人是如何进行语言处理，即语言理解和语言生成的？这个问题仍是当今科学最大的未解之迷。脑科学、认知科学领域均有一些发现和假说，比如达马西奥(Antonio Damasio)等脑科学家是这样看的[1]：人脑是由神经元组成的大规模复杂神经网络，生物信号在神经网络上不断传递，使神经网络的状态不断发生变化，不同的状态形成不同的神经表征(neural representation)。这种“神经计算”都是在人的下意识中进行的，只有一部分能够上升到意识，对应着人的思考。人的思考其实是在意识中产生表象(image)的神经计算，表象有视觉表象、听觉表象、运动表象等。

人的语言处理也一样，本质是在下意识中进行的神经计算，能意识到的只是语言理解和生成过程中产生的表象。人的语言理解过程是唤起语言相关的概念的表象，在其基础上组合产生出新的表象的过程。语言中的符号包括语音符号、文字符号，以表征的形式记忆在人脑里，在语言处理中被激活使用。人的语言处理并不是符号计算。

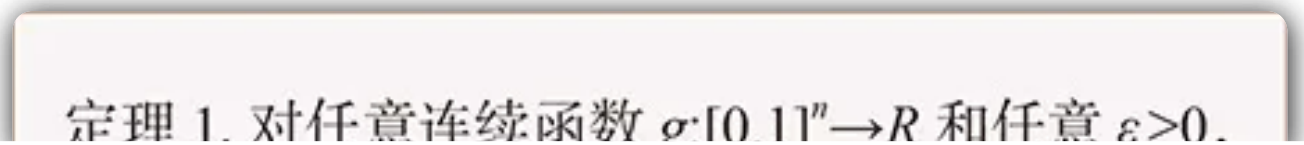
深度学习

深度学习是以复杂人工神经网络为模型的机器学习。（人工）神经网络是受生物神经网络启发而开发的，由（人工）神经元连接组成的网络，本质是数学模型。神经元是非线性函数，神经网络是由许多神经元组成的复合函数。神经网络的特点是拥有大量参数，参数的估计可以通过在数据上的目标函数优化得到。参数的学习使用反向传播算法，只要神经网络的函数可微分就可以进行。神经网络的计算能够实现某种功能，如图像识别、机器翻译。

事实证明，深度学习是实现机器智能的强大工具。以下从机器学习理论的角度总结深度学习的优缺点。

优点

深度学习的优点主要体现在三个方面。


定理 1. 对任意连续函数 $\sigma:[0,1]^n \to R$ 和任意 $\varepsilon>0$，

存在一个二层神经网络 $f(x)$，使得对于任意 $x \in [0,1]^n$，有 $|f(x)-g(x)|<\varepsilon$ 成立。

定理 2. 存在这样的布尔函数，可以由深度为 k 的多项式复杂度的逻辑门电路表示，等价的深度为 k-1 的逻辑门电路变为指数复杂度。这里深度指从输入到输出的最长路径的长度，复杂度指逻辑门的个数。

定理 3. 二层的过参数化 ReLU 神经网络，用于多分类，学习使用随机梯度下降，进行随机初始化，如果每个类的数据来自多个分布的混合模型，而且类之间的分布间距足够大，那么学习到的神经网络具有较小的泛化误差。

定理 4. 强健的学习可以定义为以下 min/max 优化问题：

$$\min_\theta E_x\left[\max_{\|x-x'\|_\infty\le\varepsilon} L(\theta, x')\right]$$

这里 L 是损失函数，x 和 x' 是样本，θ 是模型参数。考虑二类分类问题，若两类的数据来自两个不同的高斯分布，则强健的学习样本复杂度显著大于一般的学习样本复杂度。

图1 关于深度学习的一些理论结果

第一，神经网络拥有强大的函数近似能力。通用函数近似定理（定理1）指出，即使是二层神经网络，都可以以任意精度近似任意一个连续函数。假设实现某一功能的“理想”的函数存在，则有可能存在一个神经网络是这个函数的充分近似。

