深度学习综述论文

摘要:

深度学习允许由多个处理层组成的计算模型学习多层次抽象的数据表示。这些方法极大地提高了语音识别、视觉对象识别、物体检测以及药物发现和基因组学等许多其他领域的先进水平。深度学习 深度学习通过使用反向传播算法来指示机器应如何改变其内部参数，从而发现大型数据集中错综复杂的结构。机器应如何改变其内部参数，这些参数用于根据上一层的表征计算每一层的表征。的内部参数。深度卷积网络在处理图像、视频、语音和音频方面取得了突破性进展。而递归网络则在文本和语音等顺序数据方面大放异彩。

引言:

机器学习技术为现代社会的许多方面提供了动力：从网络搜索到社交网络的内容过滤，再到电子商务网站的推荐，机器学习技术越来越多地出现在相机和智能手机等消费产品中。机器学习系统用于识别图像中的物体，将语音转录为文本，将新闻条目、帖子或产品与用户的兴趣相匹配，以及选择相关的搜索结果。这些应用越来越多地使用一类名为深度学习的技术。

传统的机器学习技术处理原始形式的自然数据的能力是有限的。几十年来，构建模式识别或机器学习系统需要精心的工程设计和大量的领域专业知识，以设计一个特征提取器，将原始数据（如图像的像素值）转换为合适的内部表示或特征向量，学习子系统（通常是分类器）可以从中检测或分类输入中的模式。

表征学习是一套方法，它允许机器输入原始数据，并自动发现检测或分类所需的表征。深度学习方法是一种具有多层次表征的表征学习方法，它由简单但非线性的模块组成，每个模块将一个层次的表征（从原始输入开始）转换成更高、稍微更抽象层次的表征。只要组合足够多的此类转换，就能学习到非常复杂的功能。对于分类任务来说，较高的表征层会放大输入中对辨别很重要的方面，并抑制无关的变化。例如，图像以像素值阵列的形式出现，第一层表征中学习到的特征通常代表图像中特定位置和区域是否存在边缘。第二层通常通过发现边缘的特定排列来检测图案，而不考虑边缘位置的微小变化。第三层可能会将图案组合成更大的组合，对应于熟悉物体的各个部分，随后的各层会将物体检测为这些部分的组合。深度学习的关键在于，这些特征层不是由人类工程师设计的：它们是通过通用目标学习程序从数据中学习出来的。

深度学习在解决人工智能界多年来一直无法解决的问题方面取得了重大进展。事实证明，深度学习非常善于发现高维数据中的复杂结构，因此适用于科学、商业和政府的许多领域。除了在图像识别和语音识别方面打破记录外，它还在预测潜在药物分子的活性、分析粒子加速器数据、重建大脑回路以及预测非编码 DNA 突变对基因表达和疾病的影响等方面击败了其他机器学习技术。也许更令人惊讶的是，深度学习在自然语言理解的各种任务中，特别是在主题分类、情感分析、问题解答和语言翻译方面，取得了非常有前景的成果。

我们认为，深度学习在不久的将来会取得更多成功，因为它只需要很少的手工工程，因此可以轻松利用可用计算量和数据量的增加。目前正在为深度神经网络开发的新学习算法和架构只会加快这一进程。

正文-监督学习:

最常见的机器学习形式（无论是否深度学习）是监督学习。想象一下，我们想要构建一个系统，将图像分类为房子、汽车、人或宠物。我们首先要收集一个包含房屋、汽车、人和宠物图像的大型数据集，每张图像都标有其类别。在训练过程中，机器会看到一幅图像，并以向量的形式输出分数，其中每个类别一个分数。我们希望所需的类别在所有类别中得分最高，但这在训练之前不太可能实现。我们计算出一个目标函数，用于测量输出分数与理想分数模式之间的误差（或距离）。然后，机器会修改其内部可调参数，以减小误差。这些可调参数通常称为权重，是实数，而且可视为定义机器输入-输出功能的“旋钮”。在一个典型的深度学习系统中，可能有数以亿计的可调权重，以及数以亿计的标注示例来训练机器。

为了正确调整权重向量，学习算法会计算出一个梯度向量，该梯度向量会为每个权重指出，如果权重增加一个微小的量，误差会增加或减少多少。然后，权重向量的调整方向与梯度向量相反。

目标函数是所有训练实例的平均值，可以看作是权重值高维空间中的丘陵地貌。负梯度向量表示在该景观中最陡峭的下降方向，使其更接近最小值，在该处输出误差平均较小。

在实践中，大多数从业者使用一种称为随机梯度下降（SGD）的程序。这包括显示几个例子的输入向量，计算输出和误差，计算这些例子的平均梯度，并相应地调整权重。这个过程在训练集中的许多小例子组中重复进行，直到目标函数的平均值停止下降。之所以称其为随机过程，是因为每一小组示例都会对所有示例的平均梯度做出不准确的估计。与复杂得多的优化技术相比，这种简单的程序通常能以惊人的速度找到一组好的权重。训练结束后，系统的性能将在另一组称为测试集的示例上进行测量。测试集的作用是测试机器的泛化能力--即机器对训练期间从未见过的新输入做出合理回答的能力。

