doi:10.11920/xnmdzk.2025.03.010

深度学习在解码大脑语音信息中的应用

杨淑淇[1],谭　颖[1,2]
(1.西南民族大学计算机与人工智能学院,四川 成都 610041;2.西南民族大学计算机系统国家民委重点实验室,四川 成都 610041)

摘　要:医疗保健和神经科学界长期致力于从大脑活动中解码语言信息.脑机接口在支持获得性脑损伤患者通过大脑信号重新与周围环境建立交流方面获得了突破性进展.然而,获取脑信号的方式无法长期地维持且提取到的脑信号中存在大量噪声,因此提高数据的信噪比具有重要意义.近年来,人工智能在提取和汇聚大量原始数据特征方面展现出了卓越的性能.总结了一系列基于深度学习从人类大脑信号中解码语音的特征提取技术.首先对这些技术涉及的深度学习方法进行了介绍,接下来对这些技术的具体应用进行了列举,最后对如何更好将人工智能技术应用于解码大脑语音信息做出了展望.

关键词:神经元解码;音频和语音处理;深度学习;计算认知科学
中图分类号:TP11　　**文献标志码**:A　　**文章编号**:2095-4271(2025)03-0315-08

Deep learning in decoding brain speech information

YANG Shuqi[1], TAN Ying[1,2]
(1. School of Computer Science and Artificial Intelligence, Southwest Minzu University, Chengdu 610041, China;
2. State Ethnic Affairs Commission Key Laboratory for Computer Systems, Southwest Minzu University, Chengdu 610041, China)

Abstract: The healthcare and neuroscience communities have long worked on decoding linguistic information from brain activity. Brain-computer interfaces have made breakthroughs in supporting patients with acquired brain injury to re-establish communication with their surroundings through brain signals. However, the acquisition of brain signals cannot be sustained over a long period of time and there is a lot of noise in the extracted brain signals, so it is important to improve the signal-to-noise ratio of the data. In recent years, artificial intelligence has demonstrated excellent performance in extracting and aggregating features from large amounts of raw data. This paper summarized a series of feature extraction techniques for decoding speech from human brain signals based on deep learning. First, the paper provided an introduction to the deep learning methods involved in these techniques, then enumerated the specific applications of these techniques, and gave an outlook on how to better apply artificial intelligence techniques to decoding brain speech information.

Keywords: neuronal decoding; audio and speech processing; deep learning; computational cognitive science

沟通交流占据着人类生活的一大部分,促进着信息共享、人际交往等.一些因为事故、中风、感染和退行性神经系统疾病等原因造成获得性脑损伤的患者往往存在着语言和沟通能力的障碍,这会极大影响其生活质量[1].为了使患有语言障碍者与周围的环境进行快速交流,脑机接口(Brain-computer Interface,BCI)被用于辅助大脑与设备信息的直接交换[2].BCI 设备大致可分为无创和有创两种类型.非侵入性神经成像技术作为一种替代方法可以为获得性脑损伤患者提供有效的无声交流,它通常涉及利用脑电图(electro-

收稿日期:2024-04-25
通信作者:谭颖(1974-),男,教授,研究方向:脑科学、深度学习.E-mail:ty_edu@163.com
基金项目:中央高校基本科研业务费专项资金优秀学生培养工程项目(2023NYXXS046)

encephalogram,EEG)、脑磁图(magnetoencephalogram,MEG)来测量大脑活动.然而,这种交流通常显著慢于正常对话速率且需要持续关注提供视觉反馈的计算机屏幕[3].此外,这些技术所采集到的脑部信息中往往存在不可避免的噪声,因此,如何尽可能地提升信噪比进而提高语音解码速率具有重要意义.皮层电图(Electrocorticography,ECoG)通过获取患者大脑皮层中的微电极反馈出的信息可以直接从参与语音产生的神经元的活动中预测预期的语音[4].尽管 ECoG 的采集依赖于外科手术植入,但患者无须保持长时间的高度集中[5],这就使得其所采集到的语音信息质量较高,但距离患者沟通能力的完全恢复仍有很长的距离.因此,不论是基于 BCI 的间接或直接大脑语音编码方法,帮助获得性脑损伤患者实现有效交流的关键挑战在于提高信号质量和提升语音预测的准确率.

