Nature 重磅：Hinton、LeCun、Bengio 三巨头权威科普深度学习

借助深度学习，多处理层组成的计算模型可通过多层抽象来学习数据表征（ representations）。这些方法显著推动了语音识别、视觉识别、目标检测以及许多其他领域（比如，药物发现以及基因组学）的技术发展。利用反向传播算法（backpropagation algorithm）来显示机器将会如何根据前一层的表征改变用以计算每层表征的内部参数，深度学习发现了大数据集的复杂结构。深层卷积网络（deep convolutional nets)为图像、视频和音频等数据处理上带来突破性进展，而递归网络（recurrent nets ）也给序列数据（诸如文本、语言）的处理带来曙光。

机器学习为现代生活诸多方面带来巨大动力：从网页搜索到社交网络内容过滤再到电商网商推荐，在相机、智能手机等消费品中也越来越多见。机器学习系统被用来识别图像中的物体、将语音转为文本，根据用户兴趣自动匹配新闻、消息或产品，挑选相关搜索结果。这类被应用程序越来越多地采用的技术，叫做深度学习。

传统机器学习技术在处理原始输入的自然数据方面能力有限。几十年来，建构模式识别或机器学习系统需要利用严谨的工程学和相当丰富的专业知识设计出一个特征提取器，它能将原始数据(例如图像像素值)转化成适于内部描述或表征的向量(vector)，在提取器中，学习子系统（通常是一个分类器）可以检测或分类输入模式。

表征学习（representation learning）是这样一套学习方法：输入原始数据后，机器能够自动发现检测或分类所需的表征信息。深度学习是一种多层描述的表征学习，通

过组合简单、非线性模块来实现，每个模块都会将最简单的描述（从原始输入开始）转变成较高层、较为抽象的描述。通过积累足够多的上述表征转化，机器能学习非常复杂的函数。就分类任务来说，更高层的表征会放大输入信号的特征，而这对区分和控制不相关变量非常关键。比如，图片最初以像素值的方式出现，第一特征层级中，机器习得的特征主要是图像中特定方位、位置边沿之有无。第二特征层级中，主要是通过发现特定安排的边缘来检测图案，此时机器并不考虑边沿位置的微小变化。第三层中会将局部图像与物体相应部分匹配，后续的层级将会通过把这些局部组合起来从而识别出整个物体。深度学习的关键之处在于：这些特征层级并非出自人类工程师之手；而是机器通过一个通用（general-purpose）学习程序，从大量数据中自学得出。

某些根深蒂固的问题困扰了人工智能从业者许多年，以至于人们最出色的尝试都无功而返。而深度学习的出现，让这些问题的解决迈出了至关重要的步伐。深度学习善于在高维度的数据中摸索出错综复杂的结构，因此能应用在许多不同的领域，比如科学、商业和政府。此外，除了图像识别和语音识别，它还在许多方面击败了其他机器学习技术，比如预测潜在药物分子的活性、分析粒子加速器的数据、重构大脑回路、预测非编码 DNA 的突变对基因表达和疾病有何影响等。也许，最让人惊讶的是，在自然语言理解方面，特别是话题分类、情感分析、问答系统和语言翻译等不同的任务上，深度学习都展现出了无限光明的前景。

在不久的将来，我们认为深度学习将取得更多成就，因为它只需要极少的人工参与，所以它能轻而易举地从计算能力提升和数据量增长中获得裨益。目前正在开发的用于深层神经网络的新型学习算法和体系结构必将加速这一进程。

监督式学习

不管深度与否，机器学习最普遍的形式都是监督式学习（supervised learning）。比如说，我们想构造一个系统，它能根据特定元素对图片进行分类，例如包含一栋房子、一辆车、一个人或一只宠物。首先，我们要收集大量包含有房子、车、人或宠物的图片，组成一个数据集（data set），每张图片都标记有它的类别。在训练时，每当我们向机器展示一张图片，机器就会输出一个相应类别的向量。我们希望的结果是：指定类别的分数最高，高于其他所有类别。然而，如果不经过训练，这将是不可能完成的任务。为此，我们通过一个目标函数来计算实际输出与期望输出之间的误差或距离。接下来，为了减小误差，机器会对其内部可调参数进行调整。这些可调参数常被称为「权重」（weight），是实数，可看做定义机器输入-输出功能的「门把手」。在一个典型的深度学习系统中，可能存在着成千上亿的可调权重及用以训练机器的标记样本。

为了正确地调整权重矢量（ weight vector），学习算法会计算出一个梯度矢量（ gradient vector）。对每一个权重，这个梯度矢量都能指示出，当权重略微增减一点点时，误差会随之增减多少量。接着，权重矢量就会往梯度矢量的反方向进行调整。

从所有训练范例之上，平均看来，目标函数（ objective function）可被视为一片崎岖的山地，坐落于由权重组成的高维空间。梯度矢量为负值的地方，意味着山地中最陡峭的下坡方向，一路接近最小值。这个最小值，也就是平均输出误差最小之处。

在实践中，大多数业内人士都是用一种被称为「随机梯度下降」（SGD - Stochastic

Gradient Descent）的算法（梯度下降 Grident Descent 是「最小化风险函数」以及「损失函数」的一种常用方法，「随机梯度下降」是此类下的一种通过迭代求解的思路——译者注）。每一次迭代包括以下几个步骤：获取一些样本的输入矢量(input vector)，计算输出结果和误差，计算这些样本的平均梯度，根据平均梯度调整相应权重。这个过程在各个从整个训练集中抽取的小子集之上重复，直到目标函数的平均值停止下降。它被称做随机(Stochastic)是因为每个样本组都会给出一个对于整个训练集(training set)的平均梯度(average gradient)的噪音估值(noisy estimate)。较于更加精确的组合优化技术，这个简单的方法通常可以神奇地快速地找出一个权重适当的样本子集。训练过后，系统的性能将在另外一组不同样本（即测试集）上进行验证，以期测试机器的泛化能力（ generalization ability） ——面对训练中从未遇过的新输入，机器能够给出合理答案。

