计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(11):36−47 [doi: 10.15888/j.cnki.csa.009269]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向目标用户的深度学习模型可视化综述①

胡凯茜, 李　欣, 裴炳森

(中国人民公安大学 信息网络安全学院, 北京 100038)
通信作者: 李　欣, E-mail: lixin@ppsuc.edu.cn

摘　要: 深度学习模型在某些场景的实际应用中要求其具备一定的可解释性, 而视觉是人类认识周围世界的基本工具, 可视化技术能够将模型训练过程从不可见的黑盒状态转换为可交互分析的视觉过程, 从而有效提高模型的可信性和可解释度. 目前, 国内外相关领域缺少有关深度学习模型可视化工具的综述, 也缺乏对不同用户实际需求的研究和使用体验的评估. 因此, 本文通过调研近年来学术界模型可解释性和可视化相关文献, 总结可视化工具在不同领域的应用现状, 提出面向目标用户的可视化工具分类方法及依据, 对每一类工具从可视化内容、计算成本等方面进行介绍和对比, 以便不同用户选取与部署合适的工具. 最后在此基础上讨论可视化领域存在的问题并加以展望.

关键词: 可视化工具; 深度学习模型; 模型可解释性; 目标用户; 注意力机制

引用格式: 胡凯茜,李欣,裴炳森.面向目标用户的深度学习模型可视化综述.计算机系统应用,2023,32(11):36–47. http://www.c-s-a.org.cn/1003-3254/9269.html

Review on Visualization of Deep Learning Models for Target Users

HU Kai-Xi, LI Xin, PEI Bing-Sen

(Academy of Information Network Security, People's Public Security University of China, Beijing 100038, China)

Abstract: Deep learning models require certain interpretability in practical applications in certain scenarios, and vision is a basic tool for humans to understand the surrounding world. Visualization technology can transform the model training process from an invisible black box to an interactive and analyzable visual process, effectively improving the credibility and interpretability of the model. At present, there is a lack of review on deep learning model visualization tools in related fields, as well as a lack of research on the actual needs of different users and the evaluation of user experience. Therefore, this study summarizes the current situation of the application of visualization tools in different fields by investigating the literature related to interpretability and visualization in recent years. It proposes a classification method and basis for target user-oriented visualization tools and introduces and compares each type of tool from the aspects of visualization content, computational cost, etc., so that different users can select and deploy suitable tools. Finally, on this basis, the problems in the field of visualization are discussed and its prospects are provided.

Key words: visualization tools; deep learning model; model interpretability; target users; attention mechanism

1　研究背景

在深度学习模型的部署应用中, 可解释性的缺乏不仅降低了其可靠性, 而且限制了深度学习技术的应用范围[1]. 可视化工具的开发应用, 能够直观展示复杂模型的网络架构, 帮助用户理解模型预测或决策背后的机理, 有效提高深度学习模型的可解释性. 针对不同用户的需求, 现有工具对网络结构、训练过程、神经元隐藏状态和注意力机制等不同角度进行可视化操作. 本节讨论了

① 基金项目: 国家重点研发计划 (2020AAA0107705)
收稿时间: 2023-04-05; 修改时间: 2023-05-06; 采用时间: 2023-05-17; csa 在线出版时间: 2023-08-09
CNKI 网络首发时间: 2023-08-10

深度学习模型可解释性的研究背景, 分析了可解释性的各个维度以及重要需求, 从而引入可视化工具的实现和发展历程. 通过收集与整理深度学习模型可视化领域的历年文献, 提出了本文的研究方法和主要工作.

1.1 可解释性与可视化

解释意味着赋予或提供意义, 或以可理解的术语解释和呈现某种概念[2]. 向他人解释自己决策背后的理由的能力是人类智力的一个重要方面. 同理, 可解释模型旨在自我解释系统决策和预测背后的机理[1]. 在评估模型的性能时, 对模型的信任程度是影响用户产生积极或消极感知的重要因素. 为此, Lipton[3] 提炼了关于可解释性的论述, 并讨论了赋予模型可解释性的含义和技术. 在分析预测模型的可解释性时, Guidotti 等[4] 制定一组需要考虑的维度, 包括全局和局部可解释性、解释的时效限制与用户的专长. 不同学科的研究者共同致力于定义、设计和评估可解释系统, Abdul 等[5]、Chatzimparmpas 等[6] 和 Freitas[7] 共同指出可解释模型应满足可靠性、鲁棒性、准确性和可扩展性等重要需求.

作为信息可视化 (InfoVis)[8] 的延伸, 可视化分析 (visual analytics, VA) 的概念被 Wong 等[9] 提出, 其目标是利用交互式可视化界面对大型、复杂且抽象的数据集描述的问题进行分析推理[10]. 深度学习模型的可视化研究领域发展迅速, 开发交互式可视化工具的最重要目标是提高模型的可信性和可控性, 帮助调试和改进模型, 促进人类与人工智能高效协作. 目前已有可视化分析系统来支持模型解释、调试和改进, Tzeng 等[11] 最早为深度学习模型设计可视化工具, 采用节点链接图可视化方法显示神经网络的结构, 并根据神经节点激活的强度对给定输入进行着色. Karpathy 等[12] 最早提出利用可视化工具来理解 RNNs 模型, 通过构建热力图 (heatmap) 来衡量每个输入的字符与给定隐藏神经元激活的相关性, 并在其网站上可视化了神经网络每一层的激活, 这种方法能够在单个字符的预测应用中识别可解释的语义单位.

1.2 论文研究方式

深度学习可视化课题在过去的几年中吸引了来自机器学习、深度学习和视觉分析界等研究人员的关注. 该领域的出版物广泛分布在不同类型的期刊上, 包括计算机视觉、机器学习、图形学、可视化和人工智能等. 对这些文献进行调查和分析, 可以为深度学习模型可视化研究提供更加全面深入的理解和探索.

在第 1 轮文献搜集中, 使用结构化和迭代的方式来寻找模型可视化的相关研究, 并对文献提出的可视化工具进行分类. 在初次选择论文过程中, 检索了深度学习、模型可解释性和可视化领域的重要会议和期刊, 如计算语言学协会年会 (ACL)、IEEE 可视化会议 (VIS)、人工智能会议 (AAAI). 同时, 由于可视化是一个发展非常迅速的话题, 在 arXiv、Google Scholar、ReadPaper、百度学术等学术搜索引擎中, 以深度学习模型、可视化、模型可解释性等关键词及其组合进行检索, 筛选出在标题或摘要中提及使用可视化方法理解或分析深度学习模型的文献. 在第 2 轮文献收集中, 针对第 1 轮的领域相关文献, 通过 Google Scholar 和 ReadPaper 等搜索引擎筛选其参考文献以及被引用中的其他论文. 特别地, 重点关注了 IEEE VAST、IEEE InfoVis、EuroVis、ICML、NIPS、ACM SIGKDD、ICCV 和 CVPR 等领域期刊和会议. 随后进行主轴编码来组织文献, 对各种深度学习模型可视化技术和工具进行分类整理和对比分析.

