重庆理工大学学报(自然科学)
Journal of Chongqing University of Technology(Natural Science)
2024 年 第 38 卷 第 8 期
Vol. 38 No. 8 2024
doi: 10.3969/j. issn. 1674-8425(z). 2024. 08. 021

神经网络的拓扑解释综述

何宇楠[1],阳 蕾[2],王佳慧[3]

(1. 重庆理工大学 数学科学研究中心, 重庆 400054;
2. 重庆理工大学 理学院, 重庆 400054;
3. 重庆大学 国家卓越工程师学院, 重庆 401147)

摘 要:随着神经网络技术在医疗诊断、金融风险评估等关键领域的广泛应用,其决策过程的透明度和可解释性需求日益增加。尽管已有大量研究从不同维度探讨了神经网络的解释性,但当前的方法仍未能完全揭示其决策机制,限制了其在高可靠性和高解释性要求的场景中的广泛应用。通过系统综述拓扑学方法在神经网络解释性研究中的应用,详细分析了这些方法在揭示神经网络内部工作机制方面的优势与不足。具体探讨了拓扑工具在分析神经网络特征空间和参数空间的作用,并总结了相关研究在实际应用中面临的挑战和未来的发展方向,为进一步提升神经网络的透明度和可解释性提供了有益参考。

关键词:神经网络可解释性;拓扑数据分析;持续同调;Mapper 算法

中图分类号:TP183 **文献标识码**:A **文章编号**:1674 - 8425(2024)08 - 0182 - 09

0 引言

人工智能在计算机视觉、图像识别、自然语言处理等领域取得的成绩,得益于深度神经网络的发展。自 2012 年深度神经网络模型在 ImageNet 图片分类挑战赛上取得突破性成果以来,这类网络已逐步演进,形成了结构更复杂、参数更多、层次更深的模型。当图像、文本或语音片段输入到庞大的神经网络中时,可能有数百万个参数参与到实际计算过程。然而,由于人类难以追踪并理解神经网络中的每一次数学运算的具体含义,这导致神经网络常被视作“黑箱”模型。无法解释的模型存在严重的安全信任问题,例如恶意构造的对抗样本可以使神经网络的决策产生明显错误[1-2];医院里基于神经网络的智能分诊台因为病人通过药物得到有效缓解,从而将哮喘患者视为低风险人群而耽误了治疗[3];特斯拉汽车上的自动驾驶传感器误将一辆在公路上行驶的白色牵引车识别为天空从而造成车祸[4]。这些不安全因素可能会导致神经网络模型在需要高度信任度的领域(如医疗、国防和自动驾驶)中受到质疑,从而限制了神经网络的发展[5-8]。

近年来,为了提高神经网络的可解释性,研究人员提出了一系列方法,大部分方法试图从探索某样本的局部空间内的扰动或者对抗样本来揭示神经网络整体上的规律。面对神经网络数百万个参数以及前向传播过程中产生的高维特征映射,如何对神经网络的特征空间(包括样本空间以及特征映射空间)以及参数空间进行有效的分析仍然是一个比较棘手的问题。主要面临的挑战是数据量大以及数据维度高。而拓扑方法在处理该类问题上具有潜在价值。因为利用拓扑进行神经网络可解释性的研究过程更关注数据全局特征和拓扑结构,而不是具体的数字和几何形状,所以这种从复杂数字到简单结构的转变可以更好地解决数据量大且维度高所带来的挑战。另外拓扑方法关注的是数据本身的“形状”特征,这种特征不需要对数据进行降维就可以获得,也就不存在信息缺失的问题。即使需要降维和可视化,拓扑方法也可以帮助我们发现数据的内在结构和模式,从而进行更有效的降维和可视化。此外,由于拓扑只关注数据的全局性质,即小范围的变化很难导致整体结构变动,因而具有较好的鲁棒性。最为重要的是在进行数据结构的观测中,拓扑方

收稿日期:2023 - 12 - 07
基金项目:重庆理工大学科研启动基金项目;重庆市教委科学技术研究项目(KJQN202101108)
作者简介:何宇楠,男,博士,讲师,主要从事数据科学研究,E-mail:heyunan@ cqut. edu. cn。

本文引用格式:何宇楠,阳蕾,王佳慧. 神经网络的拓扑解释综述[J]. 重庆理工大学学报(自然科学),2024,38(8):182 - 190.
Citation format:HE Yunan, YANG Lei, WANG Jiahui. A review on topological interpretation of neural networks[J]. Journal of Chongqing University of Technology(Natural Science),2024,38(8):182 - 190.

法是从多尺度进行研究,且不依靠于度量,这使得拓扑方法不仅可以应用于更多类型的数据和领域,还可以提供更为直观和可解释性的结果。在此之前,拓扑数据分析在理论和应用方面都有很好的发展,例如文献[9-10]为拓扑工具提供了发展的理论依据;文献[11-13]分别从生物医学、经济金融领域展现了拓扑方法的重要性。

通过拓扑方法对神经网络的参数空间和特征空间进行可视化,以及对神经网络模型的行为和特性进行解释等方面已经有了一系列的工作,本文旨在总结和归纳这些工作以了解拓扑方法在解释神经网络方面的研究进展。