很多当今机器学习的实际应用都在人工设定的特征上使用「线性分类」(linear classifiers)。一个「二元线性分类器」(two-class linear classifier)可以计算出特征向量的「加权和」(weighted sum)。如果「加权和」高于阈值，该输入样本就被归类于某个特定的类别。

二十世纪六十年代以来，我们就知道线性分类只能将输入样本划分到非常简单的区域中，即被超平面切分的半空间。但是，对于类似图像及语音识别等问题，要求「输入-输出函数」(input–output function)必须对输入样本的无关变化不敏感，比如，图片中物体的位置，方向或者物体上的装饰图案，又比如，声音的音调或者口音；与此同时「输入-输出函数」又需要对某些细微差异特别敏感（比如，一匹白色的狼和一种长得很像狼

的被称作萨摩耶的狗）。两只萨摩耶在不同的环境里摆着不同姿势的照片从像素级别来说很可能会非常地不一样，然而在类似背景下摆着同样姿势的一只萨摩耶和一只狼的照片在像素级别来说很可能会非常相像。一个「线性分类器」(linear classifier)，或者其他基于原始像素操作的「浅层（shallow）」分类操作是无论如何也无法将后者中的两只区分开，也无法将前者中的两只分到同样的类别里的。这也就是为什么「浅层」「分类器(classifiers)」需要一个可以出色地解决「选择性-恒常性困境」（selectivity–invariance dilemma)的「特征提取器」(feature extractor)—— 提取出对于辨别图片内容有意义的信息，同时忽略不相关的信息，比如，动物的姿势。我们可以用一些常规的非线性特征来增强「分类器」(classifiers)的效果，比如「核方法」(kernel methods)，但是，这些常规特征，比如「高斯核」(Gaussian Kernel)所找出来的那些，很难泛化（generalize）到与训练集差别别较大的输入上。传统的方法是人工设计好的「特征提取器」，这需要相当的工程技巧和问题领域的专业知识。但是，如果好的「特征提取器」可以通过「通用学习程序(General-Purpose learning procedure)」完成自学习，那么这些麻烦事儿就可以被避免了。这就是深度学习的重要优势。

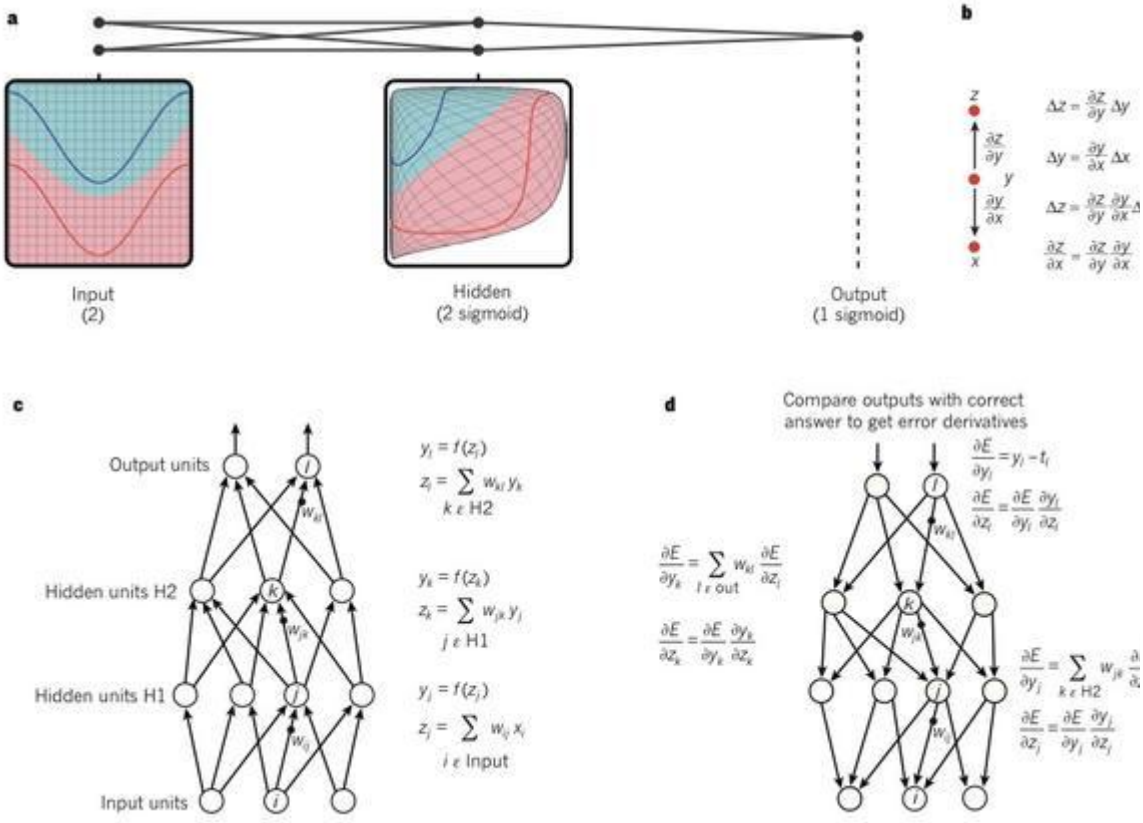

图 1| 多层神经网路和反向传播。

a. 一个多层神经网络（如图所示相互连接的点）能够整合（distort）输入空间（图中以红线与蓝线为例）让数据变得线性可分。注意输入空间的规则网格（左侧）如何转被隐藏单元（中间）转换的。例子只有两个输入单元、两个隐藏单元和一个输出单元，但事实上，用于对象识别和自然语言处理的网络通常包含了数十或成千上万个单元。(本节引用改写自 C. Olah (http://colah.github.io/).)

