

深度学习-LeCun、Bengio和Hinton的联合综述

深度学习 卷积神经网络 递归神经网络 无监督学习 自然语言处理 计算机视觉 语音识别 Geoffrey Hinton
Yoshua Bengio Yann LeCun BP算法

原文摘要：深度学习可以让那些拥有多个处理层的计算模型来学习具有多层次抽象的数据的表示。这些方法在许多方面都带来了显著的改善，包括最先进的语音识别、视觉对象识别、对象检测和许多其它领域，例如药物发现和基因组学等。深度学习能够发现大数据中的复杂结构。它是利用BP算法来完成这个发现过程的。BP算法能够指导机器如何从前一层获取误差而改变本层的内部参数，这些内部参数可以用于计算表示。深度卷积网络在处理图像、视频、语音和音频方面带来了突破，而递归网络在处理序列数据，比如文本和语音方面表现出了闪亮的一面。

机器学习技术在现代社会的各个方面表现出了强大的功能：从Web搜索到社会网络内容过滤，再到电子商务网站上的商品推荐都有涉足。并且它越来越多地出现在消费品中，比如相机和智能手机。

机器学习系统被用来识别图片中的目标，将语音转换成文本，匹配新闻元素，根据用户兴趣提供职位或产品，选择相关的搜索结果。逐渐地，这些应用使用一种叫深度学习的技术。传统的机器学习技术在处理未加工过的数据时，体现出来的能力是有限的。几十年来，想要构建一个模式识别系统或者机器学习系统，需要一个精致的引擎和相当专业的知识来设计一个特征提取器，把原始数据（如图像的像素值）转换成一个适当的内部特征表示或特征向量，子学习系统，通常是一个分类器，对输入的样本进行检测或分类。特征表示学习是一套给机器灌入原始数据，然后能自动发现需要进行检测和分类的表达的方法。深度学习就是一种特征学习方法，把原始数据通过一些简单的但是非线性的模型转变成为更高层次的，更加抽象的表达。通过足够多的转换的组合，非常复杂的函数也可以被学习。对于分类任务，高层次的表达能够强化输入数据的区分能力方面，同时削弱不相关因素。比如，一副图像的原始格式是一个像素数组，那么在第一层上的学习特征表达通常指的是在图像的特定位置和方向上有没有边的存在。第二层通常会根据那些边的某些排放来检测图案，这时候会忽略掉一些边上的一些小的干扰。第三层或许会把那些图案进行组合，从而使其对应于熟悉目标的某部分。随后的一些层会将这些部分再组合，从而构成待检测目标。深度学习的核心方面是，上述各层的特征都不是利用人工工程来设计的，而是使用一种通用的学习过程从数据中学到的。

深度学习正在取得重大进展，解决了人工智能界的尽最大努力很多年仍没有进展的问题。它已经被证明，它能够擅长发现高维数据中的复杂结构，因此它能够被应用于科学、商业和政府等领域。除了在图像识别、语音识别等领域打破了纪录，它还在另外的领域击败了其他机器学习技术，包括预测潜在的药物分子的活性、分析粒子加速器数据、重建大脑回路、预测在非编码DNA突变对基因表达和疾病的影响。也许更令人惊讶的是，深度学习在自然语言理解的各项任务中产生了非常可喜的成果，特别是主题分类、情感分析、自动问答和语言翻译。我们认为，在不久的将来，深度学习将会取得更多的成功，因为它需要很少的手工工程，它可以很容易受益于可用计算能力和数据量的增加。目前正在为深度神经网络开发新的学习算法和架构只会加速这一进程。

监督学习

机器学习中，不论是否是深层，最常见的形式是监督学习。试想一下，我们要建立一个系统，它能够对一个包含了一座房子、一辆汽车、一个人或一个宠物的图像进行分类。我们先收集大量的房子，汽车，人与宠物的图像的数据集，并对每个对象标上它的类别。在训练期间，机器会获取一副图片，然后产生一个输出，这个输出以向量形式的分数来表示，每个类别都有一个这样的向量。我们希望所需的类别在所有的类别中具有最高的得分，但是这在训练之前是不太可能发生的。通过计算一个目标函数可以获得输出分数和期望模式分数之间的误差（或距离）。然后机器会修改其内部可调参数，以减少这种误差。这些可调节的参数，通常被称为权值，它们是一些实数，可以被看作是一些“旋钮”，定义了机器的输入输出功能。在典型的深学习系统中，有可能有数以百万计的样本和权值，和带有标签的样本，用来训练机器。为了正确地调整权值向量，该学习算法计算每个权值的梯度向量，表示了如果权值增加了一个很小的量，那么误差会增加或减少的量。权值向量然后在梯度矢量的相反方向上进行调整。我们的目标函数，所有训练样本的平均，可以被看作是一种在权值的高维空间上的多变地形。负的梯度矢量表示在该地形中下降方向最快，使其更接近于最小值，也就是平均输出误差最低的地方。

