深度学习的昨天、今天和明天 *

作者：约书亚·本吉奥（Yoshua Bengio）
雅恩·乐昆（Yann LeCun）
杰弗里·辛顿（Geoffrey Hinton）
译者：马卓奇

关键词：深度学习

译者按：2018年，美国计算机学会（ACM）决定将计算机领域的最高奖项图灵奖颁给约书亚·本吉奥（Yoshua Bengio）、雅恩·乐昆（Yann LeCun）和杰弗里·辛顿（Geoffrey Hinton），以表彰他们在计算机深度学习领域的突出贡献。今年，三位获奖者再次受ACM邀请共聚一堂，共同回顾了深度学习的基本概念和一些突破性成果，讲述了深度学习的起源、发展及未来面临的挑战。

人工神经网络的研究起源于科学家的一个观察：人类智能是从相对简单的非线性神经元组成的高度并行的网络中产生的，这些神经元通过调整连接的强度来进行学习。这一观察结果引出了一个核心的计算问题：这类网络如何学习复杂任务（如目标识别或语言理解）所需的丰富特征表示？

对于这个问题，深度学习给出的答案是利用多层激活向量作为表示，并通过跟踪衡量网络性能的目标函数的随机梯度，来学习能够产生这些向量的连接强度。令人惊讶的是，概念上如此简单的方法在大型训练集上却是非常有效的，看来关键因素是“深度”，因为浅层网络根本不起作用。

我们回顾了深度学习的基本概念和几年前的一些突破性成果[63]。这里我们简要介绍深度学习的起源，描述一些近期进展，并讨论一些未来的挑战。这些挑战包括无监督学习，未见类别的测试样本，以及使用深度学习方法来解决人类需要通过一系列逻辑性的、有意识的步骤才能解决的任务，即著名心理学家丹尼尔·卡内曼（Daniel Kahneman）[56]命名为“系统2”的任务，与之对应的是一种“不费力”的“系统1”任务，比如目标识别或即时自然语言理解。

从手工编码的符号表达式到学习的分布式表示

人工智能有两种完全不同的范式。简单地说，逻辑启发范式（logic-inspired paradigm）将顺序推理视为智能的本质，其目的是使用人为设计的推理规则在计算机中实现推理，这些规则利用手工设计的符号表达式将知识形式化。大脑启发范式（brain-inspired paradigm）认为从数据中学习表示是智能的本质，通过手工设计或进化规则修改人工神经元模拟网络中的连接强度来实现学习。

在逻辑启发范式中，符号的内部结构没有意义，它的意义存在于它与其他符号的关系中，这种关系可以用一组符号表达式或者关系图来表示。相比之

* 本文译自 *Communications of the ACM*，“Deep Learning for AI”，2021, 64 (7): 58-61 一文。

下，在大脑启发范式中，用于交流的外部符号被转化为代表神经活动的内部向量，这些向量具有丰富的相似性结构。通过学习每个符号的活动向量以及学习非线性变换，以便填充与符号串缺失元素对应的活动向量，可以对一组符号串的固有结构建模。鲁姆哈特（Rumelhart）等人[74]在小型数据集上首先验证了这一点，后来本吉奥等人[14]在真实的语句上验证了这一结论。最近的一个令人印象深刻的模型是BERT[22]，它也利用自注意力机制来动态连接一组单元。

使用神经活动向量来表示概念以及使用权重矩阵来捕获概念之间的关系，其主要优点是能够带来自动泛化。如果“星期二”和“星期四”由非常相似的向量表示，它们将对神经活动的其他向量产生非常相似的因果影响。这有助于类比推理，并表明直接、直观的类比推理是我们的主要推理模式，而逻辑顺序推理是后期的发展[56]。

深度学习的兴起

在21世纪初，深度学习引入的一些元素使训练深层网络变得更加容易，为神经网络研究注入了新的活力。GPU和大型可用数据集的出现是深度学习的关键推动因素，而具有自动求导功能的开源、灵活的深度学习平台（如Theano[16]、Torch[25]、Caffe[55]、TensorFlow[1]、PyTorch[71]）也在很大程度上推动了深度学习的发展，这一发展也使得训练复杂的深度网络、重新使用最新模型及其构建模块变得更加容易。而更多层网络的组合能够允许更复杂的非线性，在感知任务中取得了意想不到的良好结果。

深度学习深在哪里? 尽管神经网络越深就越强大这种观念[82]在现代深度学习技术出现之前就存在，但正是结构和训练策略的不断进步[15, 35, 48]，带来了与深度学习兴起相关的显著进步。但是，为什么越深层次的网络对于我们建模时所关注的输入-输出关系的泛化性越好呢? 要明白，这不仅仅是参数变多了的问题，因为在同样参数量的情况下，深度网络通常比浅层网络的泛化能力更强[15]。事实也证明了这一点，计算机视觉中最流行的卷积网络体系结构是ResNet[43]，其中最常见的代表性网络ResNet-50有50层。有一些在本文中没有提到但也非常有用的元素包括图像变形、drop-out[51]和批归一化[53]。

