计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(7):14−25 [doi: 10.15888/j.cnki.csa.009585]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于神经元统计建模分析的模型不确定性度量①

雷雅婧

(复旦大学 计算机科学技术学院, 上海 200438)
通信作者: 雷雅婧, E-mail: 21210240059@m.fudan.edu.cn

摘　要: 神经网络的不确定性反映模型对自身预测结果的置信水平, 能在决策不可靠时促使及时的人工干预, 提升系统安全性. 然而, 现有度量方法常需要对模型或训练过程进行显著修改且实施复杂度高. 为此, 本文提出一种基于神经元统计建模分析的不确定性度量方法. 该方法充分利用模型单次前向传播过程中的激活值, 首先以改进的核密度估计技术构建神经元的激活分布, 模拟神经元的正常工作范围. 接着采用邻域加权密度估计方法计算异常因子, 用以量化测试样本与神经元激活分布的偏离程度. 最终通过统计方法综合各神经元的异常因子作为样本的异常统计量, 为模型不确定性的评估提供新的视角. 实验结果涵盖多个公开数据集和模型, 通过可视化特征图直观展示本文方法在区分域内外样本方面的显著效果. 此外, 本文方法在域外检测任务中表现出卓越性能, AUROC 指标在多种实验设置下均超越其他现有方法, 验证提出方法的通用性和有效性.

关键词: 不确定性分析; 深度学习; 激活分布; 异常因子; 域外检测

引用格式: 雷雅婧.基于神经元统计建模分析的模型不确定性度量.计算机系统应用,2024,33(7):14–25. http://www.c-s-a.org.cn/1003-3254/9585.html

Model Uncertainty Measurement Based on Neuron Statistical Modeling and Analysis

LEI Ya-Jing

(School of Computer Science, Fudan University, Shanghai 200438, China)

Abstract: The uncertainty of neural networks reflects the predictive confidence of deep learning models, enabling timely human intervention in unreliable decision-making, which is crucial for enhancing system safety. However, existing measurement methods often require significant modifications to the model or training process, leading to high implementation complexity. To address this, this study proposes an uncertainty measurement approach utilizing neuron statistical modeling and analysis with activation values within a single forward propagation. An improved kernel density estimation technology is employed to construct neuron activation distributions and stimulate neuron normal operating range. Subsequently, a neighborhood-weighted density estimation method is utilized to calculate anomaly factors, effectively qualifying deviations of test samples from neuron activation distribution. Finally, by statistically combining the anomaly factors of each neuron, the cumulative anomaly factors of the sample provide a new perspective in assessing model uncertainty. Experimental results across multiple public datasets and models visually demonstrate the significant effectiveness of the proposed method in distinguishing between in-domain and out-of-domain samples through visualizing feature maps. Moreover, the method exhibits exceptional performance in out-of-domain detection tasks, with AUROC exceeding other methods across various experimental setups, validating its generality and effectiveness.

Key words: uncertainty analysis; deep learning; activation distribution; anomaly factor; out-of-domain detection

① 基金项目: 国家自然科学基金 (62106051); 上海浦江计划 (21PJ1400600); 国家重点研发计划 (2022YFC3601405); 上海研究与创新功能项目 (17DZ2260900)
收稿时间: 2024-02-02; 修改时间: 2024-03-05; 采用时间: 2024-03-19; csa 在线出版时间: 2024-05-31
CNKI 网络首发时间: 2024-06-04

© 中国科学院软件研究所 http://www.c-s-a.org.cn

深度学习技术凭借其卓越的预测性能已在图像识别[1–3]、自然语言处理[4–6]等领域得到广泛应用, 取得斐然的成果[7]. 然而, 神经网络作为代表性技术, 其黑盒特性使它们缺乏解释性和透明度, 无法提供关于预测置信度的信息[8]. 这一局限性在自动驾驶[9]和医疗诊断[10]等安全性要求严苛的应用领域内, 构成了导致灾难性后果的高风险隐患. 正如特斯拉无人驾驶因算法误将白色卡车识别为天空未能进行紧急制动而导致一人死亡的事故, 暴露深度学习算法一个致命问题: 传统模型只能输出特定预测结果, 而不能表明模型本身对决策结果是否自信. 因此, 深度学习不确定性的度量显现出高度重要性, 它本质上评估了模型对自身预测结果的置信程度. 如果模型在维持高性能预测的同时, 提供不确定性度量, 有助于关键时刻及时采取人工干预, 确保决策的可靠性. 这不仅能提升现有系统的安全性[11], 还能扩展深度学习技术在高风险领域的应用范围. 此外, 深度学习不确定性的研究深化了人类对神经网络内在机制的理解, 还可以在主动学习和强化学习任务中作为策略学习的重要组成部分, 以及在风险评估和环境监控等方面有广阔的应用前景.

深度学习不确定性根据产生原因通常被分为模型不确定性 (model uncertainty) 和数据不确定性 (data uncertainty). 模型不确定性, 或称认知不确定性 (epistemic uncertainty), 源自模型本身的不足, 由训练过程错误、模型结构问题或对未知样本的知识缺乏导致[12]. 如图 1 中两端高不确定性区域所示, 模型不确定性在缺乏训练样本的情况下大幅上升, 当未见过的样本加入模型参与训练后, 模型不确定性会降低. 数据不确定性, 亦称偶然不确定性 (aleatoric uncertainty), 它描述的是数据中内在噪声, 即由数据本身的随机性造成的不可预测的误差[13]. 如图 1 的虚框部分, 数据不确定性在样本模糊或背景繁杂等噪声高的情况下较大, 通常由数据采集过程中的信息损失导致. 例如, 通过特定分辨率的图像像素表示真实世界信息, 或是标注过程中的错误, 这些情况下的信息损失无法通过增加训练样本量减轻. 需要注意的是, 数据不确定性仅在讨论域内样本时有意义, 因为它本质上是量化样本被预测为不同类别的歧义程度[14]. 本研究聚焦于探究卷积神经网络的模型不确定性, 这有助于发现模型潜在的不足和改进方向, 对于优化模型的训练过程并提高其在面对未知数据时的表现具有重要意义.

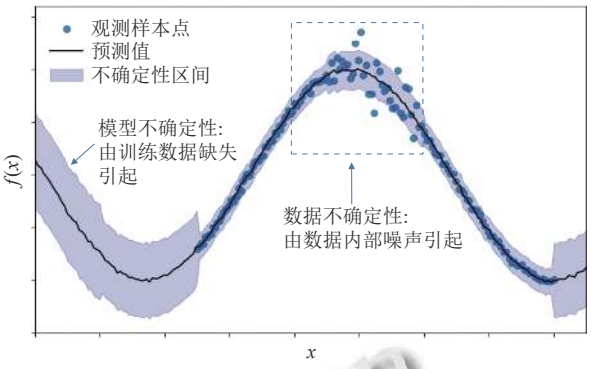

图 1 模型不确定性与数据不确定性图例

在现有不确定性度量方法中, 贝叶斯推断 (Bayesian inference)[15–18]和集成学习 (ensemble methods)[19–21]等传统方法虽然理论基础坚实, 但需要对模型架构或训练过程进行显著修改, 且实施复杂度高. 基于以上, 本文提出一种创新的基于单确定性神经网络 (single deterministic methods) 的不确定性度量框架 AFDU (anomaly factor based deterministic uncertainty). 单确定性神经网络在每次前向传递中提供相同的结果, 符合目前多数深度学习传统任务的模型构建情况. 本文提出方法通过统计建模分析, 充分利用神经网络在单次前向传播过程中的激活值, 无需更改模型结构或训练策略, 既维持原始模型的高准确度, 又实现高效的不确定性度量, 易于融入现有系统. 此外, AFDU 通过可视化分析增强模型面对未知或异常情况的可解释性.

