

可解释的深度网络抗噪音干扰性逐层评估方法^{*}

周瑞丰¹, 栾瑞鹏², 张超¹

(1. 大连理工大学 数学科学学院, 辽宁 大连 116024; 2. 91550 部队, 辽宁 大连 116023)

摘要: 在复杂应用场景中, 深度神经网络易受噪音数据干扰, 如何客观、有效、可靠地评估深度网络的抗噪音干扰性已成为智能化技术开发的关键问题之一。然而, 现有的评估方法只能对网络抗噪音干扰性进行整体评估, 且不具备可解释的理论依据。为了解决上述问题, 基于闵可夫斯基差线性可分性度量, 提出一种深度神经网络逐层抗噪音干扰性评估方法, 实现了神经网络抗噪音干扰性的可解释性评估。该方法可以对深度网络各隐层数据映射行为进行定量分析, 解析噪音干扰对网络各隐层的影响机理, 进而建立面向网络隐层的可解释抗噪音干扰性评估方法。

关键词: 抗干扰性评估; 可解释性; 神经网络

中图分类号: TP181

文献标识码: A

DOI: 10.19358/j.issn.2097-1788.2025.09.007

引用格式: 周瑞丰, 栾瑞鹏, 张超. 可解释的深度网络抗噪音干扰性逐层评估方法 [J]. 网络安全与数据治理, 2025, 44(9): 46–50.

Explainable methods for layer-wise evaluating noise resistance of deep networks

Zhou Ruifeng¹, Luan Ruipeng², Zhang Chao¹

(1. School of Mathematical Science, Dalian University of Technology, Dalian 116024, China;

2. Unit 91550, Dalian 116023, China)

Abstract: In complex application scenarios, deep neural networks are susceptible to noise data interference. How to objectively, effectively, and reliably evaluate the noise resistance of deep networks has become one of the key issues in the development of intelligent technology. However, the existing evaluation methods can only evaluate the noise resistance of the whole network, and have no explainable theoretical basis. To address the above problems, based on Minkowski difference based linear separability measure, a layer-wise noise resistance evaluation method for deep neural networks is proposed, achieving an explainable evaluation of the network's noise resistance performance. This method can quantitatively analyze the data mapping behavior of each hidden layer of the deep network, explain the influence mechanism of noise interference on each hidden layer of the network, and then establish an explainable noise resistance evaluation method for the network's hidden layers.

Key words: noise resistance evaluation; explainability; neural network

0 引言

随着神经网络在图像分类^[1]、语义分割^[2]等领域的长足发展, 其被广泛地应用于工业生产^[3]、自动驾驶^[4]、医学^[5]等新的任务场景中。然而, 这些新的任务场景往往更加复杂, 其中包含着干扰输入数据的多种噪音, 会显著影响网络的性能。同时, 在这些任务场景中, 网络判断出错所造成的后果往往要更加严重。因此, 如何评估网络对任务场景中噪音的抗干扰性, 从而判断网络在

该任务场景中的适用性, 成为了一个需要解决的问题。

目前, 通过生成让网络判断出错的最小扰动的对抗样本^[6-8]或者生成任务场景^[9-10]中常见的模拟噪音都可以作为生成噪音的方法, 并进一步得到网络在面对噪音下的数值化抗干扰性指标。但是, 现有的网络评估方法大多对网络整体进行评估, 无法分析噪音对网络输出的影响机理, 缺乏可解释性, 并且评估结果也无法对网络改进起到积极的指导作用。

为了解决上述问题, 本文引入基于闵可夫斯基差的线性可分性度量 (Minkowski Difference based Linear Sepa-

^{*} 基金项目: 国家自然科学基金面上项目 (62176040)

rability Measure, MD-LSM) 作为分类神经网络的隐层分析工具, 借助 MD-LSM 对网络隐层的数据映射行为进行分析, 得到噪音对网络各个隐层的扰动, 确定噪音对网络的影响机理, 从而实现网络的可解释性抗干扰性逐层评估。