我们认为，深度网络的优势在于它们利用了一种特殊的组合形式，即一层中的特征以多种不同的方式组合，从而在下一层中创建更为抽象的特征。

对于类似感知的任务，这种组合非常有效，并且有强有力的证据表明，生物感知系统也使用了这种组合[83]。

无监督预训练。当训练样本的标注数据较少，而实现任务所需的神经网络的复杂性相对较高时，可以使用其他信息源来创建特征提取层，然后利用有限的标注数据对特征检测网络进行微调。在迁移学习中，信息的来源是另一个具有大量标注的有监督学习任务。但是，通过堆叠自动编码器，也可以在不使用任何标注数据的情况下创建特征提取层[15, 50, 59]。

首先，我们学习一个特征提取层，它能够允许重构输入。然后学习特征提取器的第二层，它能够允许重构第一层特征提取器的活动。在用这种方法学习几个隐藏层后，我们尝试从最后一个隐藏层的响应中预测标签，并将误差反向传播到所有层，以便在不使用标注信息的情况下对最初得到的特征提取器进行微调。预训练也许会提取出与最终分类无关的各种数据结构，但是在计算成本较低而数据标注成本较高的情况下，只要预训练能够将输入转换为更容易分类的特征表示，就没有问题。

除了提高泛化能力外，无监督预训练初始化权重后，即可更容易地用反向传播对深度网络进行微调。过去，预训练对优化的影响最主要的作用是颠覆了深度网络很难训练这一观点，但是现在由于人们使用线性整流函数（Rectified Linear Unit，ReLU）和残差连接[43]，预训练能带来的影响也没有那么大了。然而，预训练对泛化能力的影响已被证明是非常重要的。它可以利用大量未标注的数据来训练非常大的模型，例如在自然语言处理中，有大量的数据可供使用[26, 32]。在迁移学习中，预训练和微调这一通用原则已经成为深度学习中的一个重要工具，甚至已经成为现代元学习的一个组成部分[33]。

线性整流函数的成功之谜。深度网络的早期成功，是因为使用了逻辑 sigmoid 非线性函数或与之密切相关的双曲正切函数，对隐藏层单元进行无监督预训练。长期以来，神经科学领域一直设想线性整流单元的概念，并已经在受限玻尔兹曼机的变体和卷积神经网络中应用。研究人员发现，非线性整流（现在被称为 ReLUs, 有许多现代变体）可以更加方便地通过反向传播和随机梯度下降训练深度神经网络，而无须逐层预训练。这是让深度学习优于以往的目标识别方法的技术进步之一。

语音和物体识别的突破进展。声学模型将声波的表示转化为音素片段的概率分布。罗宾森（Robinson）[72] 和摩根（Morgan）等人 [69] 分别使用晶片机和数字信号处理技术（DSP）芯片证明，如果有足够的处理能力，神经网络在声学建模方面可以与最先进的技术相媲美。2009 年，两名研究生 [68] 使用 Nvidia GPU，证明了预训练的深度神经网络在 TIMIT 数据集上的表现略优于 SOTA 方法。这一结果重新激发了几个主要的语音识别研究小组对神经网络的兴趣。2010 年，本质上相同的深度网络在不需要特定语音训练的情况下，在大规模语音识别方面击败了 SOTA 方法 [28, 46]。2012 年，谷歌设计了一个生产版本，显著改善了安卓系统（Android）上的语音搜索。这是深度学习颠覆性力量的早期证明。

与此同时，深度学习在 2012 年 ImageNet 竞赛中取得了令人瞩目的胜利，在自然图像中识别一千种不同类别物体的错误率几乎减半 [60]。这场胜利的关键在于李飞飞团队为训练集收集了超过了一百万张有标记的图像 [31] 所做出的重大努力，以及艾利克斯·克里泽夫斯基（Alex Krizhevsky）对多个 GPU 的高效使用。当前的硬件（包括 GPU）鼓励使用大的批尺寸，以便将从内存中获取权重的成本分摊到该权重的多次使用中。纯粹的在线随机梯度下降法的收敛速度更快，因为每个权重仅用一次。未来的硬件可能直接在原地使用权重，而不是从内存中获取。

深层卷积神经网络包含了一些新的功能，例如使用 ReLU 加快学习速度，使用 dropout 防止过度拟合，但它基本上还是乐昆团队多年来一直研究的一种前馈卷积神经网络 [64, 65]。计算机视觉社区对这一突破的反应令人钦佩。鉴于卷积神经网络无可争辩的优越性，社区迅速放弃了以前的手工设计特征的方法，转而使用深度学习。

深度学习的最新进展

我们在本文选择性地讨论了深度学习的一些最新进展，还有许多重要的课题没有介绍，如深度强化学习、图神经网络和元学习。

软注意力和 Transformer 结构。深度学习的一个重要发展，特别是在序列处理方面，是乘法交互的使用，尤其是以软注意力的形式 [7, 32, 39, 78]。这是对神经网络的一个变革性补充，因为它将神经网络从纯粹的矢量变换机器转变为能够动态选择对哪些输入进行操作的架构，并且可以将信息存储在可微联想记忆中。这种架构的一个关键特性是，它们能有效地操作不同类型的数据结构，包括集合和图。