1 神经网络的拓扑解释

研究神经网络的可解释性时,拓扑方法首先将神经网络的数据(包括特征空间和参数空间)进行拓扑表示[14-15],随后,研究者通过拓扑工具识别和分析这些空间的基本结构和模式,从而深入理解神经网络的工作机制和决策过程。这种基于拓扑工具的数据分析方法被称为拓扑数据分析(topological data analysis,TDA),它融合了拓扑学、计算机科学和数据分析的交叉学科理念。计算拓扑持续性,首先由 Edelsbrunner 等[16]和 Zomorodian 等[17]提出,随后经历了连续的改进和完善。这一方法逐渐发展成为完整的 TDA 体系,其原理和应用由 Carlsson 在文献[18]中详细阐述。在 TDA 中,持续同调和 Mapper 算法是两个重要的工具。持续同调主要用于提取数据的拓扑特征和刻画几何形状,而 Mapper 算法则主要用于高维数据可视化。利用拓扑工具解释神经网络的方法按研究对象可以分为两类:神经网络的特征空间以及参数空间。

1.1 基于特征空间的拓扑解释

特征空间是一个用于表示数据对象的空间,其中每个数据对象由其特征值来表示。特征空间中的每个点对应于一个具有一定特征的数据对象,而这些特征通常是从原始数据中提取或计算出来的。本节的研究对象既包含输入层的数据空间也包括隐藏层的特征空间,有时也不加区分地统称它们为特征空间。

研究特征空间的拓扑信息的变化可以帮助人们理解神经网络的训练过程。通过每一层的变换,包括扭曲、拉伸或弯曲等,训练良好的神经网络分类模型能够处理原始数据所构成的流形,最终实现数据在特征空间内的线性可分性[19]。当采用连续可微的激活函数(如 tanh 和 sigmoid,但不能是 ReLU)时,神经网络各层间的变换可以被视为同胚映射。Naitzat 等[20]研究了数据所在流形的拓扑结构在经过训练良好的神经网络时的演变过程。对于给定流形,研究者通过构建数据空间内的 k 近邻图来计算样本点间的测地距离,并据此形成距离矩阵。接着对测地距离 ε 进行滤过,得到滤过复形并计算其持续同调。根据计算的结果找出与原数据(数据是通过在已知拓扑结构的流形上随机采样生成的)所在流形具有相同 0 维 Betti 数、1 维 Betti 数、2 维 Betti 数的单纯复形及其所对应的参数 k 和 ε,该复形被认为是原数据所在流形的近似。然后使用这些参数以同样的方式为每个神经层的特征空间构建单纯复形并计算其 Betti 数,进而通过 Betti 数反映数据拓扑结构的演变。研究表明,不论数据所在流形的拓扑结构复杂度如何(可通过不同维度的 Betti 数之和衡量),神经网络均能通过逐层操作降低其复杂度(即减少Betti 数之和),从而简化其拓扑结构。此外,ReLU 激活函数相较于 tanh 函数能更快地降低 Betti 数,因此 ReLU 激活函数的性能更为优越。最后,通过观察 Betti 数在浅层和深层网络中的变化趋势,发现浅层网络仅在最后的层中改变拓扑结构,而深层网络则在每一层中更平缓地改变拓扑结构。因此,尽管理论上浅层网络可以逼近任何函数,但在实际应用中,深度神经网络通常能更有效地学习复杂的函数映射和表示。Shaidullah[21]的研究发现,在网络的拓扑结构变得难以识别之前,网络的行为可以近似看作是 tanh 激活函数网络前两层和 ReLU 激活函数网络第一层的连续变形。通过对神经网络各层激活函数值的拓扑分析,尤其是利用持续图来跟踪和展示数据通过网络层后的拓扑变化,研究揭示了在网络的浅层,数据的拓扑结构基本保持不变。然而,在网络的深层部分,使用 ReLU 激活函数的网络相比于使用 tanh 激活函数的网络,对数据的拓扑结构造成了更加明显的改变。Wheeler 等[22]也通过计算特征空间的持续同调来研究数据通过网络不同层时拓扑信息的变化,与使用 Betti 数不同,他们通过 persistence landscapes(基于持续图)来描述激活特征的拓扑属性,并用其范数衡量拓扑复杂度,得到了与文献[20]不太一致的结论,即数据在经过神经网络的各个层时,尽管其拓扑复杂度整体趋势是下降的,但在某些层会发生上升的情况。在图神经网络研究中,Zhao 等[23]通过分析持续同调信息评估信息流效率,并基于此提出了新的网络架构,这一架构能够在卷积过程中根据持续同调信息调整图节点间信息传递的权重,以优化节点分类性能。对于节点分类任务,新的网络架构在广泛的图基准测试中优于现有方法。

神经网络特征空间的信息可以来刻画神经网络的复杂度。Bianchini 等[24]通过拓扑学概念衡量神经网络复杂度,即在数据空间中寻找分类模型的决策边界并计算这些边界的 Betti 数,以此来评估网络的复杂度。具体而言,在数据空间中寻找神经网络分类模型的决策边界,计算这些边界的 Betti 数,再利用 Betti 数来评估神经网络的复杂性。文中为浅层网络和深层网络的复杂度分别给出了上界和下界,发现在拥有相同数量隐藏单元的情况下,深层结构相较于浅层结构具备更高拓扑的复杂度。Guss 等[25]探讨了数据集的拓扑复杂度与神经网络容量的关联。在这里神经网络容量的定义与文献[24]中的神经网络的复杂度的定义相同,即复杂度越高,表明网络的表达能力越强,容量越大。这有助于了解如何根据数据集的拓扑特征来设计更高效、更小规模的神经网络结构,从而减少计算资源和训练时间。