在实际应用中，大部分从业者都使用一种称作随机梯度下降的算法（SGD）。它包含了提供一些输入向量样本，计算输出和误差，计算这些样本的平均梯度，然后相应的调整权值。通过提供小的样本集合来重复这个过程用以训练网络，直到目标函数停止增长。它被称为随机的是因为小的样本集对于全体样本的平均梯度来说会有噪声估计。这个简单过程通常会找到一组不错的权值，同其他精心设计的优化技术相比，它的速度让人惊奇。训练结束之后，系统会通过不同的数据样本——测试集来显示系统的性能。这用于测试机器的泛化能力——对于未训练过的新样本的识别能力。

当前应用中的许多机器学习技术使用的是线性分类器来对人工提取的特征进行分类。一个2类线性分类器会计算特征向量的加权和。当加权和超过一个阈值之后，输入样本就会被分配到一个特定的类别中。从20世纪60年代开始，我们就知道了线性分类器只能把样本分成非常简单的区域，也就是说通过一个超平面把空间分成两部分。

但像图像和语音识别等问题，它们需要的输入-输出函数要对输入样本中不相关因素的变化不要过于敏感，如位置的变化，目标的方向或光照，或者语音中音调或语调的变化等，但是需要对于一些特定的微小变化非常敏感（例如，一只白色的狼和跟狼类似的白色狗——萨摩耶德犬之间的差异）。在像素这一级别上，两条萨摩耶德犬在不同的姿势和在不同的环境下的图像可以说差异是非常大的，然而，一只萨摩耶德犬和一只狼在相同的位置并在相似背景下的两个图像可能就非常类似。

一个线性分类器或者其他操作在原始像素上的浅层分类器不能够区分后两者，虽然能够将前者归为同一类。这就是为什么浅分类要求有良好的特征提取器用于解决选择性不变性困境——提取器会挑选出图像中能够区分目标的那些重要因素，但是这些因素对于分辨动物的位置就无能为力了。为了加强分类能力，可以使用泛化的非线性特性，如核方法，但这些泛化特征，比如通过高斯核得到的，并不能够使得学习器从学习样本中产生较好的泛化效果。传统的方法是手工设计良好的特征提取器，这需要大量的工程技术和专业领域知识。但是如果通过使用通用学习过程而得到良好的特征，那么这些都是可以避免的了。这就是深度学习的关键优势。