本文的基础假设为模型经过良好训练后每个神经元会形成各自的正常工作范围, 超出该范围表明神经元碰到异常状况且不确定度提升. AFDU 有两个核心组成部分: 激活分布构建模块与异常因子计算模块. 激活分布构建模块利用改进的核密度估计 (kernel density estimation, *KDE*) 技术[22]精准构建神经元的激活分布以模拟神经元的正常工作范围, 异常因子计算模块引入邻域加权密度估计方法计算异常因子以衡量样本与神经元激活分布的偏离程度. 本文在多个公开数据集和模型进行实验验证, 并对层间特征图进行可视化分析, 直观展示 AFDU 在区分域内外样本方面的显著效果. 此外, AFDU 在域外检测任务上的 AUROC 指标均超过 98%, 在多种实验设置中均超越其他方法, 验证本文提出方法在识别异常样本方面的有效性和通用性, 展现其在实际应用中的巨大潜力.

1 相关工作

在深度学习的不确定性估计领域, 主流方法包括

© 中国科学院软件研究所 http://www.c-s-a.org.cn

贝叶斯推断[15,16]、集成算法[19–21]和测试数据增强 (test-time augmentation methods)[23,24]. 贝叶斯推断通过为模型参数引入概率分布, 提供一种量化不确定性的理论框架. 然而, 由于在实际应用中参数量巨大, 严格根据贝叶斯公式计算不现实. 因此, 研究者通常采用各种形式的变分推理[17,18]或蒙特卡洛采样[25]作为近似解, 这些方法虽然有效但需要重新训练模型, 计算资源和时间成本高. 集成算法, 如深度集成学习, 通过结合多个模型的预测结果提高鲁棒性, 同时使用统计度量来表达不确定性. 尽管集成方法已被证明在多个任务中有效, 但这些方法要求保存和处理多个模型或权重版本, 极大地增加了存储和计算负担. 测试数据增强则通过在测试时对输入数据进行多样化处理, 生成多个测试变体以评估不确定性. 这要求对每个测试样本进行多次前向传递, 并且可能需要对整个数据集进行特殊预处理, 导致计算时间增加, 并引入额外的噪声.

相比之下, 基于单确定性网络的不确定性度量方法提供一种更为高效和实用的解决方案. 与贝叶斯和集成方法相比, 这种方法的优势在于其高计算效率, 并能直接利用已训练模型的潜在激活特征. 单确定性神经网络的参数固定, 确保每次前向传播提供一致结果, 这与传统深度学习任务的模型构建相符. 单确定性方法大致分为两种路径, 一种是加入不确定性度量指标对单个网络进行明确建模和重新训练[26,27], 另一种是基于已训练网络利用额外成分进行不确定性度量[28], 如引入附加的神经网络用于不确定性估计[29]. 前者需要在原始模型上进行改动, 后者对不确定性的度量与原始预测任务是分离的, 更易于集成到现有系统中, 适用性更广泛. Hendrycks 等人[30]强调了神经元激活值在评估不确定性方面的作用, 指出激活值能反映模型对特定数据点的信心程度. Lee 等人[31]和 DeVries 等人[32]的工作通过分析神经网络输出的概率分布来估计不确定性, 但这些方法通常仅依赖于输出层的 Softmax 分布分析, 无法充分捕捉模型内部的复杂动力机制, 缺乏从整体评估模型的决策过程. 一些研究工作利用特征空间距离[33,34]和特征空间密度[35,36]度量不确定性, 在单确定性方法中提供新的视角. 未见数据点在特征空间中相对于训练数据点距离大且密度小, 而加入训练过程后再次计算, 距离降低且密度增大, 符合模型加入未知样本重新训练后模型不确定性降低的定义. DUQ[33]和 SNGP[34]两种方法分别通过引入径向基函数或高斯过程, 提出距离感知输出层, 并在特征提取的过程中加入归纳偏差以保证特征空间的平滑度和灵敏性, 但它们需要修改网络结构并重新训练模型. DDU[14]以训练样本在模型最后一个卷积层的输出特征向量为基础, 构建高斯混合模型 (Gaussian mixed model, GMM)[37]作为密度估计器量化模型不确定性. DDU 虽然能达到和集成模型相近的效果, 但仅在特征空间经过规范化处理后表现较好, 如有残差结构的模型.

本文提出的基于单确定性神经网络的不确定性度量框架, 无需对原始模型进行改动, 采用基于特征空间密度度量模型不确定性的思想. 利用改进的 *KDE* 方法更精确地构建神经元激活分布, 充分挖掘激活值的潜在特征. 通过计算与统计测试样本在各神经元的异常因子, 衡量样本与分布的偏离程度, 从而能更全面地评估模型的不确定性.

2 模型的不确定性度量框架

在本文的基础假设下, 对于经过良好训练的模型, 每个神经元会识别特定的模式, 形成各自的正常工作范围, 超出该范围表明神经元碰到异常状况且不确定度提升. 神经元激活分布及样本异常因子示例如图 2 所示. 图 2 中圆形及叉形点由左至右依次表示异常因子 (anomaly factor, AF) 等于 7.85, 1.05, 1.86, 1.00, 1.02, 4.97, 11.68 和 20.74 时的对应情况. 基于训练集样本的激活值, 每个神经元形成的激活分布可以模拟其正常工作范围——正常数据点会落入神经元的工作范围内, 异常因子较低, 从而模型的不确定性较低, 如图 2 中圆形数据点; 相对地, 异常数据点会偏离神经元的工作分布, 异常因子较高, 导致模型的不确定性较高, 指示模型可能正面对之前未见过的新情况, 如图 2 中叉形数据点. 此处的正常和异常表现为数据点相对于训练集数据点的域内和域外.

本文研究的问题定义如下: 给定训练样本集 $x=\{x_1,x_2,\cdots,x_n\}$、预训练模型 M 和测试样本 t, 其中 n 代表训练集的样本总数. 记 $a_l^N(x_i)\in R^1$ 为第 i 个训练样本 x_i 在神经网络第 l 层中第 N 个神经元的输出激活值, $a_l^N(x)=\left\{a_l^N(x_1),a_l^N(x_2),\cdots,a_l^N(x_n)\right\}$ 为训练样本集在该神经元的激活值集合. 研究目标为充分挖掘并利用神经元激活值的潜在分布特征, 构建神经元正常工作范

© 中国科学院软件研究所 http://www.c-s-a.org.cn

围, 即激活分布 $f_l^N(x)$. 当测试样本 t 输入模型时, 在各神经元上的输出值 $a_l^N(t)$ 分别计算得到异常因子 $WAF_l^N(t)$, 用以衡量与激活分布的偏离程度. 最终整合所有神经元的异常因子得到测试样本的不确定性度量 $u(t)$.