第二，深的神经网络比浅的神经网络拥有更精简的表达能力和更高的样本效率(sample complexity)。存在这样的情况：深而窄的神经网络与浅而宽的神经网络是等价的。但前者的参数比后者更少，只需要较少的样本就可以学到。相反，在极端情况下，浅而宽的神经网络的宽度是指数级的，现实中并不可取。这方面的理论支持来自逻辑门电路。因为神经网络可以表示逻辑门电路，所以关于逻辑门电路的结论（定理2）也适用于神经网络。

第三，深度学习有很强的泛化能力，也就是从训练集上学到的误差小的模型在测试集上也同样有小的误差[2]。深度学习中常常不做正则化，也不产生过拟合。通常是在大规模训练数据、过参数化(over-parameterized)神经网络以及随机梯度下降(SGD)训练的条件下发生的，这里的过参数化是指网络的参数数量大于训练数据数量。已有机器学习理论尚不能很好地解释这种现象，是当前热门的研究课题。最近有理论研究对一些特殊情况下的泛化能力做出了证明（定理3）[3]。

这些事实说明深度学习具有强大的复杂模式学习的能力。

缺点

深度学习也有缺点，缺乏强健性(robustness)是一个广为人知的问题。也就是说数据中很小的扰动就会导致预测错误。这应该是深度学习强大的学习能力所致。强健的学习可以定义为min/max的优化问题。一般的机器学习目标是在平均情况下预测误差最小，而强健的学习目标是在最坏情况下预测误差最小，具体来讲，数据在某个范围内发生对自己最不利的扰动时也能保证平均预测误差最小。最近的理论研究证明[4]，在一些特殊情

况下，强健的学习比一般的学习需要更多的样本（定理4），结论对深度学习和传统机器学习都适用。这对深度学习来说不是一个好消息，意味着它需要更多的样本才能变得强健。

深度学习的另一个缺点是笔者称为恰当性(adequacy)的问题。深度学习权威约书亚·本吉奥教授(Yoshua Bengio)曾说："从现有深度学习系统的失败中我们能得出什么结论？我会说，最显著的是系统往往通过表面线索学习，而这些线索有助于帮助完成特定任务。但这些通常并不是我们认为最重要的。"由于训练数据的偏差，机器学习的特点（预测误差最小化导向，训练中的随机性）等原因，深度学习常常"学到不恰当的知识"。比如，图像识别中认为有把手的就是杯子，有轮胎的就是汽车。传统机器学习也有这个问题，但深度学习的问题更突出。深度学习就像是一个只擅长考试的学生，成绩很好，但并没有掌握好知识。

可解释性

神经网络不具备可解释性，但笔者认为这并不一定是缺点。可解释性依赖于应用，比如在金融、医疗等领域的预测需要可解释，但是其他领域的预测未必如此。人也不能解释自己是如何进行感知和认知处理的，深度神经网络未必能够解释自己的决策过程。

深度学习用于自然语言处理

自然语言处理的问题从机器学习的角度可以归结为五大类，分别是分类、匹配、转换、结构预测、序列决策过程，如表1所示。深度学习使这五大类任务的正确率都有很大提升，特别是匹配和转换[5]。

表1 自然语言处理问题

问题	定义	应用
分类	$x \rightarrow c$，给字符串 x 打标签 c	文本分类、情感分析
匹配	$x, y \rightarrow R$，对字符串 x 和 y 进行匹配	搜索、问答、单轮对话（基于检索）
转换	$x \rightarrow y$，将字符串 x 转换为字符串 y	机器翻译、摘要、单轮对话（基于生成）
结构预测	$x \rightarrow [x]$，识别字符串 x 的结构	序列标注、中文分词、语义分析
序列决策过程	$\pi: s \rightarrow a$，在多个字符串表示的状态 s 选择动作 a	任务驱动多轮对话