b. 导数的链式法则告诉我们，两个微小增量（即 x 关于 y 的增量，以及 y 关于 z 的增量）是如何构成的。x 的增量Δx 导致了 y 的增量Δy，这是通过乘以∂y/∂x 来实现的（即偏导数的定义）。同样，Δy 的变化也会引起Δz 的变化。用一个方程代替另一个方

程引出了导数的链式法则（ the chain rule of derivatives），即增量Δx 如何通过与∂y/∂x 及 ∂z/∂x 相乘使得 z 也发生增量Δz。当 x,y 和 z 都是向量时这一规律也同样适用（使用雅克比矩阵）。

c. 这个公式用于计算在包含着两个隐层和一个输出层的神经网络中的前向传输，每个层面的逆向传递梯度都构成了一个模组。在每一层，我们首先计算面向每个单元的总输入值 z，即上一层的输出单元的加权和；然后，通过将一个非线性函数 f（.）应用于 z 来得出这个单元的输出。为了简化流程，我们忽略掉一些阈值项（bias terms）。在神经网络中使用的非线性函数包含了近些年较为常用的校正线性单元(ReLU) f(z) = max(0,z) ,以及更传统的 sigmoid 函数 ,比如 ,双曲线正切函数, f(z) = (exp(z) − exp(−z))/(exp(z) + exp(−z)) 和 逻辑函数 f(z) = 1/(1 + exp(−z)).

d. 该公式用于计算反向传递。在每一个隐藏层中，我们都会计算每个单元输出的导数误差，即上述层中上一层所有单元输入的导数误差的加权总和。 然后，将关于输出的导数误差乘以函数 f(z)的梯度（gradient），得到关于输入的导数误差。 在输出层中，通过对成本函数进行微分计算，求得关于输出单元的误差导数。因此我们得出结论 yl - tl 如果对应于单元 l 的成本函数是 0.5（yl - tl）2 （注意 tl 是目标值）。一旦∂E/∂zk 已知，那么，就能通过 yj ∂E/∂zk 调整单元 j 的内星权向量 wjk。

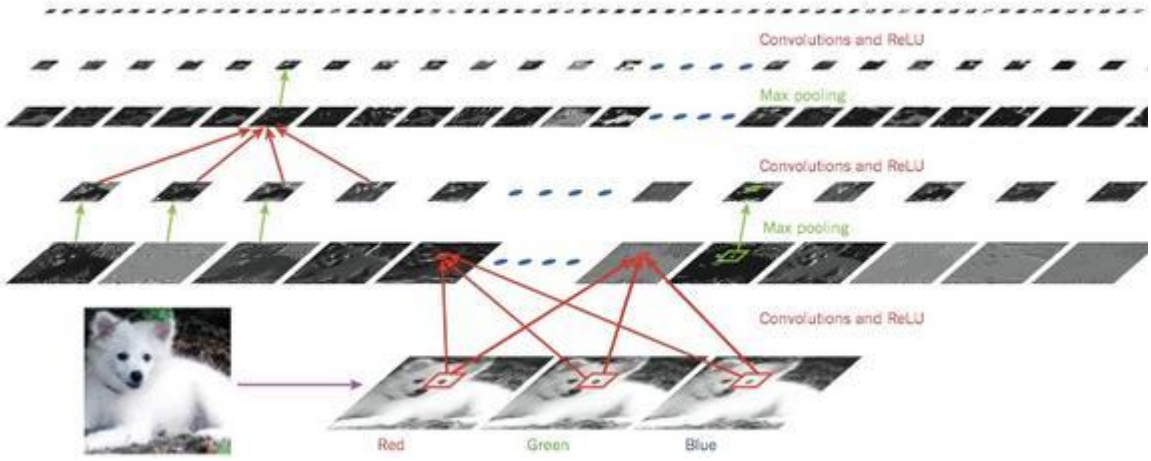

图 2 | 卷积网络的内部。一个典型的卷积网络架构的每一层（水平）输出（不是过滤器）应用到一个萨摩耶犬的图像（图 2 左下方，RGB 输入（红绿蓝），下方右侧）。每一个长方形图片都是一张对应于学习到的输出特征的特征地图，这些特征可以在图片的各个位置被侦测到。信息流是从下往上的，低级的特征充作导向性边际检测因子（ edge detectors），每个输出图像类都会被计算出一个分值。ReLU，整流线性单元。

深度学习架构由简单模组多层堆叠而成，全部（或绝大部分）模组都从事学习，大部分会计算非线性的输入输出映射。堆叠架构中的每个模组都会转换它的输入，同时增强可选择性和所呈现状态的一致性。 通过多个非线性层面（例如，深度 5-20 层），系统可以实现对于输入数据的极其微小细节的敏感性功能 --- 例如，区别开白色的狼和萨摩耶犬 --- 并且对于较大的无关变量会不敏感（例如，主体周围的环境、姿势、光照条件和周围物体。）

训练多层架构的反向传播算法

从最早的模式识别开始，研究者们就一直试图用可训练的多层网络代替人工设计特征，尽管这种解决方案很简单，直到 20 世纪 80 年代中期，它才得到人们的广泛认可。

事实证明，多层架构可以通过简单的随机梯度下降法得以训练。只要模组是由它们的输入值及其内部权重构成的相对平滑的函数（relatively smooth functions），人们就可以使用反向传播算法来计算梯度。20 世纪 70 至 80 年代，几个不同的研究小组分别发现这一方法确实可行、有用。