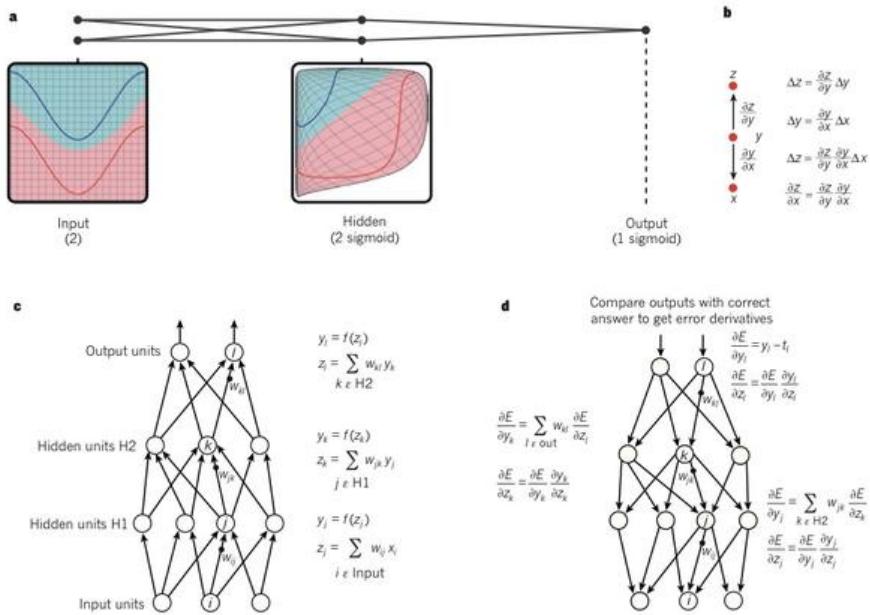


图1 多层神经网络和BP算法

1. 多层神经网络（用连接点表示）可以对输入空间进行整合，使得数据（红色和蓝色线表示的样本）线性可分。注意输入空间中的规则网格（左侧）是如何被隐藏层转换的（转换后的在右侧）。这个例子中只用了两个输入节点，两个隐藏节点和一个输出节点，但是用于目标识别或自然语言处理的网络通常包含数十个或者数百个这样的节点。获得C.Olah (<http://colah.github.io/>)的许可后重新构建的这个图。
2. 链式法则告诉我们两个小的变化（ x 和 y 的微小变化，以及 y 和 z 的微小变化）是怎样组织到一起的。 x 的微小变化量 Δx 首先会通过乘以 $\partial y / \partial x$ （偏导数）转变成 y 的变化量 Δy 。类似的， Δy 会给 z 带来改变 Δz 。通过链式法则可以将一个方程转化到另外的一个——也就是 Δx 通过乘以 $\partial y / \partial x$ 和 $\partial z / \partial y$ （英文原文为 $\partial z / \partial x$ ，系笔误——编辑注）得到 Δz 的过程。当 x ， y ， z 是向量的时候，可以同样处理（使用雅克比矩阵）。
3. 具有两个隐层一个输出层的神经网络中计算前向传播的公式。每个都有一个模块构成，用于反向传播梯度。在每一层上，我们首先计算每个节点的总输入 z ， z 是前一层输出的加权和。然后利用一个非线性函数 $f(z)$ 来计算节点的输出。简单期间，我们忽略掉了阈值项。神经网络中常用的非线性函数包括了最近几年常用的校正线性单元（ReLU） $f(z) = \max(0, z)$ ，和更多传统sigmoid函数，比如双曲线正切函数 $f(z) = (\exp(z) - \exp(-z)) / (\exp(z) + \exp(-z))$ 和logistic函数 $f(z) = 1 / (1 + \exp(-z))$ 。
4. 计算反向传播的公式。在隐层，我们计算每个输出单元产生的误差，这是由上一层产生的误差的加权和。然后我们将输出层的误差通过乘以梯度 $f'(z)$ 转换到输入层。在输出层上，每个节点的误差会用成本函数的微分来计算。如果节点l的成本函数是 $0.5 * (y_l - t_l)^2$ ，那么节点的误差就是 $y_l - t_l$ ，其中 t_l 是期望值。一旦知道了 $\partial E / \partial z_k$ 的值，节点j的内星权向量 w_{jk} 就可以通过 $y_j \partial E / \partial z_k$ 来进行调整。

深度学习的体系结构是简单模块的多层栈，所有（或大部分）模块的目标是学习，还有许多计算非线性输入输出的映射。栈中的每个模块将其输入进行转换，以增加表达的可选择性和不变性。比如说，具有一个5到20层的非线性多层系统能够实现非常复杂的功能，比如输入数据对细节非常敏感——能够区分白狼和萨莫耶德犬，同时又具有强大的抗干扰能力，比如可以忽略掉不同的背景、姿势、光照和周围的物体等。

反向传播来训练多层神经网络

在最早期的模式识别任务中，研究者的目标一直是使用可以训练的多层网络来替代经过人工选择的特征，虽然使用多层神经网络很简单，但是得出来的解很糟糕。直到20世纪80年代，使用简单的随机梯度下降来训练多层神经网络，这种糟糕的情况才有所改变。只要网络的输入和内部权值之间的函数相对平滑，使用梯度下降就奏效，梯度下降方法是在70年代到80年代期间由不同的研究团队独立发明的。

用来求解目标函数关于多层神经网络权值梯度的反向传播算法（BP）只是一个用来求导的链式法则的具体应用而已。反向传播算法的核心思想是：目标函数对于某层输入的导数（或者梯度）可以通过向后传播对该层输出（或者下一层输入）的导数求得（如图1）。反向传播算法可以被重复的用于传播梯度通过多层神经网络的每一层：从该多层神经网络的最顶层的输出（也就是改网络产生预测的那一层）一直到该多层神经网络的最底层（也就是被接受外部输入的那一层），一旦这些关于（目标函数对）每层输入的导数求解完，我们就可以求解每一层上面的（目标函数对）权值的梯度了。

很多深度学习的应用都是使用前馈式神经网络（如图1），该神经网络学习一个从固定大小输入（比如输入是一张图）到固定大小输出（例如，到不同类别的概率）的映射。从第一层到下一层，计算前一层神经元输入数据的权值的和，然后把这个和传给一个非线性激活函数。当前最流行的非线性激活函数是rectified linear unit(ReLU)，函数形式： $f(z) = \max(z, 0)$ 。过去的几十年中，神经网络使用一些更加平滑的非线性函数，比如 $\tanh(z)$ 和 $1 / (1 + \exp(-z))$ ，但是ReLU通常会让一个多层神经网络学习的更快，也可以让一个深度网络直接有监督的训练（不需要无监督的pre-train）。