我们还不知道人是如何进行语言处理的。深度学习，特别是在监督学习的场景中，实际是在用数据驱动的方法模拟人的语言处理功能，参照人如何对给定输入产生相应输出，然后进行“模仿”，如图2所示。人工的神经处理和生物的神经处理有某些相似性，但更本质的是人工神经网络作为数学模型有很强的表达能力，深度学习作为机器学习方式有很强的学习能力。

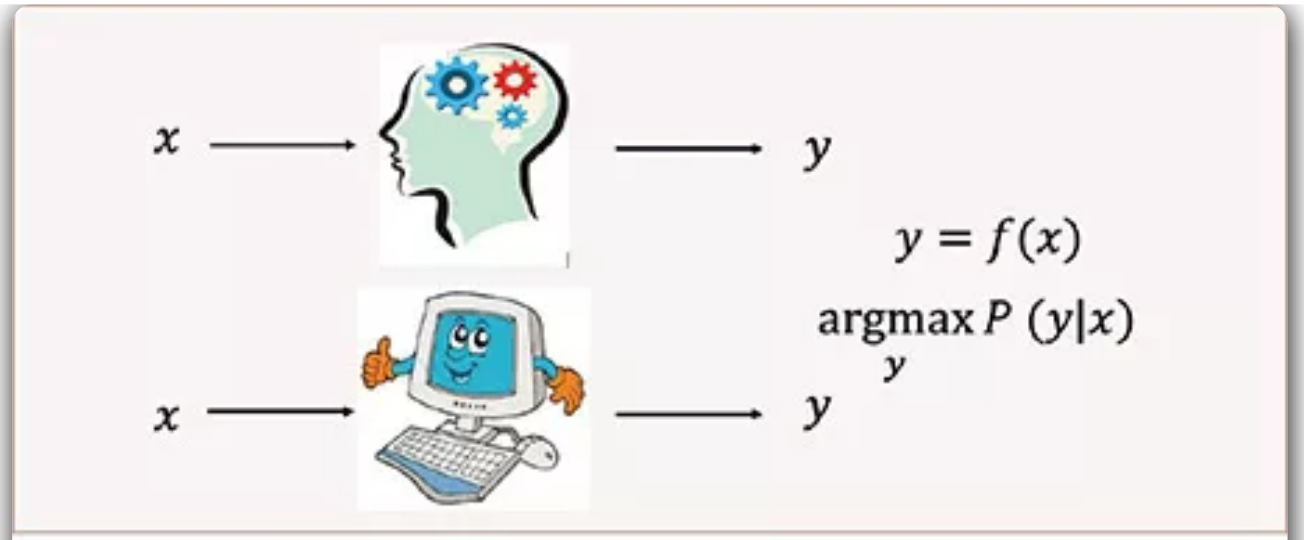


图2 自然语言处理中使用深度学习

机器学习专家戴维·麦卡莱斯特(David McAllester)认为：深度学习提供的是一种通用的可微分的编程工具集，包括各种网络、残差连接、门控、注意力、生成式对抗网络(GAN)等。人工智能的进步主要来自这些深度学习工具的使用，而这个趋势会持续下去[6]。

深度学习用于自然语言处理时，通常将单词表示为实数向量，也就是单词嵌入。将句子或文本表示为单词的向量序列，作为输入。输出是通过软最大化(softmax)得到的类别，可以针对整个输入，也可以针对每个单词。自然语言处理中常使用的模型有前馈神经网络、循环神经网络、卷积神经网络、序列对序列模型。常用的损失函数是交叉熵。注意力机制是自然语言处理中强大的工具，机器翻译中的Transformer、预训练模型BERT都使用了注意力机制。

注意力机制实际上是一种软的联想记忆机制，是键-值(key-value)数据库的一种推广。在传统的键-值数据库中，查询、键、值都是符号。给定查询(query)，找到完全匹配的键，返回相应的值。注意力机制中，查询、键、值都是实数向量。给定查询，计算查询和所有键的匹配度，以归一化匹配度为权重计算加权平均值，并返回加权平均向量，如图3所示。