目前，机器学习的许多实际应用都在手工设计的特征之上使用线性分类器。一个二类线性分类器会对特征向量的各个组成部分计算加权和，如果这个加权和超过某个阈值，那么输入就会被分类为属于某个特定类别。

自 20 世纪 60 年代起，我们就知道线性分类器只能将输入空间划分为非常简单的区域，即由超平面分隔的半空间。但是，图像识别和语音识别等问题要求输入-输出函数对输入的无关变化不敏感，如物体的位置、方向或光照变化，或语音的音调或口音变化，而对特定的微小变化（例如，白狼和一种被称为萨摩耶的类似狼的白狗之间的区别）却非常敏感。在像素级别上，两只萨摩耶犬在不同姿势和不同环境下的图像可能大相径庭，而两只萨摩耶犬和狼在相同姿势和相似背景下的图像可能非常相似。线性分类器或任何其他“浅层”分类器的工作原理是：

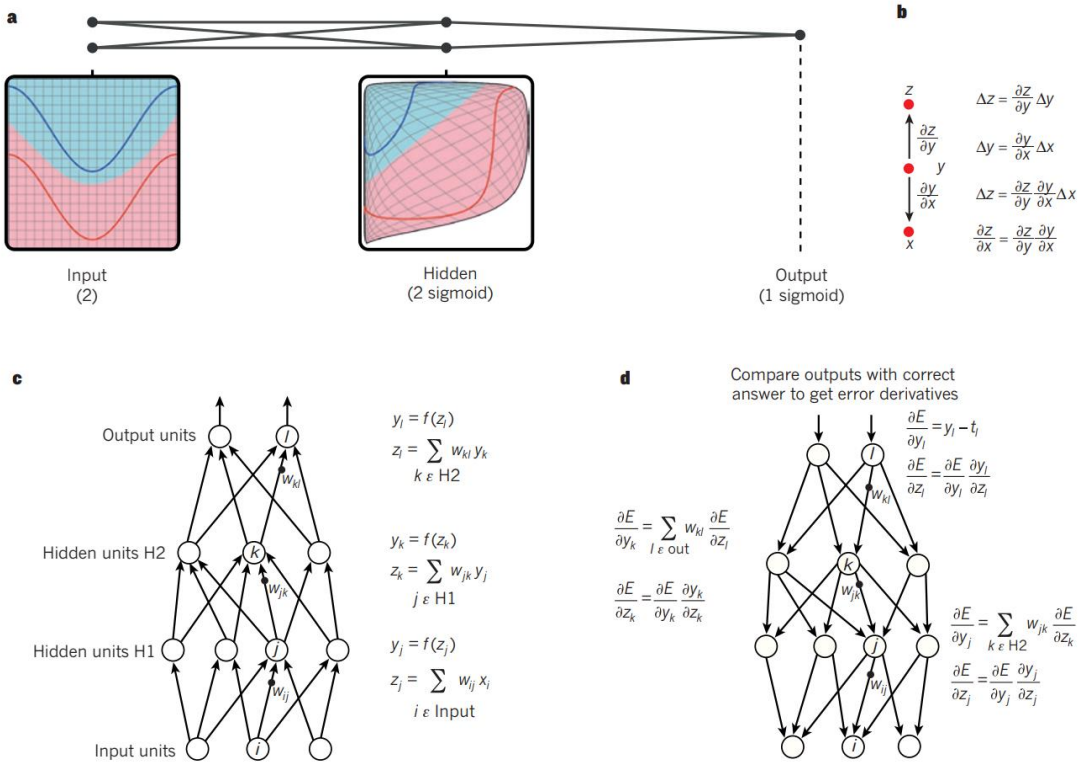

图 1 多层神经网络和反向传播

a, 多层神经网络（如连接点所示）可以扭曲输入空间，使数据类别（红线和蓝线上的示例）可线性分离。请注意，输入空间中的规则网格（如左图所示）是如何被隐藏单元转换的（如中图所示）。这是一个只有两个输入单元、两个隐藏单元和一个输出单元的示例，但用于物体识别或自然语言处理的网络包含数万或数十万个单元。经授权转载自 C. Olah (http://colah.github.io/)。

b, 导数的链式法则告诉我们两个微小的效应（x 对 y 的微小变化和 y 对 z 的微小变化）是如何构成的。x 的微小变化 Δx 通过乘以 $\partial y/\partial x$ 首先转化为 y 的微小变化 Δy（即偏导数的定义）。将一个等式代入另一个等式，就得到了导数的链式法则--Δx 如何通过乘以 $\partial y/\partial x$ 和 $\partial z/\partial x$ 的乘积而转化为 Δz。当 x、y 和 z 为向量（导数为雅各布矩阵）时，它也同样有效。

