**循环神经网络之长短期记忆网络**

**摘要**

循环神经网络的目的是用来处理序列数据，在传统的神经网络模型中，从输入层到隐含层再到输出层，层之间是全连接，层中节点是无连接，这样的全连接神经网络对于自然语言处理以及时序数据表现不好，因为像这样的数据需要对之前的数据形成依赖；另外循环神经网络在做反向传播的时候，会遇到和深度神经网络类似的问题，由于步长过长，级联的梯度计算，在传递到最后的阶段会形成梯度爆炸或者梯度消失；为了解决这些问题，长短期神经网络出现，长短期记忆（LSTM）网络已被证明对于学习包含未知长度的长期模式的序列特别有用，因为它们能够维持长期记忆。 在这样的网络中堆叠循环隐藏层还使得能够学习更高级别的时间特征，以便通过更稀疏的表示来更快地学习。 在本文中，我们从循环神经网络说起，从而深入的介绍长短期记忆网络的核心原理。

**关键词**：循环神经网络，机器学习，长短期记忆网络，异常检测

**Abstract**

The purpose of the cyclic neural network is to process the sequence data. In the traditional neural network model, from the input layer to the hidden layer to the output layer, the layers are fully connected, and the nodes in the layer are connected, such a full connection. Neural networks do not perform well for natural language processing and time series data, because data like this needs to be dependent on previous data; in addition, cyclic neural networks encounter problems similar to deep neural networks when doing backpropagation, due to The step length is too long, and the cascaded gradient calculation will form a gradient explosion or gradient disappearance in the final stage of transmission; in order to solve these problems, long-term and short-term neural networks appear, long-term and short-term memory (LSTM) networks have been proved to contain unknown for learning. Sequences of long-term patterns of length are particularly useful because they maintain long-term memory. Stacking loop hidden layers in such networks also enables learning of higher level temporal features to learn faster with more sparse representations. In this paper, we talk about cyclic neural networks, long- and short-term memory networks, and introduce some long- and short-term memory networks .

**Keywords:** cyclic neural network, machine learning, long- and short-term memory network, anomaly detection

# 1引言

我们思考所有问题不会从零开始，当我们读一篇文章，看一部电影，和一个人谈话，等等，我们都不会从零开始，我们会基于过去的事实来理解当下的场景，我们不会把过去扔在一边，我们的思考具有持久性。

传统的神经网络做不到这一点，看起来，这是传统神经网络的一个主要缺点。比如说，我们可以根据看到的电影上下文开区分电影中正在播放的部分，传统神经网络无法根据电影的上下文来推断出当前的情景属于那个部分，也就是无法做到对这种时序流数据的分类。

循环神经网络解决了这个问题，循环神经网络单元创建了自身循环的单元模型，允许信息持续存在。

循环神经网络在做反向传播的时候，会遇到和深度神经网络类似的问题，由于步长过长，级联的梯度计算，在传递到最后的阶段会形成梯度爆炸或者梯度消失；

长短期记忆是一种时间递归神经网络，论文首次发表于1997年，由于独特的设计结构，长短期记忆网络适合于处理和预测时间序列中间隔和延迟非常长的充要事件。

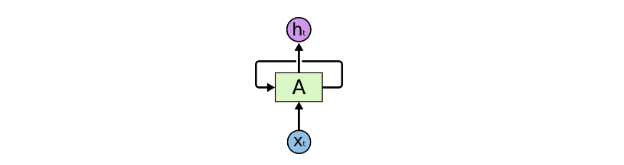
长短期记忆网络的表现通常比时间递归神经网络及隐马尔科夫模型（HMM）更好，比如用在不分段连续手写识别上。2009年，用长短期记忆网络构建的人工神经网络网络模型赢得过ICDAR手写识别比赛冠军。LSTM还普遍用于自主语音识别，2013年运用TIMIT自然演讲数据库达成17.7%错误率的记录。作为非线性模型，可作为复杂的非线性单元用于构造更大型深度神经网络。

