神经网络在社会大数据方面的应用

魏少杭 学号：20373594 学院：人工智能研究院

2022 年 12 月 30 日

注：第 1 至 5 页为正文，第 6 页为参考文献。

1 综述

1.1 经典神经网络应用简述

1. 前馈神经网络：非循环；常用于预测。2. 径向基网络：用于函数逼近、时间序列预测、分类等。3.**RNN**：序列模型。用于时间序列处理，预测或分类。LSTM 具有记忆功能，处理间隔记忆数据，处理上下文相关性。4. 自动编码器：非监督式，压缩数据、重构。5. **卷积神经网络**：卷积算子实现平移不变性等，池化算子实现减小模型参数、反映数据层次结构的作用。常用于图像、时序特征提取。6. 残差网络：将某层输入的一部分直接传入后面的某些层中，防止结果退化。7. **图神经网络**：处理图域信息的方法，具有较好性能、可解释性。适合处理非欧式距离的图数据。

1.2 神经网络处理社会大数据的优势

社会大数据是应用于社会各个方面、各个领域的大数据，涉及社会媒体信息传播、政府政策制定、人群情感行为分析、医疗公共卫生等方面，需要通过对巨大规模的数据进行挖掘，分析结论，帮助决策者作出科学合理的决策。

大数据的核心流程是数据分析。根据不同应用需求，对数据进行取舍，利用部分或全部数据分析，从而体现大数据本身的价值。随着存储等需求的硬件支持增强，大数据分析变得越来越重要。

使用神经网络处理大数据有以下原因：（1）大数据分析必须使用全体数据而非样本数据。（2）相比于精确的数据分析，更加关注对全量复杂多样数据的分析。（3）与传统的因果关系分析相比，更加关注事物之间存在的相关关系。（4）神经网络处理能够减少传统的决策中的主观因素。基于以上 4 点原因，各类神经网络往往能够发挥其提取特征、分类预测等强大功能，处理社会大数据。

神经网络处理社会大数据的优势：（1）应用范围广：深度学习能够横跨计算机科学、工程技术、统计学，并应用于政治、金融、天文、地理以及社会生活等广泛的领域。（2）模型表达能力强，能够处理具有高维稀疏特征的数据，而大数据所面临的挑战可以通过深度学习的思想、方法和技术及时有效地解决。深度学习拥有了更高效的硬件平台作为支撑，能够解决大数据分析中的若干任务。大数据时代的海量数据解决了早期神经网络由于训练样本不足出现的过拟合、泛化能力差问题。因此，大数据与深度神经网络互相成就。[3]

1.3 社会大数据与神经网络的关系

社会大数据在组织形式上分为结构化、非结构化、半结构化数据。其中，**结构化**、**非结构化大数据**是目前人们主要进行研究的方向，本文也将从结构化大数据——图数据和非结构化大数据两个角度分别对神经网络进行探讨。

结构化的社会大数据与细化的领域大数据的主要组织形式不同。如交通网络、社交网络、引用网络等等**众多社会大数据都是以图数据的形式存在**。图数据可以自然表达社会生活中的数据结构，如交通网络、万维网和社交网络等。**图神经网络**可以很好地分析这类社会网络大数据。

非结构化的社会大数据主要包含自然语言处理、语义挖掘。我们常见的有汉语分词、命名实体识别、新词发现、文本分类和聚类、话题检测、自动摘要生成、情感分析。这些应用任务主要通过**自然语言处理的各类重点神经网络模型如 RNN 及其衍生模型**，结合具体任务目标和挑战来调整技术进行处理。[4]

2 神经网络在图结构的社会大数据方面的应用

2.1 简介

深度神经网络获得巨大发展后，研究人员开始考虑将深度学习的模型引入到图数据中。最初的应用有网络嵌入方法。目前，越来越多学者将深度学习模型迁移到图数据上，进行端到端建模。图卷积神经网络在图数据应用中受到极大的关注。[14]