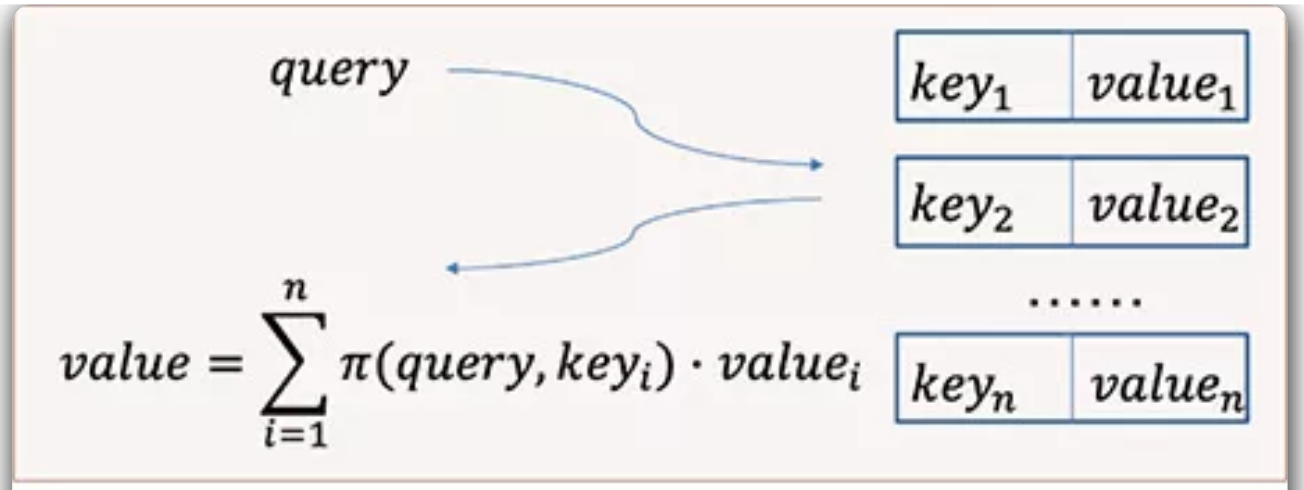


图 3 注意力机制

Transformer的编码器也是BERT的基本模型。给定一个句子，编码器可以生成其层次化的语义表示。每一个单词在每一层上有以这个单词为中心的语义表示（实数向量）。下一层的语义表示通过自我注意力(self-attention)生成上一层的语义表示。直观上每个单词的语义表示和其他所有单词的语义表示基于相似度组合成新的语义表示。比如，图4中“John”“loves”“Mary”三个单词各自有一个语义表示，在自我注意力中都成为查询、键、值。以每个单词的语义表示作为查询，以所有单词的语义表示作为键和值，通过注意力可以得到每个单词新的语义表示。这样一层层叠加，可以得到整个句子所有单词的语义表示。