同时, 可视化技术面临输入形式繁多、数据集规模大、模型架构复杂、缺乏评判标准等多种问题, Sun 等[13] 为该领域提出诸多开放性的建议, Choo 等[14] 指出目前研究的空白和机遇, 使得深度学习技术朝着可解释、高效和安全的方向发展. 综上, 本文主要工作如下: (1) 介绍可解释深度学习模型和可视化的含义, 收集与归纳可视化领域相关文献; (2) 提出可视化工具的分类方法及依据, 针对深度学习不同用户的需求进行基于时间、主题的可视化技术分析和应用情况调查; (3) 讨论可视化技术的研究热点, 提高深度学习模型的可解释性与应用价值, 促进合法高效的人工智能协作.

2 可视化工具分类

2.1 分类依据

在现有研究中, Garcia 等[15] 基于任务和架构将可视化目的分为用于网络结构可视化、用于训练过程的解释以及用于特征理解, 并提出分类依据与研究方法. Yu 等[16] 针对初学者、实践者、开发者、专家的实际需求, 将不同方法和工具按照教授深度学习概念、架构评估、调试和改进模型的工具以及可视化解释分类. Choo 等[14] 根据可解释深度学习的原理将可视化工具分为教育使用类、模型调试类和深度理解类. Sun 等[13]

从注重训练和注重测试的角度对现有的自然语言处理 (natural language processing, NLP) 任务中深度学习模型解释方法进行分类. Mohseni 等[17] 根据指定的最终用户和评估主体, 从人工智能专家、数据专家和人工智能新手 3 类用户中总结和分析了可解释人工智能的设计目标. Hohman 等[18] 为深度学习应用提供了一个全面的视觉分析工具的回顾和分类, 总结了领域数据集和可视化技术.

了解不同任务中的用户需求是模型可解释性感知的一个关键方面[15], 在文献收集整理的过程中, 如图 1 所示, 总结了不同背景知识和经验的用户对交互式可视化工具的需求: (1) 对于教育工作者, 深度学习模型可视化工具需要具有易学性和易理解性, 以便他们可以更好地传授模型的结构和原理. 此外, 还需要提供直观的可视化展示和交互式的学习体验, 以便初学者可以更好地理解模型的工作原理和应用场景; (2) 对于模型开发人员, 可视化工具应满足易用性和高效性的需求, 协助开发人员快速地构建、优化和部署深度学习模型. 这些工具还需要提供实时的可视化反馈和调试功能, 从而快速地发现和解决模型中的问题; (3) 对于研究人员, 可视化工具需要具有高度的可定制性和灵活性, 能够根据研究需求进行适当的修改和扩展. 这些工具同时需要提供丰富的可视化选项和功能, 帮助研究人员深入分析模型的内部结构和决策过程; (4) 同时, 在为科技公司设计用于实际使用和部署的可视化工具时, 工具的灵活性和可推广性是一个高度优先事项, 这些要求促使设计并研发了工业级可视化工具, 并且应能够便捷地部署在机器学习和深度学习平台中.

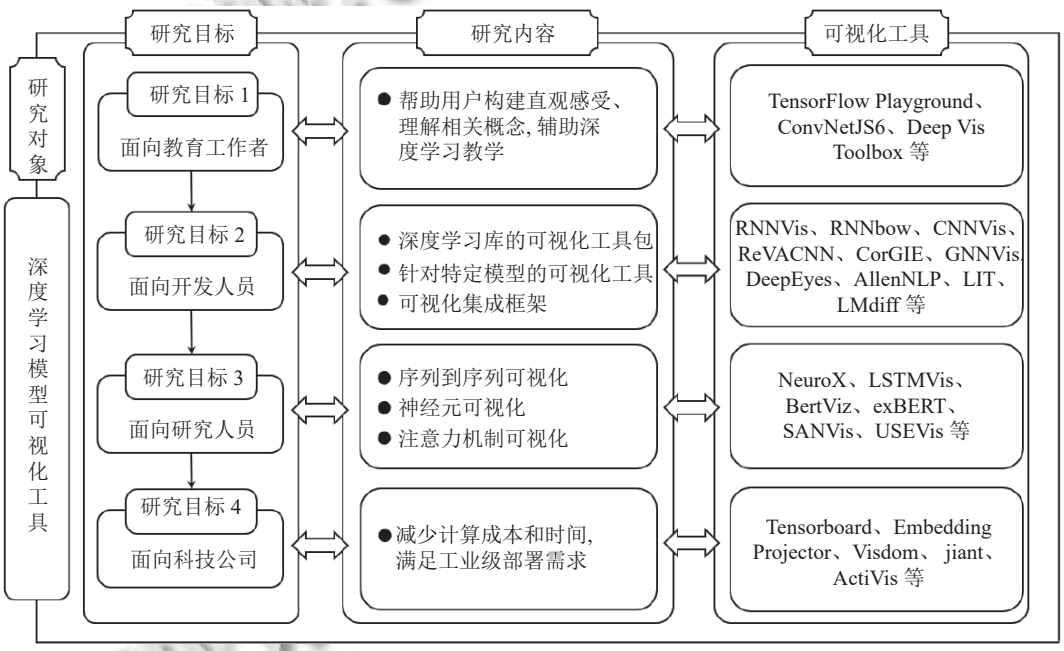

图 1 深度学习模型可视化工具分类

2.2 面向教育工作者的可视化工具

由于深度学习模型通常具有复杂的结构和参数设置, 对于初学者来说, 直接通过编码方式构建和训练模型往往存在一定难度. 因此, 可视化工具成为深度学习教学的良好选择, 它可以帮助教育工作者更好地开展模型演示与教学. Google 发布了一款基于 Web 的可视化工具 TensorFlow Playground[19], 如图 2 所示, 该工具采用 TensorFlow.js、WebGL 等关键技术, 具有界面简洁、系统开源和实时反馈的特点. TensorFlow.js 是一款基于 JavaScript 的深度学习库, 可以进行高效深度学习计算. WebGL 则是基于 Web 的图形库, 支持复杂的矩阵运算、卷积等操作, 提高模型的训练和推理速度. 在浏览器界面中, 用户可以实时调整神经网络参数, 观察损失函数和准确率曲线等指标, 直观理解深度学习模型的训练过程和结果.

相较于 TensorFlow Playground, ConvNetJS6 集成了前端可视化的关键技术[20], 且专用于卷积神经网络 (convolutional neural network, CNN)[21] 的可视化操作,

通过一系列动态视图和丰富的可视化模块, 帮助用户更好地理解 CNN 的工作原理和内部结构. 考虑到神经元之间复杂的连接方式和计算过程, 开源软件 Deep Vis Toolbox[22] 动态可视化各层滤波器的激活图, 并实时提供网络摄像头的视频输入. 该系统使用图像处理技术对神经网络进行可视化, 用户可以通过交互式界面自由探索神经网络的内部结构和训练过程.