达到之前那种有pre-train的效果。通常情况下，输入层和输出层以外的神经单元被称为隐藏单元。隐藏层的作用可以看成是使用一个非线性的方式打乱输入数据，来让输入数据对应的类别在最后一层变得线性可分。

在20世纪90年代晚期，神经网络和反向传播算法被大多数机器学习团队抛弃，同时也不受计算机视觉和语音识别团队的重视。人们普遍认为，学习有用的、

多级层次结构的、使用较少先验知识进行特征提取的这些方法都不靠谱。确切的说是因为简单的梯度下降会让整个优化陷入到不好的局部最小解。

实践中，如果在大的网络中，不管使用什么样的初始化条件，局部最小解并不算什么大问题，系统总是得到效果差不多的解。最近的理论和实验表明，局部最小解还真不是啥大问题。相反，解空间中充满了大量的鞍点（梯度为0的点），同时鞍点周围大部分曲面都是往上的。所以这些算法就算是陷入了这些局部最小值，关系也不太大。

2006年前后，CIFAR（加拿大高级研究院）把一些研究者聚集在一起，人们对深度前馈式神经网络重新燃起了兴趣。研究者们提出了一种非监督的学习方法，这种方法可以创建一些网络层来检测特征而不使用带标签的数据，这些网络层可以用来重构或者对特征检测器的活动进行建模。通过预训练过程，深度网络的权值可以被初始化为有意思的价值。然后一个输出层被添加到该网络的顶部，并且使用标准的反向传播算法进行微调。这个工作对手写体数字的识别以及行人预测任务产生了显著的效果，尤其是带标签的数据非常少的时候。

使用这种与训练方法做出来的第一个比较大的应用是关于语音识别的，并且是在GPU上做的，这样做是因为写代码很方便，并且在训练的时候可以得到10倍或者20倍的加速。2009年，这种方法被用来映射短时间的系数窗口，该系统窗口是提取自声波并被转换成一组概率数字。它在一组使用很少词汇的标准的语音识别基准测试程序上达到了惊人的效果，然后又迅速被发展到另外一个更大的数据集上，同时也取得惊人的效果。从2009年到2012年底，较大的语音团队开发了这种深度网络的多个版本并且已经被用到了安卓手机上。对于小的数据集来说，无监督的预训练可以防止过拟合，同时可以带来更好的泛化性能当有标签的样本很小的时候。一旦深度学习技术重新恢复，这种预训练只有在数据集合较少的时候才需要。

然后，还有一种深度前馈式神经网络，这种网络更易于训练并且比那种全连接的神经网络的泛化性能更好。这就是卷积神经网络（CNN）。当人们对神经网络不感兴趣的时候，卷积神经网络在实践中却取得了很多成功，如今它被计算机视觉团队广泛使用。

卷积神经网络

卷积神经网络被设计用来处理到多维数组数据的，比如一个有3个包含了像素值2-D图像组合成的一个具有3个颜色通道的彩色图像。很多数据形态都是这种多维数组的：1D用来表示信号和序列包括语言，2D用来表示图像或者声音，3D用来表示视频或者有声音的图像。卷积神经网络使用4个关键的想法来利用自然信号的属性：局部连接、权值共享、池化以及多网络层的使用。

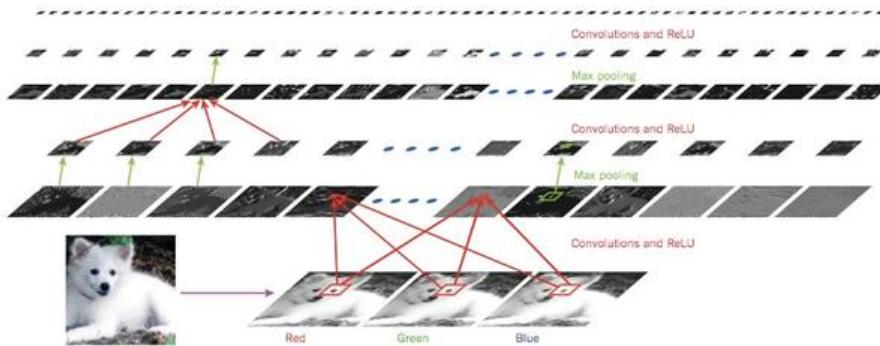


图2 卷积神经网络内部

一个典型的卷积神经网络结构（如图2）是由一系列的过程组成的。最初的几个阶段是由卷积层和池化层组成，卷积层的单元被组织在特征图中，在特征图中，每一个单元通过一组叫做滤波器的权值被连接到上一层的特征图的一个局部块，然后这个局部加权和被传给一个非线性函数，比如ReLU。在一个特征图中的全部单元享用相同的过滤器，不同层的特征图使用不同的过滤器。使用这种结构出于两方面的原因。首先，在数组数据中，比如图像数据，一个值的附近的值经常是高度相关的，可以形成比较容易被探测到的有区别的局部特征。其次，不同位置局部统计特征不太相关的，也就是说，在一个地方出现的某个特征，也可能出现在别的地方，所以不同位置的单元可以共享权值以及可以探测相同的样本。在数学上，这种由一个特征图执行的过滤操作是一个离线的卷积，卷积神经网络也是这么得名来的。