计算一个目标函数关于一个多层堆叠模组的权重梯度的反向传播过程，只不过是导数链式法则的一个实际应用。其中关键之处在于，关于模组输入值的函数的导数（或梯度），可以借助关于该模组的输出值（或序列模组的输入值）的梯度反向计算出来（图 1）。反向传播方程可在所有模组中传播梯度时被反复应用，从顶部（top)（神经网络产生预测的那一层）输出开始，一直到底部（bottom）（被接收外部输入的那一层）。这些梯度一经计算，就可直接计算出关于每个模组权重的梯度。

深度学习的许多应用 ,都会使用前馈神经网络架构(feedforward neural network architectures）（图 1）——该架构可学习将一个固定大小的输入映射到（例如，一幅图像）到一个固定大小的输出上（例如，每种分类的概率）。从一层到下一层，单元组计算其前一层输入的加权总和，并通过一个非线性函数输出结果。目前，最受欢迎的非线性函数是整流线性单元（ReLU），一个简单的半波整流器 f(z) = max(z, 0)。在过去的几十年里，神经网络使用的是更平滑的非线性，比如，tanh（z）或 1 /（1 + exp（−z））,但 ReLU 在多层网络中的学习速度通常更快 ,可允许在没有无监督预训练(without unsupervised pre-training)的情况下进行深度监督网络训练。不在输入或输出层中的单元通常被称为隐层单元（hidden units）。隐层可被看作是以非线性方式变换输入，从而使所有类别在最后一层变得线性可分（ linearly separable by the last layer）（图

1）。

20 世纪 90 年代末，神经网络和反向传播被机器学习社区大量遗弃，同时也被计算机视觉和语音识别领域忽略。人们普遍认为，学习有用的、多层级的、几乎不靠先验知识的特征提取器并不现实可行。尤其是，人们通常认为简单的梯度下降法会深陷局部极小的泥潭——在这种权重配置当中，除非进行大的改动，否则很难降低平均误差。

实践中，对大型网络而言，局部极小几乎不构成问题。无论初始条件如何，系统基本总能得到质量非常相似的解决方案。最近的理论和实证研究结果均有力地表明，总的来说，局部极小不是一个严重问题。相反，解空间（ landscape）充满了大量梯度为 0 的鞍点（ saddle points ），且在多数维度中表面向上弯曲，少数维度中表面向下弯曲。分析结果似乎表明，向下弯曲的鞍点在整体中占比相对较小，但这些鞍点的目标函数值大多相近。因此，算法陷入这些鞍点（不能继续寻优），无关紧要。

2006 年前后，加拿大高级研究所（CIFAR）聚集了一批研究人员，他们重燃了人们对深度前馈网络的兴趣。这些研究人员引入无监督学习程序——无需标记数据便可创建特征检测器层。各层特征检测器的学习目标便是在下一层重构或模拟特征检测器（或原始输入）的活动。利用这种重构学习目标来「预训练（pre-training）」几层复杂度递增的特征检测器，深层网络的权重可以被初始化为合理值。接着，最终层的输出单元可被添加到网络顶端，整个深度系统可被微调至使用标准的反向传播。在识别手写数字或检测行人时，特别是当标记的数据量非常有限的时候，这一程序非常有效。

这种预训练的方法的首次重要应用是在语音识别上，这之所以可行归功于便于编程的 GPUs 的诞生，它让研究人员可以用 10 到 20 倍的速度训练神经网络。2009 年，这个方法被用来计算一段声音采样中提取短时系数窗口对应的一系列概率值，这些概率值反映出由窗口中帧表示语音各个段落的可能性。在小词汇表的标准语音识别测试上，这种方法的训练效果打破纪录，很快它又发展到打破大词汇表的标准语音测试纪录。

到 2012 年 ,2009 年以来的各种深度网络一直的得到多个主要语音研究小组持续开发并被布局在安卓手机上。对于较小数据集来说，无监督预训练有助于防止过拟合（ overfitting），当标注数据样本小（ number of labelled examples is small）或需要迁移（ in a transfer setting ）——有很多源领域的标注数据样本但缺少目标领域的标注数据样本——的时候，深度网络的泛化（ generalization）效果显著提升。深度学习重新获得认识，人们发现，预训练阶段只是小规模数据集的必需品。

然而，还有一种特殊类型的深度前馈网络（ deep feedforward network），不仅更易训练而且泛化能力要比那些相邻两层完全相连的神经网络强大很多。这就是卷积神经网络 （ConvNet）。 在神经网络「失宠 」的日子里，卷积神经网络在实践运用中获得许多成功，最近已被计算机视觉领域广泛采用。

卷积神经网络

卷积神经网络最初是用来处理多维数组数据，比如，一张由三个 2D 数组组成、包含三个彩色通道像素强度的彩色图像。大量的数据模式都是多个数组形式：1D 用来表示信号和序列信号包括人类语言；2D 用来表示图片或声音；3D 代表视频或有声音的图

像。卷积神经网络利用自然信号特征的核心理念是：局部连接（local connections），权重共享，池化(pooling）和多网络层的使用。

典型的卷积神经网络的架构（图二）包括一系列阶段：最初的几个阶段由卷积层和池化层组成，卷积层的单元被组织到特征图（feature map）中，每个单元通过一组被称作滤波器（filter bank）的权值被连接到前一层的特征图的局部数据块。接下来，得到的局部加权和会传递至一个非线性函数，例如 ReLU。同一个特征图中的所有单元共享相同的滤波器，不同特征图使用不同滤波器。采用这种架构有两方面的原因。首先，在诸如图像这样的数组数据中，数值与附近数值之间通常是高度相关的，容易生成易被探测到的局部特征（motif）。其次，图像和其他类似信号的局部统计特征通常又与位置无关，易言之，出现在某处的某个特征也可能出现在其他任何地方，因此，不同位置的单元会共享同样的权值并且可以探测相同模式。数学上，由一个特征图完成的过滤操作是一个离线的卷积，卷积神经网络由此得名。