图卷积神经网络的构建有三个方面的主要挑战：（1）**图数据是非欧式空间数据**，各个结点局部结构互异。而传统卷积神经网络核心的基本算子：**卷积、池化，根本上依赖于数据的平移不变性**。（2）图数据具有**多样的特性**：一方面，具体的案例如社交网络中各个用户有向连接，引文网络中作者、引文异质连接等，多样的图特性给图卷积神经网络的构建带来更多信息，同时也要求图卷积神经网络的设计更复杂、精细化。（3）图数据**规模极大**。社会大数据中的图可能规模极大。例如推荐系统中的商品用户网络、社交网络中的用户网络在建模时对时间、空间均有较高的限制和要求。

图神经网络的建模应用广泛，可以分为节点级别和图级别的应用。节点级别的应用包括节点分类、链接预测等；而图级别的应用包括图的生成、分类等，如药物网络生成、蛋白质网络分类。

2.2 图卷积神经网络主要方法与算子构建

主要方法有（1）关注卷积算子构建的方法。例如基于卷积定理的图卷积神经网络、基于聚合函数的图卷积神经网络。（2）关注图上复杂信息建模的方法。例如，建模边上信息或建模高阶信息的图卷积神经网络。（3）关注训练过程优化的方法。例如深层图卷积神经网络和大规模图卷积神经网络。[14]

利用卷积神经网络对图数据建模的核心在于：如何构建适用于图数据的**图卷积算子和图池化算子**。

2.2.1 图卷积算子构建

（1）谱方法：由于图上平移不变性的缺失，节点域上定义卷积神经网络较为困难。因此考虑**从谱域定义图卷积**。主要利用了信号处理中的卷积定理，对谱空间信号乘法再**傅里叶逆变换**转换到原空间实现图卷积。如式

$$f * g = F^{-1}(F(f) \cdot F(g))$$

中，f、g 表示节点域卷积，而 F 表示傅里叶变换。这避免了因为图数据不满足平移不变性造成的卷积定义困难问题。其中，图上的傅里叶变换依赖于拉普拉斯矩阵的特征向量。通过这种方法构建的卷子算子，启发了基于卷积定理的图卷积神经网络（Spectral CNN）。Spectral CNN 利用卷积定理在每一层定义图卷积算子，并用 BP 算法学习卷积核。

谱卷积神经网络将卷积核作用在谱空间的输入信号上，并利用**卷积定理实现图卷积**，完成节点之间的信息聚合。然后将**非线性激活函数作用在聚合结果上**，**堆叠多层形成神经网络**。不过，该模型产生信息聚合的结点并不一定是临近的结点。其进一步改进为用小波变换代替傅里叶变换实现卷积定理。这带来的好处是小波变换基底的可加速计算和稀疏性使得图卷积神经网络计算复杂度大幅降低。

（2）空间方法：通过注意力机制或递归神经网络等直接从节点域学习聚合函数；或从空间角度定义图卷积神经网络的通用框架，并解释图卷积神经网络内部机制。

2.2.2 图池化算子

在传统 CNN 中，卷积会与池化相结合。池化算子作用是减少学习的参数、反映输入数据的层次结构。图卷积神经网络引入图池化操作主要是为了解决图级别的问题，目的是刻画出网络的等级结构，主要应用于图分类任务。[14]

2.3 图卷积神经网络的新进展

（1）利用**网络额外信息建模**的图卷积网络。图卷积算子忽略了除边、结点特征以外的其他重要信息，例如边上的属性、高阶网络结构信息等。相关的研究成果有符号图卷积网络、对偶图构建方法、权重重调法等。

（2）**高阶结构特征信息建模**的图卷积网络。对结点分类时，相比一阶邻居建模外，通过显示定义高阶结构如三角模体建模或更复杂、特异性的模体建模往往会取得更好的效果。