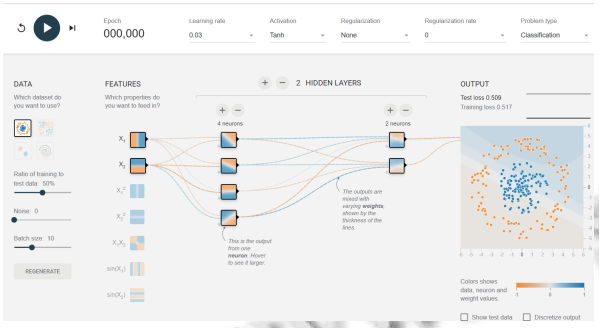

图 2 TensorFlow Playground 可视化工具

2.3 面向开发人员的可视化工具

在 Web 界面中演示网络结构的可视化工具为教育工作者提供了直观交互的操作界面, 然而深度学习模型的训练过程通常比较耗时, 且需要克服众多不确定性因素. 对于开发人员来说, 复杂的模型架构、大规模数据集以及百万量级的参数[23] 往往阻碍了对深度学习模型的理解和优化. 因此, 学者们设计了多种针对一类模型的可视化软件和针对模型群的可视化集成框架, 为探索模型的结构和训练提供了重要帮助.

2.3.1 针对特定模型的可视化工具

循环神经网络 (recurrent neural network, RNN) 适用于文本、语音等序列数据建模[24]. 通过引入不同的机制和结构, RNN 模型拥有众多变体, 如 LSTM[25] 引入记忆单元, 在长序列数据处理中保留更多的历史信息, 提高了模型的长期记忆能力, GRU[26] 优点在于参数更少、计算量更小、训练速度更快. 为探究 RNN 模型及其变体的内部结构, Yao 等[27] 提出了可视化分析方法 RNNVis. 通过联合内存芯片和词云, RNNVis 设计并实现了交互式联合聚类可视化分析方式, 使得开发人员可以灵活探索、理解和比较不同 RNN 模型的内部行为, 便于进一步优化和改进 RNNs 模型. 然而, RNNVis 不能有效解释 RNNs 模型存在的梯度消失和梯度爆炸问题[28], 为此 Cashman 等设计了基于 RNNs 的文本分类可视化工具 RNNbow[29]. 通过对文本序列中每个词语的权重进行可视化, 它可以直观展示模型在文本分类中的关键词语, 从而提升了模型的分类精度. 具体而言, RNNbow 使用堆叠条形图显示每个步骤中关于权重的梯度以及来自相邻单元的梯度贡献, 并展示反向传播训练过程中的梯度流. 实验分析表明了 RNNbow 工具在剖析梯度消失现象方面的有效性.

卷积神经网络最初由 LeCun 等[21] 提出, 并在手写数字识别任务[30] 中表现了突出的性能. Harley[31] 在研究中对卷积神经网络在 MNIST 手写数字识别数据集[30] 上的训练进行了可视化展示. 用户可以自主绘制数字作为输入, 并实时观察每个神经元对输入的响应情况, 从而探索 CNN 的工作原理和特征提取过程. 2012 年 Krizhevsky 等[32] 使用 CNNs 模型在 ImageNet 的图像分类任务[33] 上取得了优越的成绩, 随之 CNNs 在语音识别[34] 等任务中同样显著提高了识别准确性, 这些成果促进了对 CNN 模型及其变体的可视化研究. 然而, 探究 CNNs 模型内部结构的过程中面临一定的难题. 一方面, CNNs 可能由数十层或数百层组成, 每层存在数千个神经元以及神经元之间的数百万个连接, 在深度和广度上均涉及巨大规模. 为此, Liu 等搭建了视觉分析系统 CNNVis[35], 将 CNNs 建模为一个有向无环图, 并创造性提出基于双聚类的边缘捆绑算法, 减少神经元之间大量连接造成的视觉混乱. 另一方面, CNNs 由许多功能组件构成, 包括卷积层、池化层、全连接层等, 它们的价值和作用很难被充分理解[36], 为应对这一挑战, 交互式可视化分析系统 ReVACNN[37] 使用卷积网络降维的方式降低模型的复杂度. 该系统的前端由网络可视化模块和交互模块组成, 网络可视化模块用于监控卷积神经网络的底层过程, 包括卷积层、池化层和全连接层的特征提取和转换, 以及各层之间的信息流动等. 交互模块则用于对模型进行实时指导, 包括对模型的训练数据、超参数和结构进行调整和优化, 以及对模型的性能进行监控和评估.

分析图数据有助于理解图中隐藏的模式, 例如, 对社交网络的探索有助于在社交媒体中创建自适应好友推荐系统[38]. 近年来图神经网络 (graph neural network, GNN) 模型[39] 将深度学习技术扩展到图数据上并取得了重大进展. 图卷积网络 (graph convolutional network, GCN)[40] 通过在图上进行卷积操作得到更加丰富的特征表示. 与 GCN 相比, 图注意力网络 (graph attention network, GAT)[41] 通过引入注意力机制动态地计算节

点之间的权重, 并根据节点之间的相似度来对图进行卷积操作, 以便全面捕捉图中的特征信息. 上述 GNNs 涉及图的复杂拓扑结构和高维特征, 造成了训练过程中理解和调试的困难. 因此, 对 GNNs 模型可视化操作需要将拓扑图、高维特征和预测结果恰当地联系起来.

2020 年, Jin 等[42] 搭建的 GNNVis 工具具有多层次可视化、交互式探索、可扩展性和可解释性等特点. 该工具可以对 GNNs 模型的输入图、节点表示、边权重和输出结果等多个层次进行可视化展示. GNNVis 的前端由控制面板、平行集视图、投影视图等组成, 用户可以通过拖拽、缩放、选择和高亮等交互操作来探索模型的内部机制. 同时采用 Python、Flask 和 Docker 等后端技术, 实现对 GNNs 模型和图数据的分析与拓展. Liu 等[43] 在 2021 年提出 CorGIE 可视化系统, 考虑了图拓扑、节点特征和潜在嵌入之间的对应关系, 并采用 K-hop 图表揭示 GNNs 模型如何聚合关键节点的信息. 通过使用场景和 GNN 专家的案例研究, 作者验证了在 GNNs 模型开发和优化过程中引入 CorGIE 工具的有效性. 与 GNNVis 工具相比, CorGIE具有更广泛的适用范围和多维数据可视化的优势, 同时还支持多层次分析和动态可视化功能, 能够提高模型的可解释性.

2.3.2 可视化集成框架

当前, 可视化工具类型与功能日趋完善, 然而在进行多种深度学习模型的选择与对比时, 切换可视化工具可能需要花费大量时间, 从而影响了交互式分析的实际任务, 阻碍了实践者的运用. 为解决上述问题, 研究人员开发了一系列可视化集成框架, 以便调试和对比不同类模型, 便于用户根据实际任务需求选取合适的工具. Tenney 等[44] 指出, 理想的工作流应该是无缝衔接且交互的, 可视化集成框架的设计应满足灵活性、可扩展性、模块化和易于使用等要求.