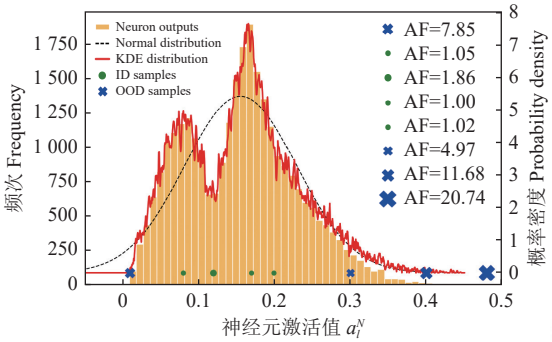

图 2 神经元激活分布及样本异常因子示例

本文提出的基于神经元统计建模分析的不确定性度量框架 AFDU 如图 3 所示. 上方虚线框部分表示卷积神经网络模型的原始分类任务过程, 其训练和测试阶段不会受到不确定性度量过程影响. 下方虚线框部分表示基于已训练模型的不确定性度量过程, 由两个核心组成部分构成: 神经元激活分布构建模块与异常因子计算模块. 神经元激活分布构建阶段, 以训练集样本在模型各神经元上的激活值 $a_l^N(x)$ 作为输入, 构建激活分布 $f_l^N(x)$ 以模拟神经元正常工作范围. 在异常因子计算阶段, 将测试样本在各神经元上的激活值 $a_l^N(t)$ 作为输入, 引入邻域加权的密度估计方法, 有效衡量各神经元中测试样本的激活值与激活分布的偏离程度, 得到异常因子 $WAF_l^N(t)$. 最终使用统计方法综合各神经元的异常因子, 层内取平均, 层间求和, 为每个样本计算得到一个综合的异常统计量, 即最终的不确定性度量 $u(t)$. AFDU 不仅增强了对模型内部决策过程的理解, 而且提供了一个有效的工具, 有助于在复杂应用中高效监控和评估模型的不确定性. 算法流程总结如算法 1.

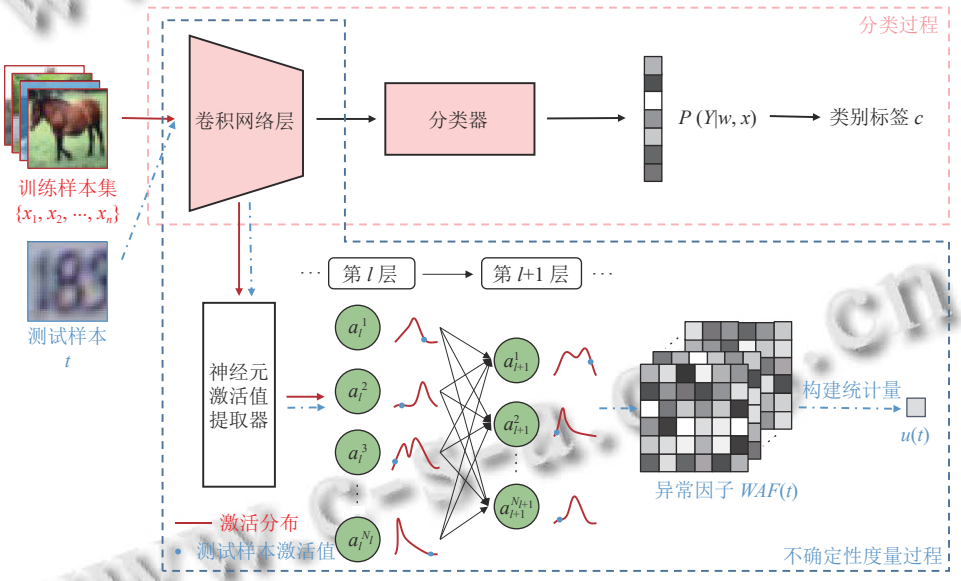

图 3 基于神经元统计建模分析的不确定性度量框架 AFDU 示意图

算法 1. 基于神经元统计建模分析的不确定性度量框架

输入: 训练数据 x, 预训练模型 M, 测试样本 t, 其中模型共 L 层, 每层 N_l 个神经元.

输出: 测试数据的不确定性度量值 $u(t)$.

1) 获得训练数据集在每个神经元上的激活值 $a_l^N(x)$.

2) 利用改进的核密度估计方法构建神经元的工作分布 $f_l^N(x)$, 见式 (3).

3) 计算测试样本在每个神经元上激活值 $a_l^N(t)$ 的局部密度 $KDE_l^N\left(a_l^N(t)\right)$, 见式 (7).

4) 计算测试样本在每个神经元上近邻集的加权邻域密度 $WDE_l^N\left(a_l^N(t)\right)$, 见式 (8).

5) 计算测试样本在每个神经元上的异常因子 $WAF_l^N\left(a_l^N(t)\right)$, 见式 (6).

6) 整合各神经元上的异常因子, 层内取平均, 层间求和, 获得一个综合的异常统计量, 即不确定性度量 $u(t)$.

$$u(t)=\sum_{l=1}^{L}\left(\frac{1}{N_l}\sum_{N=1}^{N_l}WAF_l^N\left(a_l^N(t)\right)\right)$$

2.1 神经元激活分布构建模块

在神经网络的训练过程中, 每个神经元都会学习并适应特定的模式或特征, 这些模式和特征在模型内部由神经元的激活值所反映. 为了更精确地评估模型

© 中国科学院软件研究所 http://www.c-s-a.org.cn

的不确定性, 本文首先定义和构建每个神经元的正常工作范围. 这一范围可以看作是训练样本在神经元上激活值的分布, 记为$f_l^N \sim F\left(a_l^N(x_1), a_l^N(x_2), \cdots, a_l^N(x_n)\right)$, 其中$a_l^N(x_i) \in R^1$是第 i 个训练样本在模型第 l 层的第 N 个神经元的激活值, n 代表训练数据集中的样本总数.