在研究神经网络有向加权图构建过程中,Rieck 等[26]提出了一种度量神经网络结构复杂度的方法。该方法通过将相邻两层的神经元定义为顶点,它们之间的连接定义为边,构建加权图,并通过权值滤过来计算图的持续同调。由于所得滤过复形仅包含 0 维和 1 维单形,该方法仅关注加权图的 0 维拓扑信息。持续图的 p-范数被用来定义拓扑复杂度。作者在 2 个方面展示了此方法的实用价值的研究:首先,在训练过程中,拓扑复杂度可以监测如 dropout 和 batch normalization 等正则化方法对神经网络训练过程的影响,这可以帮助理解这些正则化技术对优化过程的贡献。另外,拓扑复杂度的变化能够指示网络训练过程中的过拟合现象,从而为提前停止训练提供依据,无需完全依赖于验证数据集。

此外,Mapper 算法可以对卷积神经网络的特征空间进行可视化。Goldfarb[27]通过应用 Mapper 算法于测试数据集进行可视化,他在一个深度神经网络模型中分类了 10 个品种的猫和狗,并将测试集在神经网络上的激活作为 Mapper 算法的输入,将测试集中的样本聚成了不同的簇,每个簇构成了 Mapper 中的节点,并根据簇内样本的平均分类准确度对节点上色。通过观察该图,可以发现误分类的样本聚集成了一个由多个簇构成的密集区域。在这个区域内簇内部的样本分类效果差的原因相似,不同簇间的样本分类效果差的原因各不相同。通过分析簇内样本分类准确率低的原因,可以增加神经网络的先验知识并进一步改善神经网络的性能。比如可以对分类效果差的区域单独再训练一个模型,当新测试样本例的激活落在 Mapper 这一区域时,就可以调用该模型以增加分类准确率。然而,使用传统 Mapper 算法需手动调整多个参数,这不仅耗时而且复杂,容易因参数的不当选取影响结果的可靠性和有效性。Zhou 等[28]、Rathore 等[29]的工作则极大地提高了算法的效率,并简化了参数调整过程。通过自动化参数选择和算法优化,减少了在使用 Mapper 进行神经网络分析时的工作负担。他们将激活向量聚类成不同的簇,如果簇之间存在共同的元素,则它们之间有连接关系。激活向量聚类形成的簇及其相互连接构成了 Mapper 图的顶点和边。例如,对某些层的激活进行可视化,可以看到 Mapper 图中不同的分支所关注的特征具备不同的侧重点。Purvine 等[30]使用持续同调和 Mapper 算法探索从神经网络隐藏层的激活中提取拓扑信息的方法。首先演示了基于持续同调的 Wasserstein 距离可以用于量化层之间的差异,并且通过实验证明了这种度量方法不受模型初始化种子的影响。其次,作者展示了 Mapper 图可以帮助了解模型如何在每个层次上组织分层类别知识,最后呈现的可视化效果与[29]非常相似。

通过上述分析,得知采用拓扑学方法分析神经网络涉及对处理的数据(如特征映射和参数)进行拓扑化表征,如图 1 所示。利用持续同调和 Mapper 算法提取和可视化数据的拓扑特征。这一过程借助持续同调和 Mapper 算法来提取和可视化数据的拓扑特征。通过这些方法,可以有效地识别和分析数据空间中的基本结构和模式并进一步与神经网络的行为相关联。

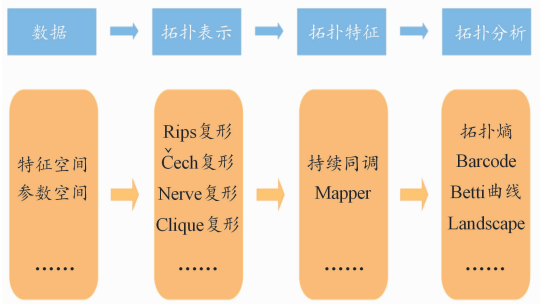

图 1 拓扑数据分析在神经网络可解释性中的应用框架

1.2 基于参数空间的拓扑解释

本文关注的参数空间包括神经网络中的可调参数,如卷积核设置、网络结构(如层数、节点数)、损失函数等。在许多机器学习算法和模型的实际应用过程中,需要通过调整一系列参数来优化模型的性能或达到特定的目标。这些参数可以影响模型的结构、行为和性能。搜索参数空间通常涉及使用各种优化技术,如梯度下降、遗传算法、贝叶斯优化等,以找到最佳参数组合,以便模型能够在给定的任务上取得最佳表现。但这些优化算法可能遇到问题,如在鞍点附近收敛速度减慢或选择最优参数困难。随着参数数量的增加,参数空间的维度也会增加,使得搜索最优参数变得更加复杂和耗时。通过数字化参数空间并分析其几何与拓扑结构变化,可以揭示参数变化的规律,也可以在搜寻最小值时,利用损失曲线的拓扑特性来帮助解决局部最小值附近收敛速度慢,亦或是找不到全局最小值等困境。