c, 神经网络中用于计算前向传递的方程，该神经网络有两个隐藏层和一个输出层，每个隐藏层构成一个模块，通过该模块可以反向传播梯度。在每一层，我们首先计算每个单元的总输入 z，它是下一层单元输出的加权和。然后将非线性函数 f(.) 应用于 z，得到单元的输出。为简单起见，我们省略了偏置项。神经网络中使用的非线性函数包括近年来常用的整流线性单元（ReLU）f(z) = max(0,z)，以及更传统的激活函数，如双曲正切函数 f(z) =(exp(z)- exp(-z))/(exp(z)+exp(-z))和对数函数 logistic，f(z) =1/(1 + exp(-z))。

d，用于计算反向传播的方程。在每个隐藏层，我们计算与每个单元输出相关的误差导数，即与上一层单元总输入相关的误差导数的加权和。然后，我们将相对于输出的误差导数乘以 f(z) 的梯度，转换成相对于输入的误差导数。在输出层，相对于单元输出的误差导数是通过微分代价函数计算得出的。如果单元 l 的成本函数为 $0.5(y_l - t_l)^2$，则得出 $y_l - t_l$，其中 t_l 为目标值。一旦 $\partial E/\partial z_k$ 已知，下一层单元 j 的连接权重 w_{jk} 的误差导数就是 $y_j \partial E/\partial z_k$。

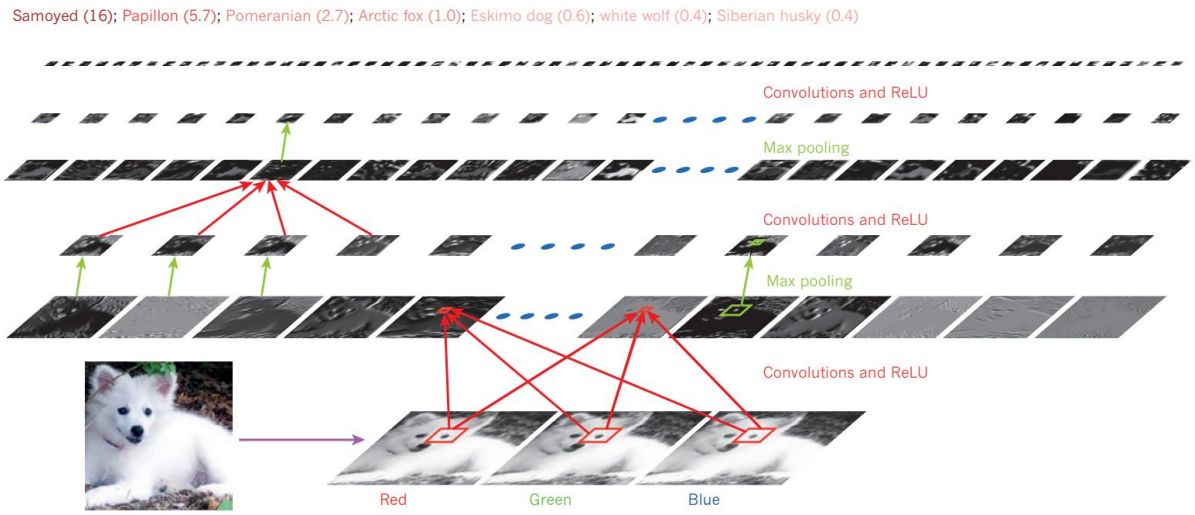

图 2 卷积网络的内部结构

图解：应用于萨摩耶犬图像（左下；RGB（红、绿、蓝）输入，右下）的典型卷积网络架构每层（水平）的输出（非过滤器）。每个矩形图像都是一个特征图，对应于在每个图像位置检测到的一个学习特征的输出。信息自下而上流动，较低层次的特征作为定向边缘检测器，并在输出中为每个图像类别计算分数。ReLU，线性整流函数。

原始像素不可能将后两者区分开来，而将前两者归为一类。这就是为什么浅层分类器需要一个好的特征提取器来解决选择性-不变性难题--它所产生的表征对图像中对判别很重要的方面有选择性，但对无关方面（如动物的姿势）不变。为了使分类器更加强大，我们可以使用通用的非线性特征，如核方法，但通用特征（如高斯核产生的特征）并不能使学习者很好地泛化到远离训练实例的地方。传统的方法是手工设计优秀的特征提取器，这需要大量的工程技术和领域专业知识。但是，如果可以使用通用目标学习程序自动学习良好的特征，这一切都可以避免。这就是深度学习的关键优势。

深度学习架构是一个由简单模块组成的多层堆栈，所有（或大部分）模块都需要学习，其中许多模块计算非线性输入-输出映射。堆栈中的每个模块都会转换其输入，以提高表征的选择性和不变性。有了多个非线性层，比如 5 到 20 层的深度，系统就能实现极其复杂的输入函数，同时对微小的细节也很敏感 -- 例如区分萨摩耶犬和白狼 -- 同时对背景、姿势、光线和周围物体等大量无关变化不敏感。

