深度神经网络中注意力模块的对抗鲁棒性分析

郭兴熠 [1]，唐晨欢 [2]，陈洁 [1]，贾世杰 [1,*]

[1] 大连交通大学自动化与电气工程学院，辽宁大连 116028
[2] 天津轨道交通集团有限公司，天津 300392

摘要: 注意力机制源自于对人类视觉的研究，通过将值得注意的部分提高权重，进而提高模型的准确率，现有的注意力有通道注意力，空间注意力以及混合模型。注意力在提高模型的准确率上的发展很充分，但在对抗攻击的影响下，注意力对模型的对抗鲁棒性的影响并未得到充分的研究。本文的目的是探究注意力对模型鲁棒性的影响。本文将常用的注意力 SENet、CBAM、ECANet、SGE、SKNet 等与常用的分类模型 ResNet 相结合，从头训练模型权重，对比结合模型的对抗鲁棒性，此外还通过与高斯、椒盐噪声的对比，体现注意力对普通噪声的影响。本文在前人的鲁棒性公式基础上，提出了对抗鲁棒性公式，对不同网络不同扰动参数的对抗样本进行统一测评。证明了注意力的结构对模型鲁棒性是有影响的，SKNet、SGE、CBAM 在 18 层模型中会提高模型的鲁棒性，随着模型深度的加深，注意力对模型鲁棒性有负向的提升，注意力对高斯、椒盐噪声有很好的鲁棒性，但也随网络的加深，体现出负向提升。

关键词: 深度神经网络; 注意力机制; 对抗样本; 对抗鲁棒性

DOI: 10.57237/j.se.2023.01.005

Adversarial Robustness Analysis of Attention Modules in Deep Neural Networks

Guo Xingyi[1], Tang Chenhuan[2], Chen Jie[1], Jia Shijie[1,*]

[1]School of Electric Engineering and Automation, Dalian Jiaotong University, Dalian 116028, China
[2]Tianjin Rail Transit Group Co., Ltd, Tianjin 300392, China

Abstract: The attention mechanism is derived from the study of human vision, which improves the accuracy of the model by increasing the weight of the noteworthy parts, and the existing attention is channel attention, spatial attention, and hybrid models. The development of attention in improving the accuracy of the model is sufficient, but under the influence of adversarial attacks, the influence of attention on the adversarial robustness of the model has not been fully studied. The purpose of this paper is to explore the effect of attention on model robustness. In this paper, the commonly used attention SENet, CBAM, ECAnet, SGE, SKNet, etc. are combined with the commonly used classification model ResNet to train the weights of the model from scratch, and compare the robustness of the combined model against confrontation. In addition, the influence of attention on ordinary noise is also reflected by comparison with Gaussian and salt and pepper noise. Based on the robustness formula of the predecessors, this paper proposes an adversarial robustness

*通信作者：贾世杰, jsj@djtu.edu.cn

收稿日期: 2023-01-05; 接受日期: 2023-02-16; 在线出版日期: 2023-03-02

formula to uniformly evaluate the adversarial samples with different disturbance parameters of different networks. It is proved that the structure of attention has an impact on the robustness of the model, and SKNet, SGE, and CBAM will improve the robustness of the model in the 18-layer model, with the deepening of the model, attention has a negative improvement on the robustness of the model. Attention has a good robustness to Gaussian and salt and pepper noise, but also reflects a negative improvement with the deepening of the network.

Keywords: Deep Neural Networks; Attention Mechanisms; Adversarial Samples; Fight Robustness

1 引言

深度神经网络现已被广泛应用到图像识别、目标检测等各类任务中[1]，但是在现实场景中，会因为雨雪雾等天气干扰导致模型出现难以预料的错误，这将直接影响模型在应用中的可靠性[2]。目前为止有很多人对如何提高模型的鲁棒性做了大量的工作，Denoising AutoEncoder通过在传统的AutoEncoder的输入层加入随机噪声来增强模型鲁棒性[3]；随机混合任意两个样本的 mixup 插值策略来增强对抗样本的鲁棒性[4]；此外，数据增强[5]、对抗训练[6]、中间层正则等方法都可以增强模型的鲁棒性。

