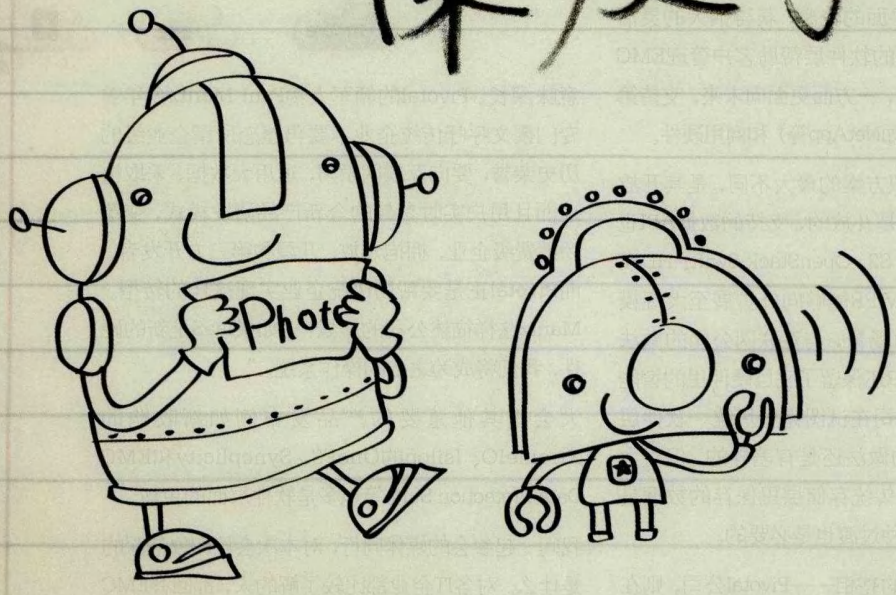
Deep Learning

深度学习



深度学习：推进人工智能的梦想

文 / 余凯，贾磊，陈雨强

2012年6月，《纽约时报》披露了Google Brain项目，吸引了公众的广泛关注。这个项目是由著名的斯坦福大学机器学习教授Andrew Ng和在大规模计算机系统方面的世界顶尖专家Jeff Dean共同主导，用16000个CPU Core的并行计算平台训练一种称为“深层神经网络”（DNN，Deep Neural Networks）的机器学习模型，在语音识别和图像识别等领域获得了巨大的成功。2012年11月，微软在中国天津的一次活动上公开演示了一个全自动的同声传译系统，讲演者用英文演讲，后台的计算机一气呵成自动完成语音识别、英中机器翻译，以及中文语音合成，效果非常流畅。据报道，后面支撑的关键技术也是DNN，或者深度学习（DL，Deep Learning）。2013年1月，在百度的年会上，创始人兼CEO李彦宏高调宣布要成立百度研究院，其中第一个重点方向就是深度学习，并为此而成立Institute of Deep Learning（IDL）。这是百度成立十多年以来第一次成立研究院。2014年4月，《麻省理工学院技术评论》杂志将深度学习列为2013年十大突破性技术（Breakthrough Technology）之首。

读者一定非常好奇，什么是深度学习？为什么深度学习受到学术界和工业界如此广泛的重视？深度学习技术研发面临什么样的科学和工程问题？深度学习带来的科技进步将怎样改变人们的生活？

机器学习的两次浪潮：从浅层学习到深度学习

在解释深度学习之前，我们需要了解什么是机器学习。机器学习是人工智能的一个分支，而在很多时候，几乎成为人工智能的代名词。简单来说，机器学习就是通过算法，使得机器能从大量历史数据中学习规律，从而对新的样本做智能识别或对未来做预测。从1980年代末期以来，机器学习的发展大致经历了两次浪潮：浅层学习（Shallow Learning）和深度学习（Deep Learning）。需要指出是，机器学习历史阶段的划分是一个仁者见仁，智者见智的事情，从不同的维度来看会得到不同的结论。这里我们是从机器学习模型的层次结构来看的。

第一次浪潮：浅层学习

1980年代末期，用于人工神经网络的反向传播算法（也叫Back Propagation算法或者BP算法）的发明，给机器学习带来了希望，掀起了基于统计模型的机器学习热潮。这个热潮一直持续到今天。人们发现，利用BP算法可以让一个人工神经网络模型从大量训练样本中学习出统计规律，从而对未知事件做预测。这种基于统计的机器学习方法比起过去基于人工规则的系统，在很多方面显示出优越性。这个时候的人工神经网络，虽然也被称作多层感知机（Multi-layer Perceptron），但实际上是一种只含有一层隐层节点的浅层模型。