正文-反向传播训练多层架构：

从模式识别的最初阶段开始，研究人员的目标就是用可训练的多层网络取代手工设计的特征，但尽管这种方法很简单，直到 20 世纪 80 年代中期才被广泛理解。事实证明，多层架构可以通过简单的随机梯度下降法进行训练。只要模块是其输入和内部权重的相对平滑函数，就可以使用反向传播程序来计算梯度。在 20 世纪 70 年代和 80 年代，几个不同的研究小组独立发现了这一想法，并证明其可行。

计算目标函数相对于多层模块权重的梯度的反向传播过程，不过是导数链式法则的实际应用而已。其关键在于，目标函数相对于模块输入的导数（或梯度），可以从相对于该模块输出（或后续模块输入）的梯度倒推计算得出（图 1）。可以反复应用反向传播方程，将梯度传播到所有模块，从顶部的输出（网络在此进行预测）一直传播到底部（外部输入）。一旦计算出这些梯度，就可以直接计算每个模块权重的梯度。

深度学习的许多应用都使用前馈神经网络架构（图 1），该架构可学习将固定大小的输入（如图像）映射到固定大小的输出（如多个类别中每个类别的概率）。从一层到下一层，一组单元计算上一层输入的加权和，并将结果通过非线性函数传递。目前，最流行的非线性函数是整流线性函数（ReLU），即简单的半波整流 f(z)=max(z,0)。在过去的几十年中，神经网络使用了更平滑的非线性函数，如 tanh(z) 或 1/(1+exp(-z))，但 ReLU 在多层网络中的学习速度通常更快，因此无需无监督预训练就能训练深度监督网络。不在输入或输出层的单元通常称为隐藏单元。隐藏层可以被视为以非线性方式扭曲输入，从而使类别在最后一层变得线性可分（图 1）。

20 世纪 90 年代末，神经网络和反向传播在很大程度上被机器学习界所抛弃，也被计算机视觉和语音识别界所忽视。人们普遍认为，利用很少的先验知识学习有用的多级特征提取器是不可行的。尤其是，人们普遍认为，简单的梯度下降会陷入不佳的局部极小值 -- 对于这种权重配置，任何微小的改变都无法降低平均误差。

实际上，在大型网络中，局部极小值不佳的问题很少出现。无论初始条件如何，系统几乎总是能获得质量非常接近的解。最近的理论和经验结果有力地表明，局部极小值在一般情况下并不是一个严重的问题。相反，在梯度为零的地方会出现大量的鞍点，曲面在大多数维度上向上弯曲，而在其余维度上向下弯曲。分析似乎表明，只有少数向下弯曲方向的鞍点大量存在，但几乎所有这些鞍点的目标函数值都非常相似。因此，算法卡在哪个鞍点并不重要。

2006 年左右，加拿大高级研究所（CIFAR）召集了一批研究人员，重新激发了人们对深度前馈网络的兴趣。研究人员引入了无监督学习程序，无需标记数据即可创建多层特征检测器。学习每一层特征检测器的目的是能够重建或模拟下一层特征检测器的活动（或原始输入）。通过使用这一重构目标“预训练”几层逐渐复杂的特征检测器，可以将深度网络的权重初始化为合理的值。最后一层输出单元可以添加到网络的顶层，然后使用标准的反向传播技术对整个深度系统进行微调。这在识别手写数字或检测行人方面效果显著，尤其是在标记数据量非常有限的情况下。

这种预训练方法的首次重大应用是在语音识别领域，它的出现得益于快速图形处理器（GPU）的出现，这种处理器编程方便，使研究人员训练网络的速度提高了 10 或 20 倍。2009 年，该方法被用于将从声波中提取的短时间系数窗口映射到窗口中心帧可能代表的各种语音片段的概率集。它在一个使用小词汇量的标准语音识别基准测试中取得了破纪录的成绩，并迅速发展到在一个大词汇量任务中取得破纪录的成绩。到 2012 年，许多主要语音小组都在开发 2009 年的深度网络版本，并已将其部署到安卓手机中。对于较小的数据集，无监督预训练有助于防止过拟合，从而在标记示例数量较少时，或在我们对某些“源”任务拥有大量示例，但对某些“目标”任务却只有极少示例的转移设置中，显著提高泛化效果。深度学习恢复后，发现预训练阶段仅适用于小型数据集。

然而，有一种特殊的深度前馈网络比相邻层之间完全连接的网络更容易训练，泛化效果也更好。这就是卷积神经网络（ConvNet）。它在神经网络不受欢迎的时期取得了许多实际成功，最近已被计算机视觉界广泛采用。

正文-卷积神经网络：

ConvNets 设计用于处理多数组形式的数据，例如由三个二维数组组成的彩色图像，其中包含三个彩色通道的像素强度。许多数据模式都是多数组形式： 1D 用于信号和序列，包括语言；2D 用于图像或音频频谱图；3D 用于视频或体积图像。ConvNets 利用了自然信号的特性，其背后有 4 个关键理念：局部连接、共享权重、池化和多层使用。