神经网络迭代过程中的卷积核空间变化过程在拓扑方向中表现为一种规律性变化。Gabrielsson 等[31]、Carlsson 等[32]首次通过使用 Mapper 算法发现参数空间中卷积核形成的点云呈现出“环”状结构。通过计算点云的持续同调,该结构得到了进一步验证。这个发现揭示了卷积神经网络的卷积核在网络学习过程中会逐渐形成一种简单拓扑结构。基于此,他们提出了一个假设:具有简单拓扑结构的参数空间可能具有更强的泛化能力。根据这个假设,他们提出通过比较一维同调类持续时间来衡量神经网络泛化能力的优劣,这个指标被他们称为“拓扑简单度”,与之前论文中提到的拓扑复杂度相对应。研究还表明,初始化时赋予卷积核“环”状结构能加速网络收敛,并提高参数优化后的泛化能力。Love 等[33]提出了拓扑深度学习方法。该方法提供了一种新的网络结构和训练算法,能够改善复杂数据的分类与表示学习性能。

通过研究神经网络结构的拓扑性质可以帮助神经网络达到更好的训练结果。Hu 等[34]提出了一种结合交叉熵损失和拓扑损失的新损失函数,其中拓扑损失主要计算网络预测结果和真实分割结果的持续图的差异。使用该损失函数训练图像分割数据集时,他们观察到在初始训练阶段,新函

数呈现交叉熵损失增加,但拓扑损失降低的状态。经过一定次数的迭代,由于分割出现错误拓扑结构的图片会被逐渐修正,这导致网络学习速度的提升。Hofer 等[35]在损失函数中引入一个可微的损失项用于单类学习,控制隐空间的拓扑结构,并发现在样本量较小的情况下,该种单类模型表现的学习性能明显优于其他方法。Moor 等[36]在此基础上提出的拓扑自编码器则更好地完成了在低维空间中保留数据的目标,他们引入的损失项可以使隐空间的持续图和原数据的持续图类似,因此隐空间可以很好地保留原数据的空间结构,减小误差损失。Pérez-Fernández 等[37]通过改变网络宽度、层数、输入顺序和标签数量研究不同网络结构之间存在相似度关系。通过加权有向图的建模和持续图分析,研究揭示了网络结构之间的拓扑相似性及其随网络深度和复杂度变化的趋势。结果显示:①对于同种网络相同任务情况下,网络结构相似度较大,随着层数和标签数量的增加,网络之间相似性也在逐渐变弱;②对于不同任务的同种网络结构,完全没有相似性。Watanabe 等[38]的研究中,将神经网络整体视为一个加权有向图,对权值进行过滤,计算其持续图,从而得到神经元的稳定性。实验表明,当数据不足时,在持续图的对角线附近会产生更多的点(同调类),这些点为拓扑噪声的可能性较大。当然,人们的探索并没有局限于此,Hofer 等[39-40]就在文献中进行了进一步拓展,开发出一种针对每个类别概率分布的新的拓扑约束和使用图神经网络进行过滤的方法,为网络的广范应用和预训练提供了依据。

为了提供一个系统的视角并梳理拓扑数据分析在神经网络解释性研究中的应用,在表 1 中总结了现有文献。该表格依据研究的主要对象、所采用的拓扑表示方法,以及不同的拓扑特征表示技术对文献进行分类。

表 1 神经网络拓扑解释方法的文献概览

数据空间	拓扑表示	拓扑特征	相关文献
特征空间	Rips 复形	Barcode	文献[20,24-25,41-45]
		Persistence landscapes	文献[22,46]
		Persistence diagram	文献[47-51]
	Nerve 复形 Rips 复形	Mapper 图	文献[27,29,52-53]
		Persistence diagram Mapper 图	文献[21,30]
	Čech 复形	Barcode	文献[54-57]
	Clique 复形 Flag 复形	Barcode Persistence diagram	文献[58-60]
	Alpha 复形	Barcode	文献[61-62]
参数空间	Rips 复形	Persistence diagram	文献[36,63]
		Barcode	文献[64-65]
	Nerve 复形	Mapper 图	文献[31-32,66]
	Clique 复形 Flag 复形	Persistence diagram Betti 曲线	文献[38,67-68]
		Persistence landscapes Barcode	文献[26,37]
	Morse 复形	Barcode	文献[69]

2 神经网络的拓扑评估

从上一节的介绍中可以看出,在介绍神经网络的可解释性时,主要涉及对数据集的特征空间和神经网络本身的参数空间的拓扑结构的分析。但除了对数据集、网络结构、超参数等偏向于减小误差、提高速度的处理外,拓扑方法在评估生成对抗网络(GANs)的性能中扮演着关键角色,这是其在神经网络模型应用中的另一重要研究方向[70]。GANs 由生成器和判别器组成,生成器负责生成逼真的图像、音频或文本等数据,而判别器则试图区分生成的数据和真实数据[71]。在对抗训练中,生成器致力于提升样本的真实性,同时判别器努力更准确地区分真伪数据。最终,生成器能够生成与真实数据相似的新样本。

最初,Feffermant 等[72]提出了一种验证真实数据是否源于低维流形的算法,该算法说明真实数据分布是从底层流形中采样得到的,评估的主要思想在于评估底层流形与生成模型中数据分布的拓扑一致性。基于复形的持续同调计算方式。Khrulkov 等[73]又引入了几何得分(geometry score)并发现,相较于其他的得分计算方法,几何得分有效捕捉了生成数据与真实数据间本质的拓扑差异,并能维持得分的一致性,即使是在迭代的后期也不会产生得分混乱的情况。它可以将真实数据分布和生成模型数据分布拓扑的相似度度量的差别通过数字清晰地展现出来,可用于评估生成模型的好坏。通过研究生成数据和真实数据的底层流形拓扑结构的差异,该方法为生成模型的参数调优提供了拓扑结构上的评估依据。在这之后,不断有人将该种方法进行推广,其中 Zhou 等[74]在以往的拓扑差异度量中引入 2 种变体,其一是监督变体,其二是基于 Wasserstein 距离提出了一个拓扑相似性准则,并将其扩展到无监督设置下生成模型的解缠评估。这种方法绕过了数据集和模型的特定条件的要求,为跨模型