卷积层的作用是探测上一层特征的局部连接，然而池化层的作用是在语义上把相似的特征合并起来，这是因为形成一个主题的特征的相对位置不太一样。一般地，池化单元计算特征图中的一个局部块的最大值，相邻的池化单元通过移动一行或者一列从块上读取数据，因为这样做就减少的表达的维度以及对数据的平移不变性。两三个这种的卷积、非线性变换以及池化被串起来，后面再加上一个更多卷积和全连接层。在卷积神经网络上进行反向传播算法和在一般的深度网络上是一样的，可以让所有的在过滤器中的权值得到训练。

深度神经网络利用的很多自然信号是层级组成的属性，在这种属性中高级的特征是通过对低级特征的组合来实现的。在图像中，局部边缘的组合形成基本图案，这些图案形成物体的局部，然后再形成物体。这种层级结构也存在于语音数据以及文本数据中，如电话中的声音，因素，音节，文档中的单词和句子。当输入数据在前一层中的位置有变化的时候，池化操作让这些特征表示对这些变化具有鲁棒性。

卷积神经网络中的卷积和池化层灵感直接来源于视觉神经科学中的简单细胞和复杂细胞。这种细胞的是以LGN-V1-V2-V4-IT这种层级结构形成视觉回路的。当给一个卷积神经网络和猴子一副相同的图片的时候，卷积神经网络展示了猴子下颞叶皮质中随机160个神经元的变化。卷积神经网络有神经认知的根源，他们的架构有点相似，但是在神经认知中是没有类似反向传播算法这种端到端的监督学习算法的。一个比较原始的1D卷积神经网络被称为时延神经网络，可以被用来识别语音以及简单的单词。

20世纪90年代以来，基于卷积神经网络出现了大量的应用。最开始是用时延神经网络来做语音识别以及文档阅读。这个文档阅读系统使用一个被训练好的卷积神经网络和一个概率模型，这个概率模型实现了语言方面的一些约束。20世纪90年代末，这个系统被用来美国超过10%的支票阅读上。后来，微软开发了基于卷积神经网络的字符识别系统以及手写体识别系统。20世纪90年代早期，卷积神经网络也被用来自然图形中的物体识别，比如脸、手以及人脸识别（face recognition）。

使用深度卷积网络进行图像理解

21世纪开始，卷积神经网络就被成功的大量用于检测、分割、物体识别以及图像的各个领域。这些应用都是使用了大量的有标签的数据，比如交通信号识别，生物信息分割，面部探测，文本、行人以及自然图形中的人的身体部分的探测。近年来，卷积神经网络的一个重大成功应用是人脸识别。

值得一提的是，图像可以在像素级别进行打标签，这样就可以应用在比如自动电话接听机器人、自动驾驶汽车等技术中。像Mobileye以及NVIDIA公司正在把基于卷积神经网络的方法用于汽车中的视觉系统中。其它的应用涉及到自然语言的理解以及语音识别中。