和卷积层用来探测前一层中特征之间的局部连接不同，池化层的作用则是对语义相似的特征进行合并。由于构成局部主题的特征之间的相对位置关系不是一成不变的，可以通过粗粒度检测每个特征的位置来实现较可靠的主题识别。一个池化层单元通常会计算一个或几个特征图中一个局部块的最大值，相邻的池化单元则会移动一列或一行从小块读取输入，这种设计不仅减少了数据表征需要的维数，而且也能对数据小规模的偏移、扭曲保持不变。两到三个卷积层，非线性层和池化层被叠加起来，后面再加上更多的卷积和全连接层。在卷积神经网络的反向传播算法和在一般深度网络上一样简单，能让所有滤波器中的权值得到训练。

多数自然信号都是分级组合而成，通过对较低层信号组合能够获得较高层的信号特征，而深度神经网络充分利用了上述特性。在图像中，线条组合形成图案，图案形成部件，部件组成物体。类似的层次结构存在于由声音到电话中的语音及文本形成过程，音素组成音节，音节组成单词，单词组成句子。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变化具有鲁棒性。

卷积神经网络中的层次的卷积和汇聚的灵感直接来源于视觉神经科学中的简单细胞和复杂细胞的经典概念，并且其整体架构让人想起视觉皮层腹侧通路的 LGN-V1-V2-V4-IT 层次结构。当向卷积神经网络模型和猴子同时展示相同的画面时，卷积神经网络的高级单元被激活，解释了猴子颞下皮层随机设置的 160 个神经元的变化。卷积神经网络有着神经认知机的基础，两者的体系结构有些类似，但是，卷积神经网络没有诸如反向传播的那种端对端的监督学习算法。原始的 1D 卷积神经网络被称为「延时神经网络（time-delay neural net ）」，用于识别语音和简单的单词。

早在 20 世纪 90 年代初，卷积网络就已有非常广泛的应用 ，最开始延时神经网络被用在语音识别和文档阅读上。文本阅读系统使用了受过训练的延时神经网络以及一个实现了语言约束的概率模型。到 20 世纪 90 年代末，该系统能够读取美国超过十分之一的支票。随后，微软发明了许多基于卷积神经网络的光学字符识别和手写识别系统。卷积神经网络在 20 世纪 90 年代初就被尝试用于包括脸、手、面部识别的自然图像目标检测中。

使用深层卷积网络进行图像识别

从 21 世纪初开始，卷积神经网络就被成功用于检测、分割和物体识别以及图像各区域。这些应用都使用了丰富的标签数据，比如，交通标志识别、生物图像（特别是神经链接组学方面）分割、面部探测、文本、行人和自然图像中的人体的检测。近些年来，卷积神经网络的一项重要成功应用就是人脸识别。

值得注意的是，图像可以在像素级别上被标记，这样就能被用于诸如自主移动机器人（autonomous mobile robots）和无人驾驶汽车等技术中。像 Mobileye 和 NVIDIA 这些公司正在将这些基于卷积神经网络的方法应用于即将面世的汽车视觉系统中。其他重要的应用程序涉及到自然语言理解和语音识别。

尽管取得了这些成就，但在 2012 年 ImageNet 竞争之前，卷积神经网络在很大程度上并未获得主流计算机视觉和机器学习团体的青睐。当深层卷积网络被应用于来源于包含 1000 个不同类型约 100 万个图像的数据集中，它们取得了惊人的成果，错误率仅是当时最佳方法的一半。该成功源于高效利用了 GPUs 和 ReLUs、一项新的被称为「dropout 」的正规化技术（ regularization technique ）以及分解现有样本产生更多训练样本的技术。成功给计算机视觉领域带来一场革命。如今，卷积神经网络几乎覆盖所有识别和探测任务，在有些任务中，其表现接近人类水平。最近一个令人震惊的例子，利用卷积神经网络结合递归网络模块来生成图像标题（ image captions ）（如图 3）。

