturbed images to lower the Mahalanobis distance (blue) and Relative Mahalanobis distance (red) OOD score. We show the perturbation strength measured by its L_2 norm (left panel) and L_∞ norm (right panel). The benefit of the Relative Mahalanobis distance to OOD adversarial robustness is not significant or not as prominted as for the near OOD CIFAR-100 vs CIFAR-10.

图 13:CIFAR-100 与 SVHN 的 AUROC，其中分布外的 SVHN 由 128 个对抗性扰动图像表示，以降低马哈拉诺比斯距离 (蓝色) 和相对马哈拉诺比斯距离 (红色)OOD 得分。我们展示了通过其 L_2 范数 (左面板) 和 L_∞ 范数 (右面板) 测量的扰动强度。相对马哈拉诺比斯距离对 OOD 对抗鲁棒性的好处并不显著，或不如近 OOD 的 CIFAR-100 与 CIFAR-10 之间的情况突出。

By studying the change in the OOD detectors' AUROC as a function of adversarial perturbation strength, we show that there are easy-to-use and generally applicable approaches to partial remedying this effect: ensembling and the Relative Mahalanobis ditance. The first approach is to ensemble several OOD detectors by averaging their predicted OOD score. The second approach is to use, instead of the standard Maximum of Softmax Probabilities or the more involved Mahalanobis distance post-processing technique, the newly proposed Relative Mahalanobis distance [15]. We also show that these approaches combine well together.

通过研究 OOD 检测器的 AUROC 随对抗扰动强度变化的情况，我们展示了有易于使用且普遍适用的方法来部分缓解这一效果: 集成和相对马哈拉诺比斯距离。第一个方法是通过平均它们预测的 OOD 得分来集成多个 OOD 检测器。第二个方法是使用新提出的相对马哈拉诺比斯距离 [15]，而不是标准的软最大概率或更复杂的马哈拉诺比斯距离后处理技术。我们还展示了这些方法能够很好地结合在一起。

We hope that by demonstrating this specific non-robustness of even the most powerful approaches to near OOD detection, more research will try to address them. We start off with proposing to use model ensembles and the Relative Mahalanobis distance where possible as an easy to use and cheap fix. However, stronger mitigation techniques will likely have to be employed to meet the frequent safety-critical nature of OOD detection.

我们希望通过展示即使是最强大的近 OOD 检测方法的这种特定非鲁棒性，能够促使更多的研究来解决这些问题。我们首先建议在可能的情况下使用模型集成和相对马哈距离作为一种易于使用且成本低廉的解决方案。然而，可能需要采用更强的缓解技术，以满足 OOD 检测的频繁安全关键性质。

Acknowledgements

致谢

We thank Jie Ren, Huiyi Hu, and Balaji Lakshminarayanan for useful comments and discussions.

我们感谢 Jie Ren、Huiyi Hu 和 Balaji Lakshminarayanan 提供的有用意见和讨论。

References

参考文献

[1] Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, and et al. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. Medical Image Analysis, 75:102274, Jan 2022.

[2] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection, 2019.

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565, 2016.

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

[5] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2016.

[6] Christopher M Bishop. Novelty Detection and Neural Network Validation. IEE Proceedings-Vision, Image and Signal processing, 141(4):217-222, 1994.

[7] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. arXiv preprint arXiv:1906.02994, 2019.

[8] Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In AISTATS, 2021.

[9] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. ECCV, 2020.

[10] Dan Hendrycks, Mantas Mazeika, and Thomas G Di-etterich. Deep anomaly detection with outlier exposure. ICLR, 2019.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR, 2021.

[12] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021.

[13] Stanislav Fort, Jie Ren, and Balaji Lakshmi-narayanan. Exploring the limits of out-of-distribution detection. arXiv preprint arXiv:2106.03004, 2021.

[14] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. NeurIPS, 2018.

[15] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshmi-narayanan. A simple fix to mahalanobis distance for improving near-ood detection, 2021.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sas-try, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[17] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2013.

[18] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Robust out-of-distribution detection for neural networks, 2020.

[19] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. Lecture Notes in Computer Science, page 430-445, 2021.

[20] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.

[21] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. arXiv preprint arXiv:1705.07057, 2017.

[22] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. JMLR, 2021.

[23] Stanislav Fort, Huiyi Hu, and Balaji Lakshmi-narayanan. Deep ensembles: A loss landscape perspective, 2019.

[24] Ian J. Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks, 2014.