鲁棒性的研究分为噪声鲁棒性和对抗鲁棒性，噪声鲁棒性包含了雨雪雾天气下的物体，加入高斯噪声、椒盐噪声的样本等一系列可见干扰，这些图像在ImageNet-C 数据集中均有体现[7]，会对模型的安全性产生一些影响。对抗鲁棒性是针对一种人为的，不为人眼所察觉的却能使网络模型判断出错的对抗样本所提出的，这种生成对抗样本的攻击算法种类繁多。2015年提出的 FGSM 是在图像的基础上添加基于预测模型梯度的噪声[8]，这种噪声能够保证对人眼的较小的可见范围，并且能够在一定程度上破坏模型对图像的预测结果，将原本置信度为 57.7%的大熊猫图像变为 99.3%置信度的长臂猿。2016 年出现了一种基于超平面分类的攻击方法 DeepFool [9]，该方法通过找到样本到决策边界的最短距离进而找到能是样本分类错误的最小噪声。基于优化的 CW 攻击算法，在使用二分法的基础上不断查找最适合的噪声扰动[10]。2018 年的 PGD 改进了 FGSM，这种算法在 FGSM 的基础上添加了迭代[11]，使得生成的扰动噪声在更小的范围内，并且对图像的攻击成功率也更高，这就使得在人眼中并无两样的图像放到神经网络中能得到截然不同的结果，这些对抗样本甚至可以应用在现实中做成对抗 T 恤干扰检测器[12]，对神经网络模型的安全构成了很大的威胁。

注意力模块最开始源自于对人类的视觉的研究，它模拟了人类视野中选择性的注意力机制[13]。在图像中，视觉对于不同物体分配的注意力是不一样的，人们总是会更多的关注容易引起视觉反应的物体，于是有一部分人将这一现象应用到视觉网络模型中，通过注意力机制使模型对特定的地方更感兴趣，增加这一部分的权重，抑制不重要的信息[14-15]。近年来，注意力机制逐渐成为深度学习中研究的热门之一[16-19]。

大量实验已经证明注意力模块对模型的准确率是有提升的，但目前为止注意力模块都是针对提高模型准确率而提出的，我们认为鲁棒性也是十分重要的一项指标，既然可以通过注意力模块去提高模型的准确率，那么是否能通过注意力模块来提高模型的鲁棒性，此处的鲁棒性主要是针对人为生成的对抗样本而言，其中也会带有一些噪声样本的对比与分析。现如今大部分注意力提供的权重由于其架构、使用的系统、库、训练策略等有很多不同，所以无法直接将他们进行一个直接的比较，我们通过一个统一的训练策略和训练方法针对现有的注意力模块搭建网络，将未加入注意力机制的模型和加入注意力机制的模型，在 cifar10 数据集上训练自己的权重并进行针对对抗样本鲁棒性的测试，使用我们提出的对抗鲁棒性的指标评定注意力对模型的影响，所以本文的目的在于通过实验探究现有的注意力机制是否会对模型的对抗样本的鲁棒性产生影响。结果证明加入 SKNet、CBAM、SGE 注意力在 18 层的模型中能够提升模型鲁棒性。

本文的主要贡献如下：

(1) 提出了一个基于对抗样本鲁棒性的评价标准，将不同扰动强度的对抗样本进行统一的指标比较；

(2) 使用了一整套完整的流程去测试常用注意力模型的对抗鲁棒性及普通鲁棒性，分析结果，验证注意力对深度神经网路的对抗鲁棒性影响。

2 相关工作

2.1 注意力机制

Mnih 等人最先提出了注意力机制的概念[20]，通过修改输入数据的权重，来提升对应部分对全局的影响。注意力模块本身独立于模型之外，其本身的目的是在不改变模型的整体结构下，作为一个外接模块，去提升模型的准确率，现如今的注意力基本是在以提升正常样本的准确率为前提进行实验和分析的，本文使用的注意力按照处理的维度可以分为通道注意力、空间注意力和混合注意力，本小节简单介绍了文中进行实验的注意力机制及原理。