90年代，各种各样的浅层机器学习模型相继被提出，比如支撑向量机（SVM，Support Vector Machines）、Boosting、最大熵方法（例如LR，Logistic Regression）等。这些模型的结构基本上可以看成带有一层隐层节点（如SVM、Boosting），或没有隐层节点（如LR）。这些模型在无论是理论分析还是应用都获得了巨大的成功。相比较之下，由于理论分析的难度，加上训练方法需要很多经验和技巧，所以这个时期浅层人工神经网络反而相对较为沉寂。

2000年以来互联网的高速发展，对大数据的智能化分析和预测提出了巨大需求，浅层学习模型在互联网应用上获得了巨大成功。最成功的应用包

括搜索广告系统（比如Google的AdWords、百度的凤巢系统）的广告点击率CTR预估、网页搜索排序（例如Yahoo!和微软的搜索引擎）、垃圾邮件过滤系统、基于内容的推荐系统等。

第二次浪潮：深度学习

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Druge Activity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。

正如文章开头所描述的，今天Google、微软、百度等知名的拥有大数据的高科技公司争相投入资源，占领深度学习的技术制高点，正是因为它们都看到了在大数据时代，更加复杂且更加强大的深度模型能深刻揭示海量数据里所承载的复杂而丰富的信息，并对未来或未知事件做更精准的预测。

大数据与深度学习

在工业界一直有个很流行的观点：在大数据条件下，简单的机器学习模型会比复杂模型更加有效。例如，在很多的大数据应用中，最简单的线性模型得到大量使用。而最近深度学习的惊人进展，促使我们也许到了要重新思考这个观点的时候。简而言之，在大数据情况下，也许只有比较复杂的模型，或者说表达能力强的模型，才能充分发掘海量数据中蕴藏的丰富信息。运用更强大的深度模型，也许我们能从大数据中发掘出更多有价值的信息和知识。

为了理解为什么大数据需要深度模型，先举一个例子。语音识别已经是一个大数据的机器学习问题，在其声学建模部分，通常面临的是十亿到千亿级别的训练样本。在Google的一个语音识别实验中，发现训练后的DNN对训练样本和测试样本的预测误差基本相当。这是非常违反常识的，因为通常模型在训练样本上的预测误差会显著小于测试样本。因此，只有一个解释，就是由于大数据里含有丰富的信息维度，即便是DNN这样的高容

量复杂模型也是处于欠拟合的状态，更不必说传统的GMM声学模型了。所以从这个例子中我们看出，大数据需要深度学习。

浅层模型有一个重要特点，就是假设靠人工经验来抽取样本的特征，而强调模型主要是负责分类或预测。在模型的运用不出差错的前提下（如假设互联网公司聘请的是机器学习的专家），特征的好坏就成为整个系统性能的瓶颈。因此，通常一个开发团队中更多的人力是投入到发掘更好的特征上去的。要发现一个好的特征，就要求开发人员对待解决的问题要有很深入的理解。而达到这个程度，往往需要反复地摸索，甚至是数年磨一剑。因此，人工设计样本特征，不是一个可扩展的途径。

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，从而最终提升分类或预测的准确性。所以“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：1. 强调了模型结构的深度，通常有5层、6层，甚至10多层的隐层节点；2. 明确突出了特征学习的重要性，也就是说，同过逐层特征变换，将样本在原空间的特征表示变换到一个新特征空间，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，更能刻画数据丰富的内在信息。所以，在未来的几年里，我们将看到越来越多的例子：深度模型应用于大数据，而不是浅层的线性模型。

深度学习的应用

语音识别

语音识别系统长期以来，在描述每个建模单元的统计概率模型时，大多采用的是混合高斯模型（GMM）。这种模型由于估计简单，适合海量数据训练，同时有成熟的区分度训练技术支持，长期以来，一直在语音识别应用中占有垄断性地位。但这种混合高斯模型本质上是一种浅层网络建模，不能充分描述特征的状态空间分布。另外，GMM建模的特征维数一般是几十维，不能充分描述特征之间的相关性。最后，GMM建模本质上是一种似然概率建模，虽然区分度训练能够模拟一些模式类之间的区分性，但能力有限。

