计算机科学
COMPUTER SCIENCE

基于SVD的深度学习模型对抗鲁棒性研究

赵子天, 詹文翰, 段翰聪, 吴跃

引用本文

赵子天, 詹文翰, 段翰聪, 吴跃. 基于SVD的深度学习模型对抗鲁棒性研究[J]. 计算机科学, 2023, 50(10): 362-368.

ZHAO Zitian, ZHAN Wenhan, DUAN Hancong, WU Yue. Study on Adversarial Robustness of Deep Learning Models Based on SVD [J]. Computer Science, 2023, 50(10): 362-368.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220800090

基于SVD的深度学习模型对抗鲁棒性研究

赵子天　詹文翰　段翰聪　吴　跃
电子科技大学计算机科学与工程学院　成都 611731
(zitianzhao_uestc@hotmail.com)

摘　要　对抗攻击的出现对于深度神经网络(DNN)在现实场景中的大规模部署产生了巨大的威胁,尤其是在与安全相关的领域。目前已有的大多数防御方法都基于启发式假设,缺少对模型对抗鲁棒性的分析。如何提升DNN的对抗鲁棒性,并提升鲁棒性的可解释性和可信度,成为人工智能安全领域的重要一环。文中提出从奇异值分布的角度分析模型的对抗鲁棒性。研究发现,模型在对抗性环境下鲁棒性的提升伴随着更加平滑的奇异值分布。通过进一步分析表明,平滑的奇异值分布意味着模型的分类置信度来源更加多样,从而也具有更高的对抗鲁棒性。基于此分析,进一步提出了基于奇异值抑制SVS(Singular Value Suppress)的对抗训练方法。实验结果表明,该方法进一步提高了模型在对抗性环境下的鲁棒性,在面对强力白盒攻击方法PGD(Project Gradient Descent)时,在CIFAR10和SVHN数据集上分别能达到55.3%和54.51%的精度,超过了目前最具有代表性的对抗训练方法。

关键词:深度学习;对抗防御;对抗训练;对抗鲁棒性;奇异值分解
中图法分类号　TP391

Study on Adversarial Robustness of Deep Learning Models Based on SVD

ZHAO Zitian,ZHAN Wenhan,DUAN Hancong and WU Yue
School of Computer Science and Engineering,University of Electronic Science and Technology of China,Chengdu 611731,China

Abstract　The emergence of adversarial attacks poses a substantial threat to the large-scale deployment of deep neural networks (DNNs) in real-world scenarios,especially in security-related domains. Most of the current defense methods are based on heuristic assumptions and lack analysis of model robustness. How to improve the robustness of DNN and improve the interpretability and credibility of the robustness has become an essential part of the field of artificial intelligence security. This paper proposes to analyze the robustness of the model from the perspective of singular values. In the adversarial environment,the improvement of model robustness is accompanied by a smoother distribution of singular values. Further analysis shows that the smooth distribution of singular values means that the model has more diverse classification confidence sources and thus has higher adversarial robustness. Based on the analysis,an adversarial training algorithm based on singular value suppress(SVS) is proposed. Experiments show that the algorithm improves the robustness of the model and can achieve accuracy of 55.3% and 54.51% respectively on CIFAR-10 and SVHN when facing the powerful white-box attack PGD(Project Gradient Descent) method,exceeding the most representative adversarial training methods at present.

Keywords　Deep learning,Adversarial defense,Adversarial training,Adversarial robustness,Singular value decomposition

1　引言

深度学习已在众多领域取得巨大成功,尤其是在计算机视觉领域,已达到甚至超越人类的表现。其已成为自动驾驶、安全监控等领域的决定性推动力量。然而,最近的研究表明,深度神经网络(Deep Neural Networks,DNN)对于对抗攻击(Adversarial Attack)毫无抵抗能力[1-6]。对图片施以极小的扰动,都可能导致模型输出结果完全改变[7]。根据是否知晓神经网络模型的信息,对抗攻击可以分为黑盒攻击[8-9]和白盒攻击[1,10]。并且,也存在专门针对各类任务的攻击,如人脸识别与检测[11-12]、自然语言处理[8,13]等。此类攻击严重制约了深度学习在线下的大规模部署,更严重威胁到基于深度模型的应用系统安全性和可信度。

针对对抗攻击,许多学者从不同的角度提出了各种方法用于提升模型的对抗鲁棒性,包括对抗训练[14-17]、对抗样本检测[18-22]。然而这些研究大多基于作者本身的启发式假设,而后利用特定的攻击手段对方法进行评价,并未从模型本身出发,为模型的对抗鲁棒性提供解释性依据。本文认为,如何正确理解并提高深度模型的对抗鲁棒性,是目前深度学习研究的重要一环。

目前,学术界也出现了一些分析模型对抗鲁棒性的工作。Li等[23]提出了基于上下网络层间梯度的关键攻击路径,从

到稿日期:2022-08-09　返修日期:2022-11-28
通信作者:詹文翰(zhanwenhan@uestc.edu.cn)

神经元的角度观察对抗攻击对模型的影响，并以此为基础，在对抗训练过程中更新梯度时，增加关键攻击路径上神经元的梯度。而 Wang 等[24]指出，由于激活函数的部分区间存在梯度消失（如 ReLU 函数中的$(-\infty,0)$区间），梯度并不能准确地反映神经元间的作用关系，会存在较大的噪声。Zhang 等[25]则直接根据神经元在面对对抗样本和干净样本时的差异来定义神经元的敏感性（Neuron Sensitivity），并发现神经元敏感性与模型的对抗鲁棒性之间存在关联。但该工作并未比较对抗训练前后神经元敏感性的变化，对模型对抗鲁棒性的解释不够明确。Gavrikov 等[26]则从卷积核的角度出发，统计并分析卷积核的权值在对抗性环境下的行为模式。他们发现模型对抗鲁棒性与卷积核的稀疏性、多样性、冗余性存在关联，然而并未能进一步提出提升对抗鲁棒性的方法。