SENet 注意力是胡杰团队于 2017 年提出的通过通道提升模型性能的注意力机制，其在 ImageNet 数据集上将 top-5 错误率从 2.991%降低到 2.251%[21]。通过 Sequeeze（压缩）和 Excitation（激发）两个操作将输入的特征图压缩成拥有全局感受野的实数，生成每个通道的特定权重，最后加权到原特征图上。

CBAM 提出于 2018 年，是一种将通道注意力与空间注意力相结合的模块[22]，相比于只关注通道的 SENet 拥有更好的效果。该方法先通过通道注意力模块 CAM，然后通过空间注意力模块 SAM，分别进行通道和空间的压缩、生成权重、缩放加权到原特征图上。

ECANet 是在 SENet 模型上改进得来的[23]，在 SENet 的基础上大幅降低了参数量，并且保持了准确率，文中证明了降维会损害通道间的交流，所以作者通过一维卷积替代了 SENet 中的两个 Fc 层。

SKNet 提出了对不同输入使用不同的卷积核[24]，因为不同感受野对于不同尺度的目标会有不同的效果，将这一思路与 SENet 注意力结合，可以自适应的对输入进行处理。

SGE 将输入的特征图分成不同组，利用局部特征和全局特征生成注意力掩码[25]，通过全局平均池化、标准化、缩放、激活函数等方法获得最终的输出特征图。

2.2 对抗攻击

目前兼具攻击成功率和攻击速度的攻击方法是 PGD 算法，该方法在是在 FGSM 算法上改进而来的。

FGSM 是一种基于梯度生成对抗样本的算法，通过反向求导梯度，并将之生成噪声叠加到原始图像上，起到攻击的效果。作者通过多次实验发现沿着梯度的方向给样本添加噪声能够提高攻击效果，并且能做到噪声的值较小，该算法公式如下所示：

$$x' = x + \varepsilon \operatorname{sgn}(\nabla_x L(\theta, x, y)) \tag{1}$$

x' 是生成之后的对抗样本，x 是原始样本，ε 是控制扰动范围的超参数，$\nabla_x L(\theta, x, y)$ 是关于 x 的梯度，为了避免生成的噪声过大，超出数字图像的像素值，使用约束函数对其进行限制。

PGD 在 FGSM 的基础上添加了迭代梯度上升，通过小步多次梯度上升得到更符合干扰方向的噪声，所以它的攻击效果要更好，其公式如下所示：

$$x^{t+1} = \Pi_{x+s}(x^t + \alpha \operatorname{sgn}(\nabla_x L(\theta, x, y))) \tag{2}$$

该公式在公式(1)的基础上，将 x' 和 x 变成了不断迭代的 x^{t+1} 和 x，并且将之前的噪声通过累乘进行积累。目前 PGD 是一阶攻击中性能最好的算法。

目前在各类防御文章中 PGD 都作为主要的攻击算法进行测试，所以本文采用的也是 PGD 攻击算法生成对抗样本。

3 设计方案

3.1 数据集的选择

本文选择的是 cifar10 数据集。在现有的针对分类网络的注意力文献中，比如提出 SENet 注意力机制的文章中，使用的数据集有 cifar10、cifar100、ImageNet1000、Places365、COCO 等。其中在分类网络中常用的就是 cifar10、cifar100、ImageNet1000。为了提高模型训练的效率，cifar10 中的图像是 32*32 大小能大幅缩短训练的时间，适合测试大量不同的网络的差异。其次其包含的种类并不单一，包含了卡车、轮船、马、蛙类、狗、鹿、猫、鸟、汽车和飞机，在测试的时候具有一定的普遍性。虽然和大型数据集相比 50000 张训练图像和 10000 张测试图像略显不足，但我们在训练模型的时候会通过数据增强的方法弥补其中的差距，训练时图像随机镜像水平翻转，周围进行像素填充，然后随机裁剪成 32*32 大小的图像，再进行标准化归一化，这些策略同样在 SENet 注意力模型数据集处理时也有同样的使用。