由于无法获取神经元激活值的先验分布, 传统的概率密度估计方法可能不够准确. 因此, 为了更好地模拟神经元的激活分布, AFDU 使用改进的核密度估计方法. *KDE* 作为一种无参数的概率密度估计方法, 能够捕捉数据的微妙结构和模式. 具体而言, *KDE* 会在每个观测数据点周围放置一个核函数, 然后将这些核函数叠加, 从而获得整体的概率密度函数. 通过调整核函数的带宽, 可以控制密度估计的平滑程度, 从而在偏差与方差之间寻找一个平衡点. 公式如下:

$$f_l^N(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - a_l^N(x)}{h}\right) \tag{1}$$

其中, $K(\cdot)$ 是核函数, 此处选择高斯核. 高斯核的形状呈正态分布, 能够为估计提供平滑且连续的结果. 此外, 由于其具有无限支持的特性, 在整个实数轴上都有定义且均不为零, 高斯核可以敏锐地捕获数据的微妙变化. 高斯核的表达式如下:

$$K(t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}} \tag{2}$$

式 (1) 中的h是带宽, 带宽的选择对 *KDE* 的结果有着至关重要的影响. 太小的带宽会导致构建分布的过度拟合, 而太大的带宽则会掩盖数据的真实结构. AFDU 采用的 improved Sheather Jones (ISJ) 算法[38]是一种自适应方法, 能够根据数据的实际分布自动确定一个近似最优的带宽, 从而使得 *KDE* 的估计既不过度拟合也不过于平滑. 同时, 在图 2 中可见, 利用 *KDE* 构建的分布 (实线) 相比于高斯分布的估计 (虚线) 表现出更高的精准度. *KDE* 方法能够更细致地捕捉数据的复杂结构, 从而生成更符合实际数据分布的曲线, 尤其在处理非标准形状或多峰分布时表现突出, 因此更适用于神经元激活先验分布未知的场景.

为了避免因训练集样本量过大而使后续数据处理变得过于复杂, 本文采取预处理措施. 具体来说, AFDU 首先对训练数据集在每个神经元上的激活值输出进行 K-means 聚类[39], 将样本空间分解为 500 个近邻簇, 即原始的激活值空间被简化为 500 个中心点$\{a_l^N(c_1), a_l^N(c_2), \cdots, a_l^N(c_{500})\}$. 相较于不使用聚类方法时的 n 个数据点 (n 一般大于 10 000), 大幅减少了数据的维度和复杂性. 在构建 *KDE* 时, 考虑到每个簇中样本的数量可能不同, 不同的簇对于整体分布的贡献程度也会有所不同. 因此, AFDU 根据每个簇的中心点的数值$a_l^N(c_i)$以及相应簇中的样本数量n_i进行加权计算, 确保在估计整体分布时每个样本簇得到适当的表示. 这种处理方法不仅简化了数据和计算复杂度, 还确保了分布估计的准确性, 兼顾了效率与准确度. 最终的构建分布公式如下:

$$f_l^N(x) = \frac{1}{\left(\sum_{i=1}^{C}\omega_i\right)h}\sum_{i=1}^{C}\omega_i K\left(\frac{x - a_l^N(c_i)}{h}\right) \tag{3}$$

其中,

$$\omega_i = n_i / n \tag{4}$$

$$C = \min(500, n) \tag{5}$$

2.2 异常因子计算模块

在得到神经元激活分布f_l^N后, 需要衡量测试样本点的异常程度. 尽管测试样本输入后在神经元上的激活值可以根据已构建的分布计算得到相应的概率密度函数 (probability density function, pdf), 但不同神经元经过 *KDE* 构建的分布尺度 (scale) 不同, 因此无法直接使用这些 pdf 值来量化异常程度. 本文受 Hu 等人[40]异常检测任务的启发, 引入“异常因子”这一指标. 不同的是, 本文关注于将该指标用于统一度量各样本在不同神经元激活分布上的异常程度.

AFDU 中的异常因子计算是基于两个主要组件的比值: 局部核密度估计 (*KDE*) 和加权邻域密度估计 (weighted neighbor density, *WDE*). 前者表示测试样本自身在该神经元上的激活值在激活分布上对应的 pdf 值, 后者表示此激活值一定范围内的邻域训练数据点对应的 pdf 值加权后的统计量. 直观地说, 正常的数据点基本位于训练数据点集的密集区域, 意味着它们局部密度高且与其邻域数据点的 pdf 值接近. 而异常数据点通常位于训练数据点集的稀疏区域, 即局部密度低且与邻域数据点的 pdf 值有较大差异. 将数据点$p = a_l^N(t)$记为测试样本在模型第 l 层的第 N 个神经元的激活值. 则该点的异常因子 (weighted anomaly factor,

© 中国科学院软件研究所 http://www.c-s-a.org.cn

WAF) 定义如下:

$$WAF_l^N(p)=\left(\frac{WDE_l^N(p)}{KDE_l^N(p)}\right) \tag{6}$$

其中, $KDE_l^N(p)$表示数据点 p 的局部核密度, 即由第 1 阶段神经元构建的分布直接计算得出:

$$KDE_l^N(p)==\frac{1}{\left(\sum_{i=1}^{C}\omega_i\right)h}\sum_{i=1}^{C}\omega_i K\left(\frac{p-a_l^N(c_i)}{h}\right) \tag{7}$$

式 (6) 中$WDE_l^N(p)$表示数据点 p 的加权邻域密度. 计算公式如下:

$$WDE_l^N(p)=\frac{\sum_{q\in N_k(p)} w_q\cdot KDE_l^N(q)}{\sum_{q\in N_k(p)} w_q} \tag{8}$$

其中, $N_k(p)=\left\{a_l^N\left(c_{j1}\right),a_l^N\left(c_{j2}\right),\cdots,a_l^N\left(c_{jk}\right)\right\}$表示该神经元上测试样本激活值$a_l^N(t)$的 k 个近邻集. 近邻集中的元素来源于训练数据集在该神经元的激活值中与$a_l^N(t)$最接近的 k 个数据点. 邻域数据点$q\in N_k(p)$的权重 w_q 定义如下:

$$w_q=\exp\left(-\frac{\left(\frac{d_k(q)}{min_k}-1\right)^2}{2\sigma^2}\right) \tag{9}$$

其中, σ为正的缩放因子, $d_k(q)$表示数据点 q 的 k 距离, 即离 q 最近的第 k 个数据点到 q 的距离. k 距离可以描述数据点的局部密度, k 距离越大, 意味该数据点所在区域的局部密度越低. min_k则为属于 p 的近邻点集$N_k(p)$中最小的 k 距离, 即:

$$min_k=\min_{q\in N_k(q)}(d_k(q)) \tag{10}$$

加权邻域密度 *WDE* 以每个数据点的 k 距离为基础, 将邻域数据点的权重设定为其 k 距离的单调递减函数. 这意味着数据点 p 的 k 个近邻点中, 局部密度越大的邻域数据点有更大的权重 w, 而局部密度越小的邻域数据点具有更小的权重 w. 如图 4 所示, 阴影部分越深的区域密度越高 (pdf 值越大, k 距离越小), 对于域内测试样本点 (p_{id}) 而言, 其 k=3 的邻域范围内, 邻域数据点越靠近数据密集区, pdf 值越大, k距离越小, 其相对于测试点的权重就越大, 即$w_{i1},w_{i2}>w_{i3}$. 该权重的设置增大了位于数据密集区域数据点的影响, 即更有效捕捉测试点和整体分布的偏差.