本文从奇异值的角度出发，尝试分析并解释模型的对抗鲁棒性。首先，对模型的中间结果以神经元为单位进行奇异值分解（Singular Value Decomposition，SVD），统计了在对抗性环境下奇异值的分布情况。通过对比发现，对抗训练在提升模型对抗鲁棒性的同时，还平滑了奇异值的分布。然后，对奇异值的分布进行了进一步的讨论和解释，指出平滑的奇异值分布意味着模型具有更丰富的的分类置信度来源；而过度集中的奇异值则说明模型依赖少数乃至单一维度的信息进行决策，对抗鲁棒性较差。在此基础上，本文提出了基于奇异值抑制的对抗训练方法，将每个样本奇异值的标准差作为正则项加入训练损失函数。实验证明，奇异值抑制能有效提升深度模型在对抗攻击下的分类精度，在面对强力的白盒攻击方法 PGD（Project Gradient Descent）[16] 时，在 CIFAR10 和 SVHN 数据集上分别能达到 55.3%和 54.51%的精度，超越了目前最具有代表性的对抗训练方法。

2 相关工作

2.1 对抗攻击

对抗攻击的概念首次由 Goodfellow 等[1]提出，其目的是在原始样本 x 的基础上加入人眼难以分辨的对抗性扰动，从而得到对抗样本 x'，诱导模型输出错误预测：

$$h_\theta(x')\neq y \quad \text{s.t.} \quad \| x-x' \|_p<\varepsilon \tag{1}$$

其中，θ 代表模型参数，$h_\theta(\cdot)$是模型输出，ε 表示攻击强度，$\| \cdot \|_p$ 表示 Lp 范数。

2.2 对抗鲁棒性的解释性研究

Li 等[23]从关键攻击路径的角度来解释模型的对抗鲁棒性。对于总共 L 层的分类模型 $\boldsymbol{F}$，输入样本 x，第 $l-1$ 层的关键攻击神经元 $\boldsymbol{\omega}_{l-1}^x$ 定义为对 l 层梯度（或损失）的贡献最大的 k 个神经元：

$$\boldsymbol{\omega}_{l-1}^x=\underset{m}{\text{top}}\ k\left(\frac{\partial \boldsymbol{F}_l}{\partial \boldsymbol{F}_{l-1}^m}\right) \tag{2}$$

其中，$\boldsymbol{F}_l^m$ 表示第 $l(l=1,2,\cdots,L)$层的第 $m(m=1,2,\cdots,K)$个神经元输出（下文相同）。如此递归地计算出每一层的关键攻击神经元后，即可得到针对样本 x 的关键攻击路径 $\boldsymbol{R}(x)=\{\boldsymbol{\omega}_1^x,\boldsymbol{\omega}_2^x,\cdots,\boldsymbol{\omega}_L^x\}$。再针对数据集 D 中每一个样本 x_i 统计每一层中的每个神经元在样本级关键路径 $\boldsymbol{R}(x_i)$中的出现频率，选出其中出现频率最高的一部分神经元作为数据集 D 上的关键路径 $\boldsymbol{R}(D)$。通过实验发现，将关键攻击路径上 30%的神经元输出替换为干净样本对应位置的激活值，几乎可以完全消除被污染样本的对抗性。并通过可视化方法发现，关键攻击路径上包含着最丰富的语义信息。

Zhang 等[25]发现神经元敏感性与对抗鲁棒性存在关联。他们提出将每个神经元在面对干净样本和对抗样本时的差异作为衡量神经元敏感性的指标，定义单个神经元的敏感性 $\Delta(\boldsymbol{F}_l^m)$为：

$$\Delta(\boldsymbol{F}_l^m)=\frac{1}{N}\sum_i^N\frac{1}{\dim(\boldsymbol{F}_l^m(x_i))}\times\| \boldsymbol{F}_l^m(x_i)-\boldsymbol{F}_l^m(x_i') \|_1 \tag{3}$$

他们还发现倒数第二层神经元的敏感性 $\Delta(\boldsymbol{F}_{L-1}^m)$与该神经元对于错误类别 y'的分类置信度贡献值具有较强的正相关性，从而推定神经元敏感性与模型对抗鲁棒性也存在关联。并基于 $\Delta(\boldsymbol{F}_l^m)$排序，抑制其中较大的 top-$k$ 个神经元的敏感性，该损失项定义为：

$$Loss(x_i,x_i')=\sum_{l=1}^{N-1}\sum_{m\in\boldsymbol{T}}\frac{1}{\dim(\boldsymbol{F}_l^m(x_i))}\times\| \boldsymbol{F}_l^m(x_i)-\boldsymbol{F}_l^m(x_i') \|_1 \tag{4}$$

其中，$\boldsymbol{T}=\{m:\boldsymbol{F}_l^m\in\text{top}\ k(\Delta\boldsymbol{F}_l)\}$，$\dim(\cdot)$表示维度。