3.2 模型的构建

本文模型选择了 ResNet[26]系列网络作为测试的

骨干网络，所有的注意力机制都是在 ResNet 的基础上添加的。ResNet 因为其独特的残差结构，应用场景十分广泛，SENet、CBAM、ECANet 等一系列注意力文章中都是以 ResNet50 网络为基础，然后不断加深模型的深度以测试注意力对分类网络的性能提升，所以本文也采用 ResNet 系列网络，由于使用的数据集是 32*32 的图像，所以最后选择了 ResNet18、ResNet32、ResNet50 三个网络来作为骨干网络，这样既能测试出不同注意力对模型鲁棒性的影响，又能测试出相同注意力在不同深度下的模型鲁棒性的影响。

本文选取的注意力有 SENet、SKNet、ECANet、SGE、CBAM 等五种基础注意力结构。注意力加入的位置参考的 SENet 一文，文献中曾对多个位置进行过测试，目的是找出能够最大限度提高模型准确率的位置，本文仅针对这一已知最大限度提升模型准确率的位置进行其鲁棒性的实验，其位置如图 1 所示。

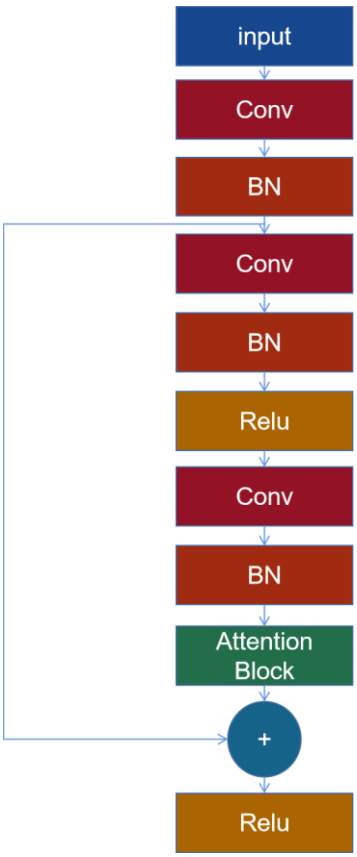

图 1 注意力加入位置

3.3 训练策略

本文采用从头训练的方法，不使用预训练权重，因为我们进行对比的注意力来源较多，并不是所有的注意力都在 cifar10 数据集上有已经训练好的模型，所以为了使结果更有对比性，本文未采用任何模型的预训练权重。我们将注意力与模型结合，通过正常样本训练出一个收敛的神经网络权重，训练策略采用 Squeeze and Excitation Networks 中的训练策略。当前模型权重训练好后，通过加深模型或者添加不同注意力以获得其他权重。loss 值在 300 代以后趋向平稳，权重从所有 epoch 的后 30%中选择，使用测试集准确度较高且训练集的 loss 值较低的代数进行测试。

3.4 对抗样本的生成

PGD 作为一阶攻击算法里效果最好的算法之一，大量的防御和鲁棒性测试的文献都选用了该攻击方法生成对抗样本，所以我们也选择该算法作为主要的攻击手段。PGD 算法中会涉及到一些超参数，这些都影响着最终的对抗样本的生成，我们采用了 2/255、4/255 和 8/255 三种扰动超参数进行实验，以体现不同扰动下注意力对模型鲁棒性的影响。为了体现对抗样本和噪声样本的区别，我们还通过添加高斯、椒盐噪声的方法生成噪声样本，去体现注意力对对抗样本和噪声样本不同的影响。

3.5 评价标准

Dan Hendrycksh 和 Thomas G.Dietterich 在他们的文章中针对模型的腐败鲁棒性提出了一个不同扰动的统一测试标准[27]。我们在此基础上，针对 PGD 算法的不同扰动，提出了对抗样本的对抗鲁棒性评价标准。

