神经网络读书报告

谢非

2017 年 6 月 4 日

目录

[1]

[1]本文为读书或读论文有所思而写

1 神经网络发展史

1.1 心理学起源

人工智能的发展或许可以追溯到公元前仰望星空的古希腊人，当亚里士多德为了解释人类大脑的运行规律而提出了联想主义心理学的时候，他恐怕不会想到，两千多年后的今天，人们正在利用联想主义心理学衍化而来的人工神经网络，构建超级人工智能，一起又一次地挑战人类大脑认知的极限。

联想主义心理学是一种理论，认为人的意识是一组概念元素，被这些元素之间的关联组织在一起。受柏拉图的启发，亚里士多德审视了记忆和回忆的过程，提出了四种联想法则：

- 邻接：空间或时间上接近的事物或事件倾向于在意识中相关联。
- 频率：两个事件的发生次数与这两个事件之间的关联强度成正比。
- 相似性：关于一个事件的思维倾向于触发类似事件的思维。
- 对比：关于一个事件的思维倾向于触发相反事件的思维。

亚里士多德描述了这些在我们意识中作为常识在起作用的法则。例如，苹果的触感、气味或味道会很自然地引出苹果的概念。令人惊讶的是，如今这些提出了超过 2000 年的法则仍然是机器学习方法的基本假设。例如，彼此靠近（在限定距离下）的样本被聚类为一个组; 经常与响应变量发生的解释变量引起模型更多的注意; 相似/不相似数据通常用潜在空间中更相似/更不相似的嵌入表示。此后两千年间，联想主义心理学理论被多位哲学家或心理学家补充完善，并最终引出了 Hebbian 学习规则，成为神经网络的基础。

1890 年，心理学家 William James 出版了第一部详细论述人脑结构及功能的专著《心理学原理》，他认为一个神经细胞受到刺激激活后可以把刺激传播到另一个神经细胞，并且神经细胞激活是细胞所有输入叠加的结果。“并且神经细胞激活是细胞所有输入叠加的结果”这一句话很重要，他的这个猜想后来得到了证实，并且我们现在设计的神经网络也是基于这个理论。

1943 年，神经病学家和神经元解剖学家 McCulloch 和数学家 Pitts 在生物物理学期刊发表文章提出神经元的数学描述和结构。并且证明了只要

有足够的简单神经元，在这些神经元互相连接并同步运行的情况下，可以模拟任何计算函数（M-P 模型）。他们所做的开创性的工作被认为是人工神经网络 (ANN) 的起点。1943 年的时候，我们已经开始用数学来描述神经元互相作用的行为。这被认为是第一个仿生学的神经网络模型，他们提出的很多观点一直沿用至今，比如说他们认为神经元有两种状态，要不就是兴奋，要不就是抑制。下节课我们学到的单层感知器就是模仿这个，单层感知器的输出要不就是 0 要不就是 1。他们最重要的贡献就是开创了神经网络这个研究方向，为今天神经网络的发展奠定了基础。

1.2 从神经网络到深度学习

Hebbian 学习规则以 Donald O. Hebb 命名。Donald O. Hebb 在 1949 年的论著《The Organization of Behavior》[3] 中提出了这一法则。Hebb 也因为这篇论文被视为神经网络之父。1949 年，Hebb 提出了那条著名的规则：

> 一起发射的神经元连在一起。

更具体的表述是：

> 当神经元 A 的轴突和神经元 B 足够接近并反复或持续激发它时，其中一个或两个神经元就会发生增长或新陈代谢的变化，例如激发 B 的神经元之一——A efficiency——会增加。

这个拗口的段落可以重写为现代机器学习的语言：

> 其中代表输入信号为的神经元的突触权重的变化。表示突触后反应，表示学习率。换句话说，“Hebbian 学习规则”指出，随着两个单位共同出现频率的增加，两个单位之间的联系会加强。