Gavrikov 等[26]则将目光转向卷积核本身，他们比较了 71 种公开模型在不同的数据集上对抗训练前后卷积核行为模式的变化。在经过对抗训练后，模型的卷积核会具有更好的多样性、更少冗余以及更小的稀疏性。然而他们也提出，这些特性并不能完全解释模型的对抗鲁棒性，因为相比简单数据集，在更复杂的数据集上，模型也往往会呈现出相似的特征。

2.3 对抗训练

在面对众多对抗攻击算法的情况下，也出现了各种提升模型对抗鲁棒性、提高深度学习在对抗性环境下可用性的方法。而其中表现出最强通用性和最优效果的方法是对抗训练。

对抗训练[1,27-28]是一种针对模型本身的对抗鲁棒性提升方法，对象是模型参数本身。对抗训练本质上可被当作一种数据增强技术，利用对抗攻击算法产生对抗样本，并将其加入训练集中进行训练。对抗训练可以被抽象地归纳为解决以下 min-max 优化问题[16]：

$$\min_\theta\frac{1}{N}\sum_{i=1}^N\max_{\| x_i'-x_i \|_p\leqslant\varepsilon}\ell(h_\theta(x_i'),y_i) \tag{5}$$

其中，N 代表所有训练集样本的数量，$\ell(\cdot)$是模型的分类损失，一般采用交叉熵损失。内层的 max 过程由对抗攻击算法负责，尽可能生成攻击性最强的样本，让模型损失最大化；外层的 min 过程则代表利用对抗样本进行训练。基于 PGD 攻击[16]的对抗训练，通过在样本邻域内寻找近似最强的攻击样本，产生了高质量的对抗样本，成为对抗训练中内层的 max 优化过程的基础算法。其外层 min 优化过程的损失为交叉熵（CrossEntropy）损失：

$$CE(p(x',\theta),y) \tag{6}$$

通过在式(6)的基础上加入不同的正则项，对抗训练演化出了其他变体。例如，ALP（Adversarial Logit Pairing）[29]在式(6)的基础上加入对抗样本和干净样本之间的输出概率的欧氏距离，其表达式为：

$$CE(p(x',\theta),y)+\lambda\cdot\| p(x',\theta)-p(x,\theta) \|_2^2 \tag{7}$$

这进一步提升了模型的对抗鲁棒性。

Yan 等[15]则假设模型中的神经元对于每个类别可被分为正相关和负相关两种。他们提出在中间层 $\boldsymbol{F}_l$ 嫁接辅助分类器，通过计算梯度的方式评估识别该层中的每个神经元 $\boldsymbol{F}_l^m$ 对每个分类置信度的贡献值 $\boldsymbol{r}_l^m(c)$ 是正向的还是负向的。

$$\boldsymbol{r}_l^m(c)=\frac{\partial logit(c)}{\partial \boldsymbol{F}_l^m} \tag{8}$$

其中，$logit(c)$ 为辅助分类器中类别 c 的分类得分。接下来选择 $logit(c)$ 中得分最高的 top-k 个类的贡献度之和作为接下来对特征的缩放依据。

$$\boldsymbol{r}_l^m=\sum_{c\in \text{top } k(logit)} \boldsymbol{r}_l^m(c) \tag{9}$$

对其中正相关的神经元进行增强，而抑制负相关的神经元输出值：

$$\boldsymbol{F}_l^m=\boldsymbol{F}_l^m \cdot mask(\boldsymbol{r}_l^m) \tag{10}$$

其中，$mask(\cdot)$ 为非负的单调递增函数，用于修饰 $\boldsymbol{r}_l^m$。该算法证明了对特征的筛选可有效提升模型的对抗鲁棒性，然而当攻击算法将辅助分类器也作为攻击对象时，对结果的提升有限。

3 基于奇异值分解的对抗鲁棒性分析

为提升关于对抗鲁棒性分析的可解释性，本章利用奇异值分解这种梯度无关的计算工具，在避免梯度噪声的情况下分别定性和定量地分析了对抗训练前后奇异值分布的变化。最后结合学术界关于对抗性噪声的前沿性研究，讨论并解释了奇异值分布与模型对抗鲁棒性间的关系。

3.1 神经元的奇异值分解

给定一个数据集 D，其中数据为 $x\in\chi$，标签为 $y\in\gamma$，深度学习模型(有监督学习)的目的在于学习将 x 映射到 y 的分类模型 $F:\chi\rightarrow\gamma$。

设定模型 F 总共包含 L 层，其中第 l 层表示为 $\boldsymbol{F}_l(l=1,2,\cdots,L)$。本文选取卷积神经网络(Convolutional Neural Network，CNN)作为图像分类的模型，因此可将 l 层的输出表示为：

$$\boldsymbol{F}_l\in R^{k*w*h} \tag{11}$$

即第 l 层输出特征图的通道数、宽、高分别为 k,w,h。本文将每个神经元的输出，即每个通道的特征图合并为一个维度 $d=w*h$，从而得到一个行数和列数分别为 k,d 的矩阵，记做 $\boldsymbol{A}_l$。对 $\boldsymbol{A}_l$ 做奇异值分解，得到：

$$\boldsymbol{A}_l=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{12}$$

其中，$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 均为标准正交矩阵，$\boldsymbol{\Sigma}=\text{diag}(\sigma_1,\sigma_2,\cdots,\sigma_d)$ 是维度为 $k*d$ 的对角阵，代表 $\boldsymbol{A}_l$ 的 d 个奇异值，且 $\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_d>0$。