PGD 攻击算法是针对梯度的迭代攻击算法，在攻击之前会对噪声有一个随机初始化，为了降低该初始化的影响，通过多次重复攻击的结果进行均值处理，得到 top-1 的平均错误率。原网络对抗样本平均 top-1 错误率如公式(3)，n 本文取 3：

$$ME_{Adv}^{Network}=\frac{1}{n}\sum_{i=1}^{n}E_{Adv}^{Network} \tag{3}$$

注意力网络对抗样本平均 top-1 错误率如公式(4)：

$$ME_{Adv}^{AttNetwork}=\frac{1}{n}\sum_{i=1}^{n}E_{Adv}^{AttNetwork} \tag{4}$$

上面两个式子中 $E_{Adv}^{Network}$ 代表了对抗样本在原始网络中的错误率，$E_{Adv}^{AttNetwork}$ 代表了对抗样本在注意力网络中的错误率，二者的区别仅在于是否加入了注意力模

块。

对于 PGD 攻击算法有不同程度的扰动，生成的对抗样本的扰动分别定义为 8/255、4/255、2/255 共三种，如果要将不同注意力模型进行统一的衡量，需要有一个统一的对比模型，所以这里加入了未加入注意力模块的原始模型，二者的比值能够很好的反映出加入注意力后二者的相关性，能够将不同注意力模型放到同一水平线进行比较，其公式如下：

$$NAE_{Adv}^{AttNetwork}=\sum_{s=1}^{c}ME_{s,Adv}^{AttNetwork}/\sum_{s=1}^{c}ME_{s,Adv}^{Network} \tag{5}$$

本文 c=3，式中 s 的取值代表了不同的扰动，当 s=0 时，PGD 的扰动 $\varepsilon=2/255$，s=1 时 $\varepsilon=4/255$，s=2 时 $\varepsilon=8/255$。

为了展现网络攻击前和攻击后的变化，仅仅有对抗样本的 top-1 错误率是不够的，还需要加入和干净样本的 top-1 错误率的差距，所以相对对抗鲁棒性公式如下：

$$RNAE_{Adv}^{AttNetwork}=\sum_{s=1}^{c}(ME_{s,Adv}^{AttNetwork}-E_{Clean}^{AttNetwork})/\sum_{s=1}^{c}(ME_{s,Adv}^{Network}-E_{Clean}^{Network}) \tag{6}$$

上述公式(6)是相同攻击算法下不同扰动的对抗鲁棒性评价标准。$E_{Clean}^{AttNetwork}$ 代表了正常样本在注意力网络的 top-1 错误率，$E_{Clean}^{Network}$ 代表了正常样本在原始网络中的 top-1 错误率，如果注意力模型能对对抗样本拥有较好的抗性，则 RNAE 指标的值会更小。

4 实验结果及分析

4.1 实验参数

我们的所有的网络都是在 Windows 系统下的 pytorch 库为基础搭建的，使用的 GPU 是英伟达 2080Ti，CPU 是 Intel i7，RAM 为 32G，训练策略中的损失函数使用的是交叉熵损失函数，优化器使用的梯度下降，学习率为 0.1，且随迭代次数的增加不断减小，衰减参数为 10^{-4}，迭代次数为 1000 次，对于所有的数据集我们采用的是中心裁剪和镜像反转的方法，并进行归一化和标准化，上述的数据集及训练策略均来自于 SENet 文献中的 cifar 数据集的训练策略。

我们攻击采用的是一阶攻击算法中效果最好的 PGD 算法，扰动大小设置为 2/255，4/255, 8/255，迭代次数为 40，迭代步长为 $2*10^{-4}$，$4*10^{-4}$，$8*10^{-4}$，迭代步长的取值源于扰动最大值对迭代次数的均分结果。

4.2 实验数据

表 1 是 cifar10 数据集在 ResNet18 网络中充分训练得到的权重的测试结果，一共使用了六个网络，包含一个骨干网络，五个注意力网络，测试一共分为四部分，第一部分为无噪声的测试集，第二部分为 PGD 对抗样本测试集，ε 代表了三种不同强度的扰动，第三部分为高斯噪声测试集，sigma 代表了不同强度的高斯噪声，第四部分为椒盐噪声测试集，amount 代表了添加椒盐噪声的数量。