微软研究院语音识别专家邓立和俞栋从2009年开始和深度学习专家Geoffery Hinton合作。2011年微软宣布基于深度神经网络的识别系统取得成果并推出产品，彻底改变了语音识别原有的技术框架。采用深度神经网络后，可以充分描述特征之间的相关性，可以把连续多帧的语音特征并在一起，构成一个高维特征。最终的深度神经网络可以采用高维特征训练来模拟。由于深度神经网络采用模拟人脑的多层结果，可以逐级地进行信息特征抽取，最终形成适合模式分类的较理想特征。这种多层结构和人脑处理语音图像信息时，是有很大的相似性的。深度神经网络的建模技术，在实际线上服务时，能够无缝地和传统的语音识别技术相结合，在不引起任何系统额外耗费情况下，大幅度提升了语音识别系统的识别率。其在线的使用方法具体如下：在实际解码过程中，声学模型仍然是采用传统的HMM模型，语音模型仍然是采用传统的统计语言模型，解码器仍然是采用传统的动态WFST解码器。但在声学模型的输出分布计算时，完全用神经网络的输出后验概率乘以一个先验概率来代替传统HMM模型中的GMM的输出似然概率。百度在实践中发现，采用DNN进行声音建模的语音识别系统相比于传统的GMM语音识别系统而言，相对误识别率能降低25%。最终在2012年11月，百度上线了第一款基于DNN的语音搜索系统，成为最早采用DNN技术进行商业语音服务的公司之一。

国际上，Google也采用了深层神经网络进行声音建模，是最早突破深层神经网络工业化应用的企业之一。但Google产品中采用的深度神经网络只有4-5层，而百度采用的深度神经网络多达9层。这种结构差异的核心其实是百度更好地解决了深度神经网络在线计算的技术难题，因此百度线上产品可以采用更复杂的网络模型。这将对于未来拓展海量语料的DNN模型训练有更大的优势。

图像识别

图像是深度学习最早尝试的应用领域。早在1989年，Yann LeCun（现纽约大学教授）和他的同事们就发表了卷积神经网络（Convolution Neural Networks，简称CNN）的工作。CNN是一种带

有卷积结构的深度神经网络，通常至少有两个非线性可训练的卷积层，两个非线性的固定卷积层（又叫Pooling Laye）和一个全连接层，一共至少5个隐含层。CNN的结构受到著名的Hubel-Wiesel生物视觉模型的启发，尤其是模拟视觉皮层V1和V2层中Simple Cell和Complex Cell的行为。在很长时间里，CNN虽然在小规模的问题上，如手写数字，取得过当时世界最好结果，但一直没有取得巨大成功。这主要原因是，CNN在大规模图像上效果不好，比如像素很多的自然图片内容理解，所以没有得到计算机视觉领域的足够重视。这个情况一直持续到2012年10月，Geoffrey Hinton和他的两个学生在著名的ImageNet问题上用更深的CNN取得世界最好结果，使得图像识别大踏步前进。在Hinton的模型里，输入就是图像的像素，没有用到任何的人工特征。

这个惊人的结果为什么在之前没有发生？原因当然包括算法的提升，比如dropout等防止过拟合技术，但最重要的是，GPU带来的计算能力提升和更多的训练数据。百度在2012年底将深度学习技术成功应用于自然图像OCR识别和人脸识别等问题，并推出相应的桌面和移动搜索产品，2013年，深度学习模型被成功应用于一般图片的识别和理解。从百度的经验来看，深度学习应用于图像识别不但大大提升了准确性，而且避免了人工特征抽取的时间消耗，从而大大提高了在线计算效率。可以很有把握地说，从现在开始，深度学习将取代“人工特征+机器学习”的方法而逐渐成为主流图像识别方法。

自然语言处理

除了语音和图像，深度学习的另一个应用领域问题是自然语言处理（NLP）。经过几十年的发展，基于统计的模型已经成为NLP的主流，但作为统计方法之一的人工神经网络在NLP领域几乎没有受到重视。最早应用神经网络的NLP问题是语言模型。加拿大蒙特利尔大学教授Yoshua Bengio等人于2003年提出用embedding的方法将词映射到一个矢量表示空间，然后用非线性神经网络来表示N-Gram模型。世界上最早的最早的深度学习用于NLP的研究工作诞生于NEC美国研究院，其研究员Ronan Collobert和Jason Weston从2008年开始采用embedding和多层一维卷积的结构，用于POS Tagging、Chunking、Named Entity Recognition、Semantic Role Labeling等四个典型NLP问题。值得注意的是，他们将同一个模型用于不同任务，都能取得与业界最前沿相当的准确率。最近以来，斯坦福大学教授Chris Manning等人在将深度学习用于NLP的工作也值得关注。

总的来说，深度学习在NLP上取得的进展没有在语音图像上那么令人影响深刻。一个很有意思的悖论是：相比于声音和图像，语言是唯一的非自然信号，是完全由人类大脑产生和处理的符号系统，但模仿人脑结构的人工神经网络却似乎在处理自然语言上没有显现明显优势？我们相信，深度学习在NLP方面有很大的探索空间。从2006年图像深度学习成为学术界热门课题到2012年10月Geoffery Hinton在ImageNet上的重大突破，经历了6年时间。我们需要有足够的耐心。