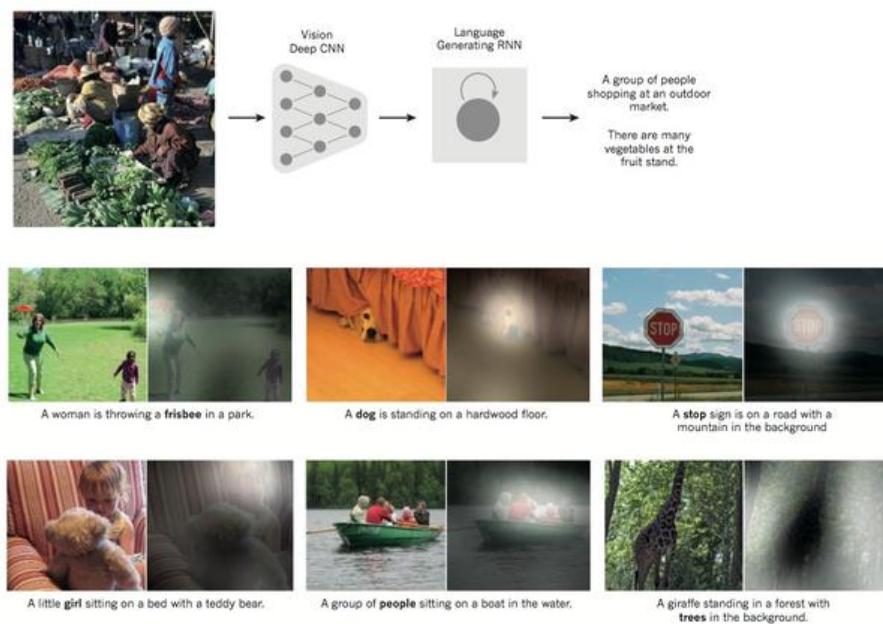


图3 从图像到文字

尽管卷积神经网络应用的很成功，但是它被计算机视觉以及机器学习团队开始重视是在2012年的ImageNet竞赛。在该竞赛中，深度卷积神经网络被用在上百张网络图片数据集，这个数据集包含了1000个不同的类。该结果达到了前所未有的好，几乎比当时最好的方法降低了一半的错误率。这个成功来自有效地利用了GPU、ReLU、一个新的被称为dropout的正则技术，以及通过分解现有样本产生更多训练样本的技术。这个成功给计算机视觉带来一个革命。如今，卷积神经网络用于几乎全部的识别和探测任务中。最近一个更好的成果是，利用卷积神经网络结合回声神经网络来产生图像标题。

如今的卷积神经网络架构有10-20层采用ReLU激活函数、上百万个权值以及几十亿个连接。然而训练如此大的网络两年前就只需要几周了，现在硬件、软件以及算法并行的进步，又把训练时间压缩到了几小时。

基于卷积神经网络的视觉系统的性能已经引起了大型技术公司的注意，比如Google、Facebook、Microsoft、IBM、yahoo!、Twitter和Adobe等，一些快速增长的创业公司也同样如是。

卷积神经网络很容易在芯片或者现场可编程门阵列（FPGA）中高效实现，许多公司比如NVIDIA、Mobileye、Intel、Qualcomm以及Samsung，正在开发卷积神经网络芯片，以使智能机、相机、机器人以及自动驾驶汽车中的实时视觉系统成为可能。

分布式特征表示与语言处理

与不使用分布式特征表示（distributed representations）的经典学习算法相比，深度学习理论表明深度网络具有两个不同的巨大的优势。这些优势来源于网络中各节点的权值，并取决于具有合理结构的底层生成数据的分布。首先，学习分布式特征表示能够泛化适应新学习到的特征值的组合（比如， n 元特征就有 2^n 种可能的组合）。其次，深度网络中组合表示层带来了另一个指数级的优势潜能（指数级的深度）。

多层神经网络中的隐层利用网络中输入的数据进行特征学习，使之更容易预测目标输出。下面是一个很好的示范例子，比如将本地文本的内容作为输入，训练多层神经网络来预测句子中下一个单词。内容中的每个单词表示为网络中的N分之一的向量，也就是说，每个组成部分中有一个值为1其余的全为0。在第一层中，每个单词创建不同的激活状态，或单词向量（如图4）。在语言模型中，网络中其余层学习并转化输入的单词向量为输出单词向量来预测句子中下一个单词，可以通过预测词汇表中的单词作为文本句子中下一个单词出现的概率。网络学习了包含许多激活节点的、并且可以解释为词的独立特征的单词向量，正如第一次示范的文本学习分层表征文字符号的例子。这些语义特征在输入中并没有明确的表征。而是在利用“微规则”（‘micro-rules’，本文中直译为：微规则）学习过程中被发掘，并作为一个分解输入与输出符号之间关系结构的好方式。当句子是来自大量的真实文本并且个别的微规则不可靠的情况下，学

习单词向量也一样能表现得很好。利用训练好的模型预测新的事例时，一些概念比较相似的词容易混淆，比如星期二（Tuesday）和星期三（Wednesday），瑞典（Sweden）和挪威（Norway）。这样的表示方式被称为分布式特征表示，因为他们的元素之间并不互相排斥，并且他们的构造信息对应于观测到的数据的变化。这些单词向量是通过学习得到的特征构造的，这些特征不是由专家决定的，而是由神经网络自动发掘的。从文本中学习得单词向量表示现在广泛应用于自然语言中。