表 1 ResNet18 注意力模型对抗鲁棒性评估

		ResNet	**SENet**	**SKNet**	**ECANet**	**CBAM**	**SGE**
无噪声	测试集	88.77%	88.86%	89.77%	89.52%	89.51%	89.12%
PGD	ε=2/255	33.63%	28.59%	35.87%	34.79%	32.36%	36.47%
	ε=4/255	24.13%	20.19%	26.58%	24.71%	25.11%	26.16%
	ε=8/255	12.10%	10.01%	14.27%	11.91%	14.94%	12.67%
	RNAE	1.000	1.058	0.998	1.004	0.998	0.978
Gauss	sigma=5	81.10%	85.53%	86.62%	84.29%	84.23%	83.23%
	sigma=15	41.65%	55.06%	58.25%	47.81%	54.02%	51.15%
	sigma=25	18.30%	29.40%	28.20%	24.98%	25.04%	27.48%
	RNAE	1.000	0.771	0.768	0.890	0.840	0.842
Pepper＆Salt	amount=0.02	55.76%	61.77%	62.48%	59.55%	63.42%	60.19%
	amount=0.04	34.88%	44.44%	40.94%	36.96%	43.31%	40.56%
	amount=0.06	21.15%	30.97%	26.18%	22.72%	27.36%	28.31%
	RNAE	1.000	0.837	0.904	0.966	0.870	0.895

在对抗样本中效果较好的注意力是 SGE、SKNet 和 CBAM。我们可以看到在加入注意力之后，原始样本的准确率都有轻微的提升，但在对抗样本的准确率中，并不是所有的模型的鲁棒性都有提升。SENet 的对抗样本准确率在各种强度扰动下，都有下降，CBAM 模型的对抗样本准确率在扰动较小的时候略低于

ResNet，但当扰动不断增大后，其对抗样本准确率实现反超，ECANet 网络与 CBAM 相反，在对抗噪声较小的时候其对抗样本准确率高于 ResNet，当噪声增大后其效果在不断降低。SKNet 和 SGE 攻击后的准确率都高于原始网络，效果较好。

在高斯和椒盐噪声样本中，所有的注意力模型其准确率都优于原始模型。

我们提出的标准 RNAE 是在综合所有强度下的评价鲁棒性的指标，数值越小证明该模型对抗鲁棒性越好，SGE、SKNet、CBAM 证明了在 ResNet18 的基础模型上，其准确率和对抗鲁棒性都有所提高，SENet 效果较差，ECANet 虽然在低强度扰动下鲁棒性微弱提升，但在所有强度下综合来看，整体的 RNAE 指标是下降的。高斯噪声样本中 RNAE 值最小的是 SKNet 注意力网络，说明该网络在高斯噪声下拥有最好的鲁棒性。椒盐噪声下鲁棒性最好的是 SENet 注意力网络。

表 2 ResNet34 注意力模型对抗鲁棒性评估

		ResNet	**SENet**	**SKNet**	**ECANet**	**CBAM**	**SGE**
无噪声	测试集	89.39%	90.39%	89.86%	90.36%	90.33%	90.30%
PGD	ε=2/255	35.38%	33.46%	34.57%	36.51%	34.22%	33.37%
	ε=4/255	25.50%	24.11%	24.45%	26.26%	24.47%	24.52%
	ε=8/255	13.05%	12.22%	12.25%	12.61%	12.29%	12.91%
	RNAE	1.000	1.037	1.021	1.008	1.029	1.030
Gauss	sigma=5	84.60%	86.29%	84.80%	85.48%	84.21%	86.19%
	sigma=15	47.61%	50.30%	57.50%	57.43%	58.19%	58.92%
	sigma=25	24.17%	24.56%	39.03%	31.00%	27.62%	32.50%
	RNAE	1.000	0.984	0.789	0.869	0.903	0.835
Pepper＆Salt	amount=0.02	53.30%	63.04%	61.83%	68.23%	64.57%	68.61%
	amount=0.04	35.05%	44.17%	46.40%	52.78%	42.51%	51.82%
	amount=0.06	25.84%	28.04%	37.48%	36.66%	23.28%	36.76%
	RNAE	1.000	0.883	0.804	0.737	0.913	0.738