1 基于闵可夫斯基差的线性可分性度量

线性可分理论认为, 当数据经过分类神经网络的隐层时, 不同类别数据集的线性可分性会逐步提升, 并最终在输出层达到很高的线性可分性, 从而实现数据分类。基于这种思想, 人们提出了多种线性可分性度量作为解释分类神经网络运行机理的工具, 并通过实验验证了度量的有效性。例如, He 和 Su^[11] 引入了 $\text{tr}(S_+^T S_-)$ 这个指标, 它是广义瑞利商 (Generalized Rayleigh Quotient, GRQ) 的一种变体, 用来衡量隐层输出的线性可分程度。Schilling 等人^[12] 引入广义判别值 (GDV) 来在训练过程中度量不同类别数据隐层输出之间的可分性。然而, 这些度量都存在一些缺陷, 例如计算复杂度高; 鲁棒性差; 均为相对度量, 无法根据指标判断其是否线性可分, 这些问题阻碍了线性可分性理论的发展。

Zhang 等人^[13] 提出了基于闵可夫斯基差的线性可分性度量, 解决了如上问题, 并实现了二分类神经网络训练过程的隐层线性可分性演变分析。对于数据集 $A = \{a_1, \dots, a_l\}_{i=1}^l$, $B = \{b_1, \dots, b_j\}_{j=1}^j$, 两个数据集间闵可夫斯基差为 $\text{MD}(A, B) = \{m_{ij} = a_i - b_j \mid i=1, \dots, l, j=1, \dots, j\}$ 。则 MD-LSM 可由下式给出:

$$\text{LS}_1(A, B) = \max_{\omega \in \mathbb{R}^n} \left\{ \frac{\left| \sum_{i \in I, j \in J} \omega^T m_{ij} \right|}{\sum_{i \in I, j \in J} \left| \omega^T m_{ij} \right|} \right\} \quad (1)$$

通过对 ω 进行快速的近似计算, 降低了 MD-LSM 计算复杂度, 从而实现了指标的快速求解。

作为一种网络隐层线性可分性分析方法, MD-LSM 可以量化地描述分类神经网络的隐层映射行为, 从而为解读神经网络隐层对神经网络整体性能的影响提供了重要工具。

2 基于 MD-LSM 的网络抗噪音干扰性评估方法

网络抗噪音干扰性评价方法通过 MD-LSM 建立了一种求解线性分类超平面的快速计算方法, 并用于数据筛选和隐含层评价准则计算。根据不同类型的数据和噪音进行抗噪音干扰性评估, 并给出相应的评估结果。

2.1 快速线性分类超平面计算方法

根据 MD-LSM 的研究内容, 针对 l 隐层求解得到的 ω^l 可以作为线性分类超平面的权重参数, 对输入数据的隐层输出进行分类。为了求得线性分类超平面, 还需要

得到偏置 b^l 的具体数值。假设隐层的输出为 $X^l = \{X_1^l, \dots, X_N^l\}_{i=1}^N$ (其中 $N = I + J$), 对应的标签为 $Y = \{y_1, \dots, y_N\}_{i=1}^N$ 。 $\pm \omega^l$ 可以作为线性分类超平面的权重, 正负号的选择可以用 ω^l 与正负类别数据的均值的内积决定。则偏置 b^l 可由下式求解得出:

$$\min_{b^l} \sum_{i=1}^N \text{Relu}((1 - 2y_i)(\omega^l X_i^l + b^l)) \quad (2)$$

其中 Relu 为线性激活函数。而 $\xi_i^l = (2y_i - 1)(\omega^l X_i^l + b^l)$ 则可以反映输入数据 X_i 的 l 隐层输出的线性可分性。根据 ξ_i^l 的正负和大小, 可以得到 X_i^l 的分类正确与否和距离线性分类超平面的距离。

2.2 网络抗噪音干扰性评估方法

在任务场景中, 网络通常会面临不同类型的噪音干扰。不同类型的噪音对网络的干扰能力不同, 并且网络的抗噪音干扰能力在不同类型输入数据上的表现也往往不同。因此, 抗噪音干扰性评价不能建立在单一的评价标准之上, 而应形成一套具有多层次评价标准的客观评价体系。

假设随机干扰函数为 γ , 可以在任务场景中模拟一种随机噪音。借助干扰函数, 将原始无噪音输入数据 X 转化为有噪音输入数据 $\gamma(X)$ 。重复这个过程 n 次, 并将噪音数据输入到网络中, 可以得到网络对干扰数据的分类精度 $\text{acc}(X, \gamma, n)$, 这可以被认为是噪音对网络干扰能力在样本 X 上的体系。为了充分衡量模拟干扰噪音 γ 的影响, 参数 n 不能太小。然而, 对所有的输入数据添加干扰会使数据集扩大 n 倍, 这将大大增加精度计算的时间成本。减少计算量的一种方法是对数据集进行筛选, 通过减小数据集的规模来简化计算。