典型 ConvNet 的架构（图 2）由一系列阶段组成。前几个阶段由两类层组成：卷积层和池化层。卷积层中的单元以特征图的形式组织，其中每个单元通过一组称为滤波器组的权值与上一层特征图中的局部斑块相连。然后，这一局部加权和的结果会通过一个非线性器（如 ReLU）。一个特征图中的所有单元共享同一个滤波器组。一个层中的不同特征图使用不同的滤波器组。采用这种结构有两个原因。首先，在图像等阵列数据中，局部数值组往往高度相关，形成独特的局部图案，易于检测。其次，图像和其他信号的局部统计与位置无关。换句话说，如果一个图案可能出现在图像的某一部分，那么它也可能出现在任何地方，因此不同位置的单元可以共享相同的权重，并在阵列的不同部分检测到相同的图案。在数学上，特征图谱进行的过滤操作是离散卷积，因此得名。

虽然卷积层的作用是检测前一层特征的局部连接，但池化层的作用是将语义相似的特征合并为一个。由于构成图案的特征的相对位置可能会有一些变化，因此可以通过粗粒度每个特征的位置来可靠地检测图案。典型的池化单元会计算一个特征图（或几个特征图）中局部单元的最大值。相邻的池化单元从偏移超过一行或一列的斑块中获取输入，从而降低了表征维度，并产生了对微小偏移和扭曲的不变性。卷积、非线性和池化的两个或三个阶段被叠加，然后是更多的卷积层和全连接层。通过 ConvNet 进行梯度反向传播与通过普通深度网络一样简单，可以训练所有滤波器组中的所有权重。

深度神经网络利用了许多自然信号都是组成层次的特性，其中较高层次的特征是由较低层次的特征组成的。在图像中，边缘的局部组合形成图案，图案组合成部件，部件形成物体。语音和文本中也存在类似的层次结构，从声音到电话、音素、音节、单词和句子。当上一层的元素在位置和外观上发生变化时，池化可以使表征变化很小。

ConvNets 中的卷积层和池化层直接受到视觉神经科学中简单细胞和复杂细胞经典概念的启发，其整体结构让人联想到视觉皮层腹侧通路中的 LGN-V1-V2-V4-IT 层次结构。当向 ConvNet 模型和猴子展示相同图片时，ConvNet 中高级单元的激活可以解释猴子颞下部皮层 160 个随机神经元组的一半差异。ConvNet 起源于新认知计算机（neocognitron），其架构与之有些类似，但 neocognitron 没有端到端的监督学习算法，如反向传播算法。一种被称为延时神经网络的原始一维 ConvNet 被用于识别音素和简单单词。

卷积网络的应用可以追溯到 20 世纪 90 年代初，首先是用于语音识别和文档阅读的延时神经网络。文件阅读系统使用的是与实现语言限制的概率模型联合训练的卷积网络。到 20 世纪 90 年代末，该系统读取了美国 10% 以上的支票。后来，微软公司部署了许多基于 ConvNet 的光学字符识别和手写识别系统。20 世纪 90 年代初，人们还尝试将 ConvNets 用于自然图像中的物体检测，包括人脸和手，以及人脸识别。

正文-利用深度卷积网络理解图像：

自 2000 年代初以来，ConvNets 已成功应用于图像中物体和区域的检测、分割和识别。这些都是标注数据相对丰富的任务，如交通标志识别、生物图像分割（特别是用于连接组学）以及自然图像中人脸、文字、行人和人体的检测。ConvNets 最近在实际应用中取得的一项重大成就是人脸识别。

重要的是，图像可以在像素级进行标注，这将应用于技术领域，包括自动移动机器人和自动驾驶汽车。应用于技术领域，包括自主移动机器人和自动驾驶汽车。Mobileye 和英伟达（NVIDIA）等公司正在即将推出的汽车视觉系统中使用这种基于 ConvNet 的方法。其他越来越重要的应用包括自然语言理解和语音识别。

.