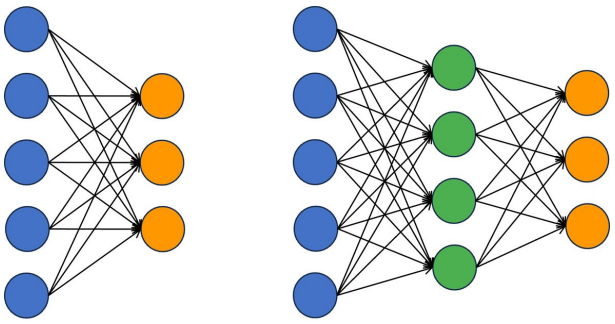
深度学习(Deep Learning,DL)是一种端到端算法,具有自动特征学习功能[6].由于采集到的生物数据信息进行分析、决策等诸多方面都具有复杂的非线性关系,因此很适用于 DL 对其进行特征计算.得益于 DL 模型的强大性能,在对原始数据进行预处理后,用不同的 DL 模型对神经元信号中隐含的特征信息进行提取,克服了获得的大脑语音信号中存在大量噪声的局限性,从而让语言障碍患者重新与周围环境进行交流成为可能.随着 DL 技术的发展和数据规模的逐渐扩大,人们逐渐希望模型能在训练过程中提取到更加普适的特征表示,进而实现无须额外训练就可以快速地将模型迁移到相似的具体任务上.预训练作为 DL 模型的一种训练策略,它在与目标任务相关的大规模数据上训练模型,获得的预训练模型可以学习到更通用普适的表示[7].在基于大脑语音解码任务中,数据资源通常具有有限性.预训练模型的出现使得其可以在通用语音识别任务上获得表现优异的初始化参数,进而在大脑语音解码任务的优化过程中较快地趋于收敛且提高预测性能,有望成为一个有潜力的发展方向.

1 方法

本文意在系统地介绍 DL 方法如何解码大脑数据中隐藏的语音表示.在这项工作中,本文将简要介绍这些算法的原理,然后回顾这些算法应用的相关文献,并以此为后续研究者提供一个研究的回顾和未来的发展方向.

1.1 多层感知机

人工神经网络旨在通过模仿生物神经网络的工作方式来模拟智能行为.最简单的人工神经网络是单层结构,由输入层和输出层组成(图 1(a)).然而,尽管在输出层使用了非线性激活函数,但单层神经网络对于复杂的数据模式往往性能较差.



(a)单层神经网络架构 (b)多层神经网络的架构

图 1 神经网络架构

Fig.1 Architecture of neural networks

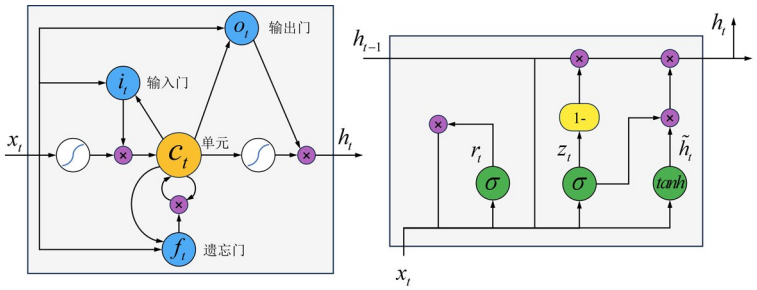
多层感知器(Multilayer Perceptron,MLP),在输入层和输出层之间纳入若干个隐藏层以突破上述限制(图 1(b)).其中,每层包含多个单元,这些单元完全连接到相邻层的单元,但同一层中的单元之间没有连接.反向传播是一种计算 MLP 中梯度的有效算法[8],它通过网络将误差值从输出层传播回输入层,一旦得到所有层的梯度向量,就更新参数.在损失函数收敛或达到预定义的迭代次数之前,更新过程停止,网络获得模型参数.

1.2 递归神经网络

递归神经网络(Recurrent Neural Network,RNN)能够从顺序和时间序列数据中学习特征和长期依赖关系.最流行的 RNN 架构是长短期记忆网络(Long Short-Term Memory,LSTM)[9],它由记忆单元 C_t、遗忘门 f_t、输入门 i_t 和输出门 O_t 组成(图 2(a)).

存储单元将相关信息一直传输到序列链,这些门控制来自各种来源的激活信号,以决定向存储单元添加和删除哪些信息.与基本的 RNN 不同,LSTM 能够通过上面介绍的门来决定是否保留现有记忆.理论上,如果 LSTM 从输入的序列数据中学习到一个重要的特征,它可以长时间保持这个特征,从而捕获潜在的长期依赖关系.一种流行的 LSTM 变体是门控循环单元(Gated recurrent Unit,GRU)(图 2(b)),它将遗

忘门和输入门合并为一个“更新门”,并将记忆单元状态和隐藏状态组合为一个状态.更新门决定添加和丢弃多少信息,重置门决定忘记多少以前的信息.这使得 GRU 比标准 LSTM 更简单[10].



