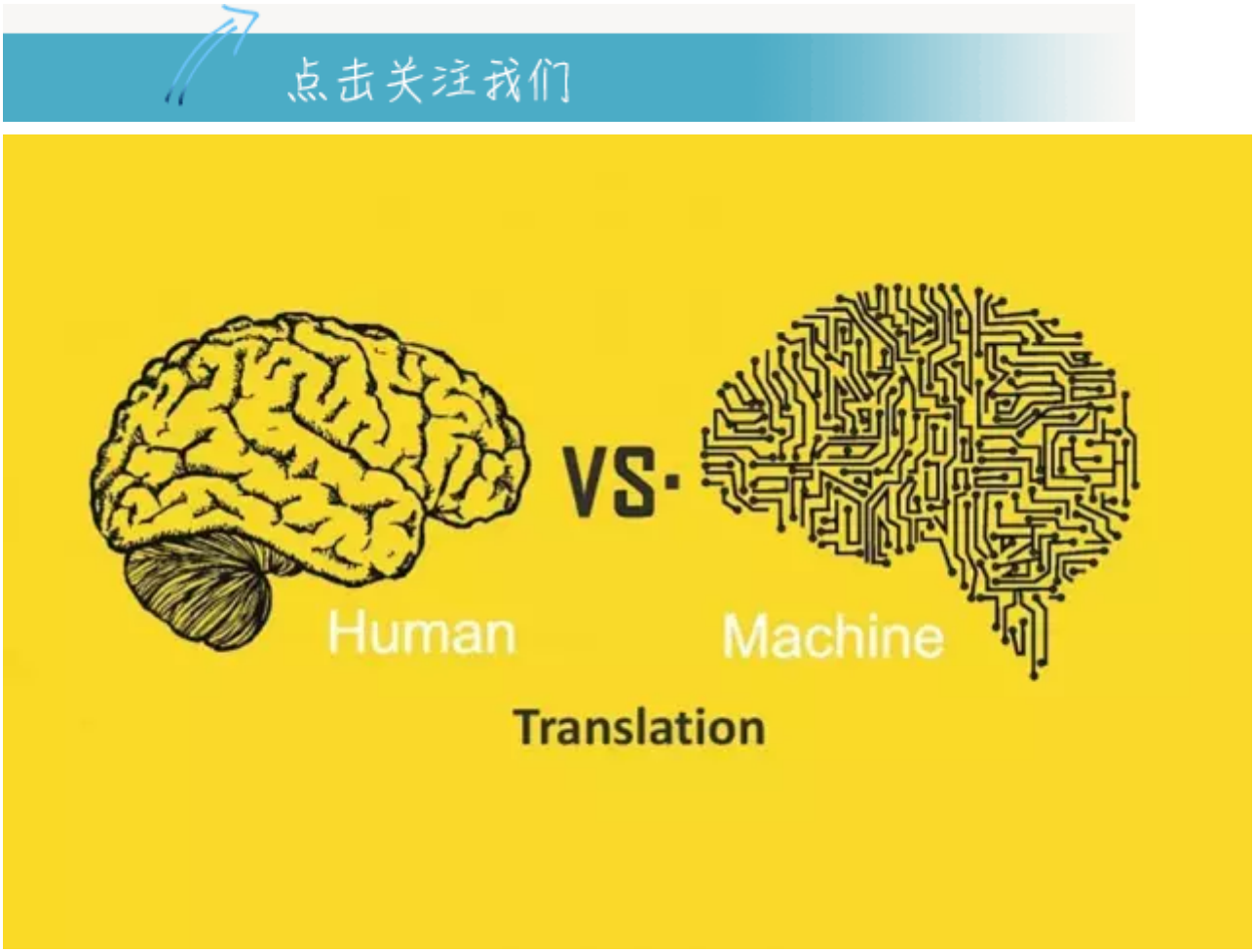
一文看尽深度学习RNN：为啥就它适合语音识别、NLP与机器翻译?