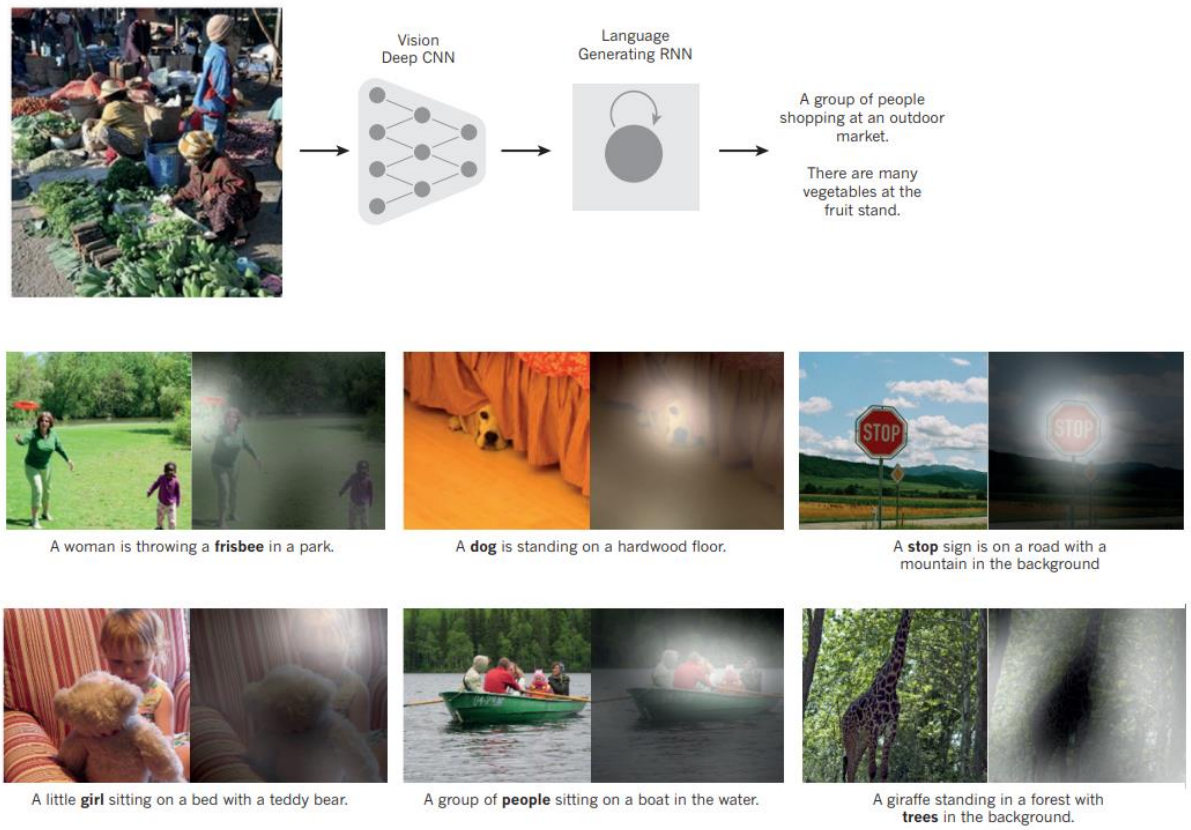

图 3 从图像到文本

图解：由递归神经网络（RNN）生成的字幕，RNN 将深度卷积神经网络（CNN）从测试图像中提取的表示作为额外输入，经过训练后可将图像的高级表示“翻译”为字幕（上图）。当 RNN 在生成每个单词（粗体）时，可以将注意力集中在输入图像的不同位置（中间和底部；浅色斑块受到更多关注），我们发现 RNN 可以利用这一点更好地将图像“翻译”为标题。

尽管取得了这些成功，卷积网络在很大程度上还是被主流计算机视觉和机器学习社区所遗弃，直到 2012 年的 ImageNet 竞赛。当深度卷积网络被应用于一个包含 1000 种不同类别的、来自网络的约一百万张图像的数据集时，它们取得了惊人的成绩，几乎将最佳竞争方法的错误率降低了一半。这一成功得益于对 GPU、ReLUs、一种名为 dropout 的新正则化技术以及通过变形现有示例生成更多训练示例的技术的有效利用。这一成功为计算机视觉领域带来了一场革命；ConvNets 现在已成为几乎所有识别和检测任务的主流方法，并在某些任务中接近人类的表现。最近一个令人惊叹的演示结合了 ConvNets 和递归网模块，用于生成图像标题（图 3）。

最新的 ConvNet 架构有 10 到 20 层 ReLU、数亿个权重和数十亿个单元之间的连接。两年前，训练如此庞大的网络可能需要数周时间，而现在，硬件、软件和算法并行化方面的进步已将训练时间缩短到几小时。

基于 ConvNet 的视觉系统的性能已促使包括谷歌、Facebook、微软、IBM、雅虎、Twitter 和 Adobe 在内的大多数大型科技公司，以及越来越多的初创公司启动研发项目，并部署基于 ConvNet 的图像理解产品和服务。

ConvNets 易于在芯片或现场可编程门阵列中进行高效的硬件实现。英伟达（NVIDIA）、Mobileye、英特尔（Intel）、高通（Qualcomm）和三星（Samsung）等多家公司正在开发 ConvNet 芯片，以便在智能手机、摄像头、机器人和自动驾驶汽车中实现实时视觉应用。

正文-分布式表征和语言处理：

深度学习理论表明，与不使用分布式表征的传统学习算法相比，深度网络具有两种不同的指数优势。这两种优势都源于组成的力量，并取决于底层数据生成分布是否具有适当的组成结构。首先，通过学习分布式表征，可以将所学特征值的新组合泛化到训练期间所见的组合之外（例如，n 个二进制特征可能有 2^n 种组合）。其次，在深度网络中组成表征层可以带来另一种指数优势（深度指数）。