(a)长期短记忆网络　　(b)门控循环单元

图 2　长期短记忆网络和门控循环单元

Fig.2　Long-term short memory network and gated recurrent unit

1.3　卷积神经网络

卷积神经网络(Convolutional Neural Networks,CNN)直接将二维或三维图像作为输入,更好地保留和利用相邻像素或体素之间的结构信息[11].然而语音信息往往以一维向量的形式存在,因此,应用 CNN 时往往需将语音信息转换为矩阵形式.

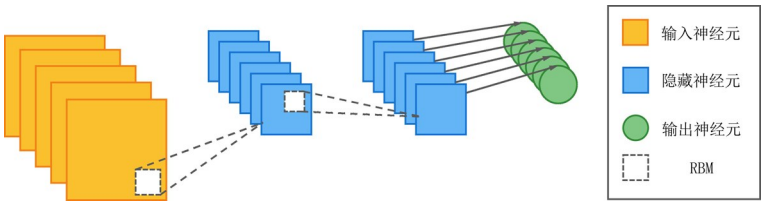


图 3　卷积神经网络的体系结构

Fig.3　Architecture of convolutional neural network

从结构上讲,CNN 是一个层的序列,每一层通过一个可微函数将一层激活转换为另一层激活.图 3 显示了 CNN 架构,它由三种类型的神经层组成:卷积层、池化层和全连接层[12].卷积层取一小块输入作为局部感受野,然后利用各种可学习的核对感受野进行卷积以生成多个特征映射.池化层执行非线性下采样,以减小下一卷积层的输入体的空间维度.全连接层将三维或二维输入映射为一维特征向量.

1.4　Transformer

RNN 网络模型及其变体往往存在长距离依赖等问题,使得 Transformer 模型[13]的自注意力结构(图 4)取代了在自然语言处理任务中常用的 RNN 网络结构.自注意力机制拥有强大的记忆力,能够记住更长距离的信息.更重要的是,自注意力机制的引入使得 Transformer 模型支持并行计算,极大地降低了运算速率.

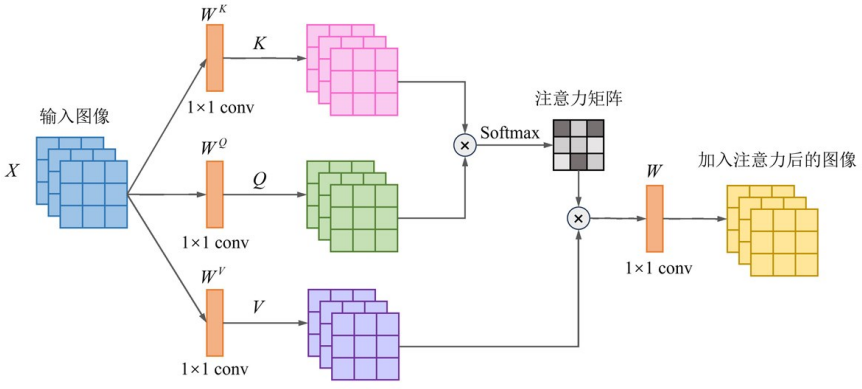


图 4　注意力机制在图像中的使用

Fig.4　Use of attention mechanisms in images

在实际中,自注意力块接收的是输入 X 或者上一个编码块的输出.如图 4 所示,X 与引入的可学习权重矩阵 W^Q、W^K、W^V 相乘得到向量 Q(查询),K(键),V(值),Q,K,V 通过自注意力块的输入经过线性变换得到.为进一提升模型性能,Transformer 还提出了多头注意力机制,该机制对多个注意力模块进行单独计算,不仅获得了更多层面的语义信息并且极大地丰富了信息的汇聚表示.

2 计算机辅助方法在解码大脑语音信息中的应用

思维在大脑中常常以对话的形式存在.因此,人类即使因为外部创伤而无法发出声音,大脑仍然可以将脑海中的信息组织为语言.随着科学技术的日益发展壮大,研究人员试图开发基于语音意象的 BCI,以帮助存在语言障碍的患者重新实现日常交流.