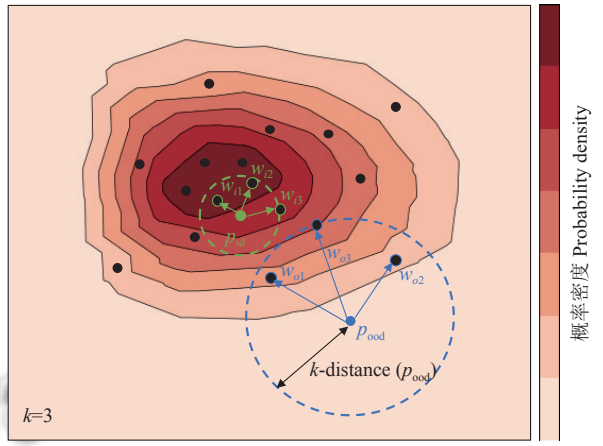

图 4 异常因子计算图

异常因子 (*WAF*) 的值反映了测试数据点的局部核密度与其邻域密度之间的关系. 局部核密度较小且邻域密度较大的数据点, 计算得到的异常因子较大. 值得注意的是, 对于大部分与群集明显分离的异常数据点, 它们的局部密度 (*KDE*) 与邻域密度 (*WDE*) 的差异明显, 即其异常因子 (*WAF*) 会远大于 1. 相反, 对于大部分正常数据点, 它们的局部密度与邻域密度非常接近, 这使得它们的异常因子围绕 1 波动. 如图 4 所示, 域内数据点 (p_{id}) 位于数据的密集区域, 其局部密度高且 k 距离小, 属于其k近邻集的数据点局部密度较高且与域内数据点 (p_{id}) 接近, 因此异常因子接近于 1; 而域外数据点 (p_{ood}) 位于数据的稀疏区域, 其局部密度低且 k 距离大, 权重高的邻域点局部密度大, 与域外数据点 (p_{ood}) 自身局部密度的比值会远大于 1. 基于此, 异常因子的计算值能有效反映测试数据点与训练数据点分布的偏差.

3 实验分析

为了验证提出的不确定性度量框架 AFDU 的有效性, 本文采取了一系列实验措施, 着重从特征图的可视化以及域外样本检测 (out-of-domain detection) 任务的角度, 评估基于模型不确定性构建的异常因子在区分域内和域外样本方面的性能. 本文选择多组公开数据集与模型进行实验验证, 数据集包括 MNIST[41]、CIFAR-10[42]、FER2013[43]、SVHN[44]、CIFAR-100[42], 模型架构包括 LeNet[41]、VGG19[23]、Wide-ResNet-28-

© 中国科学院软件研究所 http://www.c-s-a.org.cn

10[45], 以验证提出框架的通用性.

在特征图可视化分析中, 可以观察到域外样本上的异常因子显著高于域内样本, 而且更集中地出现在背景等非关键判别区域及目标主体的轮廓上. 这表明本文方法能够有效识别模型在处理未知数据时的不确定性区域.

在域外样本检测任务中, 本文综合每个样本在各神经元上计算得到的异常因子, 为每个样本生成一个异常统计量. 以该异常统计量作为区分域内外样本的关键指标分数, 实验结果表明本文方法在这一任务上的评价指标结果优于现有的其他不确定性度量方法, 体现该框架的实用性和优越性.

3.1 特征图可视化分析

为了深入理解本文的不确定性度量框架如何识别并展示模型的不确定性, 本文进行了特征图的可视化分析. 这一过程能够揭示域内外样本在不同神经元激活分布上计算得到的异常因子的分布差异.

本文在已训练模型中提取了样本在网络层间的特征图, 这些特征图代表了模型在不同阶段的理解和内部表示. 通过将异常因子映射到这些特征图上, 能够直观地观察到域内样本与域外样本在特征空间中神经元激活值的差异. 图 5 和图 6 所展示的例子为模型浅层特征图的示例, 特征图每个像素点的黑白亮度代表样本在相应神经元上激活值的大小, 彩色由紫至红表示异常因子的值由小到大. 在域内样本的特征图中, 异常因子突出的区域较为稀疏, 且异常因子值较低, 表明模型对于这些样本具有较高的置信度和较低的不确定性. 相反, 在域外样本的特征图中, 由于域外样本在多数神经元上的激活值偏离了模型在训练过程中学习到的数据分布, 异常因子突出区域更为集中和显著, 且异常因子值普遍较高, 揭示模型对这些样本的不确定性增加.

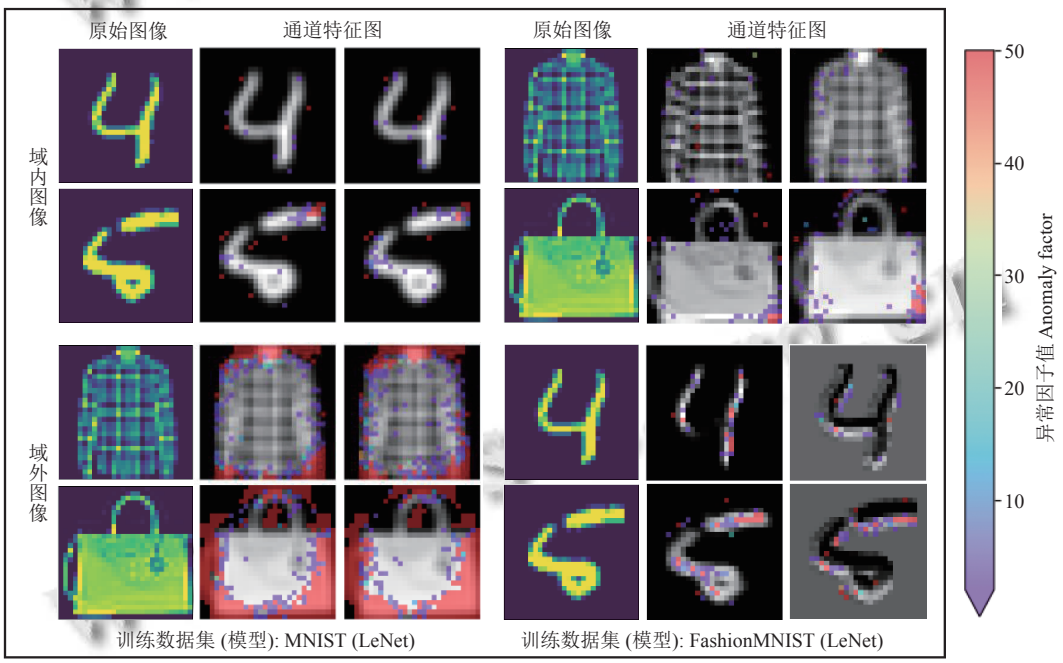

图 5 特征图的异常因子可视化分析 – LeNet 以 MNIST-FashionMNIST 作为域内外测试对

图 5 展示的结果基于的实验设置如下: LeNet 模型分别对 MNIST 数据集和 FashionMNIST 数据集进行训练, 得到两个高准确率的预训练模型, 并以对方的数据集作为域外样本进行实验. MNIST 是手写数字的简单灰度图像, 而 FashionMNIST 是更复杂的时尚产品图像, 两者明显属于不同的数据领域. 观察结果如下: 图 5 左半部分为以 MNIST 作为训练集且分类准确率达到 98.37% 的 LeNet 模型, 上半部分为域内测试样本, 下半部分为域外测试样本, 且分别展示了原始图像及浅层特征图. 在 MNIST 数据集上训练的 LeNet 模型预期在处理手写数字图像时显示出较低的不确定性, 这在实验中得到了验证. 神经元的激活模式显示, 对于 MNIST 样本, 异常因子突出的区域较为稀疏, 表明模型对这些样本的预测有较高置信度. 然而, 当同一模型