尽管 Hebbian 学习规则被视为奠定了神经网络的基础，但今天看来它的缺陷是显而易见的：随着共同出现的次数增加，连接的权重不断增加，主信号的权重将呈指数增长。这就是 Hebbian 学习规则的不稳定性（Principe et al., 1999）。幸运的是，这些问题没有影响 Hebb 作为神经网络之父的地位。Erkki Oja 扩展了 Hebbian 学习规则以避免不稳定性，并且他还表明，遵循此更新规则的神经元的行为，近似于 Principal Component Analyzer[8]（PCA）的行为 (Oja, 1982)。Frank Rosenblatt 通过引入感知器的概念进一

步实现了 Hebbian 学习规则 [9]（Rosenblatt, 1958）。像 Hebb 这样的理论家专注的是自然环境中的生物系统，而 Rosenblatt 构建了一个名为感知器的电子设备，它具有根据关联进行学习的能力。早期神经元模型和现代感知器之间的一个区别是非线性激活函数的引入。将感知器放在一起，就变成了基本的神经网络。通过并列放置感知器，我们能得到一个单层神经网络。通过堆叠一个单层神经网络，我们会得到一个多层神经网络，这通常被称为多层感知器 [6]（MLP ）（Kawaguchi, 2000）。单层神经网络具有局限性，正是这种局限性导致了相关的研究曾经一度停滞了进二十年，但同时，也正是这种局限性刺激了神经网络向更高层结构进发，渐渐迎来了如今的深度学习时代。

1.3 Universal Approximation

Universal Approximation 是深度学习的理论基石.

神经网络的一个显著特性，即众所周知的通用逼近属性，可以被粗略描述为 MLP 可以表示任何函数。可以从以下三方面探讨这一属性：

- 布尔逼近：一个隐藏层的 MLP 可以准确的表示布尔函数；
- 连续逼近：一个隐藏层的 MLP 可以以任意精度逼近任何有界连续函数；
- 任意逼近：两个隐藏层的 MLP 可以以任意精度逼近任何函数。

Universal approximation 成为如今神经网络与深度学习一片繁荣景象的重要理论基石，Universal approximation 的相关理论——一个多层神经网络具备表达任何方程的能力——已经成为深度学习的标志性特点。本章节的一个最大的贡献在于将过去在这个问题上的相关理论研究工作加以整理，分三个脉络阐释了三种不同的 Universal approximation。作者重新整理了从上世纪八十年代末期到本世纪初期的相关理论工作，把原本艰深晦涩的理论证明以形象的语言重新描述出来。如图 1 所示，无数个线性 decision boundary 组合叠加可以制造出圆形边界，而无数个圆形边界的叠加何以逼近任何一个方程。

浅层神经网络的通用逼近属性以呈几何级数增长的神经元为代价，因此是不现实的。关于如何在减少计算单元数量的同时维持网络的表达力，这

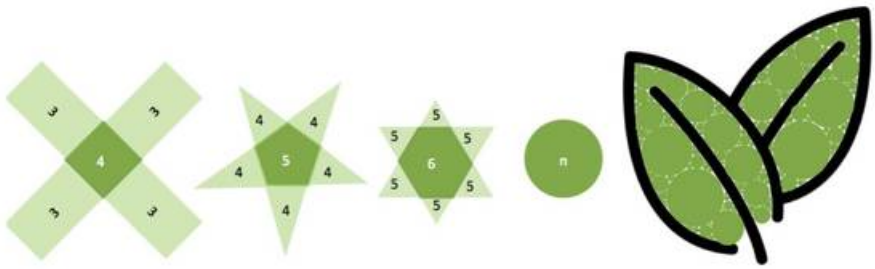


图 1: 由 decision boundary 组合出任意方程

个问题已被追问多年。从直觉出发，Bengio 和 Delalleau（2011）认为追求更深的网络是自然的，因为

1. 人类神经系统就是一个深层次的结构。

2. 人类倾向于将一个抽象层次的概念表示为较低层次的概念组合。

今天，我们的解决方案是建立更深的结构，这一方案的理论支撑是，要想达到一个具有多项式的 k 层神经网络的表达能力，如果使用 k-1 层结构，则神经元的数量需要以指数级增长。不过，理论上，这仍是一个尚未最终证明的结论。但这仍可以看出，深度学习中“深度”二字的重要性：从姚期智老师 1985 年的工作，到 Yoshua Bengio 近几年的成果，无一不在重复“深度”的价值。在深度学习炙手可热的今天，在还有很多人讨论“深度”的必要性的今天，希望相关的老师和同学们仔细审视前人的成果。