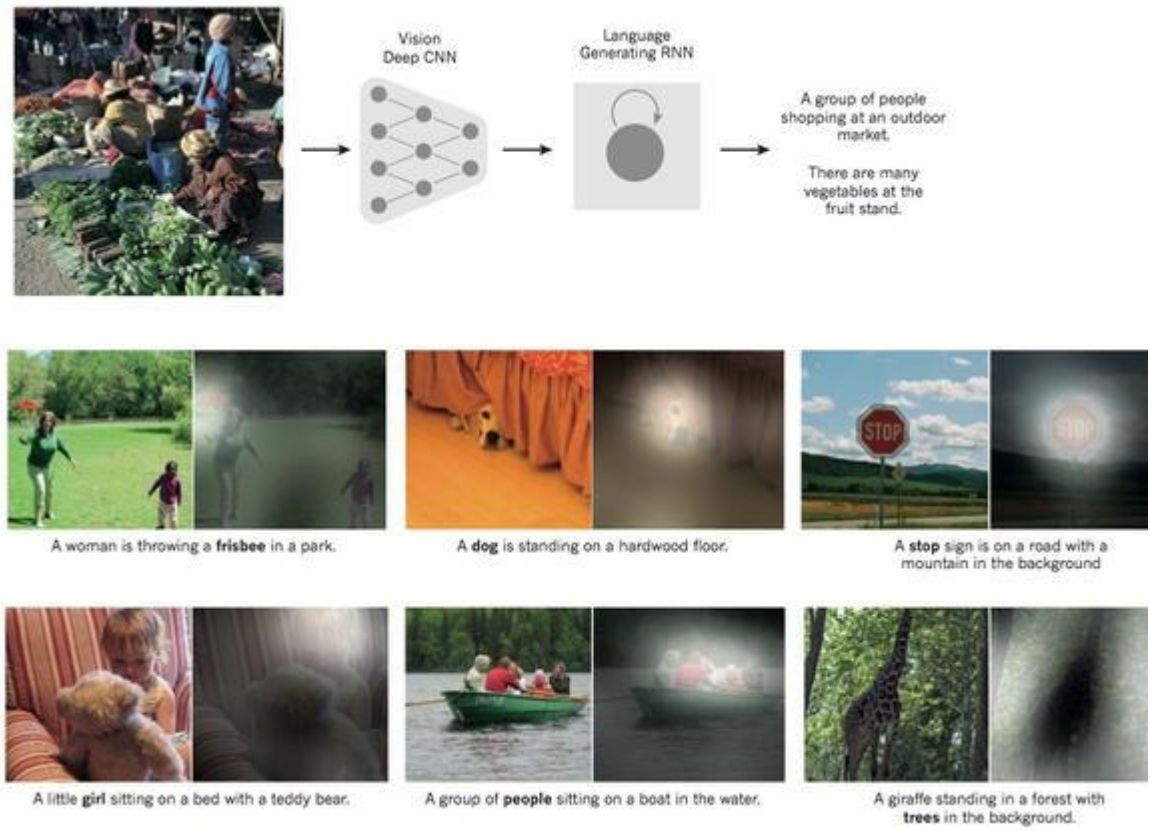

图 3 |从图像到文本。将递归神经网络（RNN）生成的标题作为额外输入，深度卷积神经网络（CNN)会从测试图片中提取表征，再利用训练好的 RNN 将图像中高级（ high-level ）表征「翻译成 」标题（上图）。当 RNN 一边生成单词（黑体所示），一边能将注意力集中在输入图像的不同位置（中间和底部；块状越亮，给予的注意力越多）的时候，我们发现，它能更好地将图像「翻译成」标题。

当前的卷积神经网络由 10~20 层 ReLUs，数百万个权值及数十亿个连接组成。两年前，训练如此庞大的网络可能需要数周时间，而随着硬件、软件和算法并行化（algorithm parallelization）的进步，训练时间已经缩短至几个小时。

卷积神经网络的视觉系统良好表现促使包括谷歌、Facebook、微软、IBM、雅虎、

推特和 Adobe 在内的多数主要科技公司以及数量激增的创业公司开始启动研发项目，部署基于卷积神经网络的图像识别产品和服务。

卷积神经网络易于在芯片或现场可编程门列阵（FPGA）中得以高效实现。为了实现智能手机、相机、机器人和无人驾驶汽车上的实时视觉应用，NVIDIA、Mobileye、因特尔、高通和三星等许多公司都正在开发卷积神经网络芯片。

分布式表征和语言处理

深度学习理论显示，与不适用分布式表征的经典学习算法相比，深度网络有两处异常明显的优势。这些优势源于节点权重（the power of composition）以及底层数据生成分布具有适当的组成结构。第一，学习分布式表征能够将通过训练而学习获得的特性值泛化为新的组合（例如，n 元特征有 2n 组合可能）。第二，深度网络中的表征层相互组合带来了另一个指数级优势的潜力（指数性的深度）。

多层神经网络的隐藏层学会以一种易于预测目标输出的方式来再现网络输入。一个很好的示范就是训练多层神经网络根据局部文本中的前述语句预测下一个词。文本的每个词表示成网络中的 N 分之一向量，也就是说，每个成分的值为 1，余下的为 0。在第一层中，每个字创建一个不同模式的激活或单词向量（如图 4 所示）。在语言模型中，网络中的其他层学习如何将输入的单词向量转化成输出单词向量来预测下一个单词，也能用来预测词汇表中单词作为文本中下一个单词出现的概率。正如学习分布表征符号文本最初展示的那样，网络学习了包含许多激活节点（active components ）、且每一个节点都可被解释成一个单词独立特征的单词向量。这些语义学特征并没有在输入时被

清晰表现出来。而是在学习过程中被发现的，并被作为将输入与输出符号结构化关系分解为微规则（micro-rules）的好方法。当词序列来自一个大的真实文本语料库，单个微规则并不可靠时，学习单词向量也一样表现良好。当网络被训练用于预测新文本中的下一个词时，一些单词向量非常相似，比如 Tuesday 和 Wednesday，Sweden 和 Norway 。这种表征被称为分布式表征，因为它们的元素（特性）并非相互排斥，且它们构造信息与观测到的数据变化相对应。这些单词向量由所习得的特性组成，这些特性并非由科学家们事先决定而是由神经网络自动发现。现在，从文本中习得的单词向量表征被非常广泛地使用于自然语言应用。

表征问题是逻辑启发与神经网络启发认知范式争论的核心问题。在逻辑启发范式中，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。它并不包含与使用相关的内部结构，而且为理解符号含义，就必须与审慎选取的推理规则的变化相联系。相比之下，神经网络使用大量活动载体（ big activity vectors）、权重矩阵和标量非线性，实现一种快速「直觉 」推断，它是轻松常识推理的基础。

