语音图文信息处理中的深度学习方法进展专刊序言

柯登峰 [1] 俞 栋 [2] 贾 珈 [3]

最近几年来, 深度学习赚足了世人的眼球. 如果说, 前几年深度学习在工业界受到狂热追捧的话, 那么今年开春以来, 深度学习则是受到普通老百姓的百般青睐. 随着阿法狗 (AlphaGo) 采用深度学习技术打败了围棋冠军李世石的消息传开, 街头巷尾男女老少都在津津乐道于深度学习技术与人工智能的未来. 深度学习技术本质上是深度神经网络技术, 是神经网络发展的重要阶段. 它在历史上经历了许多挫折和磨难, 才成就了如今的辉煌.

起初, 神经网络技术只在神经科学和生物物理学领域流行, 科学家用它来解释神经细胞的工作原理, 反对弗洛伊德学派《自我与本我》中对癔症和催眠等精神分析学的解释. 神经元工作的模型最早由神经生理学家沃伦 · 麦卡洛克 (Warren McCulloch) 和数学天才少年沃尔特 · 皮兹 (Walter Pitts) 于 1943 年合作提出. 他们认为看似神秘的精神失常是来源于大脑里的神经元的激发异常, 其工作原理是纯机械式的, 可被离散化的时间信号 ($t=0,1,2,\cdots$) 所表示, 各时刻中神经元的状态可表示为 0 或 1. 1957 年, 心理学家法兰克 · 罗森布拉特 (Frank Rosenblatt) 在此基础上建立了感知机模型, 并借鉴神经心理学家赫布 (Hebb) 提出的学习规则, 实现了可以自主学习的感知机. 他用感知机实现了一些简单的视觉处理工作, 证实了感知机的学习能力和分辨能力. 美国政府对此十分重视, 给予大力支持. 锋芒毕露的罗森布拉特招致人工智能符号逻辑学派领军人物闵斯基 (Minsky) 等的不满, 以感知机无法解决 XOR (Exclusive or) 这种最简单的数学问题进行质疑和打击, 从此政府不再对神经网络研究给予资金支持, 使得神经网络技术陷入了 20 年大饥荒时期.

直到 1974 年, 保罗 · 乌博思 (Paul Werbos) 的博士论文提出了多层感知机网络和 BP (Back propagation) 算法, 成功解决了 XOR 问题, 才使神经网络技术有了转机, 但研究的人员依然较少. 1984 年, 霍普菲尔德用模拟电路实现了自己提出的新型神经网络, 该网络可以解决模式识别问题, 还可以给出组合优化问题的近似解, 极大地振奋了神经网络领域的研究. 此时, 在诺贝尔生理学奖得主克里克 (Crick) 等的鼓励下, 开始了 "联接主义 (Connectionism)" 运动 (该运动主张将心理学、人工智能和心理哲学联接在一起). 这个运动的带头人则是深度学习鼻祖杰弗里 · 辛顿 (Geoffrey Hinton) 和两位心理学家鲁梅尔哈特 (Rumelhart) 和麦克利兰德 (McLelland). 而卷积神经网络 (Convolutional neural network, CNN) 的提出者 — 言 · 乐村 (Yann LeCun) 则是辛顿的学生. 然而, 那时候电脑的处理能力还远不能满足深度学习的要求, 这使得神经网络技术的发展十分困难.