© 中国科学院软件研究所 http://www.c-s-a.org.cn

处理 FashionMNIST 样本时, 特征图展示了显著不同的激活模式. 异常因子突出的区域数量和密集度显著增加, 表明模型在试图对时尚商品的图像进行分类时显示出更高的不确定性, 体现 FashionMNIST 图像与模型训练时的 MNIST 手写数字图像在视觉特征上有显著差异. 在交换域内和域外数据集的对比实验中, 当 FashionMNIST 数据集用于训练 LeNet 模型时, 相同的图像从域外转为域内样本, 异常因子的值和数量都大幅下降. 当 MNIST 数据集用作域外测试集时, 特征图中的异常激活区域数量明显增多, 密度也更高. 实验结果强调不同数据集之间的激活差异, 揭示本文提出的不确定性度量框架能够有效地区分域内和域外数据.

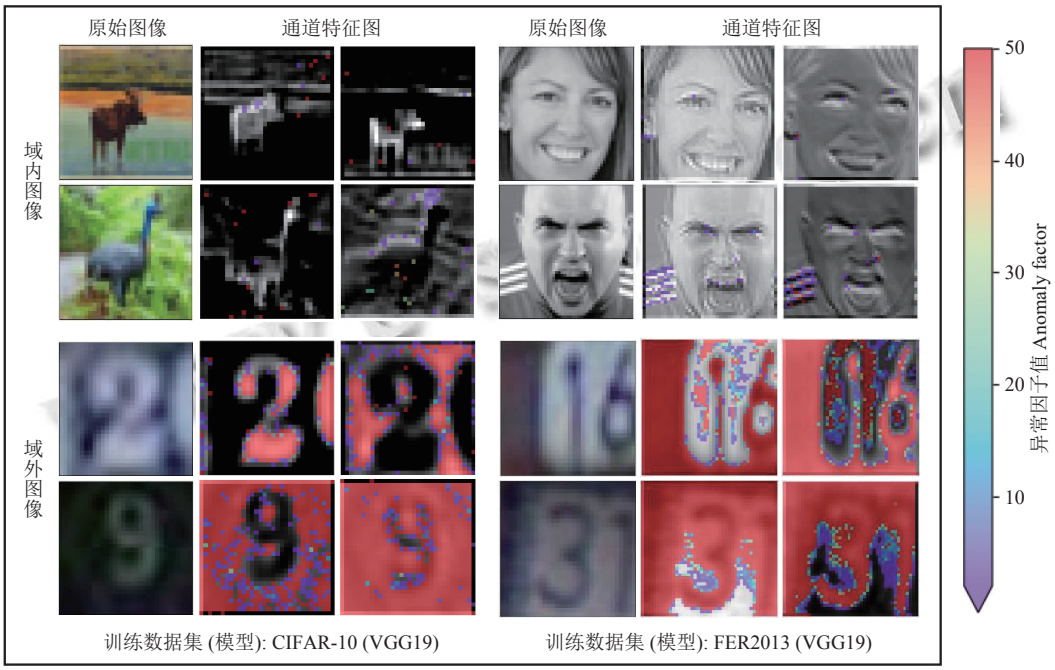

图 6 特征图的异常因子可视化分析, VGG19 以 CIFAR-10-SVHN 与 FER2013-SVHN 作为域内外测试对

为验证所提出框架的通用性, 图 6 呈现更复杂数据集和模型的实验结果. 在实验中, 本文使用 VGG19 模型分别对两个不同的数据集进行了训练: CIFAR-10 自然图像数据集, 模型分类准确率达到 92.86%; FER2013 表情图像数据集, 模型分类准确率为 73.03%. 此外, 使用 SVHN 交通标识数据集作为域外测试集. 当模型处理域内数据集的图像时, 异常因子值较低且数量较少, 表现出模型对这类数据有很好的理解和置信度. 然而, 面对 SVHN 域外数据集的图像时, 异常因子值与数量的增加再次表明模型在处理与训练分布不同的数据时遇到阻碍, 不确定性显著提高. 通过这两种情况的比较, 本文的框架在不同类型的图像数据集上均能有效展示模型的不确定性水平. 由此可见, 尽管数据的特性和模型的训练目标不同, 本文方法仍旧能够捕捉样本在神经元激活模式中的异常, 进一步验证了提出框架在不同数据域的通用性和适应性.

值得注意的是, 样本的异常因子在特征图的背景等非关键判断区域或者边缘区域更加凸显, 比如图 6 中鸟类头顶区域以及男子衣服的横条区域. 这表明模型对未见过的数据存在较高的不确定性, 以及这些区域的神经元可能对识别新颖或异常的模式更为敏感. 同时, 虽然 VGG19 模型框架在 FER2013 数据集上训练得到的模型准确度不高, 但异常因子的可视化依然能明显区分域内和域外特征图的差异.

3.2 域外样本检测

域外样本检测任务是评估模型不确定性指标有效性的关键测试场景. 面对模型未曾遇到的域外样本, 预期这些样本会展现出与域内样本截然不同的较高不确定性特征, 在本研究中特指为异常统计量. 测试样本在已训练模型各神经元上计算得到异常因子后, 通过层内取平均、层间求和的方式最终为每个样本整合生成一个异常统计量. 利用该统计量本文能够区分并识别导致模型预测不准确的未知样本, 从而验证本文所提出的不确定性度量框架的有效性.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

本项实验的设定包括: (1) LeNet 模型架构 + MNIST 训练集 + FashionMNIST 域外测试集; (2) VGG19 模型架构 + CIFAR-10 训练集+ SVHN/CIFAR-100 域外测试集; (3) WRN-28-10 模型架构 + CIFAR 训练集 + SVHN/CIFAR-100 域外测试集; (4) VGG19 模型架构 + FER2013 训练集 + SVHN 域外测试集. 该实验设定遵循由简至繁验证本文提出方法的通用性. 通过 (2)、(3) 的比较可以分析相同数据集、不同模型架构的域外检测量化结果差异; 通过 (2)、(4) 的比较可以分析不同数据集、相同模型架构的域外检测量化结果差异. 评价指标使用 AUROC (area under the receiver operating characteristic curve), 表示不同阈值下模型正确区分域内外样本的能力, 取值范围为 0–1, 值越高表示模型的识别能力越强.

