**学科前沿讲座报告**

**计算机科学与技术专业**

报告名称 大数据背景下的深度学习算法

报告时间 2018/11/02

专 业 计算机科学与技术

班 级 16060104

姓 名 杨旭

学 号 16060104117

教 师 苏小会

课程成绩

**计算机科学与技术专业学科前沿讲座报告**

1. **报告名称:****大数据背景下的深度学习算法**
2. **报告时间:2019/11/2**
3. **报告地点:教1-203**
4. **报告内容**

**摘要**随着大数据时代的到来，基于深度学习技术的机器学习方法被用于有效地分析和处理这些数据。本文详细地介绍了深度学习算法（Deep Learning）的由来，国内外发展的现状，深度学习的模型以及面临的问题。大数据深度学习的重要性和发展方向。  
关键词：大数据、深度学习Deep Learning  
**背景**

2006年，加拿大多伦多大学教授、机器学习领域泰斗——Geoffrey Hinton和他的学生Ruslan Salakhutdinov在顶尖学术刊物《科学》上发表了一篇文章，开启了深度学习在学术界和工业界的浪潮。这篇文章有两个主要的信息：1. 很多隐层的人工神经网络具有优异的特征学习能力，学习得到的特征对数据有更本质的刻画，从而有利于可视化或分类；2. 深度神经网络在训练上的难度，可以通过“逐层初始化”（Layer-wise Pre-training）来有效克服，在这篇文章中，逐层初始化是通过无监督学习实现的。   
 自2006年以来，深度学习在学术界持续升温。斯坦福大学、纽约大学、加拿大蒙特利尔大学等成为研究深度学习的重镇。2010年，美国国防部DARPA计划首次资助深度学习项目，参与方有斯坦福大学、纽约大学和NEC美国研究院。支持深度学习的一个重要依据，就是脑神经系统的确具有丰富的层次结构。一个最著名的例子就是Hubel-Wiesel模型，由于揭示了视觉神经的机理而曾获得诺贝尔医学与生理学奖。除了仿生学的角度，目前深度学习的理论研究还基本处于起步阶段，但在应用领域已显现出巨大能量。2011年以来，微软研究院和Google的语音识别研究人员先后采用DNN技术降低语音识别错误率20％~30％，是语音识别领域十多年来最大的突破性进展。2012年，DNN技术在图像识别领域取得惊人的效果，在ImageNet评测上将错误率从26％降低到15％。在这一年，DNN还被应用于制药公司的DrugeActivity预测问题，并获得世界最好成绩，这一重要成果被《纽约时报》报道。

大数据的存储、计算和分析是大数据处理中的关键技术，大数据的价值产生于对大数据的有效分析。因此，数据分析环节是大数据处理流程的核心和关键。互联网是大数据的主要来源之一。据数据分析公司 Domo 在 2011年和 2014年对每分钟新增网络数据的不完全统计，上传到YouTube 的视频时长从48 h上升为72 h，在Facebook 上分享的信息从684478条上升为2460000条，对Google的搜索请求从200万次上升为400万次等。每时每刻互联网上都有无数的网页和海量的音频、视频、图像等数据产生。现今大数据已触及并影响和改变着人们工作和生活的方方面面。从网络上的电子商务交易到对整个国家经济决策的制定，从医生诊断疾病、做出治疗方案到公路交通治理，从天气预报灾害预测到降低犯罪率、提高安全保障等，到处可见大数据的身影，但是却只有极少量的数据被分析和有效利用。这些数据呈现出非结构化和多源等特点，需要开发新的更加智能的数据挖掘和智能算法技术，从众多复杂的数据中发现规律和模式，提取出新的知识，辅助人类决策或给出可能的预测。采用人工智能和机器学习技术来分析大数据，让机器代替人工获得从数据中提取知识的能力，具有广泛的应用前景，以互联网上的应用为例：计算广告学，垃圾邮件过滤，推荐系统，时间序列等都是大数据的用武之地。大数据时代背景下，如何对纷繁复杂的数据进行有效分析，让其价值得以体现和合理的利用，是当前迫切需要思考和解决的问题。深度学习方法正是开启这扇大门的一把钥匙[1]。  
  深度学习是机器学习研究中的一个新的领域，其动机在于建立、模拟人脑进行分析学习的神经网络，它模仿人脑的机制来解释数据，例如图像，声音和文本。同机器学习方法一样，深度机器学习方法也有监督学习与无监督学习之分.不同的学习框架下建立的学习模型很是不同.例如，卷积神经网络(Convolutional neural networks, 简称CNNs)就是一种深度的监督学习下的机器学习模型，而深度置信网(Deep Belief Nets, 简称DBNs)就是一种无监督学习下的机器学习模型[2]。