在加拿大高级研究院 (Canadian Institute for Advanced Research, CIFAR) 基金支持下, 辛顿于 2006 年提出了一种快速训练深度信任网络 (Deep belief nets) 的方法. 多数人认为, 这一工作标志着神经网络进入深度学习阶段. 2009 年, 辛顿的学生默罕穆德 (Mohamed) 首次将深度学习应用于 Timit 库语音识别并取得当时全球最佳的识别率. 随后, 微软和谷歌将深度学习应用于大词汇量语音识别获得成功, 识别性能比当时最好的 GMM-HMM (Gaussian mixture model-hidden Markov model) 技术相对提升了 20 %～30 %. 紧接着在 2012 年, 辛顿的学生将深度学习技术用于图像识别国际比赛并夺得冠军, 在 ImageNet 上前 5 名候选错误率为 15.3 %, 远超第二名 26.2 % 的成绩. 值得一提的是, 在深度学习出来之前, 语音领域已经将传统的 GMM-HMM 性能发挥到极致, 各种特征优化技术、自适应技术、区分度训练技术以及时序化训练技术均被优化到极限, 识别性能提升已经非常困难. 语音领域的科研工作者们热切期待还有更好的技术出现, 以拯救语音识别的未来. 而图像识别领域的最好成绩基本上被支持向量机 (Support vector machine, SVM) 所垄断, 寻找颠覆性的新技术也成了大家心里的期望. 深度学习连续斩获语音和图像两个领域的桂冠, 引起了科研人员高度的关注, 纷纷购买设备开始了深度学习的研究.

深度学习的成就也引来了 IT 巨头的哄抢. 谷歌用 4 亿美元巨资收购了辛顿三个人的小公司, 百度宣布成立深度学习研究院, 脸谱 (Facebook) 出手抢走了卷积神经网络 (CNN) 提出者乐村 (LeCun), 推特 (Twitter) 收购了疯狂比特 (Madbits), 苹果则收购了有声智商 (Vocal IQ). 事情远不止如此, 巨头们对深度学习人才和公司的哄抢还在持续进行中. 神经网络技术沉寂了数十年时间, 终于以深度学习的方式得到了空前繁荣.

多方的投入使深度学习技术在多个方面取得重要进展. 短短几年内, 深度学习颠覆了语音、图

收稿日期 2016-06-01 Manuscript received June 1, 2016
DOI 10.16383/j.aas.2016.y000005
引用格式 柯登峰, 俞栋, 贾珈. 语音图文信息处理中的深度学习方法进展专刊序言. 自动化学报, 2016, **42**(6): 805−806
Citation Ke Deng-Feng, Yu Dong, Jia Jia. Guest editorial for special issue on deep learning for speech, text and image understanding. *Acta Automatica Sinica*, 2016, **42**(6): 805−806
1. 中国科学院自动化研究所 北京 100190 中国 2. 微软雷德蒙研究院 雷德蒙 98052 美国 3. 清华大学 北京 100084 中国
1. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China 2. Microsoft Research, Redmond 98052, USA 3. Tsinghua University, Beijing 100084, China

像、视频等众多领域的算法设计思路, 产生了一系列具有重大价值的新成果. 新型的神经网络结构和学习方法层出不穷. 先是出现了各种大数据下的并行计算方法, 如异步随机梯度下降法 (Asynchronous stochastic gradient descent, ASGD)、同步随机梯度下降法 (Synchronous stochastic gradient descent, SSGD)、分布式免海森法 (Distributed Hessian free) 等, 使得大规模数据学习成为可能; 同时还研究了训练时候批大小 (Batch size)、学习率大小、隐藏节点规模与模型性能的对应关系; 随后又融入了时序化训练算法、多任务学习算法、自适应技术等多种优化技术. 在深度学习帮助下, 很多技术达到了产业化的水平, 于是出现了深度学习模型解码时需要的 SIMD (Single instruction multiple data) 指令优化、内存换页优化、缓存命中率优化、跳帧识别技术、模型压缩技术 (无损的数据压缩和有损的 SVD (Singular value decomposition) 分解)、快速搜索技术等配套技术. 学术界对深度学习的研究也日趋多元化. 有人从数据入手, 研究海量规模数据的筛选、无标注数据对模型的优化、无监督或轻监督学习法对提升模型性能的帮助. 有人从结构入手, 研究时间序列与神经网络的结合 (如 LSTM (Long short term memory)、BLSTM (Bidirectional long short term memory) 等)、端到端解码的结构设计 (如 LSTM + CTC (Connectionist temporal classification) 等)、关注机制与神经网络的结合 (如引入 Attention 模型)、CNN 池化结构优化 (如 Max Pooling 替换成 L2-Pooling 等)、激活函数优化 (如 ReLU (Rectified linear unit) 等)、权值矩阵的函数化、循环网络的展开逼近以及隐层之间级联关系的改造等. 有人从算法入手, 研究神经网络的抗噪能力 (如 SDAE (Stacked denoising autoencoders) 等)、泛化能力 (如 Dropout 等) 以及自适应方法 (如插入线性层、特征联合自适应等). 另外, 还有许多难以总结的零碎研究 (例如, 采用波形信号作为语音识别的输入, 双说话人条件下的 DNN 建模和解码技术, 利用 DNN 的高层输出进行决策树分裂等).