在该深度下，所有对抗样本下的注意力模型效果都不如原网络。加入注意力之后，对于测试集模型的准确率依旧有所上升，SENet、SKNet、CBAM、SGE 网络的对抗样本准确率都不如不加注意力的网络，ECANet 的表现规律与 ResNet18 网络中的现象相同，在低对抗噪声的条件下，其对抗样本准确率优于骨干网络，但随着对抗噪声的增大，其对抗样本准确率逐渐不如骨干网络。

在高斯样本和椒盐样本下，所有网络的鲁棒性依旧有所提高。

在 RNAE 指标下，对于对抗样本的所有注意力网络效果都不好，ECANet 虽然在扰动较小时有优势，但在全扰动下的的鲁棒性是不如不加入注意力机制的网络的。SKNet 依旧在高斯噪声样本中拥有最好的鲁棒性，在椒盐噪声样本下，鲁棒性最好的是 ECANet 和 SGE，相差不大。

表 3 ResNet50 注意力模型对抗鲁棒性评估

		ResNet	**SENet**	**SKNet**	**ECANet**	**CBAM**	**SGE**
无噪声	测试集	90.00%	90.92%	90.04%	91.02%	90.05%	90.41%
PGD	ε=2/255	38.27%	34.90%	42.00%	35.73%	30.89%	33.56%
	ε=4/255	27.78%	25.62%	31.07%	25.53%	20.06%	22.47%
	ε=8/255	14.46%	13.73%	15.73%	11.95%	9.22%	10.20%
	RNAE	1.000	1.048	0.957	1.055	1.108	1.082
Gauss	sigma=5	84.00%	87.38%	86.75%	85.86%	84.09%	84.22%
	sigma=15	51.20%	57.56%	63.60%	56.89%	56.00%	38.85%
	sigma=25	24.62%	29.31%	39.19%	28.17%	30.07%	16.66%
	RNAE	1.000	0.894	0.731	0.927	0.908	1.194
Pepper＆Salt	amount=0.02	64.33%	66.78%	69.70%	65.08%	57.80%	59.23%
	amount=0.04	43.65%	48.28%	53.24%	45.42%	43.42%	35.89%
	amount=0.06	27.17	32.61%	37.95%	27.44%	29.35%	21.67%
	RNAE	1.000	0.928	0.810	1.002	1.035	1.145

所有注意力网络的测试集准确率都高于原网络，但在对抗样本测试集中只有 SKNet 的准确率提升了，

其余的注意力网络的准确率都在下降，SKNet 在所有强度扰动下的准确率都高于原始网络。

高斯噪声样本中，SGE 的强扰动下的准确率下降幅度很大，其余注意力模型准确率都在上升。椒盐噪声中 ECANet、CBAM、SGE 三个注意力网络在不同扰动下准确率都有不同程度的下降，只有 SENet 和 SKNet 的准确率有提高。

在 RNAE 指标下，仅有 SKNet 网络在对抗样本中的鲁棒性有了正向提升。高斯噪声和椒盐噪声样本下鲁棒性最好的也是 SKNet。

4.3 实验结果总结

在 ResNet18 网络下，有三个注意力网络能够提升模型的鲁棒性，但当网络层数的加深，除 SKNet 外的注意力网络都对模型的鲁棒性有负向的提升。在高斯噪声和椒盐噪声样本下，所有注意力模型都对模型的鲁棒性有正向提升，但当网络加深到 ResNet50 时，ECANet、CBAM、SGE 都对鲁棒性有了负向提升。综上所述，SKNet 的效果是五种注意力中最好的，我们猜测可能跟其对不同输入使用不同的卷积核相关。