2018 年, Pezzotti 等[45] 设计推出了可视化系统 DeepEyes, 该系统支持 TensorFlow[46]、PyTorch[47] 和 Keras[48] 等多种深度学习框架, 具有操作简便、数据分析速度快等特点, 并支持多种数据源和数据格式的导入和导出, 能够满足不同场景下用户的使用需求. 为满足灵活性和模块化的要求[44], Wallace 等[49] 提出的 AllenNLP 框架能够为多数自然语言处理模型提供内置的解释方法和前端可视化组件库, 并支持基于梯度的对抗攻击. 通过在预训练语言模型 BERT[50] 上进行实时演示, 验证了该工具包的实用性. 相较于工具包的安装使用, 2021 年 Tenney 等[44] 引入了基于 Web 的用户界面 LIT 以提高可视化框架的灵活性. 如图 3 所示, 通过将可视化界面集成到一个精简的网页, LIT 可实现局部快速探索和错误分析, 同时具有强大的数据交互和演示功能. 随着不同领域的多维数据集迅速增长[51], 可视化技术在数据分析和决策支持中发挥重要作用. 2019 年 Roumani 等[52] 面向生物信息学领域提出一种可视化数据挖掘软件 BioNetApp, 用于交互关联和对比分析从多个实验中获得的大量数据, 并采用交互式网络图、热力图、条形图等形式帮助用户理解和探索这些数据. 为便于开发人员对深度学习模型的深入挖掘, 2021 年 Strobelt 等[53] 搭建了 LMdiff 集成框架. 该框架支持用户对文本实例进行逐个单词的检查, 以验证模型行为的假设, 并从大型语料库中识别出最相关的短语, 从而帮助模型选择文本实例.

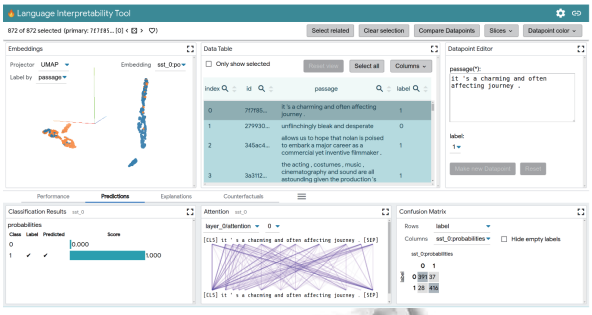

图 3 可视化集成框架 LIT

2.4 面向研究人员的可视化工具

上述可视化方法主要集中在理解模型结构和调试模型训练过程中, 为用户提供了高效直观的交互式学习和研究工具, 但未充分研究如何有效地将专家知识纳入分析过程, 以及如何理解模型的复杂参数等问题. 深度学习模型的训练往往需要消耗数小时到数十天的运算[54], 可视化工具可以将专家知识和深度学习技术结合起来, 在模型序列、注意力机制和神经元探索等领域中有潜在应用[55], 能够协助研究人员解释模型背后的工作机理[56].

2.4.1 序列到序列可视化

序列到序列模型 (Seq2Seq) 也被称为编码器-解码器模型[57], 在机器翻译[58]、自然语言生成[59] 和文本摘要[60] 等应用中表现出了较强的准确率和鲁棒性[61]. 然而, Seq2Seq 模型复杂的网络结构增加了解释的难度, 这意味着当模型的预测结果出现意外情况时, 难以通

过常规方法找出错误的原因. 为此, 研究人员设计使用可视化工具, 对训练后的序列到序列模型进行阶段性交互, 提高其在各种应用场景中的应用效果[11]. 2018 年, Strobelt 等[62] 创建了一个基于模型可解释的序列到序列可视化分析系统 Seq2Seq-Vis, 采用最近邻方式探索潜在替换、“What-if”式反事实场景、自适应调整通道数等技术. 如图 4 所示, Seq2Seq-Vis 在小规模 IWSLT'14 数据集[63] 上可视化潜在向量序列随时间的状态进展, 从而方便检测模型错误, 并使用“What-if”式提问方法对模型进行探测以增强可解释性. 此外, Seq2Seq-Vis 通过交互式工具调整深度学习模型的通道数, 实现了自适应调整通道数的功能, 帮助用户探索并选择最优的模型架构.

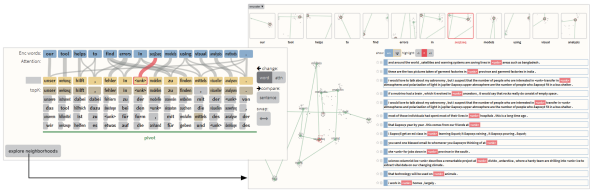

图 4 序列到序列可视化工具 Seq2Seq-Vis

2.4.2 神经元可视化

随着输入的增加, 神经元隐藏状态和参数的激增[64] 给分析模型行为带来了新的挑战. 为此, 可视化神经元隐藏状态有助于开展模型架构选择[65]、模型蒸馏[66] 和控制数据偏差[67] 等研究. 其中, Dalvi 等[68] 设计了 NeuroX 工具包, 采用了基于非线性降维算法的可视化方法, 包括基于 t-SNE 算法的二维图像和 Umap 降维的三维图像. 研究人员可以将选定的神经元进行可视化操作并对其进行消融实验, 以评估这些神经元对模型精度的影响. 然而, 隐藏状态中的语义信息高度分散, 单个输入词可能导致全部神经元隐藏状态的变化[69], 隐藏状态和单词之间的这种多对多关系, 进一步阻碍了研究人员理解神经元隐藏状态中的嵌入信息.

为解决上述问题, Strobelt 等[70] 提出了可视化分析工具 LSTMVis, 旨在解决 RNNs 模型复杂的网络结构和难以解释预测结果的问题. 如图 5 所示, 一方面, 该工具将单个神经元随时间戳或序列的激活模式呈现为线图, 向用户直观展示模型的内部状态. 另一方面, LSTMVis 还设计了隐喻识别功能, 可以突出显示神经元隐藏状态聚类的结果及其与原始数据的关系, 支持将神经元状态的聚类结果可视化为热力图. 同时, 该工具可以通过配置文件, 适应不同的模型架构, 便于研究人员交互式地探索隐藏节点的学习行为, 深入理解不同模型的内部机理.

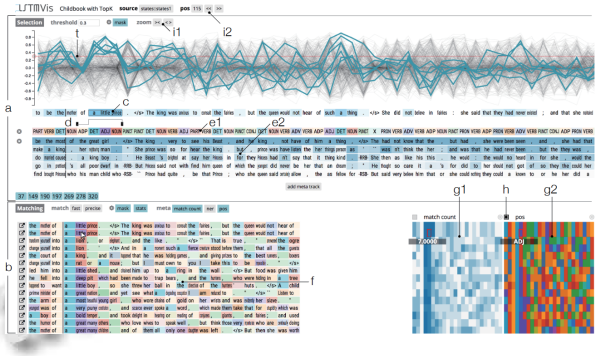

图 5 神经元隐藏状态可视化工具 LSTMVis

2.4.3 注意力机制可视化

注意力机制 (attention mechanism)[71] 本质是模拟人脑, 例如当我们阅读一页书时, 虽然可以看到整页纸的全部内容, 但人脑对这页纸的关注度本质上有一定的权重区分. 自注意力机制 (self-attention) 将注意力权重的计算范围拓宽到了整个输入序列, 更全面地捕捉句子内部的相互依赖关系[72]. 为了提高模型性能, 自注意力机制采用查询—键—值矩阵, 即 QKV 模式捕捉序列中的长距离依赖关系. “多头”指有多组 QKV 矩阵, 能够提高模型对不同位置的关注能力[73], 获取更高维度特征信息.