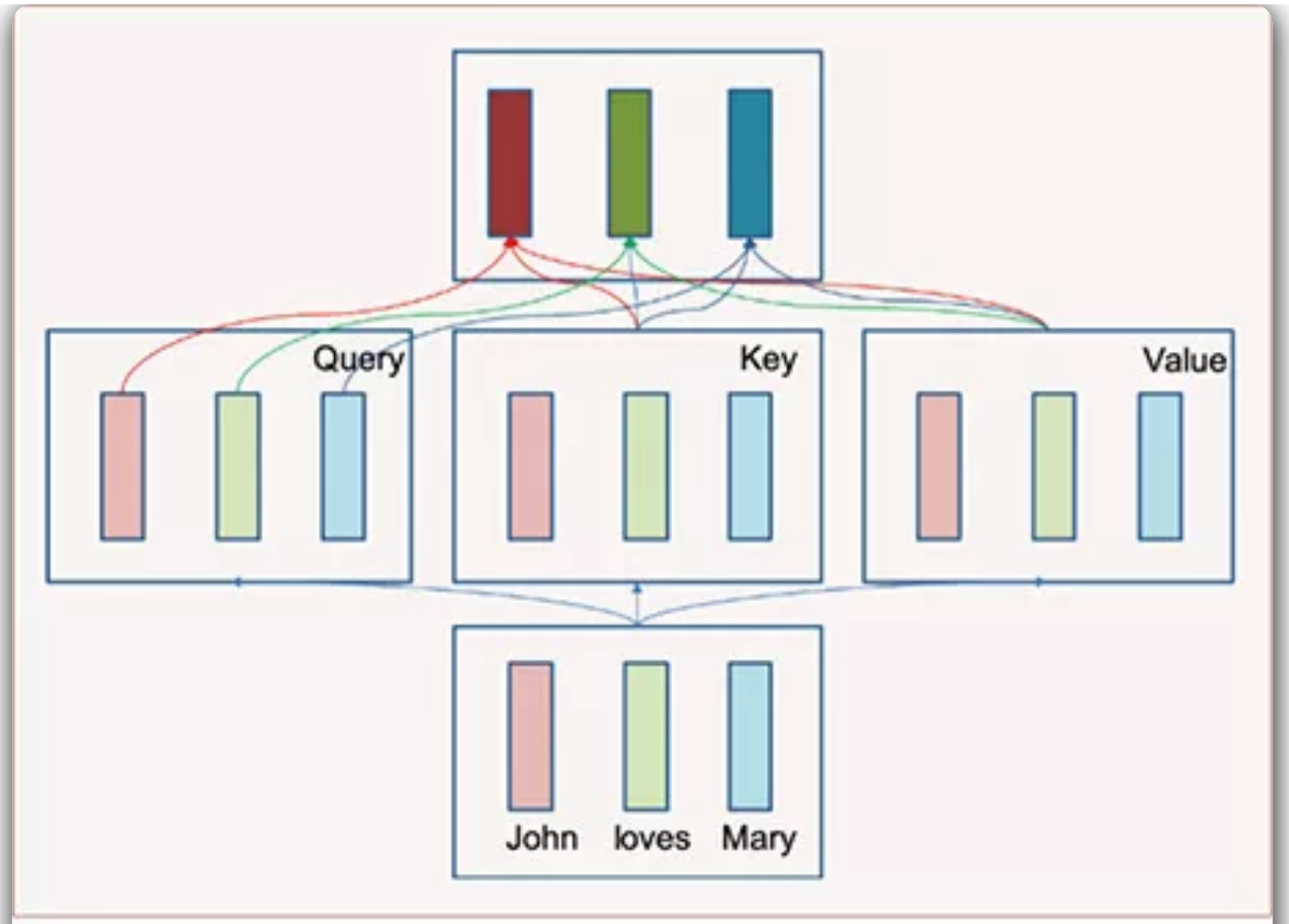


图 4 通过自我注意力从下一层语义表示生成上一层语义表示

未来研究课题

面向未来，围绕着深度学习与自然语言处理，笔者认为有几个重要的课题需要进行探索，也能带来很大的技术进步，分别是多模态、生成、预训练、神经符号处理，关键是开发相关的新的神经网络模型。

深度学习之前，图像、语音、语言几个领域的技术相对比较独立。深度学习把它们紧密地联系在一起。首先，有很多深度学习技术可以跨领域使用，比如卷积神经网络。其次，在深度学习里，图像、语音、语言的内容都用同样的实数向量来表示，可以做跨模态的信息处理，比如看图说话就是典型的例子。多模态信息处理还有很大的发展空间，可以预见今后还会产生很多新的技术及新的应用。