软注意力可以在一层的某一模块使用，以动态选择上一层传递的向量，进行组合和计算输出。这可以使输出独立于输入的呈现顺序（将它们视为一个集合），或者使用不同输入之间的关系（将它们视为一个图）。

Transformer 架构 [85] 已经成为许多应用中的主流架构，它堆叠了许多层“自注意力”（Self-attention 模块。同一层中的每个模块都使用一个标量积来计算其查询向量与该层中其他模块的键向量之间的匹配度。匹配度被归一化为总和为 1，然后使用得到的标量系数来形成上一层中其他模块产生的值向量的凸组合。得到的向量构成下一阶段计算模块的输入。模块可以是多头的，这样每个模块可以计算几个不同的查询、键和值向量，从而使每个模块可能有几个不同的输入，每个输入都以不同的方式从前一阶段的模块中进行选择。在这种操作中，模块的顺序和数量无关紧要，因此可以对一组向量进行操作，而不是像传统神经网络中那样对单个向量进行操作。例如，语言翻译系统在输出的句子中生成一个单词时，可以选择关注输入句子中相应的一组单词，而不考虑它们在文本中的位置。

尽管对于坐标变换 [44] 和循环网络 [52] 来说，乘积门限是早期就存在的想法，但它近期的形式使其成为主流。另一种看待注意力机制的方式是，它使人们能够通过适当选择的模块来动态传递信息，并且以新颖的方式组合这些模块，以提升数据分布外的泛化能力 [38]。

Transformer 已经取得了显著的性能提升，革命性地改变了自然语言处理的研究现状 [27, 32]，并且已经在工业中投入常规使用。这些系统都以自监督的方式进行预训练，以预测文本片段中的缺失单词。

也许更令人惊讶的是，Transformer 已经成功地应用于从符号角度求解积分和微分方程 [62]。最近一个非常热门的趋势是在卷积网络的基础上使用 Transformer，以取得在图像任务上的最优表现 [19]。Transformer 以求导的方式进行后处理和基于对象的推理，使系统能够进行端到端的训练。

无监督和自监督学习。监督学习虽然在各种任务中都取得了成功，但通常需要大量人工标注的数据。同样，当强化学习仅基于奖励时，它需要大量的交互。这些学习方法倾向于生成任务特定的、专门化的系统，而这些系统在它们所训练的狭窄领域之外往往是脆弱的。在低资源语言翻译、医学图像分析、自动驾驶和内容过滤等应用中，减少学习任务所需的人工标注样本数量或与环境的交互次数，对于提高系统的域外鲁棒性是至关重要的。

人类和动物似乎能够通过观察，以独立于任务的方式学习大量有关环境的背景知识。这种知识是常识的基础，使人类只需几个小时的实践就可以学会诸如驾驶等复杂的任务。**未来人工智能要考虑的一个关键问题是，人类如何从观察中学到这么多东西?**

在监督学习中，N 个类别中的某一类别的标签平均最多能传递 $\log_2(N)$ 位信息。在无模型强化学习中，奖励同样只能传递少量信息。相比之下，音频、图像和视频是高带宽的信息模态，隐含着大量关于环境结构的信息。这驱动了自监督学习的预测或重建方式，即通过预测数据中被掩盖或损坏的部分来训练系统“填补空白”。自监督学习在训练 Transformer 提取向量方面非常成功，这些向量能够捕捉单词或单词片段的上下文相关含义，并且这些向量对下游任务也非常有效。

对于文本，Transformer 能够从一组离散集的概率分布中预测缺失的单词。但在高维连续域，如视频中，某个特定视频片段的可信连续集是庞大而复杂的，如何正确地表示可信连续集的分布是一个亟待解决的问题。