基于多头自注意力机制的 Transformer 模型[74] 能够同时捕获不同的句法、语义和上下文信息. 它可以在大型语料库上预训练, 学习到通用的语言表示, 从而在下游任务中提高性能. 通过使用可视化工具, 用户可以更好地理解 Transformer 模型的内部结构和决策过程, 并探测模型错误. Li 等[75] 设计与实现了开源可视化分析工具 T3-Vis, 通过一组内置算法来计算输入序列不同部分的重要性, 采用热力图、柱状图等形式将模型的各层输出和注意力权重进行可视化, 帮助研究者了解模型的内部组件及属性. 随着 BERT[50]、RoBERTa[76]、GPT[77] 和 XLNet[78] 等预训练语言模型的发布, 深度学习领域需要更加灵活高效的可视化工具[79], 以帮助用户理解微调的原理和多头自注意力机制的特征. 为此, Vig[80] 提出 BertViz 工具, 该工具展现 3 个层次的可视化视图: 注意力头视图用于可视化模型层中单个或多

个头的注意力权重; 模型视图则采用鸟瞰图形式针对特定输入将模型所有层进行注意力可视化; 神经元视图 (如图 6 所示) 可视化了输入文本序列中神经元如何相互作用以产生注意力. 此外, BertViz工具在 GPT-2 和 BERT 等模型上进行演示和实例分析, 帮助用户掌握模型层中每个头的注意力权重分布情况, 深入理解神经元之间的相互作用.

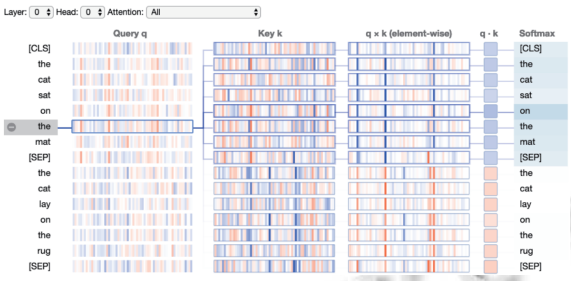

图 6 注意力机制可视化模型 BertViz 神经元视图

为解决句子或文档级别注意力可视化的问题, 可视化系统 exBERT[81] 基于相似性匹配方法, 通过将输入匹配到大型注释数据集中的相似上下文, 解释每个注意力头所习得的知识, 从而提高了模型的可解释性和可视化效果. 同时, 该系统将静态分析的稳健性与动态视图的直观性相结合, 用户可以通过交互式操作来控制输入和模型状态. 此外, 注意力机制已被广泛用于文本嵌入生成任务中[82], 其本质上是句子标记和词语之间双向关系的编码, 研究人员提出了节点链接图[83]和邻接矩阵[84] 技术来可视化这种双向关系. 例如, USEVis 系统[85] 支持用户根据自己的需求选择不同的图表类型、数据维度和页面布局, 同时使用节点链接图和邻接矩阵两种方法进行更全面的可视化分析任务, 并研究注意力机制在提取句子语义和句法方面的能力. Park 等[86] 开发了一个基于自注意力网络的可视化分析系统 SANVis, 该系统通过调整欧氏距离帮助用户识别同一层中的相似模式和不同模式, 支持用户分析多头自注意力网络的行为和特征, 以便在不同粒度层次上深入理解数据, 并且在机器翻译场景中取得较好效果.

2.5 面向科技公司的可视化工具

为满足工业级使用和部署的需求[87], 减少计算成本和时间, 可视化工具的灵活性和可推广性被视为高度优先事项[88]. 大多数深度学习库都提供基本的可视化工具包, 帮助用户调试当前模型并改善性能. TensorFlow[46] 是深度学习领域的热门框架, 其官方的可视化工具包 Tensorboard 可以读取 TensorFlow 事件工作流, 汇总数据并显示在标量仪表板、图像仪表板和直方图仪表板上, 同时支持用户可视化数据流图和观察可视化张量. 为了实现神经网络产生的 2D 和 3D 视图的可视化, TensorFlow 开发人员在 Tensorboard 的基础上集成新的可视化模块 Embedding Projector[89], 该模块提供了一个基于主成分分析和 T-分布随机近邻嵌入的 2D/3D 嵌入视图, 揭示数据点在给定层上多维表示之间的关系. 此外, Visdom[90] 是一个基于 Web 的交互式可视化工具包, 易于与 PyTorch[47] 等深度学习框架配合使用. 开源工具包 jiant[91] 常用于进行多任务学习和迁移学习, 支持用户使用最先进的模型进行模块化和配置驱动的实验, 并在 50 余个英语自然语言理解任务中得到广泛应用[3].

Kahng 等[92] 与模型架构师进行了长期研究和探讨, 设计了一个应用于 Facebook[93] 的交互式可视化系统 ActiVis, 可用于解释大规模深度学习模型的结果. 如图 7 所示, 该系统通过紧密集成多个协同视图, 例如用于解释模型架构的计算视图和用于模式发现和比较的神经元激活视图, 用户可以在实例和子集级别探索复杂的深度学习模型. 在 TREC 问答数据集[94] 上实验表明, ActiVis 可以将复杂模型的概述和结构化检查紧密集成, 扩展到各种行业规模的数据集和模型, 帮助科技公司更好地调试深度学习模型并节约计算成本.

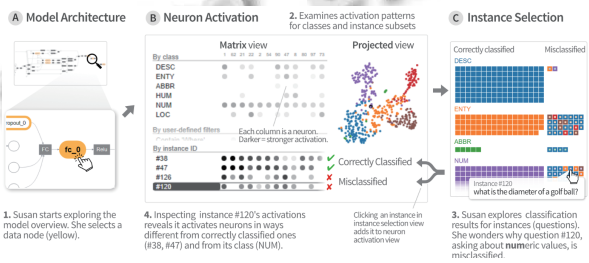

图 7 工业级可视化系统 ActiVis

3 研究热点分析

可视化工具为抽象数据和模型建立直观地交互式表达, 在增强深度学习模型的可解释性方面发挥着至关重要的作用. 可视化的关键技术包括数据降维、聚类、分类、回归等可视化算法; 图表类型、颜色、布局等可视化设计; 缩放、旋转、选择、过滤等交互技术. 同时现有工具面临大规模和多模态数据可视化、效率与成本、用户使用体验、隐私和安全问题等挑战.

本节基于现有可视化工具的关键技术和面临的挑战, 提出深度学习模型可视化的下述研究方向.

