# 论文4 大数据与深度学习综述

摘要: 大数据时代改变了基于数理统计的传统数据科学，促进了数据分析方法的创新，从机器学习和多层神经网络演化而来的深度学习是当前大数据处理与分析的研究前沿。从机器学习到深度学习，经历了早期的符号归纳机器学习、统计机器学习、神经网络和20世纪末开始的数据挖掘等几十年的研究和实践，发现深度学习可以挖掘大数据的潜在价值。本文给出大数据和深度学习的综述，特别是，给出了各种深层结构及其学习算法之间关联的图谱，给出了深度学习在若干领域应用的知名案例。最后，展望了大数据上深度学习的发展与挑战

1.为什么要用大数据？

社会生产、生活各个方面时刻都在迅速产生巨量的数据。

2.大数据为什么要用神经网络处理？

（1）要使用全生命周期数据，即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。

3.大数据的核心流程——数据的分析。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。

4.P6-9摆出了很多神经网络，可以作为之后整理神经网络-大数据应用的参考！！

5.使用神经网络分析处理大数据的优势：

（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）优点：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。

6.（2016年）基于大数据方面深度学习应用展望

（1）计算机视觉领域：目标和行为识别；

（2）深度学习与复杂推理结合。

# 论文2 大数据与大数据机器学习

1.大数据的定义

大数据是指具有海量（Ｖｏｌｕｍｅ）、多样（Ｖａｒｉｅｔｙ）、时效（Ｖｅｌｏｃｉｔｙ）、不精确（Ｖｅｒａｃｉｔｙ）和价值（Ｖａｌｕｅ）这５种特征的数据，这５种特征简称大数据的５Ｖｓ特征．

2.大数据的挑战：

多样性的挑战主要体现在多模态大数据机器学习上，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数据融合或集成学习．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．

第１种是直接对不同模态的数据在原始空间进行融合；在第１种方法中，比较有代表性的研究工作包括：基于深度神经网络的多模态大数据融合［５１－５２］和基于语义的多模态大数据融合［５３］，文献［５３］对多模态大数据融合的研究方法进行了全面而深入的综述，具有很高的参考价值。

3.现实生活中的常见大数据处理的案例（可以从里面挑选一些案例来找神经网络的案例来写。）

很多现实大数据处理问题本身就是非平衡大数据学习问题．例如**，医疗大数据处理问题、罪犯视频跟踪大数据处理问题、信用卡欺诈检测大数据处理问题**等。

# 图书3《大数据》刘鹏

1.大数据与深度学习之间有什么样的关系？

随着CPU和GPU计算能力的大幅提升，**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，大数据需要深度学习，深度学习的发展又需要大数据支撑。

# 图书4《大数据智能分析》 张华平。。。

1.非结构化社会大数据语义挖掘

（NLP方向）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。

2.深度学习算法优势：

理论上，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，

3.神经网络在处理非结构化社会大数据中的具体应用：

（1）词语统计——汉语分词：

基于双向循环神经网络与条件随机场的词法分析。

（2）情感分析：

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。评论性信息反映了人们对于某个特定对象的态度、立场和意见。无论对于个人还是机构，面对网络上海量的信息，一条一条地获取是十分耗费时间成本的。在这种情况下，如果能够对这些评论进行分析归纳，判断出哪些评论是持有积极态度，哪些评论持有消极态度，根据这些判断完成下分支的各种推理任务。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；在推荐系统中，可以对检索出来的评论进行统计分析，根据最终的正面或反面结果推荐给用户；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、销售商等机构带来好处。

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。

情感分析是一个分类任务，有两大策略：（1）一次三分策略：一次直接把待分类的文本分成中性、褒义、贬义；（2）二次二分类策略：第一次实现主客观文本分类，第二次实现对主观性文本分成褒义、贬义两类。

**情感分析在社会大数据上的应用：**

①用户评论分析与决策：

人们选择购买某产品前往往会查看产品的相关评论，并通过与其他产品的对比来做最终决策。但是人们往往没有足够精力、时间去留来所有评论信息，所以可能决策会有风险。但情感分析技术可以很好地解决这一问题。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：

互联网具有隐蔽性、开放性、虚拟性、渗透性、自由性等特点，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，散布一些偏激、虚假、煽动性言论，再通过互联网传播被无限放大，影响人们的正确判断、威胁社会和谐稳定。需要通过对文本进行情感分析来监控舆情。

③信息预测：

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。具体来说，在金融市场方面，情感分析可以适用于股票预测，这是因为股票市场是与当前股民的主观意志相关的，根据股民的情感流露可以预测未来数据。此外，政治方面，情感分析可以用于如美国总统竞选中的立场分析和结果预测，分析选举论坛上的情感倾向，可以预测谁更可能成为获胜者。

④其他领域：信息抽取、问答系统等。

**综述神经网络在情感分析中的应用：（论文6）**

**①单一神经网络下的情感分析**

**A.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。**

**B.过去的研究者利用单一神经网络进行的情感分析有如下成果：如论文[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]提出顺序卷积注意递归网络、论文[10]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。**