2 深度——神经网络

2.1 深度生成模型

从八十年代的 Self Organizing Map 到 Hopfield Network, 再到鼎鼎大名的 BoltzmannMachine 和 RestrictedBoltzmann Machine, 直到 Hinton 塑造的 Deep Belief Network。深度学习的研究一路走来, 悠长的历史之中, 一个又一个璀璨的明星在前人的基础上诞生。

图 2 总结了本节将涉及的模型。水平轴代表这些模型的计算复杂度, 而垂直轴代表表达能力。这是六个里程碑式的模型。

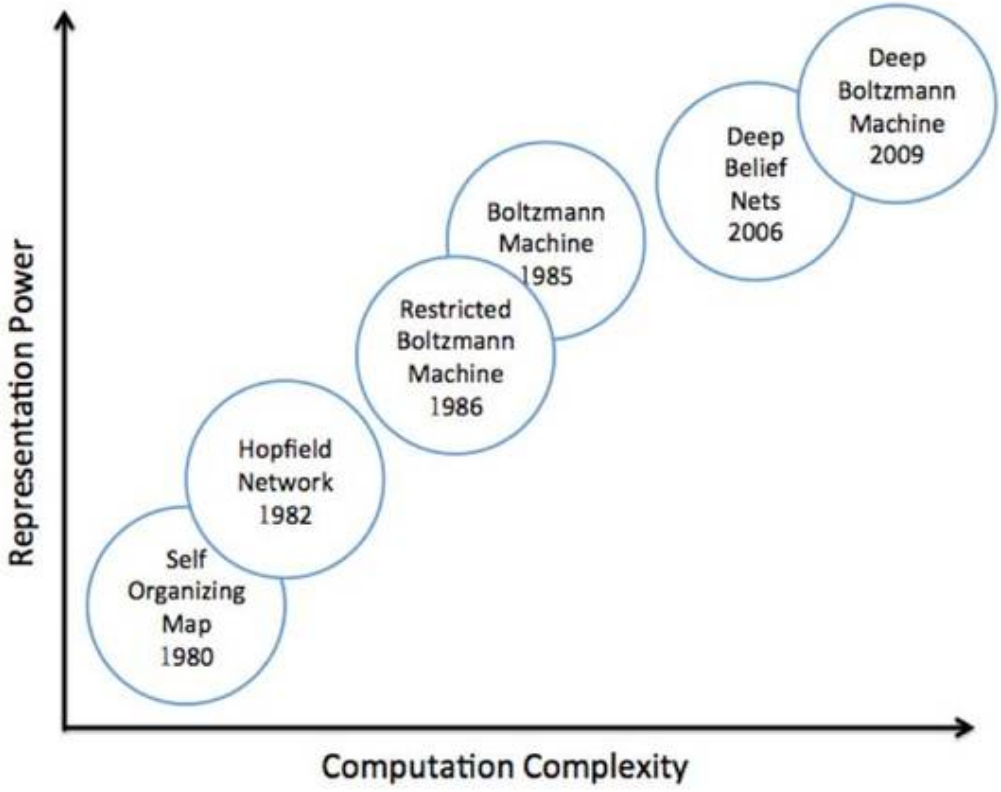


图 2: 六个里程碑式的模型

其中, 早先的模型, 比如 Self Organizing Map 和 Hopfield Network, 它

们的诞生主要基于人们对于这个世界基本的认知，因而相关的介绍也更加的浅显易懂。中期的模型，比如 Boltzmann Machine 和 Restricted Boltzmann Machine 虽然依然是前面模型的改进，当时的学者却更依赖于相关的数学和物理理论，因而本文 [13] 在此处的介绍也是理论韵味十足，用公式铺路，步步前行。当 Hinton 介绍了 Deep Belief Network 之后，深度学习更依赖于经验性的结论的特点又将本文重点转为文字性的介绍。