此处，由于做了维度合并 $d=w*h$，根据定义，对 $\boldsymbol{A}_l$ 做奇异值分解可以理解为 $\boldsymbol{A}_l$ 被投影到一组新的正交基 $\boldsymbol{V}^{\mathrm{T}}$ 上。并且这组正交基中每个向量的维度与原始神经元的维度相同，因此每个原始神经元的输出都能由 $\boldsymbol{V}^{\mathrm{T}}$ 的线性组合明确表示出，每个奇异值 σ_i 则刻画了正交基 $\boldsymbol{V}^{\mathrm{T}}$ 中每个向量的重要程度。如果 $\boldsymbol{A}_l$ 被投影到维度不同的一组基上，则 $\boldsymbol{A}_l$ 与 $\boldsymbol{V}^{\mathrm{T}}$ 会具有更加复杂的对应关系，不利于以神经元为单位解释模型行为。此外，该方法不依赖于梯度，仅利用模型的正向推理的中间结果进行计算，避免了梯度带来的噪声，有助于更准确地捕获模型的行为模式[24]。

3.2 对抗鲁棒性分析与解释

接下来，本文利用特征值分解作为工具，研究在对抗性环境下，深度神经网络模型的行为特点，并从奇异值分布的角度提出了关于模型对抗鲁棒性的解释。

选择最广泛使用的 ResNet-18 作为目标模型，数据集为 CIFAR-10，分别进行标准分类训练(Vanilla)和对抗训练 SAT(Standard Adversarial Training)。对抗训练采用无目标 PGD-L∞攻击产生对抗样本，令 $\varepsilon=8/255$，即式(1)中的约束项变为：

$$\|x'-x\|_\infty\leqslant 8/255 \tag{13}$$

迭代次数设为 20 步，步长为 0.8/255，将此攻击设置记为 PGD-20。

图 1 给出了在 PGD-20 攻击下，经过对抗攻击前后 ResNet-18 模型的不同表现。采用 PGD-20 在 CIFAR10 的验证集上生成对抗样本，并统计 ResNet-18 中倒数第二层(最后一个残差块)输出特征图的奇异值分解结果。

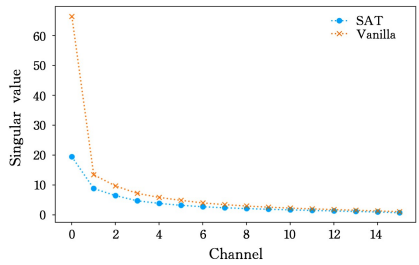


图 1 对抗样本在对抗训练与非对抗训练模型中的奇异值比较

Fig. 1 Singular value of adversarial samples on Vanilla model vs. SAT model

从图 1 可以看出，Vanilla 模型下的奇异值的分布极其不均匀。根据低秩逼近理论(Eckart-Young-Mirsky 定理)，特征值越大的特征向量 $\boldsymbol{v}_i\in\boldsymbol{V}^{\mathrm{T}}$ 包含有越多的信息量。这意味着在经过 SVD 得到特征向量构成的新坐标中，绝大部分信息量集中于极少数的几个方向。因此，模型做出分类决策的置信度绝大部分来源于极少数的几个维度方向。然而，经过对抗训练，不仅奇异值的绝对值减小了，奇异值分布的均匀性也增加了。

图 2 给出了干净数据集上的奇异值分布情况。结果表明，在面对未掺入对抗性噪声的干净样本时，Vanilla 模型和 SAT 模型也表现出了与在对抗性环境下相似的行为模式，即 SAT 模型特征图的奇异值分布更均匀。

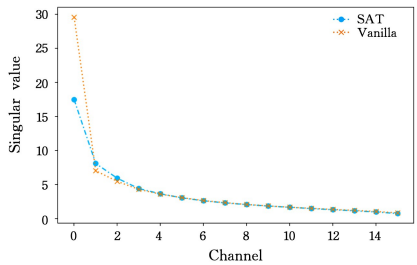


图 2 干净样本在对抗训练与非对抗训练模型中的奇异值比较

Fig. 2 Singular value of clean samples on Vanilla model vs. SAT model

为了更准确地理解神经元的行为模式，本文进一步评估了奇异值的分布，用以下两个指标来进行观察：(1)max/sum(%)，奇异值中最大值占总和的百分比；(2)STD，奇异值标准差。计算方式为先对每个样本进行计算，得出每个样本的最大值占比和标准差，再求两个统计量在所有样本间的均值。

表 1 中，cln 表示原始的验证集(干净样本)，adv 代表与 cln 对应的对抗样本集。本文比较了 3 种模型：Vanilla 模型，SAT 模型，以及加入奇异值抑制损失所训练出的 SVS 模型(具体训练算法见第 4 章)。

表 1 特征值统计

Table 1 Statistics of singular value

Models	max/sum/%		STD	
	adv	cln	adv	cln
Vanilla	49.74	42.31	12.82	6.75
SAT	30.09	30.51	4.56	5.11
SVS	16.01	17.14	0.70	0.91

由表 1 可知，无论是干净样本还是对抗样本，奇异值的最大值占比在经过对抗训练后均减小。尤其是对抗样本，最大值从接近一半(49.74%)下降为 30.09%。标准差的结果也具有类似分布规律，相比 Vanilla 模型，SAT 模型具有更紧凑的奇异值分布。SVS 则进一步抑制了奇异值在少数特征向量上的集中。这说明对抗训练会促使模型将做出分类判别所依赖的语义特征分散到更多的维度上，而不是仅仅依赖少数维度就做出判定。