一种方法是找出哪些数据的隐层输出的 ξ_i^l 较低, 尤其是对于靠后的隐层。相比于 ξ_i^l 较高的数据, ξ_i^l 较低的数据往往更容易受到噪音干扰而导致分类错误, 应将其作为重点研究对象。

假设将输出层之前的隐层用来选择生成噪音数据的原始数据集。将样本输入到网络中, 得到最终的隐层输出 $X^l = \{X_1^l, \dots, X_N^l\}_{i=1}^N$, 并计算线性可分性评价准则 $\{\xi_i^l\}_{i=1}^N$, 选择分类正确且 ξ_i^l 较小的数据作为原始数据, 构建正负原始样本集 $X^+ = \{X_1^+, \dots, X_S^+\}_{i=1}^S$, $X^- = \{X_1^-, \dots, X_S^-\}_{i=1}^S$ 和总样本集 $X^* = \{X_1^*, \dots, X_{2S}^*\}_{i=1}^{2S}$, S 是正负元素样本集包含的数据个数。将噪音 γ 添加到选定的样本集并输入到网络中进行准确率评估, 分别得到正样本、负样本和总样本的准确率:

$$\text{ACC}_\gamma^+ = \frac{1}{S} \sum_{i=1}^S \text{acc}(X_i^+, \gamma, n) \quad (3)$$

$$ACC_{\gamma}^{-} = \frac{1}{S} \sum_{i=1}^S \text{acc} (X_i^{-}, \gamma, n) \quad (4)$$

$$ACC_{\gamma} = \frac{1}{2S} \sum_{i=1}^{2S} \text{acc} (X_i^{*}, \gamma, n) \quad (5)$$

以上结果可作为评价噪音 γ 对神经网络的干扰在正负样本和总样本上的评估指标。对于包含多种不同噪音的任务场景,可以根据每种噪音的特点构造相应的近似函数,从而得到任务场景中所有的噪音生成函数组 $\Gamma = \{\gamma_k\}_{k=1}^K$ 。求解 γ_k 在相应正样本、负样本和总样本上的抗噪音干扰性评价准则为 $ACC_{\gamma_k}^{+}$ 、 $ACC_{\gamma_k}^{-}$ 和 ACC_{γ_k} 。如果能够得到任务中各个噪音的权重 $P = \{p_k\}_{k=1}^K$,则场景中所有噪音对正负样本和总样本的评估指标为:

$$ACC^{+} = \sum_{k=1}^K p_k ACC_{\gamma_k}^{+} \quad (6)$$

$$ACC^{-} = \sum_{k=1}^K p_k ACC_{\gamma_k}^{-} \quad (7)$$

$$ACC = \sum_{k=1}^K p_k ACC_{\gamma_k} \quad (8)$$

然而,上述方法仅仅只是概括地分析了网络整体受到的抗噪音干扰性,并无法给出噪音对网络的详细作用机理。如果能够将上述评估指标推广到网络的隐藏层(或是 Res 网络中的残差块),计算噪音干扰对网络隐层的扰动大小,发现噪音对网络整体影响与网络各隐层影响的关联关系,进而从网络隐层的角度解读噪音对网络的影响机理,那么就实现了可解释的网络抗噪音干扰性的逐层评估。

通过 2.1 节中的方法,可以借助原始数据快速计算出每个卷积层或每个残差块输出的分类超平面。通过类似于上述网络整体抗干扰性评估指标的计算过程,计算网络隐层的抗噪音干扰性评估指标: $ACC_{\gamma}^{l,+}$, $ACC_{\gamma}^{l,-}$, ACC_{γ}^l , $ACC_{\gamma}^{l,+}$, $ACC_{\gamma}^{l,-}$, ACC_{γ}^l 。其中 l 表示所评估的卷积层或残差块的位置。通过这些指标,可以从线性可分的角度得到噪音对隐层输出的扰动程度,从而实现对网络抗噪音干扰性的局部评价,并为噪音干扰网络输出结果的作用机理提供解释依据。

2.3 网络抗噪音干扰性评估算法流程

算法的输入参数包括两部分。第一部分是算法运行所需的基础参数,包括待检测网络 F 及所用数据和标签集 X, Y 。第二部分是可调节参数,包括噪音生成函数组 Γ 、原始数据集尺寸 S 、噪音权重数据集 P 与噪音样本重复生成次数 n 。通过调节如上参数,可以对不同的任务场景和不同的检测需求进行适配,从而能够在不调整算法框架的情况下适应不同的评估需求,增强算法的适用性。