和数据集之间的比较提供了更深的依据。Barannikov 等[69]通过构建损失函数的 Morse 复形,将梯度下降算法的局部行为与损失曲面的全局特性联系在一起,并利用条形码描述损失函数曲面的拓扑结构。研究者进一步利用条形码的 Bottleneck 距离定义了神经网络的"拓扑阻碍分数",用以量化基于梯度的优化过程中局部最小值的不良程度。他们发现:①随着网络深度和宽度的增加,拓扑障碍得分有所下降。换言之,在网络更深、更宽的情况下,损失函数的局部最小值相对更易被梯度下降算法优化;②在某些情况下,条形码中局部最小值片段的长度与局部最小值的泛化误差之间存在一定关联。这表明损失函数的条形码可能反映了网络的泛化能力,线段的长度与泛化误差相关联。具体来说,条形码中较长的线段通常暗示较高的泛化误差,而较短的线段则可能指向更优的泛化表现。Clough 等[75]详细分析了 MNIST 数据集,并发现当图片添加了拓扑噪声,仅依赖卷积的识别准确率会降低,因为卷积可能改变图像的拓扑结构。通过在卷积过程中结合拓扑损失函数,网络能够度量并减轻图像拓扑结构的变化,从而可以更好地识别和去除拓扑噪声。修改后的网络权重在测试时识别的正确率显著提高,能够恢复损坏图像并减少与原图的差别。结果显示,运用持续同调来增强拓扑特征的鲁棒性,可有效在图像分割中指导权重更新。另一方面,对神经网络进行内部的分析有助于改进各项任务,包括神经网络架构的选择、训练行为的干预和提供先验知识等,这些改进不仅可以节省计算机算力,还有提高模型的泛化能力、迭代速度和精确度等作用。此外,这一方法特别适用于医学领域的应用,如心脏磁共振成像(CMR)和超声图像分割。

除了上述介绍的研究外,还有很多相关领域的学者进行了不同程度和方向的探索。在拓扑可解释性的研究中,文献[68,76-77]中提出的结论也可以很好地阐述神经网络在拓扑尺度下的变化过程,了解这些规律对增强神经网络的可解释性至关重要。另外,文献[59,78-80]都在对抗生成网络部分提出了相关见解。尽管文献[81-83]已经对拓扑数据分析进行过很多总结,但是在拓扑有利于提高神经网络可解释性的总结仍在更新中。显然,拓扑学在神经网络的应用远不止于此,预期未来会有更多创新的方法和工具被开发,以深化对神经网络工作原理的理解。

3 结束语

神经网络的可解释性长期以来一直是人工智能领域的核心挑战之一。目前,已经有许多学者从因果模型、逻辑推理等方面对其进行研究,并取得了一些初步成果。拓扑方法,它侧重于数据的内在结构和全局几何特性,为神经网络研究和解释提供了一个独特的视角,也逐渐成为人们研究神经网络可解性的一个新兴领域。本文从特征空间和参数空间2个方面介绍了目前基于拓扑方法的研究神经网络的相关文献和成果。这些研究深化了对数据和神经网络内在结构及其几何特征的理解,取得了一系列积极的研究成果。

然而,基于拓扑方法的神经网络可解释性尚处于研究的初步阶段。一方面,尽管持续同调和 Mapper 算法为网络结构提供了拓扑视角,但对于捕捉网络更细致的结构特征仍有改进空间,人们需要更细致的拓扑和几何不变量来获取数据和网络更加精细的结构。另一方面,拓扑数据分析和神经网络的结合的相关理论尚不成熟,人们对网络的拓扑特征的理解和运用缺乏一些共识。此外,在工程方面,对复杂网络的持续同调的计算依旧具有一定挑战性,其计算难度和计算速度有待提高。鉴于这些原因,希望本文对现有拓扑方法在神经网络可解释性研究中的应用的概述能够促使更多研究人员加入这一领域,推动拓扑深度学习理论的发展,进一步揭开神经网络"黑箱"的神秘面纱。

参考文献:

[1] FINLAYSON S G, BOWERS J D, ITO J, et al. Adversarial attacks on medical machine learning[J]. Science, 2019, 363(6433): 1287-1289.

[2] XU H, MA Y, LIU H C, et al. Adversarial attacks and defenses in images, graphs and text: a review[J]. International Journal of Automation and Computing, 2020, 17(2): 151-178.

[3] CARUANA R, LOU Y, GEHRKE J, et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission[C]//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 1721-1730.

[4] YADRON D, TYNAN D. Tesla driver dies in first fatal crash while using autopilot mode[N/OL]. (2016-07-01)[2023-09-16]. https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk.

[5] 化盈盈,张岱墀,葛仕明. 深度学习模型可解释性的研究进展[J]. 信息安全学报,2020,5(3):1-12.

[6] 刘全,翟建伟,章宗长,等. 深度强化学习综述[J]. 计算机学报,2018,41(1):1-27.

[7] 吴飞,廖彬兵,韩亚洪. 深度学习的可解释性[J]. 航空兵器,2019,26(1):39-46.