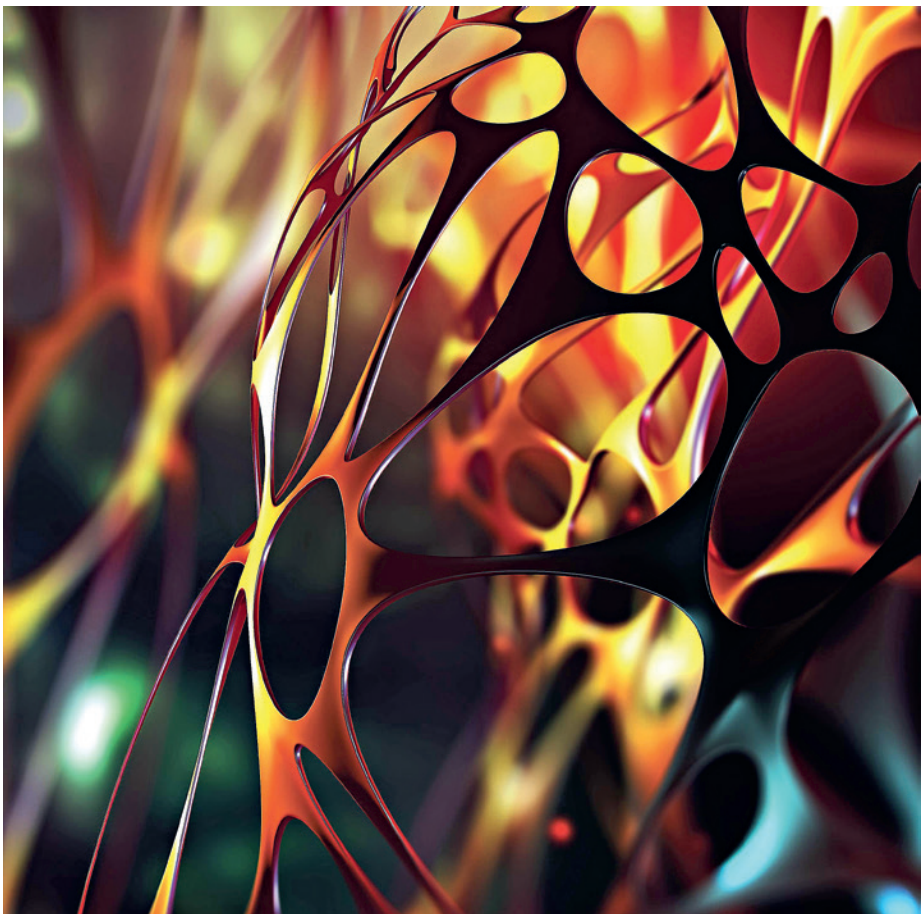
对比学习。解决该问题的一种方法是通过隐变量模型，为视频样本和可信连续集分配一个能量值 [1]。

给定一个输入视频 X 和一个连续片段 Y，我们需要一个模型，通过使用能量函数 $E(X, Y)$ 来判断 Y 是否与 X 兼容，当 X 和 Y 兼容时，该能量函数取低值，否则取高值。

对于给定的 X，$E(X, Y)$ 通过一个深度网络来计算，该神经网络以对比的方式进行训练，当 Y 与 X 兼容时，网络计算得到的 $E(X, Y)$ 具有低能量，而 Y 与 X 不兼容时，$E(X, Y)$ 具有高能量。对于给定的 X，推理过程需要找到一个使 $E(X, Y)$ 最小化的 $\check{Y}$，或者从具有低能量值的 Y 中采样。利用这种基于能量的方法来表示 Y 与 X 之间的依赖关系，使我们能够对一组不同的、多模态的可信连续集进行建模。

对比学习的关键难点在于选择好的“负面”样本：合适的样本点 Y，其能量将被推高。当可能的负面样本数量不够时，我们可以将它们全部考虑在内。这就是 softmax 所做的，因此在这种情况下，对比学习退化为有限离散符号集上的标准监督或自监督学习。但在实值高维空间中，向量 $\check{Y}$ 与 Y 有太多不同之处，为了提升模型表现，我们需要关注那些本应具有高能量但目前却能量较低的 Y。早期选取负样本的方法是基于蒙特卡罗方法，如受限玻耳兹曼机的对比散度 [48] 和噪声对比估计 [41]。生成对抗网络（Generative Adversarial Network, GAN）[36] 通过将神经网络应用于已知分布（例如高斯分布）中的潜在样本，训练生成神经网络以产生对比样本。生成器自训练以产生输出 $\check{Y}$，模型给出低能量 $E(\check{Y})$。生成器可以使用反向传播来获得 $E(\check{Y})$ 相对于 $\check{Y}$ 的梯度。生成器和模型同时进

[1] 如果能量被定义为独立系统的能量，那么在任何概率解释中，它们必须对应负对数概率。



致表征，在视觉特征学习方面取得了不错的结果。正样本对由同一幅图像的不同失真版本组成，包括裁剪、缩放、旋转、颜色偏移、模糊等方式产生的失真。负样本对是不同图像的相似失真版本，可以通过“硬负样本挖掘”的过程从数据集中巧妙地选取得到，也可以是一个迷你批处理（minibatch）中其他图像的所有失真版本。网络高层的隐向量随后作为线性分类器的输入，用有监督的方式进行训练。这种孪生神经网络（Siamese-net）方法在标准图像识别基准上取得了很好的效果[6, 21, 22, 43, 67]。最近，两种 Siamese-net 方法成功地打消了对比样本的必要性。第一个被称为 SwAV，通过量化一个网络的输出来训练另一个网络[20]；第二个被称为 BYOL，平滑两个网络其中一个的权重轨迹，能够防止模式崩溃[40]。

行训练，模型试图赋予训练样本低能量，为生成的对比样本提供高能量。

GAN 的优化有些棘手，但对抗性训练的理念已被证明卓有成效，在图像合成方面产生了令人印象深刻的结果，并在内容创造、域适应[34]以及域或风格迁移方面[87]开辟了许多新的应用领域。