2.4 图卷积神经网络在社会大数据方面的应用

在社会大数据方面的应用主要包括以下几个领域：网络分析、推荐系统、生物化学研究分析、交通预测等。不同领域包含了不同的图数据，具有较突出的特异性。**如何结合领域知识对图数据利用图卷积神经网络建模是应用层面的关键性问题**。

下面将举出几个例子来说明图卷积神经网络对于社会大数据的应用。

2.4.1 网络结构分析的应用

以引文网络为例，图数据中的结点为论文，连边关系为引用关系。对文章进行分类到领域的过程中，图卷积网络**将结点文本属性和引用网络结构**进行有效建模。[14]

图卷积网络比直接利用内容信息的 MLP 建模分类准确率明显更高。社区发现、互联网在线信息传播问题上，可以通过构建图卷积神经网络**刻画每一个用户的激活状态，关注到传播级联的全局与局部的结构特征**。[13]

2.4.2 推荐系统的应用

在商品与用户的联系建模方面，将**图卷积神经网络**用于很好地**建模图结构属性、结点特征信息，提取局部静止的特征**；而推荐系统被视为矩阵补全问题或用户和商品的链接预测问题。加入循环神经网络模块处理矩阵补全，并图卷积神经网络处理图结构的解决思路效果更好。

2.4.3 交通预测的应用

交通预测目的是利用历史交通速度和地图线路，**预测未来交通速率**。**地图线路**本身是一个**图结构**，而**交通预测又包含了时间和空间的建模**。因此，前沿的图卷积神经网络引入对时空的创新处理，如加入循环神经网络等时序模型来捕获有效存在于车辆交通网络中的复杂时空依赖。

2.5 图卷积神经网络在社会大数据应用的局限性

1. 当**网络规模非常巨大**，上亿级别时，基于**谱方法**的图卷积网络**计算特征向量矩阵的时空复杂度极高**；而基于**空间方法**的图卷积网络在更新节点时**依赖更大量的邻居节点，计算代价过大**。

2. 目前**图卷积神经网络很少关注图级别或信号级别的任务**，主要局限对结点特征分类的有效表达。

3. 目前**图卷积神经网络主要是静态性网络**，而在线社会大数据往往具有动态实时变化的特点。

3 神经网络在非结构化社会大数据方面的应用

如前文所述，情感分析、话题发现、文本分类任务、自动摘要生成、命名实体识别、新词发现、汉语分词都是非结构化的社会大数据方面的重要应用任务。

对于以上的自然语言处理相关问题，将原有的高位、系数、离散的词汇表示方法映射为分布式表示，可以有效克服传统方法中的维数灾难问题，然后再进行神经网络处理。[4] 用神经网络进行处理时，根据不同的具体任务，选择合适的神经网络类型、构建结构，是十分关键的。

下面将重点总结大数据背景下，神经网络在情感分析、话题发现等任务中的应用。

3.1 神经网络在大数据背景下的情感分析任务中的应用

3.1.1 社会大数据背景下的情感分析简介

论坛、社区、博客、购物网站上充斥了两类评论性人类自然语言的信息，包括了客观事实信息和带有人的主观情感色彩的评论性信息。无论对于个人还是机构，面对网络上海量的信息，一条一条地获取所有信息背后的态度、立场、意见是十分耗费时间成本的。我们希望建立一个模型，能够自动地对评论分为肯定或者否定的算法来实现。这些涉及对文本中主官信息进行挖掘的相关研究即为情感分析。

在**社会大数据背景下，情感分析包含的应用**有：舆情分析、推荐系统、文本过滤和电子商务等方面。舆情分析中，可以挖掘人们对某个热点事件的观点倾向，迅速掌握公众舆情；在文本过滤中，可以不仅考虑主题，而且考虑主题的正反面信息，可以准确拦截反面信息；电子商务网站提供的评论功能使得产品的市场反应得到了充分展现，我们可以对这些评论观点进行组织和分类，这将给生产厂家、销售商等带来好处。[4]

对**社会媒体数据进行情感分析的困难**：社会媒体网站所产生的文本具有长度限制、非正式表达等特点。需要通过对具体任务设计具体的神经网络进行处理，以得到理想的信息。