Jason Brownlee CDA数据分析师 2017-08-07



作者 Jason Brownlee

本文是机器学习大牛Jason Brownlee系统介绍RNN的文章，他在文中详细对比了LSTM、GRU与NTM三大主流架构在深度学习上的工作原理及各自特性。读过本文，你就能轻松GET循环神经网络在语音识别、自然语言处理与机器翻译等当前技术挑战上脱颖而出的种种原因。

循环神经网络(RNN)是一种人造神经网络，它通过赋予网络图附加权重来创建循环机制，以维持内部的状态。

神经网络拥有“状态”以后，便能在序列预测中明确地学习并利用上下文信息，如顺序或时间成分。

本文将一次性带你了解RNN在深度学习中的各种应用。

读完之后，你应该能弄懂：

- 最先进的RNN是如何进行深度学习任务的，如LSTM（长短时记忆网络）、GRU（门控循环单元）与NTM（神经图灵机）？
- 最先进的RNN同人工神经网络中更广泛的递归研究间的具体关系如何?
- 为什么RNN能在一系列有挑战的问题上表现如此出色?

我们不可能面面俱到，把所有的循环神经网络细节都讲一遍。因而，我们要把重点放在用于深度学习的循环神经网络上（LSTM，GRU和NTM），以及理解它们相关的必要背景知识。

让我们进入正题。

循环神经网络算法深度学习之旅 Photo by Santiago Medem,权利保留。

概览

我们首先来了解一下循环神经网络的研究背景。

接下来，我们会仔细研究LSTM，GRU和NTM在深度学习中的应用。

最后，我们还要了解一些同RNN用于深度学习相关的高级话题。

- 循环神经网络
  - 完全递归网络
  - 结构递归神经网络
  - 神经历史压缩机
- 长短期记忆网络(LSTM)
- 门控循环单元神经网络
- 神经图灵机

循环神经网络

首先来了解一下RNN的研究背景。

普遍的看法是，循环在拓扑上赋予网络以记忆的特性。

但还有一种理解RNN的更好角度：将其看作训练集包含了当前训练样本的一组输入样本。这就比较合乎“常规”了，比如一个传统的多层感知机。

```
X(i)->y(i)
```

但将上一组样本中的一组补充到训练样本中，则是“非常规”的。比如循环神经网络。

```
[X(i-1),X(i)]->y(i)
```

跟所有前馈网络的范式一样，这里的问题是如何将输入层连接到输出层，包括反馈激活，然后训练网络的结构以令其收敛。

现在让我们先从简单的概念开始，来了解一下不同类型的循环神经网络。

完全递归网络

这个网络保留了多层感知器的层状拓扑结构，但是网络中的每个神经元与其他神经元进行加权连接，并且有一个与其自身的反馈连接。

当然，并不是所有的连接都会被进行训练，同时由于误差导数出现极端非线性的情况，传统的反向传播将不再起作用，因此该网络采用随时间反向传播算法（BPTT）或随机梯度下降法（SGD）进行逼近。

> 更多信息请参考：Bill Wilson的Tensor Product Networks（1991）
> http://www.cse.unsw.edu.au/~billw/cs9444/tensor-stuff/tensor-intro-04.html

结构递归神经网络

结构递归神经网络是递归网络的线性架构变体。结构递归可以促进分级特征空间中的分枝，并且使得网络架构可以模仿这个进行训练。

训练过程通过子梯度方法的梯度下降实现。

> 这在R. Socher等人，Paralsing Natural Scenes and Natural Language with Recursive Neural Networks，2011中有详细描述。
> http://machinelearning.wustl.edu/mlpapers/paper_files/ICML2011Socher_125.pdf

神经历史压缩机

1991年，Schmidhuber首次发表了这个非常深度的学习机，通过无监督RNNs层次结构的预训练，能够实现数百个神经层的信用分配（即表现好的组件分配权重就大一些，有利于实现目标）。

RNN通过无监督训练预测下一次的输入。然后只有存在误差的输入才能向前传递，将新的信息传送到层次结构中的下一个RNN，然后以较慢的自组织时间尺度进行处理。

很明显这个过程中不存在信息丢失，只是进行压缩。 RNN堆栈就是数据的“深度生成模型”。 数据可以从压缩的形式进行重建。

> 参见J.Schmidhuber等人，Deep Learning in Neural Networks：An Overview，2014。
> http://www2.econ.iastate.edu/tesfatsi/DeepLearningInNeuralNetworksOverview.JSchmidhuber2015.pdf

尽管听起来不太可能,但因为误差会通过较大拓扑向后传播，增加非线性导数极值的计算量,使信用分配困难，所以反向传播仍可能失败。

长短时记忆网络

在传统的时间反向传播(BPTT)或实时循环学习(RTTL)算法中，误差信号随着时间流逝往往会爆炸或消失。

反向传播误差的时间演化指数般地依赖于权重大小。权重爆炸可能导致权重不稳定，而在梯度弥散（消失）时，会造成学习跨越长时间滞后，并且需要花费过多的时间，甚至根本不起作用。