本部分内容如下，第二部分讲述循环神经网络的原理，第三部分讲述长短期记忆网络的原理，第四部分进行总结。

# 2 循环神经网络

## 2.1 循环神经网络基本原理

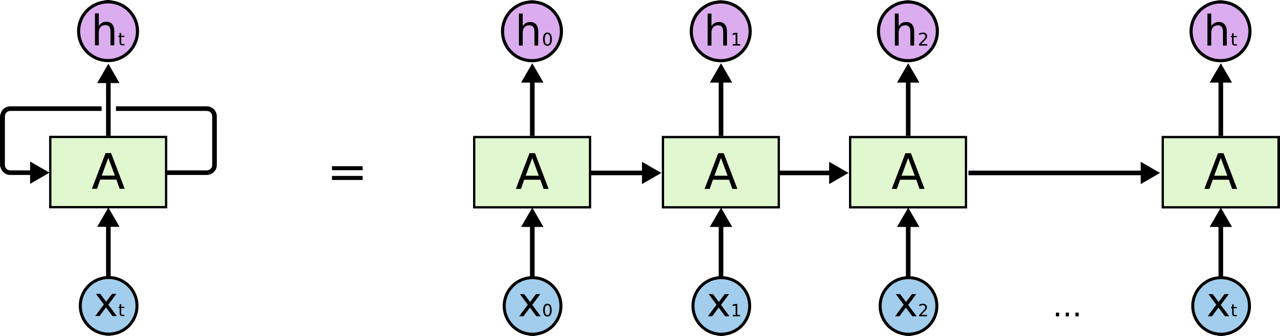
循环神经网络允许信息在自身单元中持续存在，



循环神经网络结构单元如上所示，A接受一个输入，给一个输出，循环允许这个单元能够把自己的信息传输给下一个相邻单元。

在上面定义来如何在相邻单元之间传递信息的方式之后，最基本的循环神经单元，只有一个简单的神经网络层，除了输入数据，还有前一个样本的状态，将这两个数据拼接之后，通过激活函数（tanh）的操作，输出新的状态。经过多轮次的迭代之后，就能够提取出序列间的依赖信息，但是序列的长度

循环神经网络解决了这个问题。 它们是具有循环的网络，允许信息持续存在。



这些循环使得循环神经网络看起来有点神秘。 但是，如果你多想一点，事实证明它们与普通的神经网络并没有什么不同。 可以将循环神经网络视为同一个网络单元的多个副本，每个副本都将消息传递给继承者。

这种类似链的性质表明，递归神经网络与序列和列表密切相关。它们是用于此类数据的神经网络的自然架构。

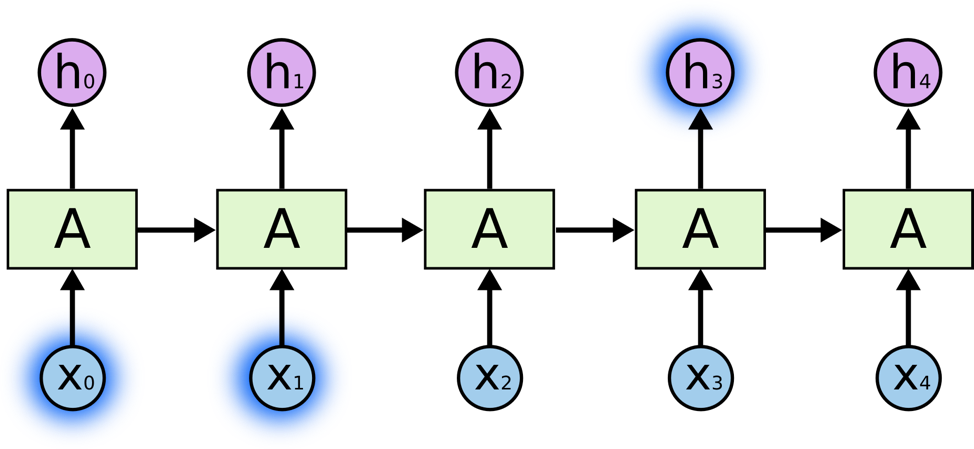
在过去几年中，将RNN应用于各个领域都取得了令人难以置信的成功：语音识别，语言建模，翻译，图像，字幕......这个应用领域还在继续扩大。