(1) 制定具体的可视化工具评价标准. 随着可视化工具分类的增多与功能的完善, 需要制定一个被广泛接受的评价标准. 目前可视化工具往往存在主观性较强的问题, 不同领域的用户有各自的需求和标准. 现有的可视化工具在解释模型时相互矛盾的现象普遍存在, 这种不一致性使得用户在理解模型的决策时感到困惑[95]. 因此, 需要制定一种全面、稳健、准确的评价标准来比较不同可解释性工具的有效性与应用价值. 在制定评价标准时应该考虑到以下方面: 可解释性、易用性、效率、可扩展性、可视化效果、交互方式、隐私和安全性等, 以便对各类可视化工具进行全面、客观的比较和评估.

(2) 可视化工具与单样本或零样本学习相结合. 通常一个深度学习模型包含数百个参数, 需要成千上万的训练实例. 在实际应用中, 如果每个特定的任务都需要单独的大规模训练样本集, 随之将产生巨大的训练成本. 单样本学习[96] 或零样本学习[97] 通过从同类模型中获得的先验知识以及领域专家的专业知识, 减少训练集的规模和训练代价, 从而降低训练成本. 因此, 未来研究的一个方向是将可视化工具与单样本学习或零样本学习相结合, 以缩小学术研究产出与现实需求之间的差距. 这种结合可以借助领域专家的先验知识直观地理解并调整模型中的参数, 从而提高模型的性能和泛化能力, 同时为不同领域的用户提供更加普适和灵活的解决方案, 具有广阔的应用前景和研究价值.

(3) 面向高级深度学习架构的可视分析. 迄今为止, 可视化工具大都面向基本的深度学习模型架构, 而许多大语言模型正在有效地用于崭新领域, 例如基于人类反馈学习的 ChatGPT、能够根据文字表述生成图像的 DALL-E 模型[98] 等. 这些模型通常由数百甚至数千个节点组成, 并涉及多层网络设计和它们之间的复杂连接结构. 这种高度复杂性为可视分析和模型可解释领域带来了前所未有的挑战. 因此, 未来的研究可以通过开发有效且高效的可视化技术, 从节点数量、层数及其连通性等方面对这类大语言模型进行直观的剖析, 帮助研究人员识别训练过程中的瓶颈和偏差.

(4) 提升深度学习模型的鲁棒性. 深度学习模型通常容易受到对抗样本[99] 扰动的干扰, 从而输出错误的预测. 开发一个高精度和高效率的深度学习模型通常需要持续的训练、评估和优化过程, 且需要根据用户的先验知识调整参数和模型结构. 在这个过程中, 可视化工具可以帮助用户理解对抗样本攻击和防御的复杂性, 并显示模型在对抗样本下的表现和弱点. 因此, 将对抗样本知识融入可视化工具是提高深度学习模型鲁棒性的一个重要研究契机, 能够帮助用户更好地理解对抗样本攻击的本质和机制, 并进一步优化深度学习模型的设计和训练策略.

4 结论与展望

深度学习模型可视化工具将深度学习模型的内部结构、训练过程和决策过程等转化为直观交互的可视化表示, 以帮助用户更好地理解、掌握和应用模型, 同时提高模型的可解释性, 改善人机交互体验. 开发可视化工具的重要目标是满足不同领域用户在使用过程中相应的需求, 促进合法高效的人工智能协作. 在本综述中, 介绍了可解释性与可视化的基本概念, 引入论文的研究方式和分类依据, 对现有的深度学习模型可视化研究成果进行整理, 总结和完善各种分类方法, 针对不同类型用户的需求, 将可视化软件按照 4 个可视化目标进行分类, 分别聚焦于教育工作者、开发人员、研究人员和科技公司的实际需求, 最后讨论了未来研究热点, 相信深度学习模型可视化工具能够朝着更准确、更高效和更安全的方向发展.

参考文献

1 雷霞, 罗雄麟. 深度学习可解释性研究综述. 计算机应用, 2022, 42(11): 3588–3602.

2 李汇来, 杨斌, 于秀丽, 等. 软件缺陷预测模型可解释性对比. 计算机科学, 2023, 50(5): 21–30. [doi: 10.11896/jsjkx.221000028]

3 Lipton ZC. The mythos of model interpretability. Communications of the ACM, 2018, 61(10): 36–43. [doi: 10.1145/3233231]

4 Guidotti R, Monreale A, Ruggieri S, *et al*. A survey of methods for explaining black box models. ACM Computing Surveys, 2018, 51(5): 93. [doi: 10.1145/3236009]

5 Abdul A, Vermeulen J, Wang D, *et al*. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018. 1–18.

6 Chatzimparmpas A, Martins RM, Jusufi I, *et al*. A survey of

surveys on the use of visualization for interpreting machine learning models. Information Visualization, 2020, 19(3): 207–233. [doi: 10.1177/1473871620904671]
7 Freitas AA. Comprehensible classification models: A position paper. ACM Sigkdd Explorations Newsletter, 2014, 15(1): 1–10. [doi: 10.1145/2594473.2594475]
8 曾悠. 大数据时代背景下的数据可视化概念研究 [硕士学位论文]. 杭州: 浙江大学, 2014.
9 Wong PC, Thomas J. Visual analytics. IEEE Computer Graphics and Applications, 2004, 24(5): 20–21. [doi: 10.1109/MCG.2004.39]
10 Lu JH, Chen W, Ma YX, *et al*. Recent progress and trends in predictive visual analytics. Frontiers of Computer Science, 2017, 11(2): 192–207. [doi: 10.1007/s11704-016-6028-y]
11 Tzeng FY, Ma KL. Opening the black box—Data driven visualization of neural networks. Proceedings of the 16th IEEE Visualization Conference. Minneapolis: IEEE, 2005. 383–390. [doi: 10.1109/VISUAL.2005.1532820]
12 Karpathy A, Johnson J, Fei-Fei L. Visualizing and understanding recurrent networks. arXiv:1506.02078, 2015.
13 Sun XF, Yang DY, Li XY, *et al*. Interpreting deep learning models in natural language processing: A review. arXiv:2110.10470, 2021.
14 Choo J, Liu SX. Visual analytics for explainable deep learning. IEEE Computer Graphics and Applications, 2018, 38(4): 84–92. [doi: 10.1109/mcg.2018.042731661]
15 Garcia R, Telea AC, Da Silva BC, *et al*. A task-and-technique centered survey on visual analytics for deep learning model engineering. Computers & Graphics, 2018, 77: 30–49. [doi: 10.1016/j.cag.2018.09.018]
16 Yu RL, Shi L. A user-based taxonomy for deep learning visualization. Visual Informatics, 2018, 2(3): 147–154. [doi: 10.1016/j.visinf.2018.09.001]
17 Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems, 2021, 11(3–4): 24. [doi: 10.1145/3387166.]
18 Hohman F, Kahng M, Pienta R, *et al*. Visual analytics in deep learning: An interrogative survey for the next frontiers. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(8): 2674–2693. [doi: 10.1109/TVCG.2018.2843369]
19 Smilkov D, Carter S, Sculley D, *et al*. Direct-manipulation visualization of deep networks. arXiv:1708.03788, 2017.
20 Chauhan JS, Wang Y. Context-aware action detection in untrimmed videos using bidirectional LSTM. Proceedings of the 15th Conference on Computer and Robot Vision (CRV). Toronto: IEEE, 2018. 222–229.
21 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]
22 Yosinski J, Clune J, Nguyen A, *et al*. Understanding neural networks through deep visualization. arXiv:1506.06579, 2015.
23 Li JW, Chen XL, Hovy E, *et al*. Visualizing and understanding neural models in NLP. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016. 681–691. [doi: 10.18653/v1/n16-1082]
24 Elman JL. Finding structure in time. Cognitive Science, 1990, 14(2): 179–211. [doi: 10.1207/s15516709cog1402_1]
25 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]
26 Cho K, van Merriënboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: ACL, 2015. 1724–1734. [doi: 10.3115/v1/D14-1179]
27 Yao M, Cao SZ, Zhang RX, *et al*. Understanding hidden memories of recurrent neural networks. Proceedings of the 2017 IEEE Conference on Visual Analytics Science and Technology (VAST). Phoenix: IEEE, 2018. 13–24. [doi: 10.1109/vast.2017.8585721]
28 Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. Proceedings of the 30th International Conference on International Conference on Machine Learning. Atlanta: JMLR.org, 2013. III-1310–III-1318.
29 Cashman D, Patterson G, Mosca A, *et al*. RNNbow: Visualizing learning via backpropagation gradients in RNNs. IEEE Computer Graphics and Applications, 2018, 38(6): 39–50. [doi: 10.1109/mcg.2018.2878902]
30 Deng L. The MNIST database of handwritten digit images for machine learning research [Best of the Web]. IEEE Signal Processing Magazine, 2012, 29(6): 141–142. [doi: 10.1109/MSP.2012.2211477]
31 Harley AW. An interactive node-link visualization of convolutional neural networks. Proceedings of the 11th International Symposium on Visual Computing. Las Vegas: Springer, 2015. 867–877. [doi: 10.1007/978-3-319-27857-5_