- LSTM是一种基于梯度学习算法训练而来的新型循环网络架构。
- LSTM旨在克服误差回流问题。它可以学习跨越超过1000步的时间间隔。
- 在有噪声、不可压缩的输入序列情况下，该网络确实不会损失短程(lag)能力。

误差的回流问题是通过一种高效的、基于梯度的算法来克服的，这种算法网络结构通过特殊神经元的内部状态让误差流变为常数（从而不会爆炸或消失）。

这些神经元会减少“输入权重冲突”和“输出权重冲突”的影响。

输入权重冲突：如果输入非零，则必须使用相同的输入权重来存储某些输入并忽略其他输入，然后就会经常收到冲突的权重更新信号。

这些信号会尝试让权重参与存储输入并保护输入。 这种冲突使得学习难度加大，并且需要一个对上下文更加敏感的机制来通过输入权重来控制“写入操作”。

输出权重冲突：只要神经元的输出不为零，来自该神经元的输出连接的权重就将吸引在序列处理期间产生的冲突权重更新信号。

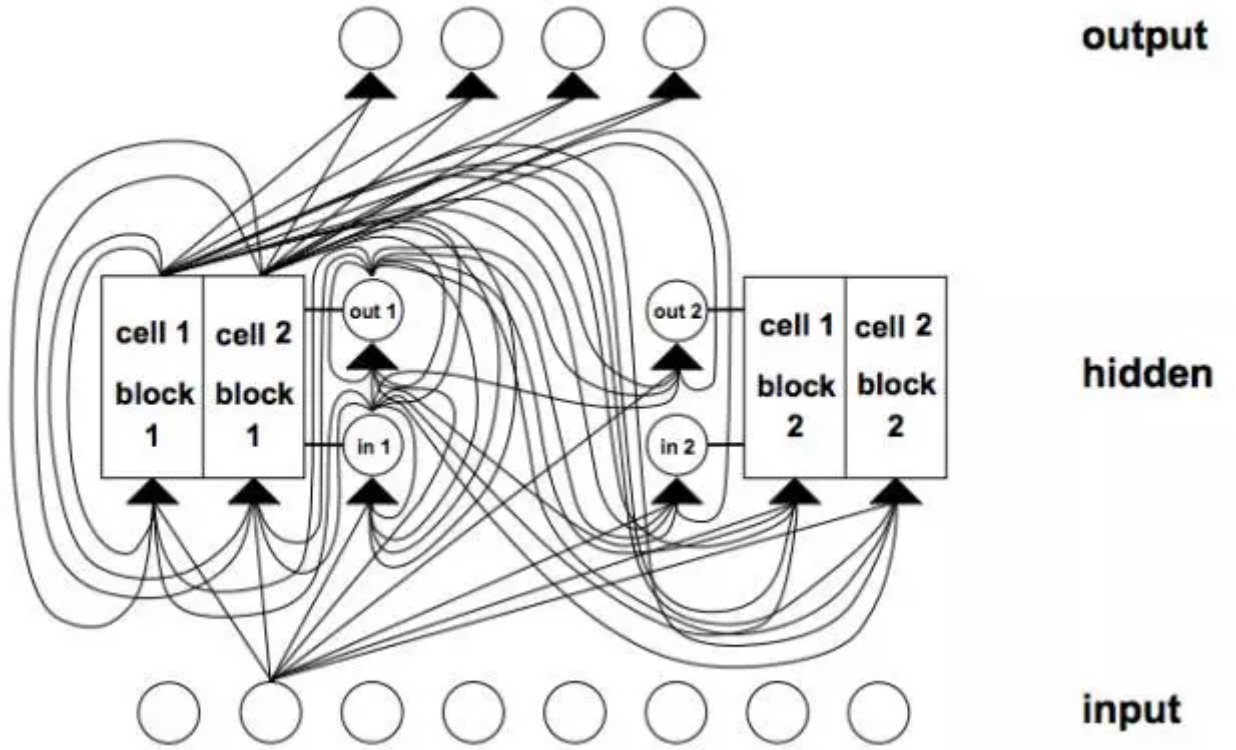
这些信号将尝试使输出权重参与访问存储在处理单元中的信息，并且在不同时间保护随后的神经元免受前馈神经元的输出的干扰。