多层神经网络的隐藏层会学习如何表示网络的输入，以便轻松预测目标输出。通过训练多层神经网络，使其能够根据先前单词的局部语境预测序列中的下一个单词，就很好地证明了这一点。上下文中的每个单词都以 1-N 向量的形式呈现给网络，也就是说，其中一个分量的值为 1，其余分量的值为 0。在第一层中，每个单词都会产生不同的激活模式或单词向量（图 4）。

在语言模型中，网络的其他层学习如何将输入词向量转换为预测下一个词的输出词向量，这可以用来预测词汇表中任何一个词作为下一个词出现的概率。网络学习的单词向量包含许多有源元件，每个有源元件都可以解释为单词的一个单独特征，这在学习符号的分布式表示时首次得到了证明。这些语义特征并没有明确出现在输入中。它们是由学习程序发现的，是将输入和输出符号之间的结构关系分解为多个“微规则”的好方法。事实证明，当单词序列来自大量真实文本语料库，而单个微规则不可靠时，学习单词向量也非常有效。

表征问题是逻辑启发式认知范式和神经网络启发式认知范式之间争论的核心。在逻辑启发范式中，符号实例的唯一属性是它与其他符号实例相同或不相同。它没有与其使用相关的内部结构；要使用符号进行推理，就必须将符号与明智选择的推理规则中的变量绑定在一起。相比之下，神经网络只是利用大活动向量、大权重矩阵和标量非线性来进行快速的“直观”推理，而这种推理正是毫不费力的常识推理的基础。

在引入神经语言模型之前，语言统计建模的标准方法并不利用分布式表征：它是基于计算长度不超过 N 的短符号序列（称为 N-grams）的出现频率。可能的 N-grams（N-语法）的数量大约为 V^N，其中 V 是词汇量的大小，因此要考虑到超过少量单词的上下文，就需要非常大的训练语料库。N-grams 将每个单词视为一个原子单位，因此无法在语义相关的单词序列中进行泛化，而神经语言模型则可以，因为它们将每个单词与实值特征的向量相关联，而语义相关的单词在该向量空间中相互靠近（图 4）。

图 4 可视化所学单词向量

图解：左侧是为语言建模而学习的单词表示，使用 t-SNE 算法将其非线性投影到 2D 以实现可视化。右图是英语-法语编码器-解码器递归神经网络学习到的短语的二维表示。我们可以观察到，语义相似的单词或单词序列被映射到附近的表征中。单词的分布式表征是通过使用反向传播来共同学习每个单词的表征和一个预测目标量的函数而获得的，例如序列中的下一个单词（用于语言建模）或整个翻译单词序列（用于机器翻译）。

正文-递归神经网络：

反向传播技术刚问世时，最令人兴奋的用途是训练递归神经网络（RNNs）。对于涉及顺序输入的任务，如语音和语言，通常最好使用 RNNs（图 5）。RNNs 一次处理一个元素的输入序列，在其隐藏单元中保持一个“状态向量”，其中隐含了序列中所有过去元素的历史信息。当我们将不同离散时间步长的隐藏单元输出视为深度多层网络中不同神经元的输出时（图 5 右），我们就可以清楚地看到如何应用反向传播来训练 RNNs。

RNN 是非常强大的动态系统，但对其进行训练已被证明是一个难题，因为在每个时间步长内，反向传播梯度要么增长要么缩小，因此在许多时间步长内，反向传播梯度通常会爆炸或消失。

由于结构和训练方法的进步，人们发现 RNN 在预测文本中的下一个字符或序列中的下一个单词方面非常出色，但它们也可用于更复杂的任务。例如，在逐字阅读英语句子后，可以训练一个英语“编码器”网络，使其隐藏单元的最终状态向量能够很好地代表句子所表达的思想。然后，该思维向量可用作联合训练的法语“解码器”网络的初始隐藏状态（或额外输入），该网络将输出法语翻译第一个单词的概率分布。如果从该分布中选择一个特定的第一个单词作为解码器网络的输入，它就会输出译文第二个单词的概率分布，依此类推，直到选择一个句号。

总之，这一过程根据取决于英语句子的概率分布生成法语单词序列。这种相当幼稚的机器翻译方式很快就能与最先进的机器翻译技术相媲美，这让人严重怀疑理解一个句子是否需要类似于使用推理规则所操作的内部符号表达。它更符合这样一种观点，即日常推理涉及许多同时进行的类比，每种类比都有助于得出结论的可信度。