深度模型，特别是卷积神经网络，总的来说可以看作是一个乘加系统，只不过复杂程度很高。在这个系统中，有极大概率会存在着从浅层到深层的“放大通路”，即浅层中数值极小的激活值经过层层计算，会对深层网络造成极大的影响。而对抗攻击算法也正是利用了这些放大通路，在原始图像上加入极小的扰动，完全颠覆模型的预测结果。这种对抗扰动的累积效应已经被之前的许多研究证实[24,30-31]。Liu 等[10]则提供了一个关于累积效应的更极端的例子：在输入图像上植入一个与模型中特定层相关联的触发器纹理，而且该纹理相对于原始图片只是一个低水平噪声。通过在训练模型时不断加强这个纹理与特定层输出的因果联系，当模型在遇到该触发器时，则会输出特定的错误结果。也就是说，在输入和输出之间建立了一个更强大的“放大通路”。

而对抗训练使得模型的奇异值分布发生了变化，最大值被强烈地抑制了，非最大值相对于最大值更大了。这说明特征图的信息来源于更多的维度，因此分类置信度的来源更加多样化，而不是少数几个维度，从而减小了上述“放大通路”存在的概率。从语义上可以理解为模型的分类不单单依赖于某些类别特有的特征，而是结合多种语义特征综合判别。例如，要判定一张图片是汽车，就需要看图中是否有轮胎、观察轮廓造型等，而不是图像中只要出现了轮子就草率地判定图中有汽车。

因此，结合以上讨论可以做出推断，仅依赖少数维度进行分类推断的模型，在面对精心设计的恶意对抗攻击时对抗鲁棒性往往会更差。而奇异值分布更加平滑的模型，具有更加优化的分类判断决策过程，具备更好的对抗鲁棒性。

4 基于奇异值抑制的对抗训练

基于第 3 章对奇异值的剖析，一种直接的想法是抑制激活值的奇异值分布，以使其分布更加平滑。换言之，迫使模型让更多的神经元参与到每个样本的特征提取中，利用更多样化的语义信息进行分类识别。通过将奇异值的标准差作为正则项加入对抗训练的损失函数中，可以很容易地实现奇异值抑制(SVS)。选定第 l 层作为奇异值抑制的对象，给定对抗样本 x'，奇异值抑制损失 $\ell_{\text{svs}}(x',\theta)$ 定义为：

$$\ell_{\text{svs}}(x',\theta)=\text{STD}(\sigma_1^l,\sigma_2^l,\cdots,\sigma_d^l)=\sqrt{\frac{1}{d-1}\sum_{i=1}^{d}(\sigma_i^l-\overline{\sigma^l})} \quad (14)$$

其中，$\overline{\sigma^l}$ 为 l 层奇异值的均值 $\sum_{i=1}^{d}\sigma_i^l$。

为了加快模型训练速度，并提高训练中的稳定性，先只采用标准对抗训练算法 SAT 对模型进行预训练，待损失稳定之后再加入式(14)作为正则项以进一步提升模型对抗鲁棒性。总的训练过程如算法 1 所示。

算法 1 利用奇异值抑制提升模型对抗鲁棒性

输入：样本数为 N 的训练集 D，初始模型 F_{init}，训练批次大小 B，总训练轮数 E，预训练轮数 E_{pre}，损失平衡项 λ_{svs}

输出：鲁棒模型 F_{robust}

1. 初始化模型参数 F_{init}
2. for e←0 to E do：

利用 PGD 攻击在 D 上产生对抗样本集 $D'=\{x_i'|i=1,2,\cdots,N\}$

if $e<E_{\text{pre}}$：$\lambda=0$ else：$\lambda=\lambda_{\text{svs}}$

/* 设置超参数 λ */

for b←0 to $\left\lfloor\frac{N}{B}\right\rfloor$ do：

$$\min_{\theta}\sum_{i\in\text{batch b}}(\text{CE}(p(x_i',\theta),y)+\lambda_{\text{svs}}(x_i',\theta))$$

/* 优化更新模型参数 */

end for

end for

3. 返回 F_{robust}，算法 1 结束

关于目标层 l 的选择，由于深度模型中不同层的行为各不相同[32]，因此如何选择合适的目标层也是较为关键的策略。本文对于在不同层进行奇异值抑制的效果进行了实验。对于 ResNet-18 中每个残差块的输出(block1－block8)进行了实验。对于每一个 block，均调整超参数 λ_{svs} 以获得最好的训练结果，并用最优结果进行对比。图 3 的实验结果表明，对最后一个残差块(block 8)进行奇异值抑制，模型会有最好的对抗鲁棒性。通过分析，其原因是越接近分类层的神经元会产生更多影响最终分类结果的高层语义信息；并且由于对抗扰动的累积效应，更深层的特征图也具有更高水平的对抗噪声，因此更容易体现出对抗噪声的特征。

另外，值得注意的是，对 block1 和 block2 进行奇异值抑制，λ_{svs} 设置过大(大于 0.4)容易导致模型不收敛。通过推测，这是因为浅层神经元更多地作为局部特征提取器，对所有类别都具有较大的激活概率；而强行抑制奇异值的标准差，容易导致浅层的神经元难以提取出足够的特征用于分类判断，从而导致模型发散。因此，选取更深的层进行奇异值抑制是更加有效的策略。