[8] 杨丽,吴雨茜,王俊丽,等. 循环神经网络研究综述[J]. 计算机应用,2018,38(S2):1-6,26.

[9] CHEN D, LIU J, WU J, et al. Persistent hyperdigraph homology and persistent hyperdigraph Laplacians[J]. Foundations of Data Science, 2023, 5(4): 558-588.

[10] YUE Y G,WU J,LEI F C. The evolution of non-degenerate and degenerate rendezvous tasks[J]. Topology and its Applications,2019,264:187 – 200.

[11] LIU J,XIA K L,WU J,et al. Biomolecular topology:modelling and analysis[J]. Acta Mathematica Sinica,2022,38(10):1901 – 1938.

[12] QIU Y C,WEI G W. Persistent spectral theory-guided protein engineering[J]. Nature Computational Science,2023,3(2):149 – 163.

[13] XU C,LIN H,FANG X. Manifold feature index:a novel index based on high-dimensional data simplification[J]. Expert Systems with Applications,2022,200:116957.

[14] CANG Z,MU L,WEI G. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening[J]. PLoS Computational Biology,2018,14(1):e1005929.

[15] BERGOMI M G,FROSINI P,GIORGI D,et al. Towards a topological-geometrical theory of group equivariant non-expansive operators for data analysis and machine learning[J]. Nature Machine Intelligence,2019,1(9):423 – 433.

[16] EDELSBRUNNER H,LETSCHER D,ZOMORODIAN A. Topological persistence and simplification[J]. Discrete & Computational Geometry,2002,28:511 – 533.

[17] ZOMORODIAN A,CARLSSON G. Computing persistent homology[C]//Proceedings of the twentieth annual symposium on Computational geometry,2004:347 – 356.

[18] CARLSSON G. Topology and data[J]. Bulletin of the American Mathematical Society,2009,46(2):255 – 308.

[19] OLAH C. Neural networks,manifolds,and topology[EB/OL]. (2014 – 04 – 06)[2023 – 09 – 16]. https://colah.github.io/posts/2014 – 03 – NN-Manifolds-Topology.

[20] NAITZAT G,ZHITNIKOV A,LIM L. Topology of deep neural networks[J]. Journal of Machine Learning Research,2020,21(184):1 – 40.

[21] SHAIDULLAH A. Topological data analysis of neural network layer representations[EB/OL]. (2022 – 07 – 01)[2023 – 09 – 16]. https://arxiv.org/pdf/2208.06438v1.

[22] WHEELER M,BOUZA J,BUBENIK P. Activation landscapes as a topological summary of neural network performance[C]//IEEE International Conference on Big Data. 2021:3865 – 3870.

[23] ZHAO Q,YE Z,CHEN C,et al. Persistence enhanced graph neural network[C]//International Conference on Artificial Intelligence and Statistics. 2020:2896 – 2906.

[24] BIANCHINI M,SCARSELLI F. On the complexity of neural network classifiers:a comparison between shallow and deep architectures[J]. IEEE Transactions on Neural Networks and Learning Systems,2014,25(8):1553 – 1565.

[25] GUSS W H,SALAKHUTDINOV R. On characterizing the capacity of neural networks using algebraic topology[EB/OL]. (2018 – 02 – 03)[2023 – 09 – 16]. https://arxiv.org/pdf/1802.04443.

[26] RIECK B,TOGNINALLI M,BOCK C,et al. Neural persistence:a complexity measure for deep neural networks using algebraic topology[EB/OL]. (2019 – 09 – 27)[2023 – 09 – 16]. https://arxiv.org/pdf/1812.09764.

[27] GOLDFARB D. Understanding deep neural networks using topological data analysis[EB/OL]. (2018 – 10 – 31)[2023 – 09 – 16]. https://arxiv.org/pdf/1811.00852.

[28] ZHOU Y,CHALAPATHI N,RATHORE A,et al. Mapper interactive:a scalable,extendable,and interactive toolbox for the visual exploration of high-dimensional data[EB/OL]. (2021 – 04 – 27)[2023 – 09 – 16]. https://arxiv.org/pdf/2011.03209.

[29] RATHORE A,CHALAPATHI N,PALANDE S,et al. TopoAct:visually exploring the shape of activations in deep learning[C]//Computer Graphics Forum,2021,40(1):382 – 397.

[30] PURVINE E,BROWN D,JEFFERSON B,et al. Experimental observations of the topology of convolutional neural network activations[EB/OL]. (2022 – 12 – 01)[2023 – 09 – 16]. https://arxiv.org/pdf/2212.00222.

[31] GABRIELSSON R B,CARLSSON G. Exposition and interpretation of the topology of neural networks[C]//2019 18th IEEE International Conference on Machine Learning and Applications,2019:1069 – 1076.

[32] CARLSSON G,GABRIELSSON R B. Topological approaches to deep learning[C]//Topological Data Analysis. Cham:Springer International Publishing,2020:119 – 146.

[33] LOVE E R,FILIPPENKO B,MAROULAS V,et al. Topological deep learning[EB/OL]. (2021 – 03 – 04)[2023 – 09 – 16]. https://arxiv.org/pdf/2101.05778.

[34] HU X,LI F,SAMARAS D,et al. Topology-preserving deep image segmentation[J]. Advances in Neural Information Processing Systems,2019:5657 – 5668.

[35] HOFER C,KWITT R,NIETHAMMER M,et al. Connectivity-optimized representation learning via persistent homology[C]//Proceedings of the 36th International Conference on Machine Learning. 2019:2751 – 2760.