运用对比学习使表征一致。对比学习提供了一种无须重新构造或生成像素就能发现良好特征向量的方法。其思想是学习一种前馈神经网络，当输入同一图像内部的两个不同图像块[10]或同一目标的两个不同视图[17]时，网络生成非常相似的输出向量；输入不同图像的图像块或不同目标的视图时，网络生成不相同的输出向量。两个输出向量之间的平方距离可以作为一种能量函数，兼容样本对的能量函数值较小，不兼容样本对的能量函数值较大。

最近发表的一系列论文使用卷积网络来提取一

变分自编码器。最近流行的一种自监督学习方法是变分自编码器（Variational Auto-Encoder，VAE）[58]。它由一个将图像映射到隐编码空间的编码器网络和一个从隐编码生成图像的解码器网络组成。VAE 通过在编码器的输出中加入高斯噪声来限制隐编码的信息容量，然后再将其传递给解码器。这类似于将小的有噪声的球体打包成半径最小的较大球体。信息的容量受限于打包球体中可以包含多少噪声球体。噪声球相互排斥，因为良好的重构误差要求不同样本的隐编码之间的重叠越小越好。从数学角度看，该系统通过在噪声分布上对隐编码边缘化而使自由能最小化。然而，使自由能相对于参数最小化是极度困难的，人们必须依赖于统计物理

中的变分近似方法，使自由能的上限最小化。

深度学习的未来

深度学习系统的性能通常可以通过简单的扩展而得到显著提高。有了更多的数据和更高的计算量，它们通常表现得更好。比如有1750亿参数的语言模型GPT-3[18]（与人脑中的神经元突触数量相比，其参数量仍然很小）生成文本的质量与只有15亿参数的GPT-2模型相比具有显著提升。聊天机器人Meena[2]和BlenderBot[73]的效果也在随着模型的扩展而不断改进。目前，人们在扩展模型方面付出了巨大的努力，将极大地改善现有系统，但当前深度学习仍然存在着根本性的不足，仅靠提升模型参数和计算量是无法解决的。

通过对比人类的学习能力和目前的人工智能，未来可以在以下几个方向取得突破：

1. 监督学习需要太多的数据标注，而无模型强化学习需要太多的实验。人类似乎能够用很少的经验也能完成某项任务的学习。

2. 目前的系统对分布变化的鲁棒性不如人类，人类只需要几个范例就能够快速适应这种类似的变化。

3. 当前深度学习最成功的是感知任务，也是所谓的系统1类任务。如何通过深度学习进行系统2类任务，即需要一系列逻辑思考才能完成的任务，这方面的研究是一个令人兴奋的领域，它仍处于起步阶段。

需要改进的地方。在早期，机器学习的理论家们就把注意力集中在独立同分布假设上，即假定测试数据应该和训练样本具有相同的分布。然而，这种假设在现实世界中并不成立。比如说，由于各种代理改变环境的行为而产生的不平稳性，或者学习代理不断学习和发现新事物，其智力界限会不断提升。现实往往是，如今最好的人工智能系统从实验室走向实地时，其性能仍然会大打折扣。我们希望在面对分布变化时，模型能够迅速适应并提升鲁棒性（不依赖于分布的泛化学习），在面对新的学习任务时，能降低样本复杂性。与人类相比，目前的有监督学习系统在学习新任务时需要更多的样例，而且这样的情况在无模型强化学习[23]中甚至更糟，因为相比标注数据，奖励机制能够反馈的信息太少了。人类可以以一种与普通独立同分布泛化的方式不同且更强大的方式学习新任务[61, 67]：我们可以正确地解释现有概念的新组合，即使这些组合在我们的训练分布下是极不可能出现的，只要它遵循我们已经学会的高级句法和语义模式即可。最近的研究向我们阐明了不同的神经网络结构在这种系统泛化能力方面的表现[8, 9]。我们该如何设计具有这些能力的未来机器学习系统，使其能更好地泛化或更快地适应分布变化呢？

从同质层到代表实体的神经元组。神经科学研究表明，相邻的神经元组（形成所谓的超列（hypercolumn））紧密相连，可能代表一种更高级别的向量单元，不仅能传递标量，而且能传递一组坐标值。这个想法是胶囊架构[47, 59]的核心，它也是软注意力机制的固有使用方法，集合中的每个元素都与一个向量相关联，从中可以读取一个键向量和一个值向量（有时也是一个查询向量）。这些向量级单元可以视为对象及其属性（如胶囊中的姿势信息）的表示。计算机视觉领域的最新论文正在探索卷积神经网络的扩展，其中，网络的顶层表示输入图像中检测到的一组候选对象，并使用类似Transformer的结构对这些候选对象进行操作[19, 84, 86]。如果神经网络为目标及其各部分指定内在参照系，并利用各部分之间的几何关系识别物体，它在面对直接的对抗攻击时就不会那么脆弱[79]，因为对抗攻击依赖于人们使用的信息之间的差异与神经网络来识别对象。