这些冲突并不是单单造成长时滞，也同样可能造成短时滞。 值得注意的是，随着滞后的增加，存储的信息必须要确保不受扰动，特别是在学习的高级阶段。

网络架构：有关当前网络状态的有用信息，可能会通过不同类型的神经元进行传递。 例如，输入门（输出门）可以使用来自其他存储单元的输入来决定是否在其存储单元中存储（访问）某些信息。

存储单元包含门结构，并将门指定到他们要调解的连接。 输入门用于消除输入权重冲突，同时输出门可以消除输出权重冲突。

门：具体来说，为了减轻输入和输出权重的冲突及扰动，需要引入乘法输入门单元以保护存储的内容不受干扰输入的扰动，乘法输出门单元通过存储与当前不相关的存储器内容保护其他单元免受干扰。



具有8个输入单元，4个输出单元和2个大小为2的存储单元块的LSTM网络示例.in1标记输入门，out1标记输出门，cell1 = block1标记块1的第一个存储单元。

摘自Long Short-Term Memory，1997年-http://dl.acm.org/citation.cfm?id=1246450

同多层感知器相比，由于包含多样的处理元件和反馈连接，LSTM的连接性更加复杂。

存储单元块：共享相同输入门和相同输出门的存储单元所形成的结构，被称为“存储单元块”。

存储单元块有助于信息存储; 跟传统的神经网络一样，在单个小区内对分布式输入进行编码并不容易。存储单元块在大小为1时将变成一个简单的存储单元。

学习：考虑可选择的实时循环学习(RTRL)的变体，由于输入和输出门引起的乘法动态特性，要确保通过存储器单元的内部状态到达“存储器单元网络输入”的反向传播的非衰减误差不会在时间上被进一步反向传播。

猜测：这种随机方法可以胜过许多时滞算法。 已经确定的是，先前的工作中使用的许多长时滞的任务，通过简单的随机权重猜测就能比通过提出的算法更快地解决问题。

> 见S.Hochreiter和J.Schmidhuber，Long-Short Term Memory，1997。
> http://dl.acm.org/citation.cfm?id=1246450

LSTM循环神经网络最有趣的应用是语言处理工作。更全面的描述请参阅Gers的论文：

> F. Gers and J. Schmidhuber, LSTM Recurrent Networks Learn Simple Context Free and Context Sensitive Languages, 2001.
> ftp://ftp.idsia.ch/pub/juergen/L-IEEE.pdf

> F. Gers, Long Short-Term Memory in Recurrent Neural Networks, Ph.D. Thesis, 2001.
> http://www.felixgers.de/papers/phd.pdf

LSTM的不足

- LSTM的高效截断版本并不能很轻松的解决“强延迟异或”类的问题。
- LSTM的每个存储单元块需要一个输入门和一个输出门，而这在其他的循环方法中并不是必需的。
- 常数误差流通过存储单元内部的“Constant Error Carrousels”，能够跟传统的前馈架构一样，产生与整个输入串相同的效果。

- LSTM与其他前馈方法一样，在“regency”的概念上有缺陷。 如果需要精密时间步长计数，可能还需要额外的计数机制。

LSTM的优势

- 内存单元中反向传播的常数误差，赋予该架构桥接长时滞的算法的能力。
- LSTM可近似于噪声问题域、分布式表示和连续值。
- LSTM概述了要考虑的问题域。这一点很重要，因为一些任务对于已经建立的循环网络来说很棘手。
- 在问题域上不需要微调网络参数。
- 在每个权重和时间步长更新的复杂性方面，LSTM基本上等同于BPTT。
- LSTM在机器翻译等领域取得了当前最先进的结果，显示出强大的能力。

门控循环单元神经网络

与LSTM一样，门控循环神经网络已成功应用在了顺序和时间数据的处理上，尤其是在语音识别、自然语言处理和机器翻译等长序列问题领域，它都表现得十分良好。

该网络在LSTM的基础上考虑了门控，并且还涉及一个生成信号的门控网络，该信号用于控制当前输入和先前存储器如何用于更新当前激活，从而更新当前的网络状态。

门自身被赋予了权重，并且在整个学习阶段根据算法选择性地更新。

门网络以增加的复杂性的形式引入增加的计算花销，因此需要进行参数化。

LSTM RNN架构使用简单RNN的计算作为内部存储器单元（状态）的中间候选。 门控循环单元(GRU)RNN将LSTM RNN模型中的门控信号减少到两个。这两个门分别称为更新门和复位门。