算法流程主要分为以下几个部分:

(1) 对网络输出层前的隐层输出进行筛选,得到用

于构建噪音数据集的原始数据。

(2) 通过生成的噪音数据,统计网络对不同类别数据和不同类型噪音的分类准确率,得到整个网络的抗噪音干扰性指标。

(3) 借助由 MD-LSM 计算得到的隐层输出分类超平面,计算隐含层的抗噪音干扰性指标。

(4) 对上述判据进行加权平均,得到综合性的网络整体与隐层抗干扰性评估指标。

算法得到的抗噪音干扰性评价指标还包括对整个网络和隐含层的评价。前者作为评估网络抗干扰性的主要指标,负责分析整个网络在不同类别数据和不同类型噪音下的扰动;后者则侧重于评估网络隐层的抗噪音干扰性,并为前者的评估结果提供可解释性依据。

3 实验结果

在本实验中,使用 Cifar-10 数据集中飞机和汽车类别的数据对两个不同的卷积神经网络 VGG-11 和 ResNet-18 进行了训练和测试。由于数据集的规模和二分类问题的相对简单性,对原始网络结构进行了一些调整,包括删除部分池化层、改变第一层的卷积核大小以及减少整个网络的通道数。同时,去掉了网络中的所有 BN 层,因为在训练二分类问题的过程中,BN 层的增加会增加网络的训练难度。利用训练后的网络评估网络抗噪音干扰性。

3.1 网络整体的抗噪音干扰性指标

通过设置五种不同的噪音仿真函数,构建具有多种噪音干扰的仿真任务场景,并对 VGG-11 和 Res-18 网络的抗噪音性能进行测试。这五种噪音的具体形式为高斯噪音 γ_G 、椒盐噪音 γ_S 、运动模糊噪音 γ_M 、雾噪音 γ_F 和遮挡噪音 γ_O 。将这些噪音视为任务场景中可能的噪音来评估网络的抗噪性,基于 VGG 网络和 Res 网络的训练集得到受噪音干扰的图像数据。评估结果如表 1、表 2 所示。

表 1 VGG 网络的整体抗噪音干扰性测试结果

类型	γ_G	γ_S	γ_M	γ_F	γ_O	Γ
飞机	0.414	0.226	0.974	1.000	0.868	0.696
汽车	0.999	0.994	0.475	0.010	0.781	0.652
总体	0.707	0.610	0.725	0.505	0.824	0.674

表 2 Res 网络的整体抗噪音干扰性测试结果

类型	γ_G	γ_S	γ_M	γ_F	γ_O	Γ
飞机	0.833	0.599	0.935	1.000	0.817	0.837
汽车	0.979	0.964	0.682	0.006	0.846	0.696
总体	0.907	0.781	0.808	0.503	0.832	0.766

从表中的数据可以看出,VGG 网络和 Res 网络的抗

噪声干扰性在除遮挡噪声外的其他噪声上都表现出明显的类别不平衡性。雾噪声是对网络干扰程度最大的噪声，网络几乎将所有添加雾噪声的数据均识别为飞机。此外，可以发现与 VGG 网络相比，Res 网络中的噪声对网络判断准确性的干扰要小于 VGG 网络。因此，可以判断在当前噪声环境下，Res 网络的抗噪声干扰性优于 VGG 网络。

3.2 网络隐层的抗噪声干扰性指标

针对上述通过网络整体的抗噪声干扰性评估指标的评估结果，借助线性分类面对网络隐层的分类准确率这一网络隐层的抗噪声干扰性评估指标（即图中的纵坐标 ACC），分析噪声对网络隐层的干扰能力，定位那些受到噪声干扰较为严重的网络隐层，从而为网络整体的抗噪声干扰性提供可解释的支撑。评估结果如图 1~图 4 所示。