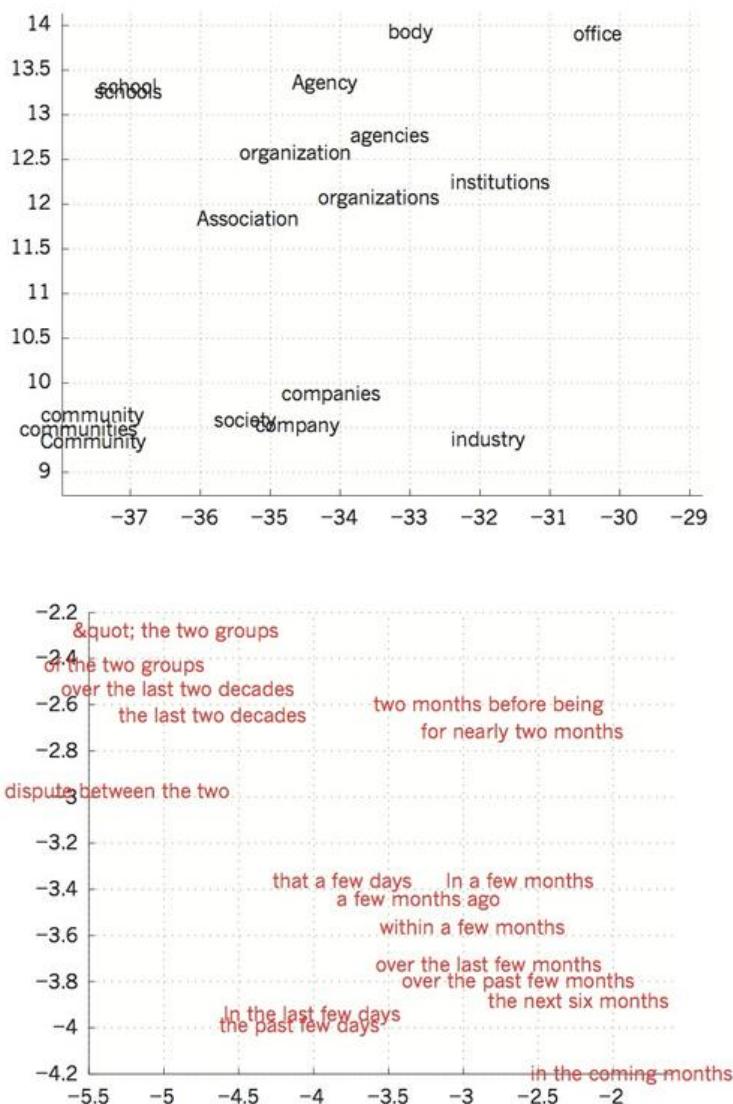


图4 词向量学习可视化

特征表示问题争论的中心介于对基于逻辑启发和基于神经网络的认识。在逻辑启发的范式中，一个符号实体表示某一事物，因为其唯一的属性与其他符号实体相同或者不同。该符号实例没有内部结构，并且结构与使用是相关的，至于理解符号的语义，就必须与变化的推理规则合理对应。相反地，神经网络利用了大量活动载体、权值矩阵和标量非线性化，来实现能够支撑简单容易的、具有常识推理的快速“直觉”功能。

在介绍神经语言模型前，简述下标准方法，其是基于统计的语言模型，该模型没有使用分布式特征表示。而是基于统计简短符号序列出现的频率增长到N（N-grams，N元文法）。可能的N-grams的数字接近于 V^N ，其中V是词汇表的大小，考虑到文本内容包含成千上万个单词，所以需要一个非常大的语料库。N-grams将每个单词看成一个原子单元，因此不能在语义相关的单词序列中一概而论，然而神经网络语言模型可以，是因为他们关联每个词与真是特征值的向量，并且在向量空间中语义相关的词彼此靠近（图4）。

递归神经网络

首次引入反向传播算法时，最令人兴奋的便是使用递归神经网络（recurrent neural networks，下文简称RNNs）训练。对于涉及到序列输入的任务，比如语音和语言，利用RNNs能获得更好的效果。RNNs一次处理一个输入序列元素，同时维护网络中隐式单元中隐式的包含过去时刻序列元素的历史信息的“状态向量”。如果是深度多层网络不同神经元的输出，我们就会考虑这种在不同离散时间步长的隐式单元的输出，这将会使我们更加清晰怎么利用反向传播来训练RNNs（如图5，右）。