多时间尺度适应。大多数神经网络只有两个时间尺度：权重在许多样本中适应得非常缓慢，而行为对于每个不同的新输入适应得非常迅速。通过添加快速适应和快速衰减的“快速权重”[49]的叠加层，会引入有趣的新计算能力。尤其是，它创建了一个高容量的短期记忆[4]，可以允许神经网络执行真正的递归，其中相同的神经元可以在递归调用中重复使用，因为它们在更高级别调用中的活动向量可以在稍后使用快速权重中的信息进行重构。多时间尺度适应的功能在元学习[12, 33, 75]中也逐渐被采纳。

更高层次的认知。当我们考虑一个新的挑战时，例如在一个具有不寻常的交通规则的城市中驾驶，甚至想象在月球上驾驶车辆，我们可以利用已经掌握的知识和通用技能，以新的方式动态地重新组合它们。这种形式的系统泛化使人类能够很好地概括那些在他们训练分布的新环境中不太可能发生的情况。然后，我们可以通过练习、微调和编译这些新技能来进一步提高技能，使它们不再需要有意识的关注。我们怎样通过重用已知的知识片段，才能赋予神经网络快速适应新环境的能力，从而避免对已知技能的干扰？Transformer[32] 和循环独立机制 [38] 朝这个方向做了初步探索。

似乎我们的内隐（系统 1）处理能力让我们在计划或推理时能够猜测潜在的好处或者危险的未来。这就提出了一个问题，即系统 1 网络如何在更高（系统 2）的层次上指导搜索和规划，也许是基于 AlphaGo 蒙特卡罗树搜索的价值函数 [77]。

机器学习研究依赖于归纳偏置或先验知识，以鼓励学习朝着与环境假设兼容的方向进行。系统 2 处理的性质及其认知神经科学理论基础 [5, 30] 表明了一些归纳偏置和架构 [11, 45]，这些可以用来设计新的深度学习系统。我们应该如何设计包含这种归纳偏置的深度学习架构和训练框架？

幼儿展现出的因果发现能力 [37] 表明，这可能是人脑的一个基本特性。最近的研究表明，在干预性变化下优化分布外的泛化能力可以用来训练神经网络，以发现因果依赖或因果变量 [3, 13, 57, 66]。我们应该如何构造和训练神经网络，能够让它们发现这个世界的一些潜在因果特性呢？

在 20 世纪提出的符号人工智能项目与这些开放性问题指出的方向有何关联呢？显然，符号人工智能项目旨在实现系统 2 的能力，例如推理，能够将知识分解成可以在一系列计算步骤中易于重组的片段，并且能够处理抽象变量、类型和实例。我们希望设计的神经网络能在处理实值向量的同时完成这些所有工作，从而保持深度学习的优势，包括使用可微计算和基于梯度自适应进行高效的大规模学习，在低水平的感知和行动中建立高水平的概念，处理不确定数据，以及使用分布式表示。 ■

作　者：

约书亚・本吉奥（Yoshua Bengio）

加拿大蒙特利尔大学计算机科学与运筹系教授，魁北克人工智能研究所 Mila 社区创始人、科学主任，CIFAR 机器与大脑学习项目联合主任。

雅恩・乐昆（Yann LeCun）

脸书（Facebook）副总裁兼首席人工智能科学家。美国纽约大学数学科学研究所和数据科学中心教授。

杰弗里・辛顿（Geoffrey Hinton）

加拿大多伦多矢量研究所（Vector Institute）首席科学顾问。谷歌（Google）副总裁和工程研究员。多伦多大学计算机科学荣誉教授。

译　者：

马卓奇

西安电子科技大学计算机科学与技术学院讲师。主要研究方向为计算机视觉与图像处理。

zhuoqima@xidian.edu.cn

（本文责任编委：**苗启广**）

参考文献

[1] Abadi, M. et al. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symp. Operating Systems Design and Implementation*, 2016, 265-283.

[2] Adiwardana, D., Luong, M., So, D., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al. Towards a human-like open-domain chatbot 2020; arXiv preprint arXiv:2001.09977.

[3] Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization, 2019; arXiv preprint arXiv:1907.02893.

[4] Ba, J., Hinton, G., Mnih, V., Leibo, J., and Ionescu, C. Using fast weights to attend to the recent past. *Advances in Neural Information Processing Systems*, 2016, 4331-4339.

[5] Baars, B. *A Cognitive Theory of Consciousness*. Cambridge University Press, Cambridge, MA, 1993.

[6] Bachman, P., Hjelm, R., and Buchwalter, W. Learning representations by maximizing mutual information across views. *Advances in Neural Information Processing Systems*, 2019, 15535-15545.

[7] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine

translation by jointly learning to align and translate, 2014; arXiv:1409.0473.

[8] Bahdanau, D., Murty, S., Noukhovitch, M., Nguyen, T., Vries, H., and Courville, A. Systematic generalization: What is required and can it be learned? 2018; arXiv:1811.12889.

[9] Bahdanau, D., de Vries, H., O'Donnell, T., Murty, S., Beaudoin, P., Bengio, Y., and Courville, A. Closure: Assessing systematic generalization of clever models, 2019; arXiv:1912.05783.