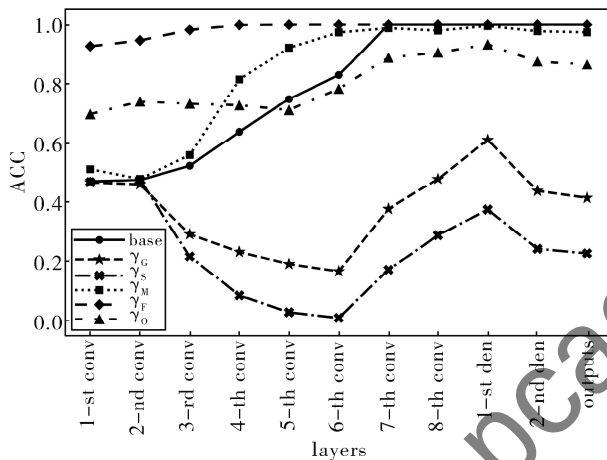


图1 VGG 网络和飞机数据集的网络隐层抗噪声干扰性测试结果

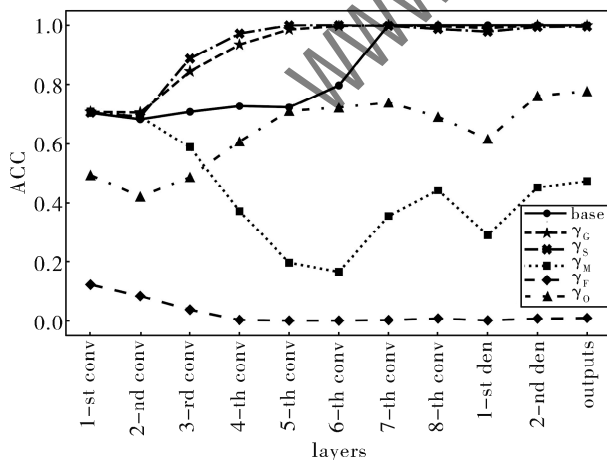


图2 VGG 网络和汽车数据集的网络隐层抗噪声干扰性测试结果

例如，对于高斯噪声和椒盐噪声，飞机数据干扰后的网络的浅层抗噪声干扰性指标明显低于原始数据，而汽车数据的浅层指标则高于原始数据，雾噪声和运动模

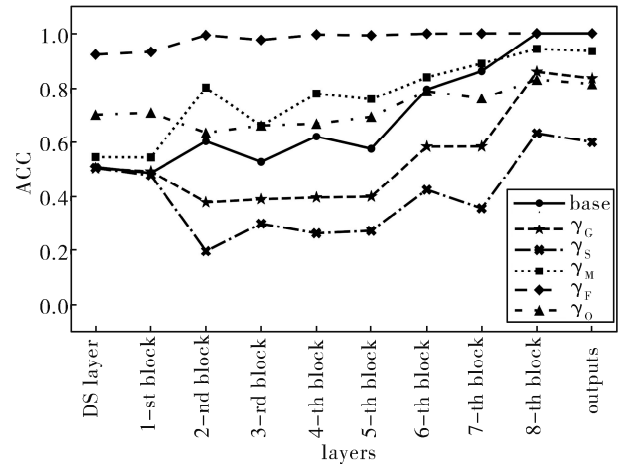


图3 Res 网络和飞机数据集的网络隐层抗噪声干扰性测试结果

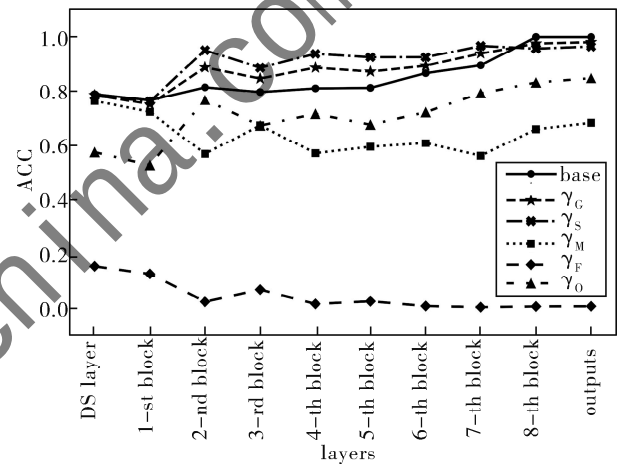


图4 Res 网络和汽车数据集的网络隐层抗噪声干扰性测试结果

糊噪声则相反。相较于网络的较深隐层，网络的较浅隐层受到了噪声更强的干扰，这可能说明抗噪声干扰性类不平衡的原因出现在网络的浅层。

相较于其他噪声，雾噪声对数据的破坏是最严重的，从雾噪声的网络隐层的抗噪声干扰性指标可以看出，网络的各个隐层都无法将叠加雾噪声的汽车数据与飞机数据真正分离。而其他噪声对于数据的破坏虽然可以借助网络后面的隐层来恢复，但是这种恢复是有限度的。网络隐层输出的指标始终无法恢复到未受到干扰的程度，最终导致网络输出的正确性受到了噪声的显著干扰。