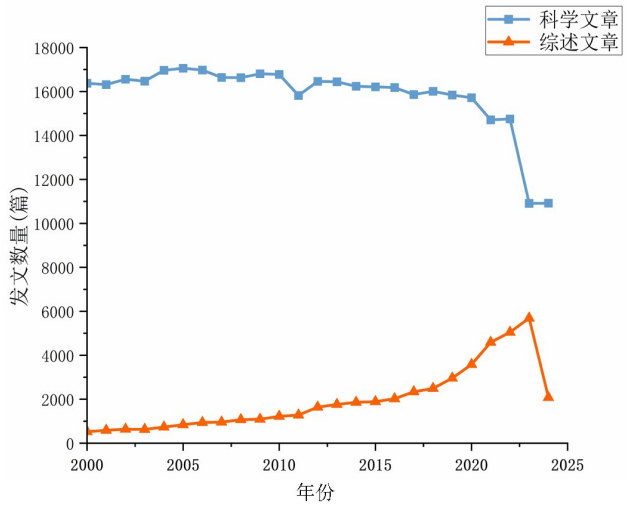


图 5 2000-2024 年(截止至 2024 年 5 月)研究型文章发表数量

Fig.5 Number of research based articles published 2000-2024 (as of May 2024)

如图 5 所示,本文对 2000-2024 年(截止至 2024 年 5 月)在相关学术引擎上以“脑机接口”“语音解码”为关键词的研究型文章发表数量进行了统计.从 2000 年开始,将 BCI 用于解码大脑中的语音信号的科学文章发文量保持稳定趋势,综述性文章呈现上升趋势.值得注意的是,尽管 2023 年科学文章发文量有所下降,但综述性文章仍呈现上升趋势.此外,截止至 2024 年 5 月,科学文章的发文量已经与 2023 年全年持平,说明该领域的研究热度正在持续上涨.

2.1 大脑语音信息公开数据集

由于大脑数据的获取成本和道德伦理限制,当前有关大脑语音解码相关的公开数据集数量较少.本文主要介绍领域内较为广泛使用的 Zuco 数据集以及两个近期公开的代表性数据集.

Zuco(Zurich Cognitive Language Processing Corpus)[14]是一个结合了 EEG 和眼动追踪记录的数据集,记录了受试者阅读自然句子时的生理数据.该数据集共收集了包含 12 名成年受试者的记录,每个受试者参与研究的时间为 4 到 6 小时,共涵盖 21 629 个单词,分布在 1 107 个句子和 154 173 次注视中.研究共包含三个不同的任务:受试者阅读来自电影评论的句子、受试者阅读传记句子和受试者进行基于句子的理解练习.在这个过程中,眼动追踪数据包括注视时间、总阅读时间、首次注视时间、单次注视时间和回视时间.此外,该数据集还记录了受试者的瞳孔大小和注视次数.脑电图数据则使用 128 通道的 EEG Geodesic Hydrocel 系统以 500 Hz 的采样率记录,并通过 Automagic MATLAB 工具进行预处理.

Wilson 等人[15]收集了包含 12 名参与者在感知和想象三种语义概念(花、企鹅、吉他)时的脑电数据,共采集视觉图像、视觉正字法、听觉三种感官模态,每种任务中的语义概念均通过不同复杂度的视觉和听觉刺激进行呈现.每个参与者通过 124 个 EEG 通道记录脑电活动,采样率为 1 024 Hz.参与者均完成了想象力生动性评估测试,并在一个隔音和遮光的房间中进行实验.数据处理包括坏通道检测、重新参考、滤波、伪迹移除和数据分段.技术验证包括事件相关电位、跨试验一致性和平均功率谱密度的计算.

MEG-MaSC 数据集[16]包含 27 名受试者(包括 15 名女性、12 名男性,平均年龄 24.8 岁)在聆听自然语言故事时的高质量 MEG 记录.每位参与者进行了两次独立的会话,每次会话中聆听四个不同的虚构故事,数据集标记了每个单词和音素的起始和结束时间,并按照“脑成像数据结构”的标准组织.所使用的四个虚构故事选自手工标注子语料库,分别为 LW1、Cable Spool Boy、Easy Money 和 The Black Willow.每个故事的音频由文本转语音技术生成,并使用不同的语速和声音来减少语言特征与声学表征之间的相关性.在每个故事之间插入了随机的单词列表和理解问题,

以确保参与者的注意力.

2.2　深度学习在解码大脑语音信息中的应用

近年来,DL 被广泛应用于复杂大脑神经活动过程的神经解码[17].DL 最基础的应用为 MLP,其强大的自适应学习能力和抗干扰的鲁棒性使得其被应用于脑语音信号解码领域.Sereshkeh 等人[18]使用 MLP 对跨多个会话的 EEG 隐蔽语音进行分类,并实现了 75.7%的平均分类准确率.由于语言认知过程需要大脑半球多个皮质结构相互协调活动,因此 Kiroy 等人[19]使用脑电图相干值对其进行刻画,并利用 MLP 对思维中存在的语音片段进行单词检测,以证明 MLP 所生成的口头和大脑内部语音空间连贯模式的显著相似性.然而,MLP 较为简单的结构使得其面临着特征提取能力有限、数据依赖性强等限制,这可能导致大脑信号中的空间模式无法得到充分捕捉.