在介绍神经语言模型前，语言统计模型的标准方法并没有使用分布式表征：它是基于计算短符号序列长度 N（称为 N-grams，N 元文法）出现的频率。N-grams 可能出现的次数与 VN 一致，这里的 V 指的是词汇量的大小，考虑到词汇量大的文本，因此需要更庞大的一个语料库。N-grams 把每一个词作为一个原子单位，因此它不能在语义紧密相关的单词序列中，一概而论，但是，神经语言模型可以实现上述功能，因为它们将每个单词与真实特征值的向量关联起来，并且语义相关的单词在该向量空间中更为贴近。（如图 4）。

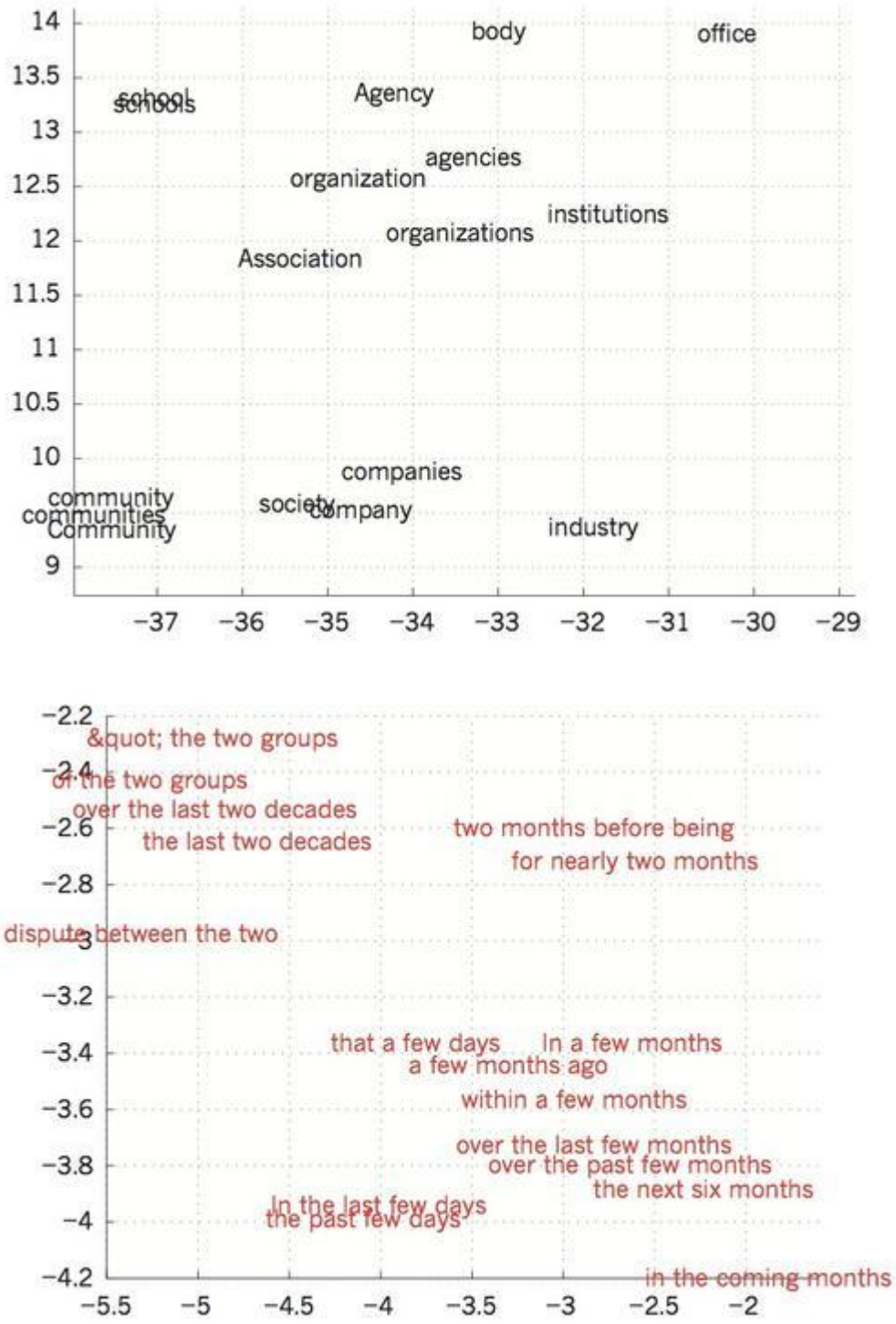

图 4|已完成学习的单词向量的可视化展现。左边介绍了为了建模语言而习得的词汇表征，通过使用 t-SNE 算法[103]非线性映射至二维空间中以便于观察。右边是一个由实现英-法互翻的递归神经网络学习短语的二维空间表示。由图可知，语义或排序相似的单词表征映射较为接近 。词汇的分布式表征通过使用反向传播获得，以此来学习每个单

词的表征形式及预测目标数量的功能，比如序列中的后续单词（如语言建模）或者翻译文字的全部序列（机器翻译）。

递归神经网络

最初引入反向传播时，最令人激动的应用便是训练递归神经网络（简称 RNNs）。对于那些需要序列连续输入的任务（比如，语音和语言），RNNs 是上乘之选（图 5）。RNNs 一次处理一个输入序列元素，同时维护隐式单元中隐含着该序列过去所有元素的历史信息的「状态向量」。当我们考虑隐式单元在不同的离散时间步长的输出，就好像它们是在多层网络深处的不同神经元的输出（图五，右）如何利用反向传播训练 RNNs，一目了然。