[10]Becker, S. and Hinton, G. Self-organizing neural network that discovers surfaces in random dot stereograms. Nature 355, 6356 (1992), 161-163.

[11]Bengio, Y. The consciousness prior, 2017; arXiv:1709.08568.

[12]Bengio, Y., Bengio, S., and Cloutier, J. Learning a synaptic learning rule. In Proceedings of the IEEE 1991 Seattle Intern. Joint Conf. Neural Networks 2.

[13]Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. A meta-transfer objective for learning to disentangle causal mechanisms. In Proceedings of ICLR'2020; arXiv:1901.10912.

[14]Bengio, Y., Ducharme, R., and Vincent, P. A neural probabilistic language model. NIPS'2000, 2001, 932–938.

[15]Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. Greedy layer-wise training of deep networks. In Proceedings of NIPS'2006, 2007.

[16]Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. Theano: A CPU and GPU math expression compiler. In Proceedings of SciPy, 2010.

[17]Bromley, J., Guyon, I., LeCun, Y., Säkinger, E., and Shah, R. Signature verification using a "Siamese" time delay neural network. Advances in Neural Information Processing Systems, 1994, 737–744.

[18]Brown, T. et al. Language models are few-shot learners, 2020; arXiv:2005.14165.

[19]Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In Procedings of ECCV'2020; arXiv:2005.12872.

[20]Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments, 2020;. arXiv:2006.09882.

[21]Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations, 2020; arXiv:2002.05709.

[22]Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning, 2020; arXiv:2003.04297.

[23]Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T., and Bengio, Y. Babyai: First steps towards grounded language learning with a human in the loop. In Proceedings in ICLR'2019; arXiv:1810.08272.

[24]Chopra, S., Hadsell, R., and LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the 2005 IEEE Computer Society Conf. Computer Vision and Pattern Recognition 1, 539–546.

[25]Collobert, R., Kavukcuoglu, K., and Farabet, C. Torch7: A matlab-like environment for machine learning. In Proceedings of NIPS Worskshop BigLearn, 2011.

[26]Collobert, R. and Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of ICML'2008.

[27]Conneau, A. and Lample, G. Cross-lingual language model pretraining. Advances in Neural Information Processing Systems 32, 2019. H. Wallach et al., eds. 7059–7069. Curran Associates, Inc.; http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.pdf.

[28]Dahl, G., Yu, D., Deng, L., and Acero, A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio, Speech, and Language Processing 20, 1 (2011), 30–42.

[29]Dayan, P. and Abbott, L. Theoretical Neuroscience. The MIT Press, 2001.

[30]Dehaene, S., Lau, H., and Kouider, S. What is consciousness, and could machines have it? Science 358, 6362 (2017, 486–492.

[31]Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of 2009 IEEE Conf. Computer Vision and Pattern Recognition, 248–255.

[32]Devlin, J., Chang, M., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of ACL'2019; arXiv:1810.04805.

[33]Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks, 2017; arXiv:1703.03400.

[34]Ganin, Y and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Proceedings of Intern. Conf. Machine Learning, 2015, 1180–1189.

[35]Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of AISTATS’2011.

[36]Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Advances in Neural Information Processing Systems, 2014, 2672–26804.

[37]Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., and Danks, D. A theory of causal learning in children: causal maps and bayes nets. Psychological Review 111, 1 (2004).

[38]Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. Recurrent independent mechanisms, 2019; arXiv:1909.10893.

[39]Graves, A. Generating sequences with recurrent neural networks, 2013; arXiv:1308.0850.

[40]Grill, J-B. et al. Bootstrap your own latent: A new approach to self-supervised learning, 2020; aeXiv:2006.07733.

[41]Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Proceedings of the 13th Intern. Conf. Artificial Intelligence and Statistics, 2010, 297–304.

[42]He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of CVPR’2020, June 2020.

[43]He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In Proceedings of CVPR’2016, 770–778.

[44]Hinton, G. A parallel computation that assigns canonical object-based frames of reference. In Proceedings of the 7th Intern. Joint Conf. Artificial Intelligence 2, 1981, 683–685.

[45]Hinton, G. Mapping part-whole hierarchies into connectionist networks. Artificial Intelligence 46, 1–2 (1990), 47–75.

[46]Hinton, G. et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing 29, 6 (2012), 82–97.

[47]Hinton, G., Krizhevsky, A., and Wang, S. Transforming auto-encoders. In Proceedings of Intern. Conf. Artificial Neural Networks. Springer, 2011, 44–51.

[48]Hinton, G., Osindero, S., and Teh, Y-W. A fast-learning algorithm for deep belief nets. Neural Computation 18 (2006), 1527–1554.

[49]Hinton, G. and Plaut, D. Using fast weights to deblur old memories. In Proceedings of the 9th Annual Conf. Cognitive Science Society, 1987, 177–186.

[50]Hinton, G. and Salakhutdinov, R. Reducing the dimensionality of data with neural networks. Science 313 (July 2006), 504–507.