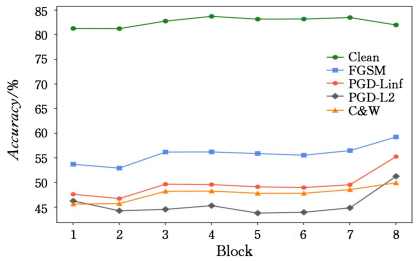


图 3 对不同层进行奇异值抑制的效果

Fig. 3 Results of SVS on different layers

5 实验与分析

5.1 实验设置

为了验证 SVS 算法的有效性，本文分别在 CIFAR-10 和 SVHN 数据集上设计了实验。目标模型为 ResNet-18，采用其最后一个残差块（倒数第二层）的输出 block8 作为 SVS 损失的目标层。

训练中均采用 PGD-L∞作为对抗样本的生成算法，$\epsilon=8/255$，迭代次数设为 10，步长为 1.6/255，实现来自于 AdverTorch[33]。预训练轮数 $E_{pre}=30$，加入奇异值抑制项后再训练 30 轮，总训练轮数 $E=60$。采用带动量的 SGD 作为模型优化器，动量设置为 0.9；初始学习率为 0.1，分别在第 30 和第 50 轮时衰减为原来的 1/10。

5.2 防御白盒攻击

白盒攻击的设定为攻击方知晓模型的全部结构和参数，是最具代表性的模型对抗鲁棒性评估方法。本文选取了几种使用最为广泛的白盒攻击算法作为模型对抗鲁棒性的评估手段，如表 2 所列，包括 FGSM[34]，PGD[16] 和 C&W[35] 攻击算法。

本文攻击算法的具体参数设置与其他经典的对抗训练算法保持一致[15-16]。其中 FGSM 设置扰动限制为 $\epsilon=8/255$；PGD-L∞设置扰动上限为 $\epsilon=8/255$（L∞范数），攻击最大迭代步数为 20，步长为 0.8/255；C&W 攻击采用 $\epsilon=8/255$（L∞范数）。为避免攻击强度过高或过低导致实验中各算法差距不明显，本文新增加的 PGD-L2 攻击设置扰动上限，对于 CIFAR-10 和 SVHN 分别设定为 $\epsilon=200/255$ 和 $\epsilon=150/255$（L2 范数），步长分别为 20/255 和 15/255，可攻击最大迭代步数均为 20。

作为对比，本文共测试了另外 3 种具有代表性的对抗训练算法在上述攻击下的表现。SAT 与第 3 章中所用的模型相同。ALP[29] 模型的训练也采用与 SVS 训练相似的过程，即先用式(6)预训练 30 轮之后再用式(7)作为损失函数进行训练；超参数 λ 与原文一致，设置为 0.5。CIFS-L4 是基于 CIFS[15] 对抗训练算法训练的模型，后缀 L4 表示模型对最后一个残差块的输出进行基于通道的激活值抑制。所有模型的学习率策略设置均相同。

表 2 列出了在 CIFAR-10 和 SVHN 数据集上的实验结果。在 CIFAR-10 数据集上，本文提出的 SVS 算法在面对 4 种最具代表性的白盒攻击算法时均取得了最好的表现。不仅在面对训练中的同种类攻击算法 PGD-L∞攻击时能持续提高模型的对抗鲁棒性（55.30%），在面对与训练过程不同的 L2 范数攻击算法时，测试结果也较其他对抗训练模型有明显提升，正确率至少提升 6%以上。在 SVHN 数据集上，除了面对 PGD-L2 攻击的结果（44.02%）略低于 CIFS-L4（44.48%），SVS 在防御其余攻击方法时依然有最优的表现。值得注意的是，SVS 算法在面对 PGD-L∞攻击时，相比 CIFS-L4 的防御成功率在 CIFAR-10 和 SVHN 数据集上分别提升了 4.09%和 2.39%。这说明 SVS 算法相比目前最先进的 CIFS-L4 算法能确切地提升模型的对抗鲁棒性。并且，总的来看，SVS 算法的综合表现优于其他所有对比算法。

表 2 白盒攻击下鲁棒性精度

Table 2 Robust accuracy under white-box attacks

（单位：%）

Models	CIFAR-10				SVHN			
	FGSM $\epsilon=8/255$	PGD-L2 $\epsilon=200/255$	PGD-L∞ $\epsilon=8/255$	C&W $\epsilon=8/255$	FGSM $\epsilon=8/255$	PGD-L2 $\epsilon=150/255$	PGD-L∞ $\epsilon=8/255$	C&W $\epsilon=8/255$
Vanilla	18.23	5.95	0.00	0.00	13.04	6.97	0.45	0.39
SAT	56.73	47.04	51.06	49.77	60.91	40.41	50.36	52.32
ALP	57.21	46.42	51.12	49.73	60.36	40.20	52.31	53.02
CIFS-L4	57.83	45.30	51.21	49.94	58.01	**44.48**	52.12	51.35
SVS	**59.28**	**51.31**	**55.30**	**49.98**	**62.98**	44.02	**54.51**	**53.36**

结合表 1，奇异值的最大值占比和标准差越小，精度就越高。而主动抑制奇异值分布的标准差，能够提高模型的对抗鲁棒性。这充分说明了 SVS 算法可以通过抑制奇异值向少数特征向量的过度集中，从而抑制激活值信息在少数维度上的聚集，进而有效优化模型的特征分布，以获得更强的对抗鲁棒性。