Res 网络和 VGG 网络在抗噪声干扰性上的差异也体现在网络的隐层上。可以看出，相较于 VGG 网络的卷积层，Res 网络的卷积层能够较好地过滤数据上叠加的高斯噪声、椒盐噪声和运动模糊噪声，从而让隐层的抗噪声干扰性指标并没有出现类似于 VGG 网络的较为严重的下降，从而实现了较好的抗噪声干扰性。

通过可解释地分析噪声对网络隐层的作用机理，可

以发现,网络的靠前隐层是受到噪音影响的主要区域,它们是噪音对网络分类性能的破坏和噪音干扰的不平衡性的主要产生原因。隐层的噪音过滤能力有所差异,这是造成网络抗噪音干扰性不同的主要原因。通过这些分析结果,本文找到了噪音对网络干扰的产生机理,实现了抗噪音干扰评估的可解释化。

4 结论

利用 MD-LSM 度量,本文按照以下方式构造了一种可解释的抗噪音干扰性评估方法,实现了对网络整体和隐层针对不同噪音抗干扰性的评估。

(1) 分析网络在面对不同噪音时网络对不同数据集的分类准确率,并以此作为网络抗噪音干扰性的评估指标。

(2) 利用线性可分性度量,对数据进行筛选并计算隐层的抗噪音干扰性指标,从而实现对网络抗噪音干扰性的可解释化评估。

在 VGG-11 和 Res-18 网络上验证了评估算法的有效性,并给出了对整个网络和网络隐层的评估结果。基于实验结果,衡量了不同网络抵抗噪音干扰的能力,分析了不同种类噪音对网络的影响方式,以及噪音干扰的类别不平衡性质,展现出了本文可解释性分析方法在分析网络抗噪音干扰能力的适用性。

参考文献

- [1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [C]//Advances in Neural Information Processing Systems, 2012, 25: 1097-1105.
- [2] BADRINARAYANAN V, KENDALL A, GIPOLLA R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (12): 2481-2495.
- [3] LINDA O, VOLLMER T, MANIC M. NEURAL network based intrusion detection system for critical infrastructures [C]//2009 International Joint Conference on Neural Networks. IEEE, 2009: 102-109.
- [4] SHAF AEI S, KUGELE S, OSMAN M H, et al. Uncertainty in machine learning: a safety perspective on autonomous driving [C]//Computer Safety, Reliability, and Security: SAFECOMP 2018

Workshops, 2018: 458-464.

- [5] SANZ J A, GALAR M, JURIO A, et al. Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system [J]. Applied Soft Computing, 2014, 20: 103-111.
- [6] ZHU M, CHEN T, WANG Z. Sparse and imperceptible adversarial attack via a homotopy algorithm [C]//International Conference on Machine Learning. PMLR, 2021: 12868-12877.
- [7] JANDIAL S, MANGLA P, VARSHNEY S, et al. AdvGAN ++: harnessing latent layers for adversary generation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019: 2045-2048.
- [8] ZHOU M, WANG L, NIU Z, et al. Adversarial attack and defense in deep ranking [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2024, 46 (8): 5306-5324.
- [9] HENDRYCKS D, DIETTERICH T. Benchmarking neural network robustness to common corruptions and perturbations [C]//International Conference on Learning Representations, 2019.
- [10] MAHARANA K, MONDAL S, NEMADE B. A review: data pre-processing and data augmentation techniques [J]. Global Transitions Proceedings, 2022, 3 (1): 91-99.
- [11] HE H, SU W J. A law of data separation in deep learning [J]. Proceedings of the National Academy of Sciences, 2023, 120 (36): e2221704120.
- [12] SCHILING A, MAIER A, GERUM R, et al. Quantifying the separability of data classes in neural networks [J]. Neural Networks, 2021, 139: 278-293.
- [13] ZHANG C, CHEN X, LI W, et al. Understanding deep neural networks via linear separability of hidden layers [J]. arXiv preprint arXiv: 2307.13962, 2023.

(收稿日期: 2025-06-29)

作者简介:

周瑞丰 (2000-), 男, 博士研究生, 主要研究方向: 人工智能可解释性理论。

栾瑞鹏 (1981-), 男, 硕士, 高级工程师, 主要研究方向: 装备试验数据工程与人工智能。

张超 (1981-), 通信作者, 男, 博士, 教授, 主要研究方向: 人工智能可解释性理论与工程应用。E-mail: chao.zhang@dlut.edu.cn。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部

www.pcachina.com