GRU(和LSTM)RNN中的选通机制与RNN的参数化相似。 使用BPTT随机梯度下降来最小化损失函数,以更新对应于这些门的权重。

每个参数更新都将涉及与整个网络的状态有关的信息。 这可能会产生不利影响。

该网络在门控的概念上进行了进一步的探索，并扩展了三种新的变量门控机制。

已经考虑的三个门控变量分别是：

- GRU1，其中每个门仅使用先前的隐藏状态和偏差来计算；
- GRU2，其中每个门仅使用先前的隐藏状态计算；
- GRU3，其中每个门仅使用偏置来计算。

参数的显着降低可以通过GRU3产生的最小数量观察出来。

这三个变量和GRU RNN均使用来自MNIST数据库的手写数字和IMDB电影评论数据集的数据进行了基准测试。

结果从MNIST数据集生成了两个序列长度，从IMDB数据集生成了一个。

门的主要驱动信号似乎是（循环）"状态"，因为"状态"包含了和其他信号有关的基本信息。

随机梯度下降的使用隐含地携带有相关网络状态的信息。 这可能可以解释在门信号中单独使用偏置的相对成功，因为其自适应更新携带了有关网络状态的信息。

门控变量对门控机制进行了探索和扩展，并对拓扑结构进行了有限的评估。

更多相关信息，请参阅：

> R. Dey and F. M. Salem, Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks, 2017. https://arxiv.org/ftp/arxiv/papers/1701/1701.05923.pdf

> J. Chung, et al., Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, 2014. https://pdfs.semanticscholar.org/2d9e/3f53fcdb548b0b3c4d4efb197f164fe0c381.pdf

神经图灵机

神经图灵机通过将神经网络耦合到外部存储器资源来扩展神经网络的能力，它们可以通过attention 的过程进行交互。所谓的NTM，其实就是使用NN来实现图灵机计算模型中的读写操作。其模型中的组件与图灵机相同。

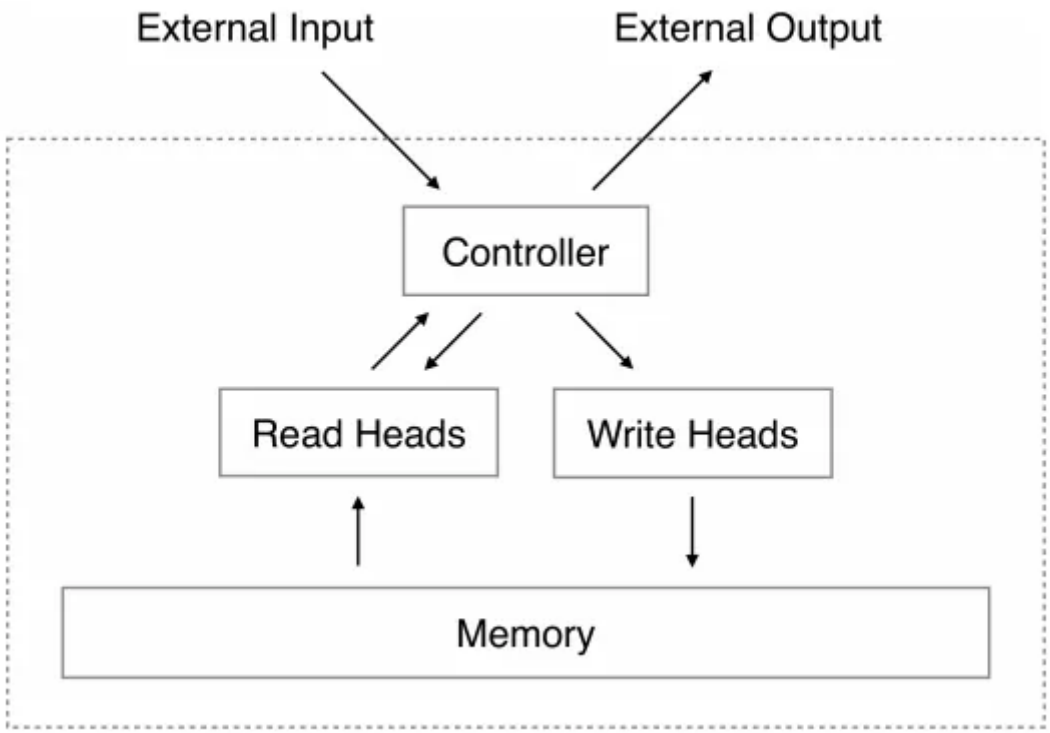
这个组合的系统类似于图灵机(Turing Neumann)或冯诺依曼(Von Neumann)结构，但是它可以端对端，通过梯度下降进行有效的训练。