5.3 超参数 λ_{svs} 对于对抗鲁棒性的影响

表 3 和表 4 分别列出了在 CIFAR-10 和 SVHN 数据集上超参数损失平衡项 λ_{svs} 对模型对抗鲁棒性的影响。在 CIFAR-10 数据集上，当 λ_{svs} 设置为 1.0 时，在面对 C&W，PGD-L2，PGD-L∞攻击时模型具有最高的对抗鲁棒性。虽然 FGSM 攻击下，$\lambda_{svs}=0.6$ 时对抗性精度最高，但是与其他设置的差别极小。在 SVHN 数据集上，λ_{svs} 为 0.4 时，除了面对 FGSM 攻击的精度略低于 $\lambda_{svs}=0.3$ 的模型外，其余攻击下模型的对抗鲁棒性最高。

可以看到，在 λ_{svs} 设定值较小时，奇异值抑制损失 $\ell_{svs}(x',\theta)$ 对模型的约束性较低，从而不能发挥奇异值抑制的全部效能；而过大的 λ_{svs} 会加剧模型的欠拟合，导致 Clean 精度（未经篡改的原始的数据集上的分类精度）和对抗鲁棒性衰减。另外，可以注意到，在 Clean 精度与模型对抗鲁棒性之间具有

一定的妥协性，即对抗鲁棒性升高，会导致 Clean 精度有一定程度的下降。

表 3 CIFAR-10 上 λ_{svs} 对模型对抗鲁棒性的影响

Table 3 Impact of λ_{svs} to model's robustness on CIFAR-10

（单位：%）

λ_{svs}	0.4	0.6	0.7	0.8	0.9	1.0	1.1	1.2
Clean	83.43	82.49	83.57	81.42	83.24	82.02	81.86	76.93
FGSM	59.23	**59.92**	58.99	59.97	59.28	59.28	59.87	57.89
PGD-L2	49.7	50.64	49.41	51.29	49.87	**51.31**	50.76	50.91
PGD-L∞	54.24	54.86	53.86	55.14	54.42	**55.30**	54.95	54.66
C&W	49.66	49.47	49.62	48.92	49.54	**49.98**	49.04	47.08

表 4 SVHN 上 λ_{svs} 对模型对抗鲁棒性的影响

Table 4 Impact of λ_{svs} to model's robustness on SVHN

（单位：%）

λ_{svs}	0.3	0.4	0.5	0.6	0.7	0.8
Clean	91.02	89.77	88.26	88.99	91.08	90.48
FGSM	**63.76**	62.98	61.50	61.82	63.60	63.03
PGD-L2	40.85	**44.02**	43.87	42.24	39.87	40.35
PGD-L∞	53.60	**54.51**	54.17	53.48	53.35	52.99
C&W	52.11	**53.36**	53.14	52.36	51.89	51.62

因此，综上所述，在 CIFAR-10 和 SVHN 数据集上分别设置 $\lambda_{svs}=1.0$ 和 $\lambda_{svs}=0.4$ 是综合考虑原始精度与模型对抗鲁棒性的最优策略。

结束语 本文以 SVD 为工具，观察并分析了对抗环境中模型的对抗鲁棒性与奇异值的关系。经过对比发现，模型的对抗鲁棒性提升伴随着更加平滑的奇异值分布，也意味着模型将更多维度的信息用于分类决策。本文进一步讨论解释了奇异值分布、模型的语义集中程度和对抗鲁棒性三者的关系。基于此推断，进一步提出了 SVS 对抗训练算法，实验表明该算法在面对众多强力的白盒攻击时能进一步提升深度模型的对抗鲁棒性。值得注意的是，表 1 显示，经过对抗训练之后的模型，干净样本的最大值占比和标准差反而比对抗样本的略大，这是一个值得进一步分析的有趣现象，初步猜测是由于训练集中只有对抗样本，模型过拟合了对抗噪声，对于奇异值的抑制反而略大于干净样本。未来的研究将致力于进一步地研究这个现象，并设计更多实验，利用不同的统计分析方法探究奇异值分布与模型对抗鲁棒性之间的关系。

参 考 文 献

[1] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and Harnessing Adversarial Examples[C]// Proceedings of the International Conference on Learning Representations. OpenReview. net, 2015: 1-11.

[2] CHEN M X, ZHANG J Y, JI S L, et al. Survey of Research Progress on Adversarial Examples in Images[J]. Computer Science, 2022, 49(2): 92-106.

[3] WANG C, WEI X L, TIAN Q, et al. Feature Gradient-based Adversarial Attack on Modulation Recognition-oriented Deep Neural Networks[J]. Computer Science, 2021, 48(7): 25-32.

[4] CHERNIKOVA A, OPREA A. FENCE: Feasible Evasion Attacks on Neural Networks in Constrained Environments[J]. ACM Transactions on Privacy and Security, 2022, 25(4): 1-34.

[5] CHEN J Y, ZHANG D J, HUANG G H, et al. Adversarial Attack and Defense on Graph Neural Networks: A Survey[J]. Chinese Journal of Network and Information Security, 2021(3): 1-28.

[6] LIU X L, LUO Y H, SHAO L, et al. Survey of Generation, Attack and Defense of Adversarial Examples[J]. Application Research of Computer, 2020, 37(11): 3201-3205, 3212.

[7] WANG Z, SONG M, ZHENG S, et al. Invisible Adversarial Attack against Deep Neural Networks: An Adaptive Penalization Approach[J]. IEEE Transactions on Dependable and Secure Computing, 2021, 18(3): 1474-1488.

[8] WANG Q, ZHENG B, LI Q, et al. Towards Query-Efficient Adversarial Attacks Against Automatic Speech Recognition Systems[J]. IEEE Transaction on Information Forensics and Security, 2021, 16: 896-908.