77]
32 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 3079–3087.
33 Dai AM, Le QV. Semi-supervised sequence learning. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 3079–3087.
34 Mohamed AR, Dahl GE, Hinton G. Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 14–22. [doi: 10.1109/tasl.2011.2109382]
35 Liu MC, Shi JX, Li Z, *et al*. Towards better analysis of deep convolutional neural networks. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 91–100. [doi: 10.1109/TVCG.2016.2598831]
36 Jarrett K, Kavukcuoglu K, Ranzato MA, *et al*. What is the best multi-stage architecture for object recognition? Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto: IEEE, 2010. 2146–2153. [doi: 10.1109/iccv.2009.5459469]
37 Chung S, Suh S, Park C, *et al*. ReVACNN: Real-time visual analytics for convolutional neural network. ACM SIGKDD Workshop on Interactive Data Exploration and Analytics (IDEA). San Francisco: ACM, 2016. 7.
38 Chen L, Xie YZ, Zheng ZB, *et al*. Friend recommendation based on multi-social graph convolutional network. IEEE Access, 2020, 8: 43618–43629. [doi: 10.1109/access.2020.2977407]
39 Dosovitskiy A, Brox T. Generating images with perceptual similarity metrics based on deep networks. Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 658–666.
40 Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. Proceedings of the 5th International Conference on Learning Representations. Toulon: ICLR, 2017.
41 Veličković P, Cucurull G, Casanova A, *et al*. Graph attention networks. arXiv:1710.10903, 2017.
42 Jin ZH, Wang Y, Wang QW, *et al*. GNNVis: A visual analytics approach for prediction error diagnosis of graph neural networks. arXiv:2011.11048, 2020.
43 Liu ZP, Wang Y, Bernard J, *et al*. Visualizing graph neural networks with CorGIE: Corresponding a graph to its embedding. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(6): 2500–2516. [doi: 10.1109/TVCG.2022.3148197]
44 Tenney I, Wexler J, Bastings J, *et al*. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. ACL, 2020. 107–118. [doi: 10.18653/v1/2020.emnlp-demos.15]
45 Pezzotti N, Höllt T, van Gemert J, *et al*. DeepEyes: Progressive visual analytics for designing deep neural networks. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 98–108. [doi: 10.1109/tvcg.2017.2744358]
46 Abadi M, Agarwal A, Barham P, *et al*. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467, 2015.
47 Paszke A, Gross S, Massa F, *et al*. PyTorch: An imperative style, high-performance deep learning library. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 721.
48 Ramasubramanian K, Singh A. Deep learning using Keras and TensorFlow. Machine Learning Using R. Apress: Springer, 2018. 667–688. [doi: 10.1007/978-1-4842-4215-5_11]
49 Wallace E, Tuyls J, Wang JL, *et al*. AllenNLP interpret: A framework for explaining predictions of NLP models. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. Hong Kong: ACL, 2019. 7–12. [doi: 10.18653/v1/d19-3002]
50 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/n19-1423]
51 Shahid MLUR, Molchanov V, Mir J, *et al*. Interactive visual analytics tool for multidimensional quantitative and categorical data analysis. Information Visualization, 2020, 19(3): 234–246. [doi: 10.1177/1473871620908034]
52 Roumani AM, Madkour A, Ouzzani M, *et al*. BioNetApp:

An interactive visual data analysis platform for molecular expressions. PLoS One, 2019, 14(2): e0211277. [doi: 10.1371/journal.pone.0211277]

53 Strobelt H, Hoover B, Satyanaryan A, *et al*. LMdiff: A visual diff tool to compare language models. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. ACL, 2021. 96–105. [doi: 10.18653/v1/2021.emnlp-demo.12.]

54 Liao XQ, Nazir S, Zhou YB, *et al*. User knowledge, data modelling, and visualization: Handling through the fuzzy logic-based approach. Complexity, 2021, 2021: 6629086. [doi: 10.1155/2021/6629086]

55 Chatzimparmpas A, Martins RM, Jusufi I, *et al*. The state of the art in enhancing trust in machine learning models with the use of visualizations. Computer Graphics Forum, 2020, 39(3): 713–756. [doi: 10.1111/cgf.14034]

56 Matveev SA, Oseledets IV, Ponomarev ES, *et al*. Overview of visualization methods for artificial neural networks. Computational Mathematics and Mathematical Physics, 2021, 61(5): 887–899. [doi: 10.1134/s0965542521050134]

57 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2014.

58 Kalchbrenner N, Blunsom P. Recurrent continuous translation models. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL, 2013. 1700–1709.

59 Gatt A, Krahmer E. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research, 2018, 61(1): 65–170. [doi: 10.1613/jair.5477]

60 Liu Y, Lapata M. Text summarization with pretrained encoders. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 3730–3740. [doi: 10.18653/v1/d19-1387]

61 Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 3104–3112.

62 Strobelt H, Gehrmann S, Behrisch M, *et al*. Seq2Seq-Vis: A visual debugging tool for sequence-to-sequence models. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 353–363. [doi: 10.1109/TVCG.2018.2865044]

63 Cettolo M, Girardi C, Federico M. WIT3: Web inventory of transcribed and translated talks. Proceedings of the 16th Annual Conference of the European Association for Machine Translation. Trento: ACL, 2012. 261–268.