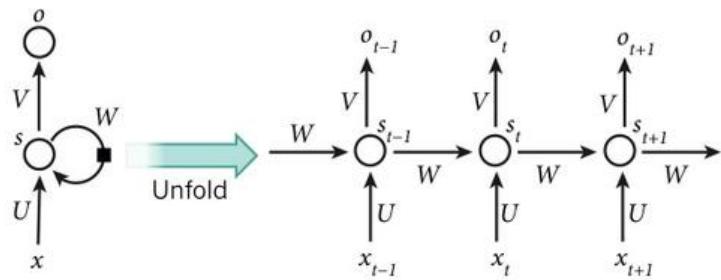


图5 递归神经网络

RNNs是非常强大的动态系统，但是训练它们被证实存在问题的，因为反向传播的梯度在每个时间间隔内是增长或下降的，所以经过一段时间后将导致结果的激增或者降为零。

由于先进的架构和训练方式，RNNs被发现可以很好的预测文本中下一个字符或者句子中下一个单词，并且可以应用于更加复杂的任务。例如在某时刻阅读英语句子中的单词后，将会训练一个英语的“编码器”网络，使得隐式单元的最终状态向量能够很好地表征句子所要表达的意思或思想。这种“思想向量”（thought vector）可以作为联合训练一个法语“编码器”网络的初始化隐式状态（或者额外的输入），其输出为法语翻译首单词的概率分布。如果从分布中选择一个特殊的首单词作为编码网络的输入，将会输出翻译的句子中第二个单词的概率分布，并直到停止选择为止。总体而言，这一过程是根据英语句子的概率分布而产生的法语词汇序列。这种简单的机器翻译方法的表现甚至可以和最先进的（state-of-the-art）的方法相媲美，同时也引起了人们对于理解句子是否需要像使用推理规则操作内部符号表示质疑。这与日常推理中同时涉及到根据合理结论类推的观点是匹配的。

类比于将法语句子的意思翻译成英语句子，同样可以学习将图片内容“翻译”为英语句子（如图3）。这种编码器是可以在最后的隐层将像素转换为活动向量的深度卷积网络（ConvNet）。解码器与RNNs用于机器翻译和神经网络语言模型的类似。近来，已经掀起了一股深度学习的巨大兴趣热潮（参见文献[86]提到的例子）。

RNNs一旦展开（如图5），可以将之视为一个所有层共享同样权值的深度前馈神经网络。虽然它们的目的是学习长期的依赖性，但理论的和经验的证据表明很难学习并长期保存信息。

为了解决这个问题，一个增大网络存储的想法随之产生。采用了特殊隐式单元的LSTM (long short-term memory networks) 被首先提出，其自然行为便是长期的保存输入。一种称作记忆细胞的特殊单元类似累加器和门控神经元：它在下一个时间步长将拥有一个权值并联接到自身，拷贝自身状态的真实值和累积的外部信号，但这种自联接是由另一个单元学习并决定何时清除记忆内容的乘法门控制的。

LSTM网络随后被证明比传统的RNNs更加有效，尤其当每一个时间步长内有若干层时，整个语音识别系统能够完全一致的将声学转录为字符序列。目前LSTM网络或者相关的门控单元同样用于编码和解码网络，并且在机器翻译中表现良好。

过去几年中，几位学者提出了不同的提案用于增强RNNs的记忆模块。提案中包括神经图灵机，其中通过加入RNNs可读可写的“类似磁带”的存储来增强网络，而记忆网络中的常规网络通过联想记忆来增强。记忆网络在标准的问答基准测试中表现良好，记忆是用来记住稍后要求回答问题的事例。

除了简单的记忆化，神经图灵机和记忆网络正在被用于那些通常需要推理和符号操作的任务，还可以教神经图灵机“算法”。除此以外，他们可以从未排序的输入符号序列（其中每个符号都有与其在列表中对应的表明优先级的真实值）中，学习输出一个排序的符号序列。可以训练记忆网络用来追踪一个设定与文字冒险游戏和故事的世界的状态，回答一些需要复杂推理的问题。在一个测试例子中，网络能够正确回答15句版的《指环王》中诸如“Frodo现在在哪？”的问题。

深度学习的未来展望

无监督学习对于重新点燃深度学习的热潮起到了促进的作用，但是纯粹的有监督学习的成功盖过了无监督学习。在本篇综述中虽然这不是我们的重点，我们还是期望无监督学习在长期内越来越重要。无监督学习在人类和动物的学习中占据主导地位：我们通过观察能够发现世界的内在结构，而不是被告知每一个客观事物的名称。

人类视觉是一个智能的、基于特定方式的利用小或大分辨率的视网膜中央窝与周围环绕区域对光线采集成像的活跃的过程。我们期望未来在机器视觉方面会有更多的进步，这些进步来自那些端对端的训练系统，并结合ConvNets和RNNs，采用增强学习来决定走向。结合了深度学习和增强学习的系统正处在初期，但已经在分类任务中超过了被动视频系统，并在学习操作视频游戏中产生了令人印象深刻的效果。

在未来几年，自然语言理解将是深度学习做出巨大影响的另一个领域。我们预测那些利用了RNNs的系统将会更好地理解句子或者整个文档，当它们选择性地学习了某时刻部分加入的策略。

最终，在人工智能方面取得的重大进步将来自那些结合了复杂推理表示学习（representation learning）的系统。尽管深度学习和简单推理已经应用于语音和手写字识别很长一段时间了，我们仍需要通过操作大量向量的新范式来代替基于规则的字符表达式操作。