[9] WEI X, GUO Y, LI B. Black-box Adversarial Attacks by Manipulating Image Attributes[J]. Information Sciences, 2021, 550: 285-296.

[10] LIU Y, MA S, AAFER Y, et al. Trojaning Attack on Neural Networks[C]// Proceedings of the Network and Distributed System Security Symposium. Reston: Internet Society, 2018: 1-15.

[11] ZHONG Y, DENG W. Towards Transferable Adversarial Attack Against Deep Face Recognition[J]. IEEE Transaction on Information Forensics and Security, 2021, 16: 1452-1466.

[12] JING H Y, ZHOU C, HE X. Security Evaluation Method for Risk of Adversarial Attack on Face Detection[J]. Computer Science, 2021, 7(48): 17-24.

[13] HAO Z Y, CHEN L, HUANG J C. Class Discriminative Universal Adversarial Attack for Text Classification[J]. Computer Science, 2022, 49(8): 323-329.

[14] WANG D N, CHEN W, YANG Y, et al. Defense Method of Adversarial Training based on Gaussian Enhancement and Iterative Attack[J]. Computer Science, 2021, 48(6A): 509-513, 537.

[15] YAN H, ZHANG J, NIU G, et al. CIFS: Improving Adversarial Robustness of CNNs via Channel-wise Importance-based Feature Selection[C]// Proceedings of the International Conference on Machine Learning. New York: PMLR, 2021: 1-11.

[16] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[C]// Proceedings of the International Conference on Learning Representations. OpenReview. net, 2018: 1-28.

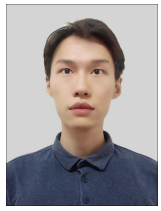
[17] WANG D, LI C, WEN S, et al. Defending Against Adversarial Attack towards Deep Neural Networks via Collaborative Multi-Task Training[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(2): 953-965.

[18] CRECCHI F, MELIS M, SOTGIU A, et al. FADER: Fast Adversarial Example Rejection[J]. Neurocomputing, 2022, 470: 257-268.

[19] XU W, EVANS D, QI Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks[C]// Proceedings of the Network and Distributed System Security Symposium. Reston: Internet Society. 2018: 1-15.

[20] WANG Y, SONG X, XU T, et al. From RGB to Depth: Domain Transfer Network for Face Anti-Spoofing[J]. IEEE Transaction

on Information Forensics and Security, 2021, 16: 4280-4290.
[21] JIN K, ZHANG T, SHEN C, et al. Can We Mitigate Backdoor Attack Using Adversarial Detection Methods? [J]. IEEE Transactions on Dependable and Secure Computing, 2022, Early Access: 1-15.
[22] WEI Z C, FENG H, ZHANG X Q et al. Research on Physical Adversarial Sample Detection Method based on Attention Mechanism[J]. Application Research of Computer, 2022, 39(1): 254-258.
[23] LI T, LIU A, LIU X, et al. Understanding Adversarial Robustness via Critical Attacking Route[J]. Information Sciences, 2021, 547: 568-578.
[24] WANG H, WANG Z, DU M, et al. Score-CAM: Score-weighted Visual Explanations for Convolutional Neural Networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. New York: IEEE Press, 2020: 111-119.
[25] ZHANG C, LIU A, LIU X, et al. Interpreting and Improving Adversarial Robustness of Deep Neural Networks with Neuron Sensitivity[J]. IEEE Transactions on Image Processing, 2021, 30: 1291-1304.
[26] GAVRIKOV P, KEUPER J. Adversarial Robustness through the Lens of Convolutional Filters[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. New York: IEEE Press, 2022: 1-9.
[27] ZHU C, CHENG Y, GAN Z, et al. FreeLB: Enhanced Adversarial Training for Natural Language Understanding[C]// Proceedings of the International Conference on Learning Representations. OpenReview. net, 2020: 1-12.
[28] ZHANG D, ZHANG T, LU Y, et al. You Only Propagate Once: Accelerating Adversarial Training via Maximal Principle[C]// Advances in Neural Information Processing Systems. New York: Curran Associates, Inc., 2019: 1-12.
[29] KANNAN H, KURAKIN A, GOODFELLOW I. Adversarial Logit Pairing[J]. arXiv: 1803.06373, 2018.
[30] MA S, LIU Y, TAO G, et al. NIC: Detecting Adversarial Samples with Neural Network Invariant Checking[C]// Proceedings of the Network and Distributed System Security Symposium. Reston: Internet Society, 2019: 1-15.
[31] LIAO F, LIANG M, DONG Y, et al. Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2018: 1778-1787.
[32] SHAHAM U, YAMADA Y, NEGAHBAN S. Understanding Adversarial Training: Increasing Local Stability of Supervised Models through Robust Optimization [J]. Neurocomputing, 2018, 307: 195-204.
[33] DING G W, WANG L, JIN X. {AdverTorch} v0.1: An Adversarial Robustness Toolbox based on PyTorch[J]. arXiv: 1902.07623, 2022.
[34] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing Properties of Neural Networks [C]// International Conference on Learning Representations. OpenReview. net, 2014: 1-10.
[35] CARLINI N, WAGNER D. Towards Evaluating the Robustness of Neural Networks[C]// Proceedings of the IEEE Symposium on Security and Privacy. New York: IEEE Press, 2016: 39-57.



ZHAO Zitian, born in 1993, Ph.D. His main research interests include AI security and voice print recognition.

ZHAN Wenhan, born in 1987, Ph.D, senior experimentalist. His main research interests include cloud computing, edge computing, distributed systems and AI.

(责任编辑:喻藜)