[36] MOOR M,HORN M,RIECK B,et al. Topological autoencoders[C]//Proceedings of the 37th International Conference on Machine Learning. 2020:7045 – 7054.

[37] PÉREZ-FERNÁNDEZ D, GUTIÉRREZ-FANDIÑO A, ARMENGOL-ESTAPÉ J, et al. Characterizing and measuring the similarity of neural networks with persistent homology [EB/OL]. (2021 - 03 - 31) [2023 - 09 - 16]. https://arxiv.org/pdf/2101.07752.

[38] WATANABE S, YAMANA H. Topological measurement of deep neural networks using persistent homology [EB/OL]. (2021 - 06 - 06) [2023 - 09 - 16]. https://arxiv.org/pdf/2106.03016.

[39] HOFER C, GRAF F, NIETHAMMER M, et al. Topologically densified distributions [C]//International Conference on Machine Learning. 2020:4314 - 4323.

[40] HOFER C, GRAF F, RIECK B, et al. Graph filtration learning [C]//International Conference on Machine Learning. 2020:4314 - 4323.

[41] SUN S, CHEN W, WANG L, et al. On the depth of deep neural networks: a theoretical view [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2016: 2066 - 2072.

[42] PETRI G, LEITÃO A. On the topological expressive power of neural networks [EB/OL]. (2020 - 11 - 01) [2023 - 09 - 16]. https://openreview.net/pdf? id = I44kJPuvqPD.

[43] YANG J, SANG L, CREMERS D. Dive into layers: neural network capacity bounding using algebraic geometry [EB/OL]. (2021 - 11 - 04) [2023 - 09 - 17]. https://arxiv.org/pdf/2109.01461.

[44] MASDEN M. Algorithmic determination of the combinatorial structure of the linear regions of ReLU neural networks [EB/OL]. (2022 - 07 - 15) [2023 - 09 - 17]. https://arxiv.org/pdf/2207.07696.

[45] LIU B, SHEN M. Some geometrical and topological properties of DNNs' decision boundaries [J]. Theoretical Computer Science, 2022, 908:64 - 75.

[46] KIM K, KIM J, ZAHEER M, et al. Pllay: efficient topological layer based on persistent landscapes [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020:15965 - 15977.

[47] BIRDAL T, LOU A, GUIBAS L J, et al. Intrinsic dimension, persistent homology and generalization in neural networks [C]//Proceedings of 35th Conference on Neural Information Processing Systems, 2021:6776 - 6789.

[48] CARRIÈRE M, CHAZAL F, GLISSE M, et al. Optimizing persistent homology based functions [C]//International Conference on Machine Learning, 2021:1294 - 1303.

[49] LIU W, GUO H, ZHANG W, et al. Toposeg: topology-aware segmentation for point clouds [C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, 2022:1201 - 1208.

[50] MAGAI G, AYZENBERG A. Topology and geometry of data manifold in deep learning [EB/OL]. (2022 - 04 - 19) [2023 - 9 - 19]. https://arxiv.org/pdf/2204.08624.

[51] SMITH A D, CATANZARO M J, ANGELORO G, et al. Topological Parallax: a geometric specification for deep perception models [EB/OL]. (2023 - 10 - 27) [2023 - 11 - 01]. https://arxiv.org/pdf/2306.11835.

[52] CARRIÈRE M, MICHEL B. Statistical analysis of mapper for stochastic and multivariate filters [J]. Journal of Applied and Computational Topology, 2022, 6(3):331 - 369.

[53] ZHOU Y, ZHOU Y, DING J, et al. Visualizing and analyzing the topology of neuron activations in deep adversarial training [C]//Proceedings of the 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning at the 40th International Conference on Machine Learning. 2023: 134 - 145.

[54] VARSHNEY K R, RAMAMURTHY K N. Persistent topology of decision boundaries [C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing. 2015:3931 - 3935.

[55] RAMAMURTHY K N, VARSHNEY K, Mody K. Topological data analysis of decision boundaries with application to model selection [C]//International Conference on Machine Learning. 2019:5351 - 5360.

[56] LI W, DASARATHY G, RAMAMURTHY K N, et al. Finding the homology of decision boundaries with active learning [C]//Proceedings of the 34th Conference on Neural Information. 2020:8355 - 8365.

[57] LEE S, YE J C. Data topology-dependent upper bounds of neural network widths [EB/OL]. (2023 - 05 - 25) [2023 - 11 - 01]. https://arxiv.org/pdf/2305.16375.

[58] CORNEANU C A, MADADI M, ESCALERA S, et al. What does it mean to learn in deep networks? And how does one detect adversarial attacks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:4757 - 4766.

[59 GEBHART T, SCHRATER P, HYLTON A. Characterizing the shape of activation space in deep neural networks [C]//IEEE International Conference on Machine Learning and Applications. 2019:1537 - 1542.

[60] ALONI L, BOBROWSKI O, TALMON R. Joint geometric and topological analysis of hierarchical datasets [C]//Machine Learning and Knowledge Discovery in Databases. 2021:478 - 493.

[61] ALHELFI L M, ALI H M. Using persistence barcode to show the impact of data complexity on the neural network architecture[J]. Iraqi Journal of Science, 2022, 63(5): 2262 - 2278.

[62] VANDAELE R, KANG B, LIJFFIJT J, et al. Topologically regularized data embeddings[EB/OL]. (2022 - 03 - 07)[2023 - 08 - 24]. https://arxiv.org/pdf/2110.09193.