**综述神经网络在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（论文[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。**

# 图书16 在线社会关系网络中信息传播建模与预测研究

1.随着社会媒体、社交网络等在线社会关系网络逐渐成为互联网服务和应用的主流，出现了信息产生社会化、信息内容碎片化和信息传播网络化等问题。在许多社会媒体平台上的信息传播，在宏观层面展现了信息的流行度或热度，例如一条微博被转发、点赞次数等，一张图片被分享的次数等。为了保证研究结果的稳定性和代表性，我们往往需要大规模数据而非小规模样本来分析信息传播和人类行为。在线社会关系网络汇聚了大量可计算的网络数据，通过研究在线社会关系网络中的信息传播，有利于对国家政策制定、企业活动、个人网络使用的准确判断。

2.社会大数据中的在线社会关系网络信息传播建模面临用户交互的动态性、传播过程复杂性、时序尺度非均匀等挑战。

3.为解决以上挑战，有学者从以下方面提出利用在线的网络社会大数据进行建模。

（1）基于社区发现时空卷积网络的流行度预测。

（2）基于图神经网络的流行度预测。

A.基于图注意力时空网络的流行度预测

B.基于耦合图神经网络的流行度预测

（3）基于图注意力网络的信息级联外源因素建模

（4）基于时空学习的多视角谣言检测框架

4.使用神经网络的原因：由于社会关系网络这种社会大数据信息在实际传播场景中的复杂性、动态性，严重依赖领域经验的特征工程费时费力、预测能力有限。依赖深度神经网络的预测方法成为主流。**深度神经网络具有强大的特征学习和特种表达能力，可以自动高效地从信息传播原始数据中提取得到有价值的特征信息。**弥补了需要手动进行特征提取的不足。这类方法可以根据不同类型的再现内容拟合模型参数，能够较好地适用于不同的信息传播预测场景。

5.神经网络在信息传播预测中主要使用端到端的方式，自动地从信息传播的历史数据中提取特征，然后通过全连接层或多层感知机映射到在线内容的未来流行度。

具体来说，自动学习特征最常用的网络是RNN模型。RNN利用其建模时序信息的预测优势，可以自动学习不同历史事件的影响力表示；也可以用于提取用户信息序列中用户的内容传播上的先后依赖关系。

注意力机制作为自然语言处理的一种找到更感兴趣、使其更好地完成任务的方法，将其引入对RNN对于不同序列特征的提取中，可以使得特征提取、组合更加具有代表性。这有利于进一步预测在线内容的传播。

CNN与RNN的融合也在传播过程建模方面起着重要作用。此外，还可以通过加入强化学习，来指导传播过程的建模。

6.图卷积神经网络因其强大的建模能力引起了广泛关注，极大推动了基于数据挖掘和社会计算等相关领域的研究和发展。

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。**

# 论文14 图卷积神经网络综述

卷积神经网络因其强大的建模能力，在自然语言处理、计算机视觉等领域获得成功应用．然而，传统的卷积神经网络只能处理欧氏空间数据，而社会数据网络，如交通网络、社交网络、引用网络等，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。由于图数据各个结点局部结构不同，这使得卷积神经网络最大的特点——平移不变性不能得到利用。在图数据上使用卷积神经网络因而较为困难。

1.深度学习之前，研究人员常用统计分析方法、图谱理论知识来做社区分析或者人群聚类等。深度神经网络崛起后，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，即先根据节点的邻近性为每个结点学习固定长度的表达，然后再将节点表达作为输入，训练分类模型。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。图卷积神经网络在深度学习模型的图数据应用中受到极大的关注。

2.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。

4.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。**

（1）最初，研究人员关注如何在图上构建卷积算子，逐渐形成了图卷积的谱方法。这种方法具有时空复杂度较高的弊端。可以通过对卷积核进行参数化，以大幅降低时空复杂度。

（2）谱方法中也蕴含了空间角度定义结点的权重矩阵思想。空间方法产生后，考虑在节点域使用注意力机制、序列化模型等建模节点间的权重。随着卷积算子的完善，人们开始考虑多样的图特性，建模图上的高阶信息。

5.图卷积神经网络的主要方法有

（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。

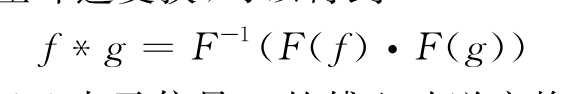
（2）关注图上复杂信息建模的方法。例如，建模边上信息或建模高阶信息的图卷积神经网络。

（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。

6.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。

7.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。



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。

8.图池化操作：

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。

9.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。

10.图卷积神经网络的应用：

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。

（1）网络分析：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。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。

（2）推荐系统：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。图卷积神经网络基于图结构的解决思路效果更好。

（3）交通预测：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。

11.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；而空间方法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。