RNN 的输出不仅仅依靠当前的输入,而且还依赖前一步的神经元状态,因此 RNN 往往也被用于处理和提取语音中的时间上下文信息.Dash 等人[20]首先利用支持向量机对神经信号中语音前、语音和语音后段进行分类,然后通过 LSTM 的顺序模式学习机制有效解码每个时间点的语音活动.为了从皮层电信号中解码出隐藏的语音文本信息,Metzger 等人[21]将大脑信号中提取的特征输入到一个双向 RNN 模型中,并通过 CTC 损失函数确定最可能的句子,以实现跨三种互补的语音相关输出模式的高性能实时解码.Liu 等人[22]试图通过设计一个语音脑机接口用于从患有沟通障碍者恢复说话时的语调,该模型通过一个模块化的 CNN-LSTM 网络并行的独立解码词汇声调和基本音节,直接从颅内记录中合成患者语调.Kuruvila 等人[23]提出了一个联合 CNN-LSTM 的模型来推断听觉注意力,以定量评估人脑在听取特定说话者语音时的抗干扰隔离能力.大脑语音信息的时序特性使得 RNN 能够捕捉数据序列中的依赖关系.然而,RNN 中每个时间片段的隐藏状态都依赖于前一个时间片段的隐藏状态和当前时间片段的输入.这也导致其关键挑战在于其难以较好地处理时序特征间的长距离依赖关系,这显然不利于大脑语音任务中句子的整体理解.

随着 DL 技术的不断发展,CNN 因其在处理空间结构数据方面的卓越表现,成为神经解码领域的另一种强大工具.Cooney 等人[24]使用嵌套交叉验证方法对超参数进行优化,训练了三个不同的专门用于解码 EEG 中语音信息的 CNN.Akashi 等人[25]利用 CNN 估算 EEG 中获得的皮层电流信号内的语音信息,该模型为“呈现”和“回忆”两种不同状态的语音信息之间的神经表征差异提供了新的见解.Kamble 等人[26]采用时间频率表示技术从被试 EEG 信号中捕获时间和频谱信息,并将其输入到 CNN 网络中,以揭示 EEG 数据中与语音相关的空间和时间模式.然而,大脑活动是一个动态的时序过程,这使得 CNN 在处理时序依赖性方面效果较差.

不同于神经网络将输入数据构建为向量或矩阵的表示,Transformer 模型用一种全新的方式来进行表示学习.Xu 等人[27]提出了一种用于听觉注意检测的 Transformer 数据驱动编码器-解码器架构,该模型包含时间自注意和通道注意模块,可以根据脑电图的时间和通道注意机制通过动态分配权重来重建语音网络.Komeiji 等人[28]将 Transformer 编码器整合到“序列到序列”模型中,以解码 EEG 中的语音信息.Chen 等人[29]采用 3D Swim Transformer 作为解码器,将皮层脑电信号转换为可解释语音参数,并将其映射到频谱图的可区分语音合成器中,该框架展现出了高度可重现性.总之,Transformer 模型在大脑语音信号解码任务中展现出强大的潜力和优势,但在实际应用中仍需解决数据需求、计算资源和噪声处理等方面的挑战,以进一步提升应用效果.

综上,巧妙结合不同神经网络模型各自的独特优势,有望从多维度、多角度深入解码大脑中的语音信息,能够更全面地揭示大脑处理语音的复杂机制.尽管 DL 方法被广泛应用于解码大脑语音领域,但仍然存在一些限制.首先,DL 模型通常需要大量的数据来训练,而且在解释性方面可能存在不足.其次,DL 模型的复杂性可能导致过拟合问题,尤其是在数据量有限的情况下.此外,DL 方法的计算资源需求较高,这限制了其在实际当中的应用.开发能够在相对较低的计算资源下实现较好的性能,并且能够满足在各种应用场景下的需求的 DL 模型亟待实现.