RNNs 是非常强大的动力系统，但训练它们也被证实存在一些问题，因为反向传播梯度在每个时间间隔内或增长或下降，因此，一段时间之后通常会导致结果激增或者降为零。

因先进的架构和训练的方式，RNNs 不仅被证实擅长预测文本中下一个字符或句子中下一个单词，还可应用于更加复杂的任务。例如，某时刻阅读英文句子中的单词后，一个英语的「编码器」网络将被生成，从而帮助隐式单元的最终状态向量很好地表征句子所传达的思想。这种「思想向量（thought vector）」可以作为一个集大成的法语「编码器」网络的初始化隐式状态（或额外的输入），其输出为法语翻译首单词的概率分布。如果从概率分布中选择一个特定首单词作为编码网络的输入，将会输出翻译句子中第二个单词的概率分布，依此类推，直到停止选择为止。总体而言，这一过程是根据英语句

子的概率分布而生成的法语单词序列。这种近乎直接的机器翻译方法的表现很快和最先进（state-of-the-art）的方法不相上下，同时引发人们对于理解句子是否需要使用推理发掘内部符号表示质疑。这与日常推理中涉及到根据合理结论类推的观点是匹配的。

除了将法语句子翻译成英语句子，还可以学习将图片内容「翻译」为英语句子（如图 3）。编码器是一种在最后隐层将像素转换为活动向量的深度卷积网络。解码器是一种类似机器翻译和神经网络语言模型的递归神经网络。近年来，引发了人们对深度学习该领域的热议。RNNs 一旦展开（如图 5），可被视作是所有层共享同样权值的深度前馈神经网络。虽然它们的主要目的是长期学习的依赖性，但有关理论和经验的例证表明很难学习并长期储存信息。

为了解决这一问题，一个扩展网络存储的想法出现。第一种方案是采用了特殊隐式单元的 LSTM，该自然行为便是长期的保存输入。一种类似累加器和门控神经元的称作记忆细胞的特殊单元：它通过在下一个时间步长拥有一个权值并联接到自身，从而拷贝自身状态的真实值和累积外部信号，但这种自联接是另一个学习并决定何时清除记忆内容的单元的乘法门所操控。

LSTM 网络最终被证明比传统的递归神经网络（RNNs）更为有效，尤其是，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地将声学转录为字符序列。目前，LSTM 网络及其相关形式的门控单元同样也用于编码与解码网络，并在机器翻译中表现良好。

过去几年里，几位学者提出一些不同的方案来增强 RNNs 存储器模块。这些建议包括，神经图灵机——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。

除了简单记忆化、神经图灵机和记忆网络被用于通常需要推理和符号操作的任务以外，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。

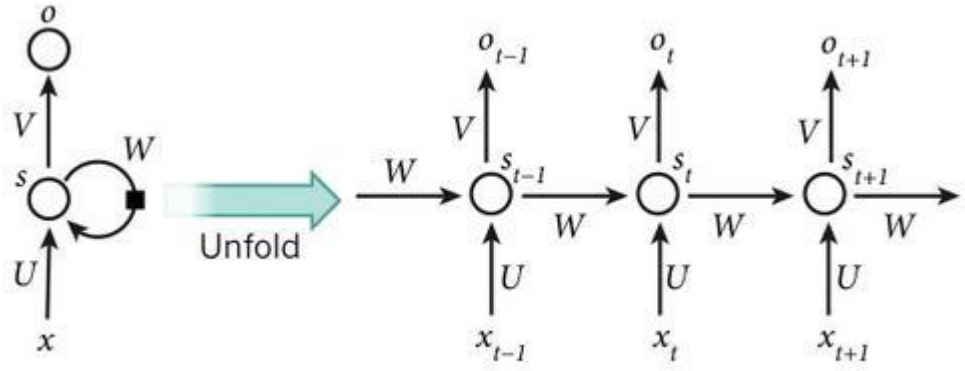

图 5 |一个递归神经网络在时间中展开的计算和涉及的相关计算。人工神经元(例如，隐式样单元分组节点在时间 t 的标准值下）获得其他神经元的输入——在之前的步骤中(黑色区域呈现，代表一步延迟，如左)。这样，一个递归神经网络可由 xt 的输入序列元素，映射到一个输出序列与元素 ot，每次 ot 值取决于所有前面的 xt´（t´≤t）。相同的参数(U,V 矩阵 W)在每步中使用。许多其他结构是可行的，包括一个变体的网络可以生成的输出序列(例如，词语)，每一个都作为下次的输入步骤。反向传播算法(图 1)可以直

接应用于计算机图形展开网络，并对所有的标准陈述和参数，计算其总误差的导数(例如，生成正确的输出序列的对数概率)。

深度学习的未来

无监督学习促进了人们重燃对深度学习的兴趣，但是，有监督学习的成功盖过了无监督学习。虽然我们没有关注这方面的评论，但是，从长远来看，我们还是期望无监督学习能够变得更加重要。（因为）人类和动物的学习方式大多为无监督学习：我们通过观察世界来发现它的结果，而不是被告知每个对象的名称。

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。我们希望机器视觉能够在未来获得巨大进步，这些进步来自于那些端对端的训练系统，并集合卷积神经网络（ConvNets）和递归神经网络（RNNs），利用强化学习来决定走向。结合了深度学习和强化学习的系统尚处在婴儿期，但是，在分类任务上，它们已经超越了被动视觉系统，并在尝试学习操作视频游戏方面，产生了令人印象深刻的结果。

未来几年，理解自然语言会是深度学习产生巨大影响的另一个领域。我们预测，当它们学习了某时刻选择性地加入某部分的策略，那些使用递归神经网络（RNNs）的系统将会更好地理解句子或整个文档。

最终，人工智能的重大进步将来自将表征学习与复杂推理结合起来的系统。尽管深度学习和简单推理已经用于语音和手写识别很长一段时间了，我们仍需要通过大量向量

操作的新范式替换基于规则的字符表达操作。