Self Organizing Map[7] 由 Kohonen 发明。SOM 是一种强大的技术，主要用于减少数据维度，通常减少到一维或二维（Germano, 1999）。在降低维度的同时，SOM 还保留了数据点的拓扑相似性。它也可以被看作是用于聚类的工具，同时把拓扑强加在聚类表示上。

Hopfield Network[4] 历史上被描述为一种 recurrent3 形式的神经网络，由 Hopfield 创造（1982）。此处的“recurrent”是指连接神经元的权重是双向的。Hopfield Network 因其内容可寻址的存储器属性而被广泛认可。Boltzmann Machine[1]，由 Ackley 等人发明，是随机隐藏单元版的 Hopfield Network。它的名字来自 Boltzmann 分布。

Restricted Boltzmann Machine（RBM）[11]，最初称为 Harmonium，由 Smolensky 发明，是一种带有限制的 Boltzmann Machine，其限制实在可见单元或隐藏单元之间没有连接。

深度置信网络由 Hinton 等人创造，他指出 RBM 可以以贪婪的方式进行堆叠和训练。最后给大家介绍的是深度生成模型谱系里的最后一块里程碑，Deep BoltzmannMachine[10]，由 Salakhutdinov and Hinton 创造。DBM 和上文提到的 DBN 的区别在于，DBM 允许在底层中双向连接。因此，用 DBM 表示堆叠的 RBM，比用 DBN 好得多。当然如果 DBM Deep Restricted Boltzmann Machine 可能更清楚。

在 DBM 之后，又出现了很多研究论文，未来的工作则更多。

2.2 卷积神经网络和人类视觉

卷积神经网络的谱系主要是从对人类视觉皮层的认识演变而来。卷积神经网络的视觉问题的成功原因之一是：复制人类视觉系统的仿生设计。本部分主要介绍了深度学习在计算机视觉角度上的发展，也就是卷积神经网络的发展，侧重于各个在 ImageNet 比赛中所有作为的神经网络模型。以介绍人类的视觉神经网络开始，所有的后续介绍都将围绕着人类的视觉神经网络展开。