[63] FU D, NELSON B J. Topological regularization for dense prediction[C]//IEEE International Conference on Machine Learning and Applications. 2022: 45 - 52.

[64] ZHENG S, ZHANG Y, WAGNER H, et al. Topological detection of trojaned neural networks[C]//Proceedings of 35th Conference on Neural Information Processing Systems. 2021: 17258 - 17272.

[65] ZHAO D. Nonparametric topological layers in neural networks[EB/OL]. (2021 - 11 - 27)[2023 - 08 - 24]. https://arxiv.org/pdf/2111.14829.

[66] GABRIELSSON R B, NELSON B J, DWARAKNATH A, et al. A topology layer for machine learning[C]//International Conference on Artificial Intelligence and Statistics. 2020: 1553 - 1563.

[67] GUTIÉRREZ-FANDIÑO A, PÉREZ-FERNÁNDEZ D, ARMENGOL-ESTAPÉ J, et al. Persistent homology captures the generalization of neural networks without a validation set[EB/OL]. (2021 - 05 - 31)[2023 - 08 - 24]. https://arxiv.org/pdf/2106.00012.

[68] ZHANG B, DONG Z, ZHANG J, et al. Functional network: a novel framework for interpretability of deep neural networks[J]. Neurocomputing, 2023, 519: 94 - 103.

[69] BARANNIKOV S, VORONKOVA D, TROFIMOV I, et al. Topological obstructions in neural networks learning[EB/OL]. (2022 - 02 - 05)[2023 - 09 - 19]. https://arxiv.org/pdf/2012.15834.

[70] HENSEL F, MOOR M, RIECK B. A survey of topological machine learning methods[J]. Frontiers in Artificial Intelligence, 2021, 4: 681108.

[71] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139 - 144.

[72] FEFFERMAN C, MITTER S, NARAYANAN H. Testing the manifold hypothesis[J]. Journal of the American Mathematical Society, 2016, 29(4): 983 - 1049.

[73] KHRULKOV V, OSELEDETS I. Geometry score: a method for comparing generative adversarial networks[C]//International Conference on Machine Learning, 2018: 2621 - 2629.

[74] ZHOU S, ZELIKMAN E, LU F, et al. Evaluating the disentanglement of deep generative models through manifold topology[EB/OL]. (2021 - 03 - 17)[2023 - 09 - 19]. https://arxiv.org/pdf/2006.03680.

[75] CLOUGH J R, BYRNE N, OKSUZ I, et al. A topological loss function for deep-learning based image segmentation using persistent homology[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(12): 8766 - 8778.

[76] CARRIÈRE M, OUDOT S. Structure and stability of the one-dimensional mapper[J]. Foundations of Computational Mathematics, 2018, 18(6): 1333 - 1396.

[77] ZHOU C, DONG Z, LIN H. Learning persistent homology of 3D point clouds[J]. Computers & Graphics, 2022, 102: 269 - 279.

[78] LACOMBE T, IKE Y, CARRIÈRE M, et al. Topological uncertainty: monitoring trained neural networks through persistence of activation graphs[EB/OL]. (2021 - 05 - 07)[2023 - 08 - 24]. https://arxiv.org/pdf/2105.04404.

[79] KOCHAR Y, VENGALIL S K, SINHA N. Using topological framework for the design of activation function and model pruning in deep neural networks[EB/OL]. (2021 - 09 - 03)[2023 - 08 - 24]. https://arxiv.org/pdf/2109.01572.

[80] WANG F, LIU H, SAMARAS D, et al. Topogan: a topology-aware generative adversarial network[C]//The European Conference on Computer Vision. 2020: 118 - 136.

[81] ACKERMAN J, CYBENKO G. Formal languages, deep learning, topology and algebraic word Problems[C]//IEEE Security and Privacy Workshops. 2021: 134 - 141.

[82] MORONI D, PASCALI M A. Learning topology: bridging computational topology and machine learning[J]. Pattern Recognition and Image Analysis, 2021, 31: 443 - 453.

[83] ALICIOGLU G, SUN B. A survey of visual analytics for explainable artificial intelligence methods[J]. Computers & Graphics, 2022, 102: 502 - 520.

A review on topological interpretation of neural networks

HE Yunan[1], YANG Lei[2], WANG Jiahui[3]

(1. Mathematical Science Research Center, Chongqing University of Technology, Chongqing 400054, China;
2. School of Science, Chongqing University of Technology, Chongqing 400054, China;
3. National Elite Institute of Engineering, Chongqing University, Chongqing 401147, China)

Abstract: As neural network technology finds extensive applications in critical fields such as medical diagnosis and financial risk assessment, the demand for transparency and interpretability in decision-making processes has increasingly grown. Although numerous studies have explored the interpretability of neural networks from various perspectives, current methods have yet to fully elucidate their decision mechanisms, which limits their deployment in scenarios requiring high reliability and interpretability. This paper systematically reviews the application of topological methods in neural network interpretability research, providing a detailed analysis of the strengths and limitations of these methods in revealing the inner workings of neural networks. The study specifically examines the role of topological tools in analyzing the feature space and parameter space of neural networks and summarizes the challenges and future directions faced by related research in practical applications. This review offers valuable insights for further enhancing the transparency and interpretability of neural networks.

Key words: neural network interpretability; topological data analysis; persistent homology; Mapper algorithm

(责任编辑 王 欢)