本文使用模型的 Softmax 熵作为鉴别模型不确定性的基准方法, 并与以下无需在域外数据上训练或微调的方法进行对比, 方法包括 Energy-based[36]、DDU[14]、DUQ[33]、SNGP[34]、5-Ensemble, 前四者属于单确定性方法, 后者属于目前不确定性评估效果最好的集成方法. Energy-based 模型以未经归一化的 Softmax 密度作为模型不确定性, 即最后一层卷积输出的 LogSumExp. DDU 利用高斯混合模型在最后一层卷积输出上构建分布, 根据密度度量分辨域内外样本. DUQ 和 SNGP 分别引入径向基函数或高斯过程构建距离感知输出层, DUQ 以核估计距离度量模型不确定性, SNGP 以预测熵作为度量指标. 5-Ensemble 使用 5 个网络组成的集成模型预测熵作为模型不确定性的度量.

表 1 为简单数据集 MNIST 手写图像和简单模型 LeNet 在域外样本检测任务的定量指标结果, 表 2 为更复杂且接近真实应用场景的数据集和模型的相关实验结果. 表格中的实验数据展示模型对测试数据集的分类准确度以及在域外数据集上的 AUROC 分数. 实验结果表明, 本文提出的 AFDU 在不影响分类准确度的同时, 基于已训练模型, 通过采用异常统计量, 实现对域内外样本的高精度鉴别, AUROC 评价指标超越了所有对比方法, 并接近于 1. 实验结果验证本文提出框架 AFDU 的有效性以及在多样化数据集及模型架构中的广泛适用性. 尽管 5-Ensemble 方法因涉及多个预测模型在分类准确度上均达到最高值, 但单确定性方法仅需基于已训练模型便可获得不确定性度量, 分类准确率同样保持在较高的水平. 同时, 本文提出的 AFDU 在 AUROC 评价指标上的表现也超过了 5-Ensemble 集成方法, 表明本文提出框架在保证原模型分类准确度的同时, 能更高效获得不确定性度量.

表 1 域外样本检测定量指标结果比较 (MNIST+LeNet)(%)

方法	模型不确定性	分类准确度	AUROC (FashionMNIST)
Softmax	Softmax熵	98.37	95.46
Energy-based	Logits LogSumExp		95.84
DDU	GMM密度		98.67
AFDU (本文)	**KDE+异常因子**		**99.94**
DUQ	核估计距离	97.86	96.35
SNGP	预测熵	98.51	97.18
5-Ensemble	集成预测熵	**98.94**	98.82

表 2 基于 CIFAR-10 数据集的实验, 使用不同的模型架构 VGG19 和 WRN-28-10, 观察各个不确定性度量方法的表现差异. 实验结果发现, 本文提出的 AFDU 不确定性度量框架在 AUROC 指标的表现上均超过其他方法, 在不同的模型中都能准确识别域内外样本. 尤其在 VGG19 的模型架构表现上, 在其他方法 AUROC 的指标只能在 90% 的情况下, AFDU 仍能达到超过 96% 的水平. 虽然 DDU 在 WRN-28-10 的表现上也能达到较高的 AUROC 值, 但是在 VGG19 的表现差强人意, 这是由于 DDU 方法受特征崩塌 (feature collapse) 的限制, 需要在模型中引入残差连接缓解该问题.

表 2 基于 VGG19 模型架构的实验, 比较不同数据集 CIFAR-10 自然图像和 FER2013 人脸表情图像的表现. 实验结果展示本文提出的 AFDU 不确定性度量框架在不同类型的图像数据集上均能有效区分域内样本和域外样本, 且在两个数据集上的 AUROC 指标都达到了所有方法中的最高值, 进一步验证了提出框架在不同数据域的通用性和适应性.

在深入分析样本的异常统计量时, 注意到域外样本的异常统计量显著高于域内样本, 超出量达到 10^5 个数量级. 这一观察验证了以下重要观点: 未受训练的域外样本在模型多数神经元上的激活值会明显偏离它们的激活分布, 导致异常统计量的显著提升. 换言之, 样本的异常统计量揭示了样本与模型激活分布的偏离程度, 反映模型面对未知数据时的不确定性水平. 这一点不仅加强了本文方法的合理性, 也凸显其在识别不确定、异常情境时的有效性.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

表 2 域外样本检测定量指标结果比较 (CIFAR-10/FER2013+VGG19/WRN-28-10)(%)

训练数据集 (模型)	方法	模型不确定性	分类准确度	AUROC (SVHN)	AUROC (CIFAR-100)
CIFAR-10 (VGG19)	Softmax	Softmax熵	92.86	86.33	83.68
	Energy-based	Logits LogSumExp		85.46	83.87
	DDU	GMM密度		90.07	87.54
	AFDU (本文)	**KDE + 异常因子**		**97.31**	**96.58**
	DUQ	核估计距离	92.24	88.97	84.06
	SNGP	预测熵	93.51	87.55	85.21
	5-Ensemble	集成预测熵	**94.27**	93.43	90.36
CIFAR-10 (WRN-28-10)	Softmax	Softmax熵	95.64	94.51	89.24
	Energy-based	Logits LogSumExp		94.86	89.32
	DDU	GMM密度		97.76	91.47
	AFDU (本文)	**KDE + 异常因子**		**99.17**	**98.86**
	DUQ	核估计距离	94.17	94.86	88.43
	SNGP	预测熵	95.56	95.20	91.05
	5-Ensemble	集成预测熵	**96.53**	97.23	93.01
FER2013 (VGG19)	Softmax	Softmax熵	73.03	89.83	—
	Energy-based	Logits LogSumExp		90.94	—
	DDU	GMM密度		96.22	—
	AFDU (本文)	**KDE + 异常因子**		**98.47**	—
	5-Ensemble	集成预测熵	**75.12**	95.16	—

4 结论与展望

本文针对深度学习模型由于透明度和解释性低导致决策不可信、在安全性要求高领域应用受限的问题, 提出一种新颖的基于神经元统计建模分析的不确定性度量方法. 该方法能够在保证原始预测任务高性能的前提下, 无需修改模型结构或训练过程, 高效且准确评估模型不确定性. 该框架由神经元激活分布构建和异常因子计算两大核心部分构成, 为每个样本构建异常统计量以作为模型不确定性的度量指标. 算法具体利用改进的核密度估计技术, 挖掘训练样本在神经元激活值中的潜在分布特征, 构建激活分布. 引入邻域加权密度估计方法计算异常因子, 精准量化测试点与激活分布偏离程度. 通过可视化实验, 本研究展示异常因子在区分域内外样本方面的显著效果. 在域外检测任务中, 本文提出方法在多种实验设置下性能均超越其他已有方法, 验证本文方法的通用性和有效性. 未来研究计划将进一步探索和改进现有工作在更复杂应用场景的性能和效率, 具体考虑以下几个方向的拓展. (1) 探究神经元间的复杂关联性并构建多元分布: 现有工作基于神经元相互独立的假设, 需改进对单神经元构建激活分布而未考虑神经元间联系的局限性. (2) 实时性提升: 尽管与其他方法相比, 本文提出方法已经降低了存储和计算资源的消耗, 但在实际应用中, 面对巨大的网络和数据规模时, 快速准确度量不确定性依旧复杂而充满挑战. (3) 跨领域应用: 将本文提出的不确定性度量方法应用到更多领域, 如自然语言处理等其他任务, 验证和优化方法在不同任务中的通用性和有效性.