初步结果表明，神经图灵机可以从输入和输出示例中推导出简单的算法，如复制、排序和联想性回忆。

RNN对长时间数据学习和进行数据转换的能力让他们从其他机器学习方法中脱颖而出。此外，因为RNN已经被证明是图灵完备的，因此只需适当地布线就能模拟任意程序。

扩展标准RNN的能力可以简化算法任务的解决方案。 因此，这种扩展主要是通过一个庞大的可寻址记忆，通过类比图灵通过无限的存储磁带来扩展有限状态机，并被称为“神经图灵机”(NTM)。

与图灵机不同，NTM是可以通过梯度下降进行训练的可微分计算机，这为学习程序提供了非常实用的机制。



NTM架构如上所示。 在每个更新周期期间，控制器网络接收来自外部环境的输入并作为响应发出输出。 它还通过一组并行读写头读取和写入存储器矩阵。 虚线表示NTM电路与外界的划分。摘自Neural Turing Machines，2014年

至关重要的是，架构的每一个组成部分都是可微分的，直接用梯度下降训练。 这是通过定义“模糊”读写操作来实现的，这些操作与内存中的所有元素或多或少地相互作用（而不是像正常的图灵机或数字计算机那样处理单个元素）。

有关更多信息，请参阅：

> A. Graves, et al., Neural Turing Machines, 2014. - https://arxiv.org/pdf/1410.5401.pdf

> R. Greve, et al., Evolving Neural Turing Machines for Reward-based Learning, 2016. - http://sebastianrisi.com/wp-content/uploads/greve_gecco16.pdf

NTM实验

复制任务可以用来测试NTM是否可以存储和调用长序列的任意信息。该测试中，网络被以随机二进制向量的输入序列，后跟分隔符标志的形式呈现。

网络需要进行训练来复制8位随机向量的序列，其中序列长度为1和20之间的随机数。目标序列仅仅是输入序列的拷贝（没有分隔符标志）。

通过要求网络将复制的序列输出指定次数来重复复制任务扩展副本，然后发出序列结束标记。该过程的主要目的是看看NTM是否可以学习一个简单的嵌套函数。

网络接收随机二进制向量的随机长度序列，随后接受出现在单独输入信道上的表示所需份数的标量值。

联想性回忆任务涉及到组织“间接”产生的数据，即一个数据项指向另一个数据项。构建项目列表来使用其中一个项目查询网络返回后续项目的要求。

接下来定义一个由分隔符符号左右限制的二进制向量序列。在将多个项目传播到网络之后，通过显示随机项目查看网络，并查看网络是否可以产生下一个项目。

动态N-gram任务用来测试是否NTM可以通过使用内存作为可重写表来快速适应新的预测分布，它可以用于保持转换统计数据，从而模拟常规的N-Gram模型。

考虑二进制序列中所有可能的6-gram分布的集合。给定所有可能的长度五位二进制的历史,每个6-gram分布可以表示为32个数字的表格，分别指定下一位将为1的概率。通过使用当前查找表绘制200个连续位来生成特定的训练序列。网络一次观察一位序列，然后会预测下一位。

优先排序任务测试NTM的排序能力。 首先将随机二进制向量序列与每个向量的标量优先级一起输入到网络中。 优先级在[-1,1]范围内均匀分布。 目标序列包含了根据优先级排序后的二进制向量。

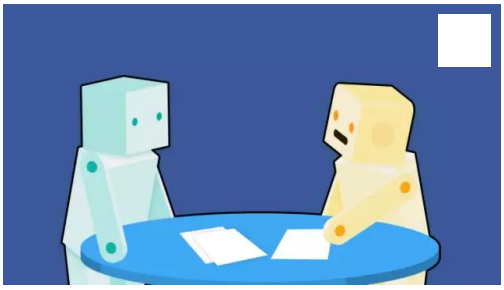
NTM有一个组件正是LSTM的前馈架构。

总结

读完本文，你应该已经理解了循环神经网络在深度学习上的用法，具体来说有以下几点：

- LSTM、GRU与NTM这些最先进的循环神经网络，是如何进行深度学习任务的
- 这些循环神经网络同人工神经网络中更广泛的递归研究间的具体关系是怎样的
- RNN能在一系列有挑战的问题上表现如此出色的原因在哪里

> 原文链接
> http://machinelearningmastery.com/recurrent-neural-network-algorithms-for-deep-learning/



Facebook中止聊天机器人项目是因为恐慌AI会自创语言了？其实你想多了