搜索广告CTR预估

搜索广告是搜索引擎的主要变现方式，而按点击付费（Cost Per Click，CPC）又是其中被最广泛应用的计费模式。在CPC模式下，预估的CTR（pCTR）越准确，点击率就会越高，收益就越大。通常，搜索广告的pCTR是通过机器学习模型预估得到。提高pCTR的准确性，是提升搜索公司、广告主、搜索用户三方利益的最佳途径。

传统上，Google、百度等搜索引擎公司以Logistic Regression（LR）作为预估模型。而从2012年开始，百度开始意识到模型的结构对广告CTR预估的重要性：使用扁平结构的LR严重限制了模型学习与抽象特征的能力。为了突破这样的限制，百度尝试将DNN作用于搜索广告，而这其中最大的挑战在于当前的计算能力还无法接受10^{11}级别的原始广告特征作为输入。作为解决，在百度的DNN系统里，特征数从10^{11}数量级被降到了10^{3}，从而能被DNN正常地学习。这套深度学习系统已于2013年5月开始服务于百度搜索广告系统，每天为数亿网民使用。

DNN在搜索广告系统中的应用还远远没有成熟，其

中DNN与迁移学习的结合将可能是一个令人振奋的方向。使用DNN，未来的搜索广告将可能借助网页搜索的结果优化特征的学习与提取；亦可能通过DNN将不同的产品线联系起来，使得不同的变现产品不管数据多少，都能互相优化。我们认为未来的DNN一定会在搜索广告中起到更重要的作用。

深度学习研发面临的重大问题

理论问题

理论问题主要体现在两个方面，一个是统计学习方面的，另一个是计算方面的。我们已经知道，深度模型相比较于浅层模型有更好的对非线性函数的表示能力。具体来说，对于任意一个非线性函数，根据神经网络的Universal Approximation Theory，我们一定能找到一个浅层网络和一个深度网络来足够好地表示。但深度网络只需要少得多的参数。但可表示性不代表可学习性。我们需要了解深度学习的样本复杂度，也就是我们需要多少训练样本才能学习到足够好的深度模型。从另一方面来说，我们需要多少计算资源才能通过训练得到更好的模型？理想的计算优化方法是什么？由于深度模型都是非凸函数，这方面的理论研究极其困难。

建模问题

在推进深度学习的学习理论和计算理论的同时，我们是否可以提出新的分层模型，使其不但具有传统深度模型所具有的强大表示能力，还具有其他的好处，比如更容易做理论分析。另外，针对具体应用问题，我们如何设计一个最适合的深度模型来解决问题？我们已经看到，无论在图像深度模型，还是语言深度模型，似乎都存在深度和卷积等共同的信息处理结构。甚至对于语音声学模型，研究人员也在探索卷积深度网络。那么一个更有意思的问题是，是否存在可能建立一个通用的深度模型或深度模型的建模语言，作为统一的框架来处理语音、图像和语言？

工程问题

需要指出的是，对于互联网公司而言，如何在工程上利用大规模的并行计算平台来实现海量数据训练，是各家公司从事深度学习技术研发首先要解决的问题。传统的大数据平台如Hadoop，由于数据处理的Latency太高，显然不适合需要频繁迭代的深度学习。现有成熟的DNN训练技术大都是采用随机梯度法（SGD）方法训练的。这种方法本身不可能在多个计算机之间并行。即使是采用GPU进行传统的DNN模型进行训练，其训练时间也是非常漫长的，一般训练几千小时的声学模型所需要几个月的时间。而随着互联网服务的深入，海量数据训练越来越重要，DNN这种缓慢的训练速度必然不能满足互联网服务应用的需要。Google搭建的DistBelief，是一个采用普通服务器的深度学习并行计算平台，采用异步算法，由很多计算单元独立地更新同一个参数服务器的模型参数，实现了随机梯度下降算法的并行化，加快了模型训练速度。与Google采用普通服务器不同，百度的多GPU并行计算平台，克服了传统SGD训练的不能并行的技术难题，神经网络的训练已经可以在海量语料上并行展开。可以预期，未来随着海量数据训练的DNN技术的发展，语音图像系统的识别率还会持续提升。

总结

深度学习带来了机器学习的一个新浪潮，受到从学术界到工业界的广泛重视，也导致了“大数据+深度模型”时代的来临。在应用方面，深度学习使得语音图像的智能识别和理解取得惊人进展，从而推动人工智能和人机交互大踏步前进。同时，pCTR这样的复杂机器学习任务也得到显著提升。如果我们能在理论、建模和工程方面，突破深度学习技术面临的一系列难题，人工智能的梦想将不再遥远。

本文作者：
余凯，百度技术副总监，千人计划国家特聘专家。
贾磊，百度主任架构师，语音技术负责人。
陈雨强，百度商务搜索部资深研发工程师，负责搜索广告CTR预估。