情感分析在社会大数据方面的应用主要有：

（1）用户评论分析与决策：消费者往往没有足够精力、时间去获取所有评论信息来作出对产品的决策。情感分析技术可以很好地解决这个问题。情感分析技术首先自动获取大量相关评论，进而挖掘出产品属性、评价词语，最终通过统计归纳推理给用户提供该产品各个属性的评价意见，方便用户做最终的决策。

（2）舆情监控：随着越来越多网民在互联网上表达观点，网络的隐蔽性、自由性可能会被人利用，以谋取利益或表达敌对政治立场，危害社会稳定。需要通过对文本进行情感分析来监控舆情。

（3）情感预测：情感分析可以帮助用户通过对互联网上的大量新闻、帖子等信息源进行整合分析，预测某一事件的未来状况。例如，在金融市场方面，股票市场与股民过去和当前的主观意志高度相关，根据股民的情感流露可以预测未来数据；美国总统选举为代表的选举预测则是通过情感分析选民情感立场而实现。

其他的领域还有信息抽取、问答系统设计等。

3.1.2 神经网络在情感分析中的应用

情感分析中的神经网络按照神经网络的结构，可以分为单一神经网络和混合（组合、融合）神经网络。

（1）**单一神经网络下的情感分析**

最初的一种模型是**三层前馈神经网络**，由输入层、隐藏层、输出层构成。输入层的每一个神经元代表一个特质，隐藏层层数和隐藏层神经元是由人工设定的，输出层代表分类标签个数。

目前，利用单一神经网络进行的情感分析已经有如下成果：

论文 [7] 在利用 LSTM 处理长序列数据和学习长期依赖性的文本情感分析问题中，进一步提出了不同参数的精简 LSTM 模型实现情感分析，证明了不同参数设置、模型层设置对结果产生影响。

也有学者通过社会媒体、论坛信息进行情感分析，获取公众意见，比如论文 [8] 通过 **LSTM 对 COVID-19 的相关网络品论进行情感分类**，有助于对新冠疫情应对的指导或决策。

传统的 CNN 方法具有忽略文本的潜在主题的问题。**CNN** 在情感分析中最主要的作用就是**挖掘相关特征进行情感分析**。此外，**递归神经网络和 CRF 结合的联合模型可以实现对方面词和意见词的特征提取**。

（2）**混合（组合、融合）神经网络下的情感分析：**

如上文单一神经网络的研究成果所示，我们可以充分考虑 RNN 和 CNN 各自的优点，将二者结合起来。论文 [9] 提出联合 RNN 和 CNN 的多层网络模型 H-RNN-CNN，实现对长文本下比单独 RNN 或 CNN 更佳的分类效果。也有研究者提出基于 CNN-LSTM 的模型，其对含有隐含语义的短文本评论的情感挖掘识别具有很好的效果。进一步地还可以引入注意力机制优化分类结果。

以上神经网络方法的情感分析比传统机器学习、情感词典等方法更能够主动学习特征，保留词语信息及其语义信息，从而更有效地进行情感分类。

3.2 神经网络在社会媒体大数据的话题发现任务中的应用

随着互联网发展，各大社交媒体蓬勃发展，社会媒体信息量呈现爆炸式增长。微博能够实现在一个话题爆发式增加热度时，作出迅速反应，总结热点话题并呈现热搜，即为话题发现算法的一个典型例子。最传统的算法是词对主题模型 BTM，该模型特点是可以通过构建词对来解决短文本的稀疏性问题。由于 BTM 无法直接应用于短文本热点话题发现的问题，进一步地，有人使用突发词对主题模型 BBTM，将词对的突发概率作为模型的先验知识以解决问题。[10]