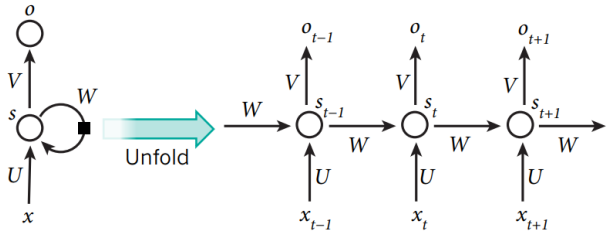

图 5 递归神经网络及其前向计算所涉及的计算在时间上的展开

图解：人工神经元（例如，节点 s 下的隐藏单元，在时间 t 时的值为 s_t）从其他神经元获得前一时间步的输入（左侧的黑色方块代表一个时间步的延迟）。这样，递归神经网络就能将一个包含元素 x_t 的输入序列映射成一个包含元素 o_t 的输出序列，而每个 o_t 都取决于之前的所有 x_t'（对于 $t' \leqslant t$）。每个时间步都使用相同的参数（矩阵 U、V、W）。还可以采用许多其他架构，包括网络可以生成一系列输出（例如单词）的变体，每个输出都将用作下一个时间步骤的输入。反向传播算法（图 1）可直接应用于右侧展开网络的计算图，计算总误差（例如，生成正确输出序列的对数概率）相对于所有状态 s_t 和所有参数的导数。

与其将法语句子的意思翻译成英语句子，不如学习将图像的意思“翻译”成英语句子（图 3）。这里的编码器是一个深度 ConvNet，它在最后一个隐藏层将像素转换成活动向量。解码器是一个 RNNs，类似于用于机器翻译和神经语言建模的 RNNs。最近，人们对这类系统的兴趣大增。

RNNs 一旦在时间上展开（图 5），就可以看作是非常深的前馈网络，其中所有层共享相同的权重。虽然它们的主要目的是学习长期依赖关系，但理论和经验证据表明，很难学会长时间存储信息。

为了纠正这种情况，一种方法是在网络中加入显式记忆。这类网络的第一个方案是长短期记忆（LSTM）网络，它使用特殊的隐藏单元，其自然行为是长期记忆输入。一个被称为记忆单元的特殊单元的作用类似于累加器或门控漏神经元：它在下一个时间步与自身有一个权重为 1 的连接，因此它可以复制自身的实值状态并累加外部信号，但这种自连接由另一个单元进行乘法门控，该单元学会决定何时清除记忆内容。

后来的实践证明，LSTM 网络比传统的 RNNs 更为有效，尤其是当 LSTM 网络在每个时间步长上都有若干层时，整个语音识别系统就可以从声学一直延伸到转录中的字符序列。LSTM 网络或相关形式的门控单元目前也被用于在机器翻译中表现出色的编码器和解码器网络。

在过去的一年中，多位学者提出了用记忆模块增强 RNNs 的不同建议。这些建议包括神经图灵机（Neural Turing Machine）和记忆网络（Memory Networks），在神经图灵机中，RNNs 可以选择读取或写入一个 “类似磁带 ”的存储器来增强网络。记忆网络在标准问题解答基准测试中表现出色。记忆网络用于记住网络随后要回答问题的故事。

除了简单的记忆，神经图灵机和记忆网络还被用于通常需要推理和符号操作的任务。神经图灵机可以学习“算法”。其中包括，当输入是一个未排序的序列时，神经图灵机可以学会输出一个排序的符号列表；在这个序列中，每个符号都有一个实值，表示其在列表中的优先级。记忆网络可以被训练成在类似文字冒险游戏的环境中追踪世界的状态，在阅读故事后，它们可以回答需要复杂推理的问题。在一个测试示例中，记忆网络看到了 15 句版本的《魔戒》，并正确回答了“弗罗多现在在哪里”等问题。

正文-深度学习的未来：

无监督学习对唤起人们对深度学习的兴趣起到了催化剂的作用，但后来被纯监督学习的成功所掩盖。虽然我们在本《综述》中没有重点讨论这个问题，但我们预计，从长远来看，无监督学习将变得更加重要。人类和动物的学习在很大程度上是无监督的：我们通过观察世界来发现世界的结构，而不是被告知每个物体的名称。

人类的视觉是一个主动的过程，它以一种智能的、针对特定任务的方式，通过一个小的、高分辨率的眼窝和一个大的、低分辨率的环绕，依次对光学阵列进行采样。我们预计，未来视觉领域的进步将主要来自端到端训练的系统，这些系统结合了 ConvNets 和 RNNs，利用强化学习来决定观察的方向。深度学习与强化学习相结合的系统还处于起步阶段，但它们已经在分类任务上超越了被动视觉系统，并在学习玩许多不同的视频游戏方面取得了令人印象深刻的成果。

自然语言理解是深度学习有望在未来几年产生巨大影响的另一个领域。我们预计，当使用 RNNs 理解句子或整个文档的系统学会了有选择地同时关注一个部分的策略后，它们将变得更加出色。

最终，人工智能的重大进步将来自于将表征学习与复杂推理相结合的系统。虽然深度学习和简单推理已在语音和手写识别中应用了很长时间，但仍需要新的范式来取代基于规则的符号表达式操作，对大型向量进行运算。