2.3　预训练模型在解码大脑语音信息中的应用

随着研究的不断深入,人们时常设想机器是否能像人类一样,可以轻松地将所学习到的知识和常识进行推理,从而在一个未知的任务中做出决策.由此衍

生而发展出的预训练模型就是将依赖于大量标注数据训练的DL模型,进一步推广至大规模、可复制的工业生产阶段.当前,自然语言处理和语言神经科学研究之前缺乏有效的连接[30].在解码大脑语音信息任务中引入预训练模型,对其"大脑—模型",有望填补这一空缺.Tang等人[31]将候选文字序列限制为标准格式的英语,提出生成性预训练转换器对单词序列引发脑波记录的可能性进行评分,进而生成可理解的单词序列,以恢复患者脑海中感知语音、想象语音甚至无声视频的含义.Défossez等人[32]结合Transformer和CNN提出了一种对比学习模型,将预训练好的"语音模块"和引入的"大脑模块"进行对齐,进而从健康个体的非侵入性录音中解码感知语音的自我监督表示.Li等人[33]使用预训练的深度神经网络来对大脑中的时序信息进行卷积,接着利用注意力机制学习时序信息的上下文.结果表明,该神经编码模型可以有效建模和评估听觉皮层的神经编码.然而,由于人类大脑会不断预测跨越多个时间尺度的表征层次,通用语言预训练模型在解码大脑语音任务上的迁移性表现欠佳.因此,Caucheteux等人[34]将大语言预训练模型的线性激活映射到大脑对语言的反应上,通过建立多个时间尺度的跨越来改善预训练模型的大脑映射,揭示了人类认知计算基础的协同作用.尽管上述预训练模型同样面临着某些任务中可能出现不期望的偏差、受试者隐私和安全保护等不同方面的限制,但其展现出强大的从"大脑"到"文本"的推理解码能力仍是未来值得深挖的方向之一.

综上,表1展示了上文所提到的所有用于解码大脑语音信息的DL模型.

表1 用于解码大脑语音信息模型一览表

Table 1 List of models used to decode brain speech information

文献	年份	数据集	模态	基线模型
Sereshkeh等人[18]	2017	共12人,每人两次并重复执行三种不同的任务	EEG	MLP
Kiroy等人[19]	2022	共10人,来自14个通道的音频	EEG	MLP
Dash等人[20]	2020	共8人,使用五个常用短语作为刺激	MEG	SVM、LSTM
Metzger等人[21]	2023	共1人,529-phrase-AAC:包含529个句子,由372个独立单词组成;50-phrase-AAC:从529-phrase-AAC中选择50个由119个独立单词组成的句子来创建.1024-word-General:包含9 655个句子,采样的1 024个独立单词组成	ECoG	RNN
Liu等人[22]	2023	共5人,受试者共发音八个指定的声调音节	ECoG	CNN、LSTM
Kuruvila等人[23]	2021	共27人,共听取5个演示文稿,共计30分钟共18人,共听取50分钟的丹麦语有声读物.共16人,共听取48分钟的荷兰故事和12分钟的演讲	EEG	CNN、LSTM
Cooney等人[24]	2020	共15人,受试者针对每个元音和单词进行想象,共计50次	EEG	CNN
Akashi等人[25]	2021	共10人,受试者听取和回忆原音字符的脑电信号	EEG	CNN
Kamble等人[26]	2023	共15人,均为男性,受试者受执行了特定动作的言语想象,每个特定单词执行15次想象,每次想象持续5秒,两个单词间具有3秒的间隔时间	EEG	CNN
Xu等人[27]	2022	-	EEG	Transformer
Komeiji等人[28]	2022	共7人(男性:4人,女性:3人),共朗读80个句子	ECoG	Transformer
Chen等人[29]	2024	共48人,每人参与包括听觉重复、听觉命名、句子补全、单词阅读和图片命名五项任务.共提供50个重复的独立单词,共400次实验	ECoG	CNN、LSTM、Transformer
Tang等人[31]	2023	受试者共听取16小时的自然语言叙事故事,并记录特定刺激短语的语义特征.	MEG	Transformer
Défossez等人[32]	2023	共175人,受试者聆听短篇故事句子作为刺激	MEG、EEG	Transformer、CNN
Li等人[33]	2023	共9人,使用TIMIT语料库中的一组599个英语句子来评估早期和晚期听觉系统的神经反应.	ECoG	Transformer、CNN
Caucheteux等人[34]	2023	共304人,27个故事,时长从7到56分钟不等;总共4.6小时的独特刺激,平均每位参与者26分钟,时长从7到99分钟不等	MEG	Transformer

3 结论

本文首先介绍了部分经典的DL算法.然后,针对DL算法在解码大脑神经元中的语音信息中的应用展开了讨论.在不久的将来,利用人工智能辅助研究人

员理解人类各种认知行为模式将是一种至关重要的方法.