[51]Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Improving neural networks by preventing co-adaptation of feature detectors. In Proceedings of NeurIPS’2012; arXiv:1207.0580.

[52]Hochreiter, S. and Schmidhuber, J. Long short-term memory. Neural Computation 9, 8 (1997), 1735–1780.

[53]Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 2015.

[54]Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. What is the best multi-stage architecture for object recognition? In Proceedings of ICCV’09, 2009.

[55]Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM Intern. Conf. Multimedia, 2014, 675–678.

[56]Kahneman, D. Thinking, Fast and Slow. Macmillan, 2011.

[57]Ke, N., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Pal, C., and Bengio, Y. Learning neural causal models from unknown interventions, 2019; arXiv:1910.01075.

[58]Kingma, D. and Welling, M. Auto-encoding variational bayes. In Proceedings of the Intern. Conf. Learning Representations, 2014.

[59]Kosiorek, A., Sabour, S., Teh, Y., and Hinton, G. Stacked capsule autoencoders. Advances in Neural Information Processing Systems, 2019, 15512–15522.

[60]Krizhevsky, A., Sutskever, I., and Hinton, G. ImageNet classification with deep convolutional neural networks. In Proceedings of NIPS’2012.

[61]Lake, B., Ullman, T., Tenenbaum, J., and Gershman, S. Building machines that learn and think like people. Behavioral and Brain Sciences 40 (2017).

[62]Lample, G. and Charton, F. Deep learning for symbolic mathematics. In Proceedings of ICLR’2020; arXiv:1912.01412.

[63]LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. Nature 521, 7553 (2015), 436–444.

[64]LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. Backpropagation applied to handwritten zip code recognition. Neural Computation 1, 4 (1989), 541–551.

[65]LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. In Proceedings of the IEEE 86, 11 (1998), 2278–2324.

[66]Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B., and Bottou, L. Discovering causal signals in images. In Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition, 2017, 6979–6987.

[67]Misra, I. and Maaten, L. Self-supervised learning of pretext-invariant representations. In Proceedings of CVPR’2020, June 2020; arXiv:1912.01991.

[68]Mohamed, A., Dahl, G., and Hinton, G. Deep belief networks for phone recognition. In Proceedings of NIPS Workshop on Deep Learning for Speech Recognition and Related Applications. (Vancouver, Canada, 2009).

[69]Morgan, N., Beck, J., Allman, E., and Beer, J. Rap: A ring array processor for multilayer perceptron applications. In Proceedings of the IEEE Intern. Conf. Acoustics, Speech, and Signal Processing, 1990, 1005–1008.

[70]Nair, V. and Hinton, G. Rectified linear units improve restricted Boltzmann machines. In Proceedings of the ICML’2010.

[71]Paszke, A., et al. Automatic differentiation in pytorch. 2017.

[72]Robinson, A. An application of recurrent nets to phone probability estimation. IEEE Trans. Neural Networks 5, 2 (1994), 298–305.

[73]Roller, S., et al. Recipes for building an open domain chatbot, 2020; arXiv:2004.13637.

[74]Rumelhart, D., Hinton, G., and Williams, R. Learning representations by back-propagating errors. Nature 323 (1986), 533–536.

[75]Schmidhuber, J. Evolutionary principles in self-referential learning. Diploma thesis, Institut f. Informatik, Tech.Univ. Munich, 1987.

[76]Shepard, R. Toward a universal law of generalization for psychological science. Science 237, 4820 (1987), 1317–1323.

[77]Silver, D., et al. Mastering the game of go with deep neural networks and tree search. Nature 529, 7587 (2016), 484.

[78]Sukhbaatar, S., Szlam, A., Weston, J., and Fergus, R. End-to-end memory networks. Advances in Neural Information Processing Systems 28, 2015, 2440–2448. C. Cortes et al., eds. Curran Associates, Inc.; http://papers.nips.cc/paper/5846-end-to-end-memory-networks.pdf.

[79]Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In Proceedings of ICLR’2014; arXiv:1312.6199.

[80]Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. Web-scale training for face identification. In Proceedings of CVPR’2015, 2746–2754.

[81]Thrun, S. Is learning the n-th thing any easier than learning the first? In Proceedings of NIPS’1995. MIT Press, Cambridge, MA, 640–646.

[82]Utgoff, P. and Stracuzzi, D. Many-layered learning. Neural Computation 14 (2002), 2497–2539, 2002.

[83]Van Essen, D. and Maunsell, J. Hierarchical organization and functional streams in the visual cortex. Trends in Neurosciences 6 (1983), 370–375.

[84]van Steenkiste, S., Chang, M., Greff, K., and Schmidhuber, J. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions, 2018; arXiv:1802.10353.

[85]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, T., and Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 5998–6008.

[86]Zambaldi, V., et al. Relational deep reinforcement learning, 2018; arXiv:1806.01830.

[87]Zhu, J-Y., Park, T., Isola, P., and Efros, A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the 2017 IEEE Intern. Conf. on Computer Vision, 2223–2232.