参考文献

1 Cheng G, Lai PJ, Gao DC, *et al*. Class attention network for image recognition. Science China Information Sciences, 2023, 66(3): 132105. [doi: 10.1007/s11432-021-3493-7]

2 Bharadiya JP. Convolutional neural networks for image classification. International Journal of Innovative Science and Research Technology, 2023, 8(5): 673–677. [doi: 10.5281/zenodo.8020781]

3 李文静, 白静, 彭斌, 等. 图卷积神经网络及其在图像识别领域的应用综述. 计算机工程与应用, 2023, 59(22): 15–35. [doi: 10.3778/j.issn.1002-8331.2302-0273]

4 Khurana D, Koli A, Khatter K, *et al*. Natural language processing: State of the art, current trends and challenges. Multimedia Tools and Applications, 2023, 82(3): 3713–3744. [doi: 10.1007/s11042-022-13428-4]

5 Bharadiya JP. A comprehensive survey of deep learning techniques natural language processing. European Journal of Technology, 2023, 7(1): 58–66. [doi: 10.47672/ejt.1473]

6 谢宇鹏, 骆昱宇, 冯建华. Navi: 基于自然语言交互的数据分析系统. 软件学报, 2024, 35(3): 1194–1206. [doi: 10.13328/

© 中国科学院软件研究所 http://www.c-s-a.org.cn

j.cnki.jos.007074]
7 Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press, 2016. 800.
8 Leibig C, Allken V, Ayhan MS, *et al*. Leveraging uncertainty information from deep neural networks for disease detection. Scientific Reports, 2017, 7(1): 17816. [doi: 10.1038/s41598-017-17876-z]
9 崔冰艳, 李贺, 崔哲, 等. 智能网联汽车换道决策安全性研究综述. 交通信息与安全, 2023, 41(4): 1–13. [doi: 10.3963/j.jssn.1674-4861.2023.04.001]
10 姚琼, 王觅也, 师庆科, 等. 深度学习在现代医疗领域中的应用. 计算机系统应用, 2022, 31(4): 33–46. [doi: 10.15888/j.cnki.csa.008411]
11 Raji ID, Dobbe R. Concrete problems in AI safety, revisited. Proceedings of the 2020 International Conference on Learning Representations (ICLR). Washington: ICLR, 2020. 1–6. [doi: 10.48550/ARXIV.2401.10899]
12 Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning. New York: JMLR.org, 2016. 1050–1059.
13 Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017). Long Beach: Curran Associates Inc., 2017. 5580–5590.
14 Mukhoti J, Kirsch A, van Amersfoort J, *et al*. Deep deterministic uncertainty: A new simple baseline. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023. 24384–24394.
15 Blundell C, Cornebise J, Kavukcuoglu K, *et al*. Weight uncertainty in neural networks. Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR.org, 2015. 1613–1622.
16 Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature, 2015, 521(7553): 452–459. [doi: 10.1038/nature14541]
17 Zhang C, Bütepage J, Kjellström H, *et al*. Advances in variational inference. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 2008–2026. [doi: 10.1109/TPAMI.2018.2889774]
18 Mostafa B, Hassan R, Mohammed H, *et al*. A review of variational inference for Bayesian neural network. In: Masrour T, Ramchoun H, Hajji T, *et al*. eds. Artificial Intelligence and Industrial Applications: Algorithms, Techniques, and Engineering Applications. Cham: Springer, 2023. 231–243.
19 Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6405–6416.
20 Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2261–2269.
21 Ganaie MA, Hu MH, Malik AK, *et al*. Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 2022, 115: 105151. [doi: 10.1016/j.engappai.2022.105151]
22 Venables WN, Ripley BD. Modern Applied Statistics with S-PLUS. 3rd ed., New York: Springer, 2013.
23 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego, 2015.
24 Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929–1958.
25 Osband I, Aslanides J, Cassirer A. Randomized prior functions for deep reinforcement learning. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 8626–8638.
26 Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 3183–3193.
27 Malinin A, Gales M. Predictive uncertainty estimation via prior networks. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 7047–7058.
28 Raghu M, Blumer K, Sayres R, *et al*. Direct uncertainty prediction for medical second opinions. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 5281–5290.
29 Ramalho T, Miranda M. Density estimation in representation space to predict model uncertainty. Proceedings of the 3rd International Workshop Engineering Dependable and Secure

© 中国科学院软件研究所 http://www.c-s-a.org.cn

Machine Learning Systems. New York: Springer, 2020. 84–96.

30 Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. Proceedings of the 5th International Conference on Learning Representations (ICLR). Washington: ICLR, 2019. 1–12.

31 Lee K, Lee K, Lee H, *et al*. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 7167–7177.

32 DeVries T, Taylor GW. Improved regularization of convolutional neural networks with cutout. arXiv:1708.04552, 2017.

33 van Amersfoort J, Smith L, Teh YW, *et al*. Uncertainty estimation using a single deep deterministic neural network. Proceedings of the 37th International Conference on Machine Learning. JMLR.org, 2020. 898.

34 Liu JZ, Lin Z, Padhy S, *et al*. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 7498–7512.

35 Postels J, Blum H, Strümpler Y, *et al*. The hidden uncertainty in a neural networks activations. Proceedings of the 2021 International Conference on Machine Learning (ICML)—2021 Workshop on Uncertainty & Robustness in Deep Learning. New York: 2021. 1–18.

36 Liu WT, Wang XY, Owens JD, *et al*. Energy-based out-of-distribution detection. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1802.

37 Maugis C, Celeux G, Martin-Magniette ML. Variable selection for clustering with Gaussian mixture models. Biometrics, 2009, 65(3): 701–709. [doi: 10.1111/j.1541-0420.2008.01160.x]

38 Botev ZI, Grotowski JF, Kroese DP. Kernel density estimation via diffusion. The Annals of Statistics, 2010, 38(5): 2916–2957.

39 MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: Statistical Laboratory of the University of California, 1967. 281–297.

40 Hu WM, Gao J, Li B, *et al*. Anomaly detection using local kernel density estimation and context-based regression. IEEE Transactions on Knowledge and Data Engineering. 2020, 32(2): 218–233.

41 Lecun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]

42 Krizhevsky A. Learning multiple layers of features from tiny images [Master's thesis]. Toronto: University of Toronto, 2009. 7.

43 Goodfellow IJ, Erhan D, Carrier PL, *et al*. Challenges in representation learning: A report on three machine learning contests. Proceedings of the 20th International Conference on Neural Information Processing. Daegu: Springer, 2013. 117–124.

44 Netzer Y, Wang T, Coates A, *et al*. Reading digits in natural images with unsupervised feature learning. Proceedings of the 2011 NIPS Workshop on Deep Learning and Unsupervised Feature Learning. 2011. 7.

45 Zagoruyko S, Komodakis N. Wide residual networks. Proceedings of the 2016 British Machine Vision Conference. York: BMVA Press, 2016. 35–67.

(校对责编: 张重毅)

© 中国科学院软件研究所 http://www.c-s-a.org.cn