近年来,随着医学影像技术的不断发展,其所采集到的大脑神经元信息的不断丰富,越来越多的 DL 方法被应用到脑信号信息提取任务中.虽然 DL 不需要特征选择,但基于 DL 的医疗数据解码往往需要对采集到的生理信号进行各种预处理,无法实现真正的端到端学习.此外,由于数据集的样本量不够大,DL 模型容易过拟合,模型的泛化能力不够强.而且,DL 对超参数的配置依赖性很强,可能会导致其性能波动较大,有时经验是影响结果的一大因素.如何更好地将 DL 方法与理论证明相结合,是未来需要努力的方向之一.随着 DL 方法的发展,预训练模型也逐渐走进人们的视野.通过微调已经训练好的模型参数,预训练模型可以推广至更加通用的领域.然而,这种推广能力仍然有限,难以应对特定专业领域的任务.

大脑的不同区域因其功能分区上的整合与分离呈现出网络结构.图卷积神经网络(Graph Convolutional Network,GCN)[35]是近年来最新提出的 DL 方法之一.它借鉴 CNN 的思想设计图形卷积网络的体系结构,可以用于处理复杂的网络拓扑结构[36].因此,将 GCN 应用于分析复杂大脑网络,可以提取和汇聚大脑网络结构数据中的特征信息.近年来,许多研究者也试图将 GCN 与 Transformer 相结合,以期从非欧几里得结构中获取大脑特征的长距离依赖关系[37-39].目前,在解码大脑神经元中的语音信息领域,结合 GCN 的研究仍处于发展的初期.鉴于其在上述基于 DL 的大脑网络分析研究中展现出的优异性能,本文合理推测 GCN 与 Transformer 的结合可以为大脑语音解码领域带来新的活力.

参考文献

[1] MONROE P,HALAKI M,KUMFOR F,et al.The effects of choral singing on communication impairments in acquired brain injury:A systematic review[J].InternationalJournal of Language & Communication Disorders,2020,55(3):303-319.

[2] 刘迎欣,李明,于扬,等.混合脑机接口在人机交互领域的应用综述[J].控制理论与应用,2023,40(12):2077-2089.

[3] RAUSA V C,SHAPIRO J,SEAL M L,et al.Neuroimaging in paediatric mild traumatic brain injury:a systematic review[J].Neuroscience and Biobehavioral Reviews,2020,118:643-653.

[4] BRUMBERG J S,NIETO-CASTANON A,KENNEDY P R,et al.Brain - computer interfaces for speech communication[J].Speech Communication,2010,52(4):367-379. DOI:10. 1016/J.specoM.2010. 01. 001.

[5] BOCQUELET F,HUEBER T,GIRIN L,et al.Key considerations in designing a speech brain-computer interface[J].Journal of Physiology-Paris,2016,110(4):392-401.

[6] LECUN Y,BENGIO Y,HINTON g.Deep learning[J].Nature,2015,521(7553):436-444.

[7] 宫丽娜,周易人,乔羽,等.预训练模型在软件工程领域应用研究进展[J/OL].软件学报,1-26[2024-09-08].https://doi.org/10. 13328/J.cnki.jos.007143.

[8] RUMELHART D E,HINTON G E,WILLIAMS R J.Learning representations by back-propagating errors[J].Nature,1986,323(6088):533-536.

[9] GRAVES a.Long short-term memory[J].Supervised sequence labelling with recurrent neural networks,2012:37-45.

[10] DEY R,SALEM F M.Gate-variants of gated recurrent unit (GRU) neural networks[C]//2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS).August 6-9,2017.Boston,Ma.IEEE,2017:1597-1600.

[11] 刘帅师,程曦,郭文燕,等.深度学习方法研究新进展[J].智能系统学报,2016,11(5):567-577.

[12] KRIZHEVSKY A,SUTSKEVER I,HINTON G e.ImageNet classification with deep convolutional neural networks[J].Communications of the ACM,2017,60(6):84-90.

[13] VASWANI A,SHAZEER N,PARMAR N,et al.Attention is all you need[J].Advances in neural information processing systems,2017,30.

[14] HOLLENSTEIN N,ROTSZTEJN J,TROENDLE M,et al.ZuCo,a simultaneous EEG and eye-tracking resource for natural sentence reading[J].Scientific Data,2018,5(1):1-13.

[15] WILSON H,GOLBABAEE M,PROULX M J,et al.EEG-based BCI dataset of semantic concepts for imagination and perception tasks[J].Scientific Data,2023,10(1):386.

[16] GWILLIAMS L,FLICK G,MARANTZ A,et al.Introducing MEG-MASC a high-quality magneto-encephalography dataset for evaluating natural speech processing[J].Scientific Data,2023,10(1):862.