相对于语音和图像领域的突飞猛进发展, 深度学习对自然语言处理的作用似乎不是很突出. 采用深度学习技术对语言模型进行建模可以获得比 N-gram 模型更好的预测效果, 但由于计算速度太慢, 通常只能用于语言概率重打分, 或者转成普通的 N-gram 模型后再使用. 采用深度学习进行机器翻译, 通常只能获得与短语模型相近的翻译质量, 却无法超越短语模型, 且受到命名实体和未登录词的影响过大, 即便引入关注机制、预处理和后处理技术, 依然无法超越传统的短语模型和层次短语模型的翻译效果, 翻译质量有待提高.

即便如此, 我们依然看好深度学习技术未来的发展. 深度学习在自然语言处理领域终有突破的一天, 而深度学习在语音和图像领域的前进脚步也不会停止. 正因如此, 我们组织了 “语音图文信息处理中的深度学习方法进展” 专刊, 以期对当前国内外最前沿的深度学习方法进行全方位多角度报道, 帮助国内科研工作者快速获取最有效的参考资料. 同时, 我们也希望发现一些富有创意的新模型和新方法, 促进语音图文多个领域间科研人员的信息沟通.

本次专刊收到了 60 余篇投稿, 其中不乏有新意的学术思想和方法, 限于出版时间和篇幅, 本刊只收录了其中 16 篇, 包括 4 篇综述文章, 9 篇行业应用文章以及 3 篇基础理论文章. 综述内容主要涵盖了知识库问答、语音分离、视频目标追踪以及人体行为识别 4 个领域; 行业应用文章包括了人脸性别识别、面部表情识别、图像美感分类、视频事件监控、视频人群计数、复杂场景下目标识别、微博实体链、发音器官运动合成以及基音检测等 9 个问题; 基础理论文章从采样算法优化、学习率调整以及人脑仿生学网络结构改进等三个方面对深度学习技术进行改进.

在此我们对作者们的辛勤工作和无私奉献表示深深的感谢, 也希望大家更多地无保留地交流最新技术成果, 共同促进科学技术的向前发展。

客座编委

柯登峰 中国科学院自动化研究所数字内容技术与服务研究中心副研究员. 主要研究方向为语音语言信息处理技术, 深度学习技术.
E-mail: dengfeng.ke@ia.ac.cn
(**KE Deng-Feng** Associate professor at Digital Content Technique and Service Research Center, Institute of Automation, Chinese Academy of Sciences. His research interest covers speech and language processing, deep learning.)

俞 栋 微软雷德蒙研究院首席研究员. 主要研究方向为语音识别, 自然语言处理, 深度学习.
E-mail: dongyu@microsoft.com
(**YU Dong** Principal researcher at Microsoft Research, Redmond. His research interest covers speech recognition, natural language processing, and deep learning.)

贾 珈 清华大学计算机科学与技术系副教授. 主要研究方向为人机语音交互, 情感计算, 深度学习.
E-mail: jjia@mail.tsinghua.edu.cn
(**JIA Jia** Associate professor in Department of Computer Science and Technology, Tsinghua University. Her research interest covers human computer speech interaction, affective computing, and deep learning.)