**主要应用**语音识别、图像识别、自然语言处理、搜索广告CTR预测   
**深度学习模型**  
  深度学习的实质，是通过构建具有很多隐层的机器学习模型和海量的训练数据，来学习更有用的特征，从而最终提高分类或预测的准确性。因此,“深度模型”是手段,“特征学习”是目的。  
  区别于传统的浅层学习，深度学习的不同在于:突出强调模型结构的深度，通常有5层、6层，甚至10多个隐层:明确特征学习的重要性，通过逐层特征变换，将样本集从原始入空间变换到一个新的特征空间，使分类或预测更加容易。  
  与人工规则构造特征的方法相比，利用大数据来学习特征，更能够刻画数据的丰富内在信息。深度学习是机器学习研究中的一个新的领域，其动机在于建立、模拟人脑进行分析学习的神经网络，它模仿人脑的机制来解释数据，例如图像、声音和文本。深度学习是无监督学习的一种。深度学习的概念源于人工神经网络的研究。含多隐层的多层感知器就是一种深度学习结构。  
  深度学习通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示。  
  相同之处，deep learning 采用了与神经网络类似的分层结构，系统由包括输入层、隐层(多层)、输出层组成的多层网络，只有相邻层节点之间有连接，同一层以及跨层节点之间相互无连接，每一层可以看作是一个logistic regression模型；这种分层结构比较接近人类的大脑的结构   
  为了克服神经网络训练中的问题，DL 采用了与神经网络很不同的训练机制。  
  传统神经网络采用back propagation迭代算法来训练整个网络:随机设定初值,计算当前网络的输出，然后根据当前输出和label之间的差去改变前面各层的参数，直到收敛。  
  整体是一个梯度下降算法。  
  Deep Learning整体上是一个layer-wise (逐层)训练机制。  
  这样做的原因是，如果采用back  propagation机制，对于一个deep network (7层以上),残差传播到最前面层已经变得太小，出现所谓的gradient diffusion (梯度扩散)。  
  Deep learning 训练过程:  
  采用自下上升非监督学习(就是从底层开始，一层一层的往顶层训练) :采用无标定数据(有标定数据也可)分层训练各层参数。这一步可看作是一个无监督训练过程，是和传统神经网络区别最大的部分(这一过程可看作feature learning 过程) :  
  先用无标定数据训练第- -层，训练时先学习第- -层参数(这- -层可以看作是得到一个使得输出和输入差别最小的三层神经网络的隐层)。  
  由于模型capacity的限制以及稀疏性约束，使得得到的模型能够学习到数据本身的结构,从而得到比输入更具有表示能力的特征:在学习得到第n-1层后，将n-1层的输出作为第n层的输入，训练第n层，由此分别得到各层的参数:  
  自顶向下的监督学习(就是通过带标签的数据去训练，误差自顶向下传输，对网络进行微调)。  
**深度学习面临的问题**  
  随着数据规模的不断增大，需要采用更为复杂的深度模型去捕获其中更丰富的信息和包含的模式，因而对计算能力提出了更高的要求。已有的学习算法如随机梯度下降，其自身是序列化的，如何通过异步更新模式更有效地改进为并行计算模式，并充分利用GPU 的并行处理能力等问题是提高模型训练效率的关键。此外，在训练深度结构模型时，需要选择大量的超参数，如模型的深度、每层包含的隐单元数、学习率等，往往某个超参数不适合就会造成模型性能的大幅下降，但对于如何选择这些超参数，仍没有一个好的指导方法，这进-步增加了训练深度模型的难度。另-方面，当前取得的突出进展都是基于经验性的，还没有很好的理论支撑，需要进一步的研究和探索，完善理论，使之更好地指导实践[1]。  
  探索新的特征提取模型是值得深入研究的内容。此外有效的可并行训练算法也是值得研究的-一个方向。当前基于最小批处理的随机梯度优化算法很难在多计算机中进行并行训练。通常办法是利用图形处理单元加速学习过程。然而单个机器GPU对大规模数据识别或相似任务数据集并不适用。在深度学习应用拓展方面，如何合理充分利用深度学习在增强传统学习算法的性能仍是目前各领域的研究重点[6]。

**参考文献**  
[1]余滨,李绍滋,徐素霞,纪荣嵘.深度学习:开启大数据时代的钥匙[J].工程研究-跨学科视野中的工程, 2014, 03:233-243.  
[2]深度学习. C:\Users\aa\AppData\Local\Temp\%W@GJ$ACOF(TYDYECOKVDYB.pnghttp://baike. baidu. com/view/9964678. htm.  
[3]李艺颖.人工智能的新浪潮一深度学习[J]. 网友世界, 2013, Z2:32.

[4]余凯,贾磊,陈雨强,徐伟.深度学习的昨天、今天和明天[J].计算机研究与发展, 2013, 09: 1799-1804.  
[5]百度深度学习实验室. C:\Users\aa\AppData\Local\Temp\%W@GJ$ACOF(TYDYECOKVDYB.pnghttp://idl. baidu. com/IDL-about. html.

[6] Zouxy. Deep. Learning (深度学习)学习笔记整理(网上资料汇总).C:\Users\aa\AppData\Local\Temp\%W@GJ$ACOF(TYDYECOKVDYB.pnghttp://blog. csdn. net/zouxy09

1. **报告心得及收获**

  在讲座老师讲解数据挖掘和机器学习的相关报告后，对于“机器学习”、“人工智能”这方面产生了很浓厚的兴趣，在听过讲座后我去图书馆查看相应的书籍和在网上搜索资料对于“机器学习”、“深度学习”这方面有了初步的认识，结合自己所学的相关内容书写了这篇读书报告。在完成本次报告内容后，我想自己仍会进行进-一步的学习，特别是有关于大数据处理、云计算与深度学习的结合，我会查阅相关资料在寒假期间搭建完成Hadoop分布式计算平台,应用深度学习模型，通过Mat lab与Java混合编程，来做一些demo, 为自己将来的研究方向和论文的书写奠定一定的基础。