我们改进的 RNAE 指标，能够将不同的模型，不同的扰动强度下的鲁棒性进行一个综合的评判，不用通过准确率的逐行对比，更容易体现出一个模型的鲁棒性是否有提升。

5 结论

针对测试深度学习模型的鲁棒性，我们以注意力模型为例，提出了一整套实验的流程和测试的评价标准，为测试不同的模型的鲁棒性提供思路。

本文通过大量实验证明了对抗攻击算法 PGD，对五种注意力模型的鲁棒性影响，针对现有的大量注意力网络和不同的攻击算法进行鲁棒性测试可以作为接下来工作的一个方向。

参考文献

[1] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification [C] // Proceedings of the IEEE international conference on computer vision. 2015: 1026-1034.

[2] Wang S, Pei K, Whitehouse J, et al. Efficient formal safety analysis of neural networks [J]. Advances in Neural Information Processing Systems, 2018, 31.

[3] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders [C] // Proceedings of the 25th international conference on Machine learning. 2008: 1096-1103.

[4] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization [J]. arXiv preprint arXiv: 1710.09412, 2017.

[5] Wang H, Huang Z, Wu X, et al. Toward Learning Robust and Invariant Representations with Alignment Regularization and Data Augmentation [J]. arXiv preprint arXiv: 2206.01909, 2022.

[6] Liu K, Liu X, Yang A, et al. A robust adversarial training approach to machine reading comprehension [C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34 (05): 8392-8400.

[7] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database [C] // 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[8] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [J]. arXiv preprint arXiv: 1412.6572, 2014.

[9] Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2574-2582.

[10] Carlini N, Wagner D. Towards evaluating the robustness of neural networks [C] // 2017 ieee symposium on security and privacy (sp). Ieee, 2017: 39-57.

[11] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks [J]. arXiv preprint arXiv: 1706.06083, 2017.

[12] Xu K, Zhang G, Liu S, et al. Adversarial t-shirt! evading person detectors in a physical world [C] // European conference on computer vision. Springer, Cham, 2020: 665-681.

[13] Chen J, Zhang H, He X, et al. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention [C] // Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval. 2017: 335-344.

[14] Jhamb Y, Ebesu T, Fang Y. Attentive contextual denoising autoencoder for recommendation [C] // Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval. 2018: 27-34.

[15] Tay Y, Luu A T, Hui S C. Multi-pointer co-attention networks for recommendation [C] // Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018: 2309-2318.

[16] 朱张莉, 饶元, 吴渊, 等. 注意力机制在深度学习中的研究进展 [J]. 中文信息学报, 2019, 33 (6): 1-11.

[17] Zhang S, Yao L, Sun A, et al. Deep learning based recommender system: A survey and new perspectives [J]. ACM Computing Surveys (CSUR), 2019, 52 (1): 1-38.

[18] 黄立威, 江碧涛, 吕守业, 等. 基于深度学习的推荐系统研究综述 [J]. 计算机学报, 2018, 41 (7): 1619-1647.

[19] 方钧婷, 谭晓阳. 注意力级联网络的金属表面缺陷检测算法 [J]. 计算机科学与探索, 2021, 15 (7): 1245.

[20] Mnih V, Heess N, Graves A. Recurrent models of visual attention [J]. Advances in neural information processing systems, 2014, 27.

[21] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks. arXiv e-prints [J]. arXiv preprint arXiv: 1709.01507, 2017.

[22] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module [C] // Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[23] Wang Q, Wu B, Zhu P, et al. Supplementary material for 'ECA-Net: Efficient channel attention for deep convolutional neural networks [C] // Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Seattle, WA, USA. 2020: 13-19.

[24] Wu W, Zhang Y, Wang D, et al. SK-Net: Deep learning on point cloud via end-to-end discovery of spatial keypoints [C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34 (04): 6422-6429.

[25] Li X, Hu X, Yang J. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks [J]. arXiv preprint arXiv:1905.09646, 2019.

[26] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[27] Hendrycks D, Dietterich T G. Benchmarking neural network robustness to common corruptions and surface variations [J]. arXiv preprint arXiv: 1807.01697, 2018.