64 Seifert C, Aamir A, Balagopalan A, *et al*. Visualizations of deep neural networks in computer vision: A survey. Transparent Data Mining for Big and Small Data. Cham: Springer, 2017. 123–144. [doi: 10.1007/978-3-319-54024-5_6]

65 Maitra C, Seal DB, De RK. NeuroDAVIS: A neural network model for data visualization. arXiv:2304.01222, 2023.

66 Jose A, Shetty SD. DistilledCTR: Accurate and scalable CTR prediction model through model distillation. Expert Systems with Applications, 2022, 193: 116474. [doi: 10.1016/j.eswa.2021.116474]

67 Yan JQ, Rong RC, Xiao GH, *et al*. HiddenVis: A hidden state visualization toolkit to visualize and interpret deep learning models for time series data. bioRxiv, 2020. [doi: 10.1101/2020.12.11.422030.]

68 Dalvi F, Nortonsmith A, Bau A, *et al*. NeuroX: A toolkit for analyzing individual neurons in neural networks. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2018. 9851–9852.

69 潘旭东, 张谧, 杨珉. 基于神经元激活模式控制的深度学习训练数据泄露诱导. 计算机研究与发展, 2022, 59(10): 2323–2337. [doi: 10.7544/issn1000-1239.20220498]

70 Strobelt H, Gehrmann S, Pfister H, *et al*. LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 667–676. [doi: 10.1109/TVCG.2017.2744158]

71 黄立威, 江碧涛, 吕守业, 等. 基于深度学习的推荐系统研究综述. 计算机学报, 2018, 41(7): 1619–1647. [doi: 10.11897/SP.J.1016.2018.01619]

72 朱张莉, 饶元, 吴渊, 等. 注意力机制在深度学习中的研究进展. 中文信息学报, 2019, 33(6): 1–11. [doi: 10.3969/j.issn.1003-0077.2019.06.001]

73 任欢, 王旭光. 注意力机制综述. 计算机应用, 2021, 41(S1): 1–6. [doi: 10.11772/j.issn.1001-9081.2020101634]

74 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

75 Li R, Xiao W, Wang LJ, *et al*. T3-Vis: A visual analytic framework for training and fine-tuning Transformers in NLP. arXiv:2108.13587, 2021.

76 Liu YH, Ott M, Goyal N, *et al*. RoBERTa: A robustly

optimized BERT pretraining approach. arXiv:1907.11692, 2019.
77 Radford A, Wu J, Child R, *et al*. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1.8: 9.
78 Yang ZL, Dai ZH, Yang YM, *et al*. XLNet: Generalized autoregressive pretraining for language understanding. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 517.
79 Xu K, Ba JL, Kiros R, *et al*. Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille: JMLR.org, 2015. 2048–2057.
80 Vig J. BertViz: A tool for visualizing multi-head self-attention in the BERT model. Proceedings of the 2019 International Conference on Learning Representations. ICLR, 2019.
81 Hoover B, Strobelt H, Gehrmann S. exBERT: A visual analysis tool to explore learned representations in transformer models. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. ACL, 2019. 187–196. [doi: 10.18653/v1/2020.acl-demos.22.]
82 Le Q, Mikolov T. Distributed representations of sentences and documents. Proceedings of the 31st International Conference on International Conference on Machine Learning. Beijing: JMLR.org, 2014.
83 Koren Y. Drawing graphs by eigenvectors: Theory and practice. Computers & Mathematics with Applications, 2005, 49(11–12): 1867–1888. [doi: 10.1016/j.camwa.2004.08.015]
84 Görg C, Liu ZC, Stasko J. Reflections on the evolution of the Jigsaw visual analytics system. Information Visualization, 2014, 13(4): 336–345. [doi: 10.1177/1473871613495674]
85 Ji XN, Tu YM, He WB, *et al*. USEVis: Visual analytics of attention-based neural embedding in information retrieval. Visual Informatics, 2021, 5(2): 1–12. [doi: 10.1016/j.visinf.2021.03.003]
86 Park C, Na I, Jo Y, *et al*. SANVis: Visual analytics for understanding self-attention networks. Proceedings of the 2019 IEEE Visualization Conference (VIS). Vancouver: IEEE, 2019. 146–150. [doi: 10.1109/visual.2019.8933677]
87 Covington P, Adams J, Sargin E. Deep neural networks for YouTube recommendations. Proceedings of the 10th ACM Conference on Recommender Systems. Boston: ACM, 2016. 191–198. [doi: 10.1145/2959100.2959190]
88 任磊, 杜一, 马帅, 等. 大数据可视分析综述. 软件学报, 2014, 25(9): 1909–1936. [doi: 10.13328/j.cnki.jos.004645]
89 Smilkov D, Thorat N, Nicholson C, *et al*. Embedding projector: Interactive visualization and interpretation of embeddings. arXiv:1611.05469, 2016.
90 Liu SX, Wang XT, Liu MC, *et al*. Towards better analysis of machine learning models: A visual analytics perspective. Visual Informatics, 2017, 1(1): 48–56. [doi: 10.1016/j.visinf.2017.01.006]
91 Pruksachatkun Y, Yeres P, Liu HK, *et al*. jiant: A software toolkit for research on general-purpose text understanding models. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. ACL, 2020. 109–117.
92 Kahng M, Andrews PY, Kalro A, *et al*. ActiVis: Visual exploration of industry-scale deep neural network models. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 88–97. [doi: 10.1109/tvcg.2017.2744718]
93 Ellison NB, Steinfield C, Lampe C. The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. Journal of Computer-mediated Communication, 2007, 12(4): 1143–1168. [doi: 10.1111/j.1083-6101.2007.00367.x]
94 Li X, Roth D. Learning question classifiers. Proceedings of the 19th International Conference on Computational Linguistics. Taipei: ACL, 2002. 1–7. [doi: 10.3115/1072228.1072378]
95 窦慧, 张凌茗, 韩峰, 等. 卷积神经网络的可解释性研究综述. 软件学报, 1–27. https://doi.org/10.13328/j.cnki.jos.006758. [2023-05-18].
96 Chen ZP, Zheng YJ, Li XJ, *et al*. Interactive trimap generation for digital matting based on single-sample learning. Electronics, 2020, 9(4): 659. [doi: 10.3390/electronics9040659]
97 王格荣. 基于语义对齐的零样本图像分类研究 [硕士学位论文]. 西安: 西安电子科技大学, 2022.
98 Ramesh A, Pavlov M, Goh G, *et al*. Zero-shot text-to-image generation. Proceedings of the 38th International Conference on Machine Learning. ICML, 2021. 8821–8831.
99 梁俊杰, 韦舰晶, 蒋正锋. 生成对抗网络 GAN 综述. 计算机科学与探索, 2020, 14(1): 1–17. [doi: 10.3778/j.issn.1673-9418.1910026]

(校对责编: 牛欣悦)