深度学习给自然语言处理带来的主要变革在于生成。序列对序列模型，特别是Transformer，大幅度提高了机器翻译的正确率。在训练语料充分的领域使用机器翻译，比如新闻的翻译，已经可以实用。生成式的对话系统之前基本不存在，现在只要有足够多的对话数据，训练一个序列对序列模型，在一定范围内可以进行表面自然的单轮对话。围绕着生成，还有很多可以研究的课题，技术创新还有很大的潜力。

预训练是指用大量的无标注数据，事先训练语言表示模型，然后用于各种语言处理任务。图5显示了预训练与语言处理任务的关系，x′表示无标注数据，h(x′)表示通过预训练学到的表示模型，可以帮助任务y=f(x)完成得更好。甚至只需要有少量数据精调任务模型即可。最近预训练的语言表示模型BERT用于不同的语言处理任务，都带来了正确率的大幅度提升，让人惊叹。无标注数据是大量存在的，预训练技术促进自然语言处理发展的前景非常可观。深度学习权威杰弗里·辛顿教授(Geoffrey Hinton)等一直强调无监督学习是深度学习未来的发展方向，理由是机器需要像人一样只用少量标注数据就可以学到知识。他在最近的Google I/O 2019大会上阐述了同样的观点[1]。

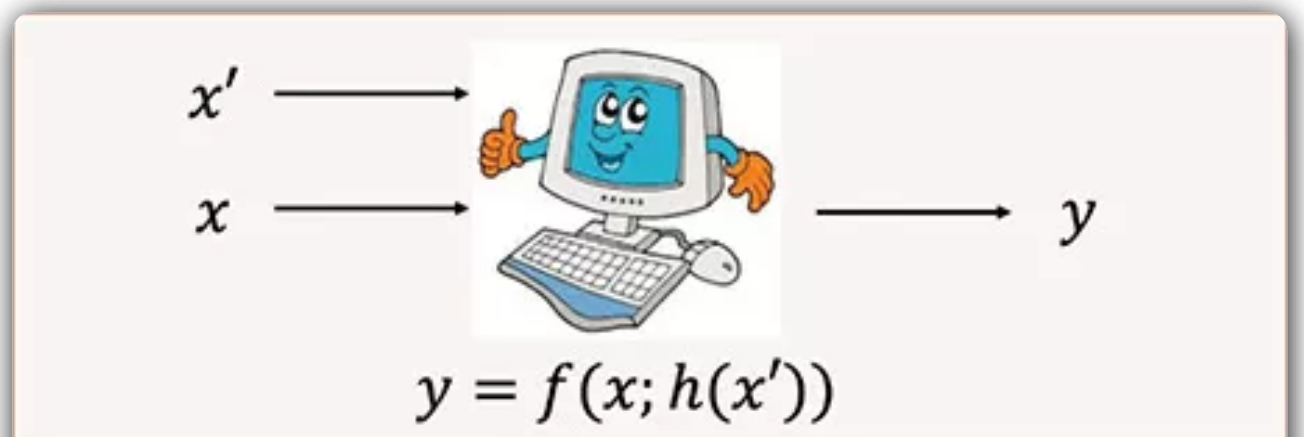


图5 深度学习中的预训练

神经符号处理是指将神经处理（深度学习）与符号处理（传统方法）进行结合，实现更强大的语言处理能力。神经符号处理并没有严格定义，图6显示了一种基本情况，s表示有结构的符号， g(s)表示从符号中得到的神经表征。可以认为神经处理与符号处理各自对应着人的下意识和意识层面的信息处理。两种处理拥有完全不同的特性，其结合不是一件简单的事情。但如果有突破，会给自然语言处理带来革命性的进步。在神经信息处理系统大会(NeurIPS 2018)上，向辛顿请教了他对神经符号处理的看法。辛顿没有正

面回答这个问题，而是用一个比喻作了说明。他说：“假设你有一个很好的电动汽车，丰田汽车来找你，问是否可以把它和传统的燃油汽车结合到一起。你问的就像是这样一个问题。”从这个回答可以感觉到辛顿不是很看好这个方向。