神经网络在话题发现任务中，如论文 [11] 结合了**注意力机制和双向 LSTM 网络**用于**检测中文事件**，这是利用了**双向 LSTM 能够充分挖掘词语间语义信息**的优势。还有论文如 [12] 提出将**神经网络和主题模型进行融合来实现短文本主题挖掘**，这使得模型挖掘文档集合的主题的效率得以提高。此外，还可以考虑 **CNN 提取特征的优势**，与**双向 LSTM 融合**，能够很好挖掘文本特征、充分理解上下文信息。

4 展望

虽然目前社会大数据背景下的神经网络应用已经取得了较多的成果，但是仍然存在一些挑战：

1. **多样性挑战**：多样性的挑战主要体现在**多模态大数据深度学习**上，数据呈现为视频、音频、文本等多种模态，针对这种挑战可能的解决方法是进行数据融合。又因为不同模态的数据之间既相互区别，又相互联系，所以如何有效地融合不同模态的大数据，以实现高效的多模态大数据学习是解决这种挑战的关键。[2]

2. **大数据的海量性挑战**：神经网络的使用过程都属于内存算法。用这些算法从容量巨大的社会大数据训练集进行学习时，往往会受限于计算机的内存容量。我们可以考虑更好地从**并行计算**的角度或从**数据约简**的角度进行解决。并行计算方法可以考虑根据大数据处理任务分解为多个解耦化的更小型的神经网络训练问题；而数据约简则是从大数据集中选择有代表性的子集，代替大数据集进行学习。

3. **社会大数据的时效性挑战**：大量的社会大数据是动态变化的，同时，我们某些任务对于时效性要求非常高，要求避免神经网络方法得到的结论能够保有时效价值。有**在线学习**、**流式学习**等具体的思路解决这个挑战。

此外，还可以结合强化学习方法，来应对一些新的交互式社会大数据问题。

参考文献

[1] 马世龙, 乌尼日其其格, 李小平. 大数据与深度学习综述 [J]. 智能系统学报, 2016, 11(6): 728-742.

[2] 张素芳，翟俊海，王聪，沈矗，赵春玲. 大数据与大数据机器学习 [J]. 河北大学学报（自然科学版),2018,38(03):299-308+336.

[3] 刘鹏主编；张燕，张重生，张志立副主编. 大数据. 北京：电子工业出版社, 2017.01.

[4] 张华平，商建云，刘兆友. 大数据智能分析. 北京：清华大学出版社, 2019.10.

[5] 范招娣. 基于改进 BiLSTM-CRF 的汽车领域热点发现研究 [D]. 合肥工业大学,2020.DOI:10.27101/d.cnki.ghfgu.2020.000805.

[6] 王婷, 杨文忠. 文本情感分析方法研究综述 [J]. 计算机工程与应用,2021,57(12):11-24.

[7] GOPALAKRISHNAN K，SALEM F M.Sentiment analy sis using simplified long short-term memory recurrent neural networks[J].arXiv：2005.03993，2020.

[8] JELODAR H，WANG Y，ORJI R，et al.Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions：NLP using LSTM recurrent neural network approach[J]. IEEE Journal of Biomedical and Health Informatics，2020，24（10）：2733-2742.

[9] MADASU A，RAO V A.Sequential learning of convolutional features for effective text classification[J]. arXiv：1909.00080，2019.

[10] 向卓元, 吴玉, 陈浩, 张芙玮. 基于 BBTM 改进算法的微博热点话题发现研究 [J/OL]. 情报杂志:1-9[2022-12-29].

[11] 沈兰奔, 武志昊, 纪宇泽, 林友芳, 万怀宇. 结合注意力机制与双向 LSTM 的中文事件检测方法 [J]. 中文信息学报,2019,33 (9):79-87.

[12] Li L S,Gan S J,Yin X D. Feedback recurrent neural network - based embedded vector and its application in topic model [J]. EURASIP Journal on Embedded Systems,2017,2017(1):5.

[13] 鲍鹏，沈华伟，程学旗作. 在线社会关系网络中信息传播建模与预测研究. 北京交通大学出版社有限责任公司, 2021.12.

[14] 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.