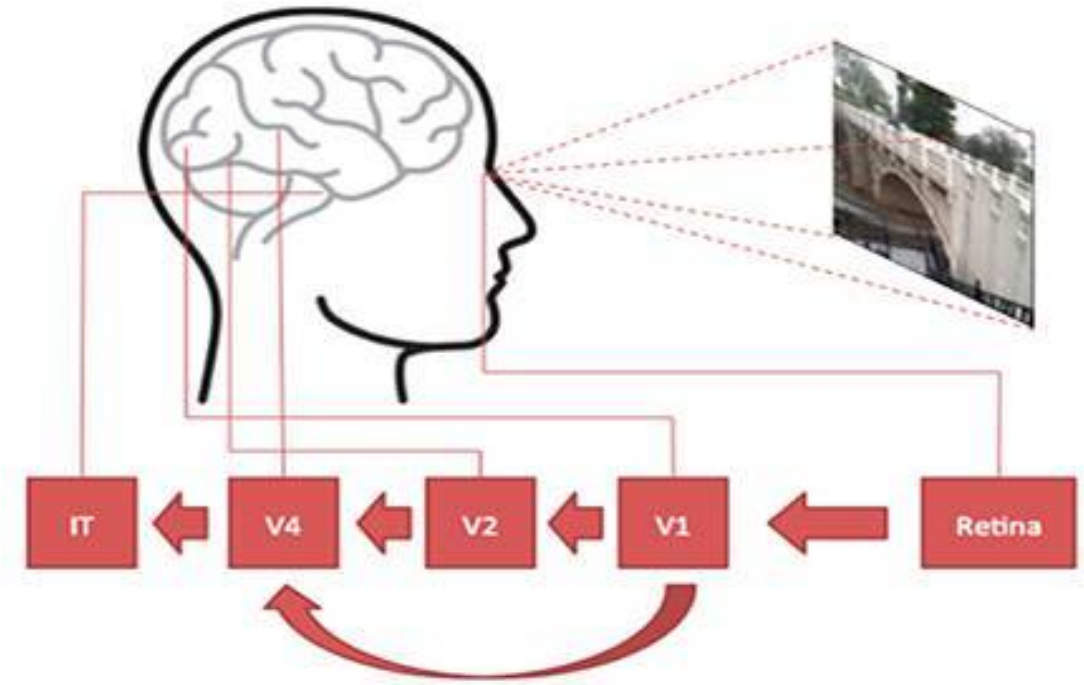


图 3: 基于人类视觉的视觉神经网络

在介绍了视觉神经网络之后，作者重点介绍了卷积在计算机视觉问题上的巨大意义。这些意义对于资深研究计算机视觉的老师同学来说可能已是陈词滥调，但是对于刚刚从深度学习时代开始接触计算机视觉问题的同学来说却可能至关重要。卷积作为一个非常有效的视觉特征提取工具，几乎是深度学习在计算机视觉问题上如此成功的基石。在介绍了卷积的意义之后，作者先介绍了 LeNet，进而带领读者回顾了近几年在 ImageNet 上有所

作为的重要模型，包括 AlexNet，VGG，和 ResNet。

值得一提的是，即便是在人们开始用计算机模仿人类视觉神经网络之后近四十年的今天，即便是在一个模型是否与人类视觉神经网络相似已经不再重要的今天，ResNet 的成功的重要创新模块依然可以在四十年前人类视觉神经网络的研究中找到影子。由此可见本文重新审视前人工作的重要性。本节的最后，作者介绍了一些在计算机视觉领域内的非常有趣的关键性问题，这些问题看似是计算机视觉问题的死穴，然而值得庆幸的是，人类视觉系统同样有这种盲点。卷积神经网络与人类视觉神经网络如此相似，甚至连缺陷都如此相似，究竟是喜是悲，还待更多后续工作揭晓。

2.3 递归网络

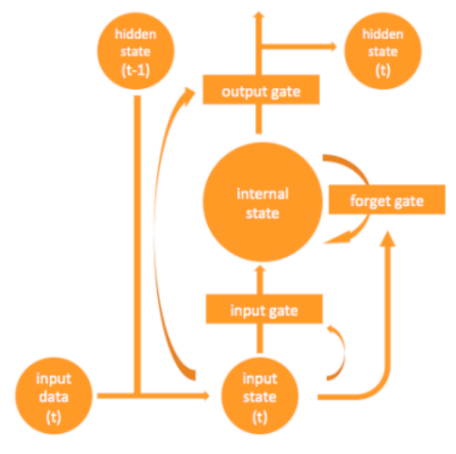
递归神经网络（RNN）是神经网络的一种，其单位的连接形成了有向循环; 这种性质赋予了其处理时间数据的能力。

“递归（recurrent）”的现代定义由 Jordan [2]（1986 年）最早提出：

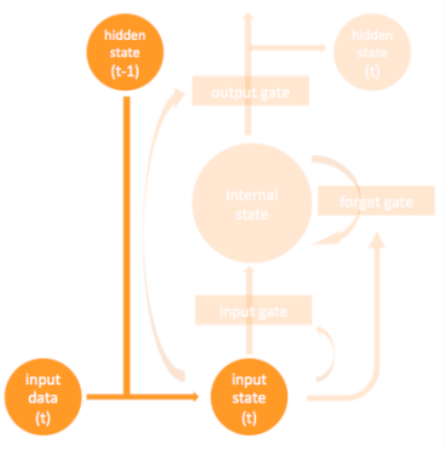
> 如果网络具有一个或多个 cycle，即，如果可以遵循从单元返回到其自身的路径，则网络被称为循环（一般将 Recurrent Neutral Network 译为“递归”神经网络）。非递归网络没有 cycle。