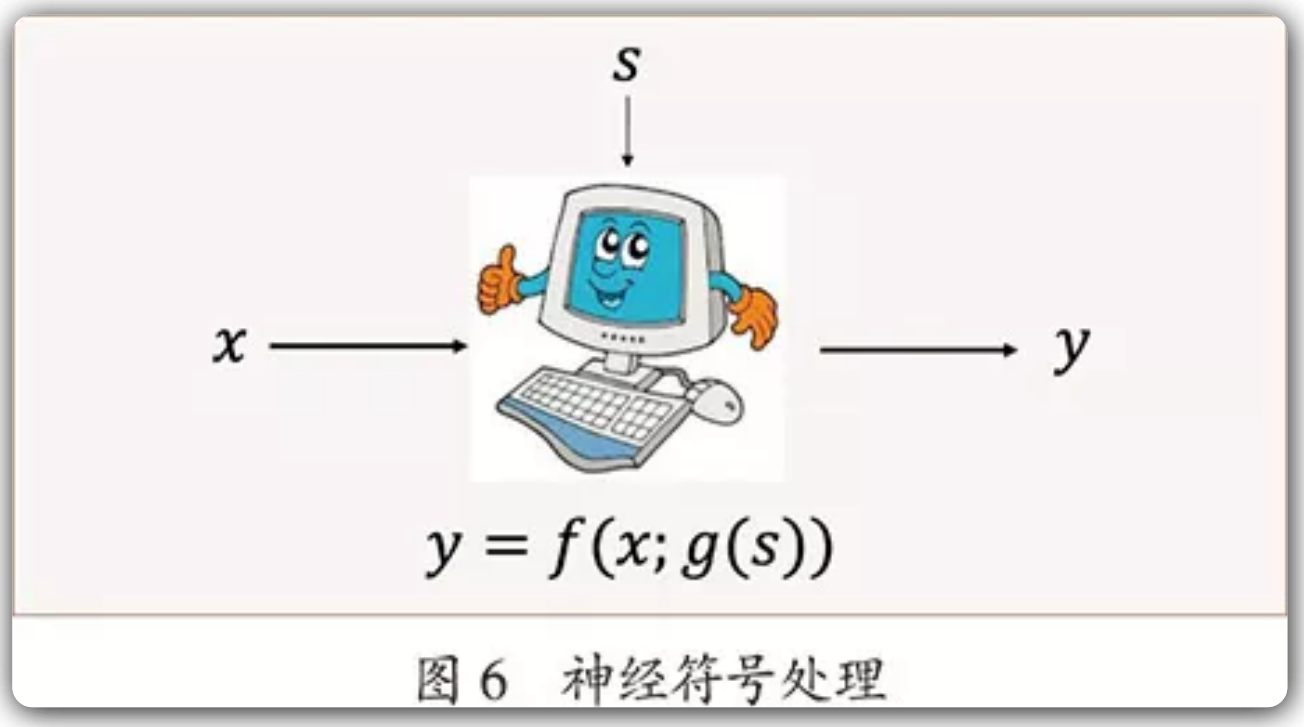


图 6 神经符号处理

结语

深度学习的成功依赖于大数据和大算力。过去30年，计算速度、通信速度、内存容量均提高了100万倍。可以预见，未来计算能力的提升有可能减缓，但增长态势不会变化。所以随着硬件技术的发展，深度学习技术本身也会不断进步，为自然语言处理领域带来巨大变革。

另一方面，语言本身是认知现象，涉及到知识和推理。人的语言使用乃至思考的机制还不清楚，但从现象上看，涉及到意识（表象）和下意识（表征）。看来仅仅利用深度学习和神经计算可能无法完全实现（这一点与属于感知的图像和语音不同）。现实中，知识主要靠人工定义，虽然有很大局限，但仍是最可行且最有效的手段。

未来的自然语言处理是基于神经计算，符号计算，还是两者的结合？现在还不是很清楚。需要长时间的探索和研究。