[17] LIVEZEY J A,GLASER J i.Deep learning approaches for neural decoding across architectures and recording modalities[J].Briefings in Bioinformatics,2021,22(2):1577-1591.

[18] REZAZADEH SERESHKEH A,TROTT R,BRICOUT A,et al.EEG classification of covert speech using regularized neural networks[J].ACM Transactions on Audio,Speech,and Language Processing,2017,25(12):2292-2300.

[19] KIROY V N,BAKHTIN O M,KRIVKO E M,et al.Spoken and inner speech-related EEG connectivity in different spatial direction[J].Biomedical Signal Processing and Control,2022,71:103224.

[20] DASH D,FERRARI P,DUTTA S,et al.NeuroVAD:Real-time voice

activity detection from non-invasive neuromagnetic signals[J].Sensors, 2020,20(8):2248.

[21]METZGER S L,LITTLEJOHN K T,SILVA A B,et al.A high-performance neuroprosthesis for speech decoding and avatar control[J].Nature, 2023,620(7976):1037-1046.

[22]LIU Y,ZHAO Z H,XU M P,et al.Decoding and synthesizing tonal language speech from brain activity[J].Science Advances,2023,9(23):eadh0478.

[23]KURUVILA I,MUNCKE J,FISCHER E,et al.Extracting the auditory attention in a dual-speaker scenario from EEG using a joint CNN-LSTM model[J].Frontiers in Physiology,2021,12:700655.

[24]COONEY C,KORIK A,FOLLI R,et al.Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG[J].Sensors,2020,20(16):4629.

[25]AKASHI W,KAMBARA H,OGATA Y,et al.Vowelsound synthesis from electroencephalography during listening and recalling[J].Advanced Intelligent Systems,2021,3(2):2000164.

[26]KAMBLE A,GHARE P H,KUMAR V,et al.Spectral analysis of EEG signals for automatic imagined speech recognition[J].IEEE Transactions on Instrumentation and Measurement,2023,72:1-9.

[27]XU Z H,BAI Y R,ZHAO R,et al.Decoding selective auditory attention with EEG using a transformer model[J].Methods,2022,204:410-417.

[28]KOMEIJI S,SHIGEMI K,MITSUHASHI T,et al.Transformer-basedestimation of spoken sentences using electrocorticography[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP).Singapore:IEEE,2022:1311-1315.

[29]CHEN X P,WANG R,KHALILIAN-GOURTANI A,et al.A neural speech decoding framework leveraging deep learning and speech synthesis[J].Nature Machine Intelligence,2024,6(4):467-480.

[30]YU S Y,GU C Y,HUANG K X,et al.Predicting the next sentence (not word) in large language models:What model-brain alignment tells us about discourse comprehension[J].Science Advances,2024,10(21):eadn7744.

[31]TANG J,LEBEL A,JAIN S,et al.Semantic reconstruction of continuous language from non-invasive brain recordings[J].Nature Neuroscience,2023,26(5):858-866.

[32]DÉFOSSEZ A,CAUCHETEUX C,RAPIN J,et al.Decoding speech perception from non-invasive brain recordings[J].Nature Machine Intelligence,2023,5(10):1097-1107.

[33]LI Y N,ANUMANCHIPALLI G K,MOHAMED A,et al.Dissecting neural computations in the human auditory pathway using deep neural networks for speech[J].Nature Neuroscience,2023,26(12):2213-2225.

[34]CAUCHETEUX C,GRAMFORT A,KING J R.Evidence of a predictive coding hierarchy in the human brain listening to speech[J].Nature Human Behaviour,2023,7(3):430-441.

[35]KIPF T N,WELLING M.Semi-supervised classification with graph convolutional networks[EB/OL].2016:1609.02907.https://arxiv.org/abs/1609.02907v4.

[36]徐冰冰,岑科廷,黄俊杰,等.图卷积神经网络综述[J].计算机学报,2020,43(5):755-780.

[37]CAI H J,GAO Y,LIU M H.Graph transformer geometric learning of brain networks using multimodal MR images for brain age estimation[J].IEEE Transactions on Medical Imaging,2023,42(2):456-466.

[38]CHEN D D,ZHANG L C.FE-STGNN:Spatio-Temporal Graph Neural Network withFunctional andEffective Connectivity Fusion forMCI Diagnosis[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2023.Cham:Springer Nature Switzerland,2023:67-76.

[39]ZHU Q,LI S R,MENG X S,et al.Spatio-temporal graph hubness propagation model for dynamic brain network classification[J].IEEE Transactions on Medical Imaging,2024,43(6):2381-2394.

(责任编辑:张阳,殷锋,付强,和力新,肖丽;英文编辑:周序林,郑玉才)