他的模型（Jordan, 1986 年）后来被称为 Jordan Network。几年后，Elman（1990）发明了另一个形式略有不同的 RNN 。从 Jordan 网络到 Elman 网络的变化引人注目，因为它引入了从隐藏层传递信息的可能性，这显著提高了后来工作中结构设计的灵活性。Hochreiter 和 Schmidhuber（1997）为 RNN 谱系引入了一个新的神经元，称为 Long Short-Term Memory（LSTM）。这一术语“LSTM”最早用于指称借助于特殊设计的存储器单元，设计用来克服消失梯度问题的算法。如今，“LSTM”广泛用于表示具有该存储器单元的任何递归网络，其现在被称为 LSTM 单元。LSTM 被用于克服 RNN 不能长期依赖的问题（Bengio et al.，1994）。为了克服这个问题，它需要专门设计的存储单元，如图 4（a）所示。LSTM 包括几个关键组件：

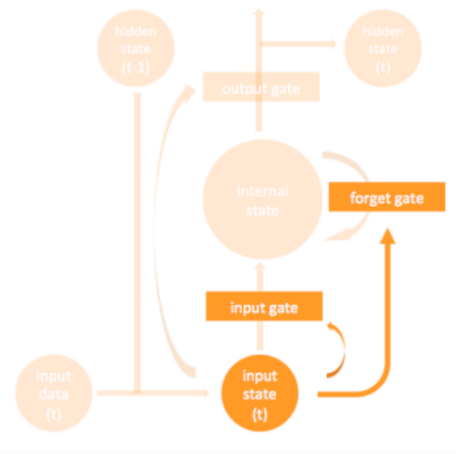
- states: 用于为输入提供信息的数值。
- gates: 用于决定 states 信息流的数值。



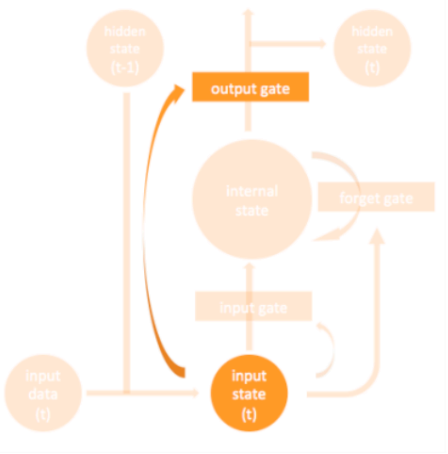
(a) LSTM “memory” cell



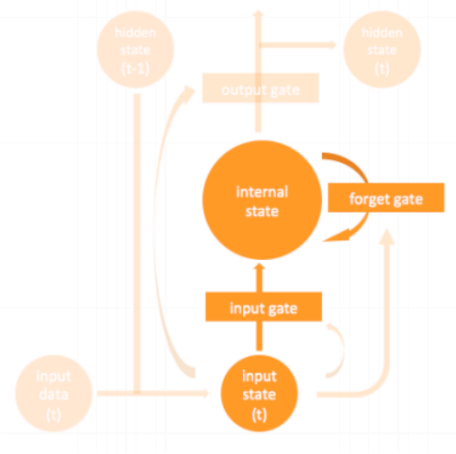
(b) Input data and previous hidden state form into input state



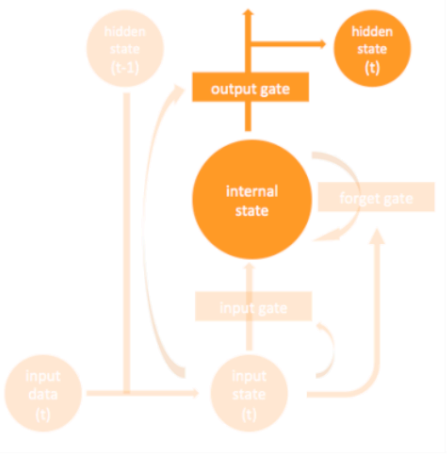
(c) Calculating input gate and forget gate



(d) Calculating output gate



(e) Update internal state



(f) Output and update hidden state

图 4: LSTM 单元和功能

作者用 6 张图详细解释了 LSTM 的多个 gate 与 state 之间复杂的相互协同作用，强大而又神秘的 LSTM 在 6 张深入浅出的图片中一目了然。这或许是迄今能找到的最清晰的 LSTM 解析了！

所有的权重都是需要在训练期间学习的参数。因此，理论上，如果必要的话，LSTM 可以学习记住长时间依赖，并且可以学会在必要的时忘记过去。这使它成为一个强大的模型。

2.4 神经网络的优化

优化是深度学习发展历史上不可回避的课题。目前存在的优化方式有：

- 梯度法：尽管神经网络已经发展了 50 多年，神经网络的优化仍然严重依赖于反向传播算法内的梯度下降法。
- 剔除（dropout）法 [12]：剔除法由（Hinton et al., 2012;Srivastava et al., 2014）创造。这种技术很快拥有了影响力，不仅因为它具有良好的性能，而且实施简单。这个想法很简单：在训练时随机剔除一些单位。更正式的表述是：在每次训练中，每个隐藏单元以概率 p 随机地被从网络中省掉了。
- Batch Normalization[5]：由 Ioffe 和 Szegedy（2015）发明的 Batch Normalization 是深度神经网络优化的另一个突破。他们解决了他们称为内部协变量移位的问题。直观上看，问题可以理解为以下两个步骤：1）如果输入改变（在统计学中，函数的输入有时被表示为协变量），则学习的函数几乎无用；2）每层都是一个函数，下层参数的变化改变了当前层的输入。这种变化可能很剧烈，因为它可能改变输入的分布。

3 启示

回顾深度学习和神经网络的发展史，从方法论的层面，我们可以得到如下启示：

- 奥卡姆剃刀定律：人们一方面将结构层层叠加，另一方面希望反向传播可以找到最佳参数。看上去他们有追求更复杂模型的倾向。但历史表明，大道至简。比如 dropout 被广泛认可，不仅因为它表现出色，更多是因为它的推理简单而直观。
- 要有野心：如果一个模型提出时具有比同时期更多的参数，它必须能解决掉一个其他模型不能漂亮解决的问题。LSTM 比传统的 RNN 复杂得多，但它出色地解决了消失梯度的问题。DBN 之所以出名并不是因为它是第一个提出将一个 RBM 放到另一个 RBM 的网络，而是因为他们提出了一个算法，使得深层架构能够被有效地训练。
- 跨学科阅读学习：许多模型受机器学习或统计学科以外的领域知识的启发。比如人类视觉皮层极大地启发了卷积神经网络的发展。甚至最近流行的残差网络也可以在人类视觉皮层中找到相应的机制。

Reference

[1] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[2] Jonathan D Cohen, David Servan-Schreiber, and James L McClelland. A parallel distributed processing approach to automaticity. *The American journal of psychology*, pages 239–269, 1992.

[3] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[4] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[6] Kiyoshi Kawaguchi. A multithreaded software model for backpropagation neural network applications. 2000.

[7] Teuvo Kohonen. The self-organizing map. *Neurocomputing*, 21(1):1–6, 1998.

[8] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

[9] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[10] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Artificial Intelligence and Statistics*, pages 448–455, 2009.

[11] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, DTIC Document, 1986.

[12] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[13] Haohan Wang, Bhiksha Raj, and Eric P Xing. On the origin of deep learning. *arXiv preprint arXiv:1702.07800*, 2017.