这些成功的关键在于使用“LSTM”，这是一种非常特殊的递归神经网络，对于许多任务而言，它比标准版本好得多。 几乎所有基于递归神经网络的令人兴奋的结果都是用它们实现的。 这篇论文将探讨这些LSTM。

## 2.2 长期依赖问题

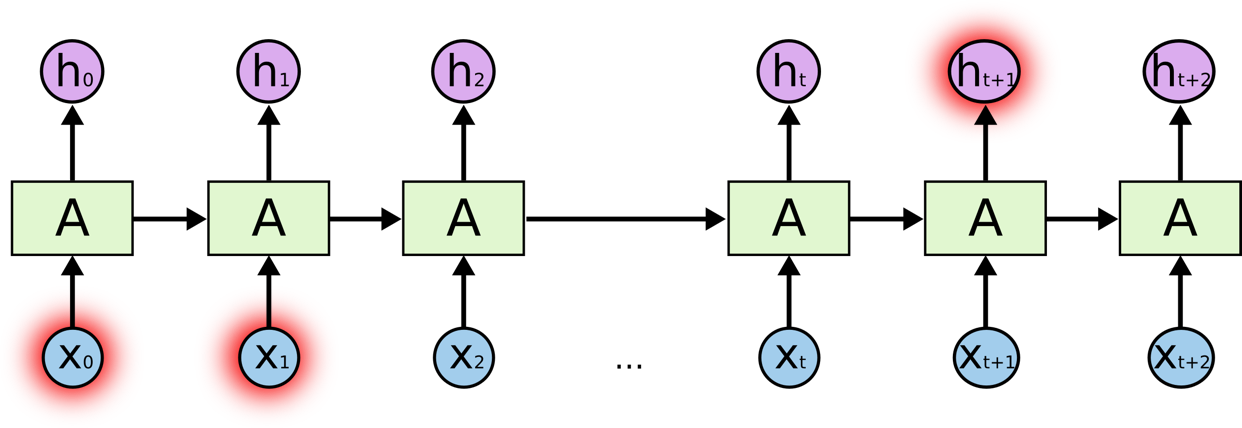
RNN的一个吸引力是他们可能能够将先前信息用于当前节点的处理，例如使用先前的视频帧可以通知对当前帧的加深理解。 如果RNN可以做到这一点，它们将非常有用。 但他们可以吗？

有时，我们只需要查看最近的信息来执行当前任务。 例如，考虑一种语言模型，试图根据之前的单词预测下一个单词。 如果我们试图预测“云在天空中”的最后一个词，我们不需要任何进一步的背景 - 很明显，下一个词将是天空。 在这种情况下，如果相关信息与所需地点之间的差距很小，RNN可以学习使用过去的信息。



但也有一些情况需要更多的背景。考虑尝试预测文本中的最后一个词“我在法国长大......我说流利的法语。”最近研究表明，下一个词可能是一种语言的名称，但如果我们继续缩小范围，我们需要更进一步的背景。 相关信息与需要变得非常大的点之间的差距完全有可能。

不幸的是，随着差距的扩大，RNN无法从很长的序列数据中学到有效的知识。



从理论上讲，RNN绝对能够处理这种“长期依赖性”。人类可以仔细挑选参数来解决这种形式的问题。但是在实践中，RNN似乎无法学习它们。 Hochreiter（1991）[德国]和Bengio等人对该问题进行了深入探讨。（1994）他找到了一些非常根本的原因，为什么RNN学习长期序列问题如此的困难。

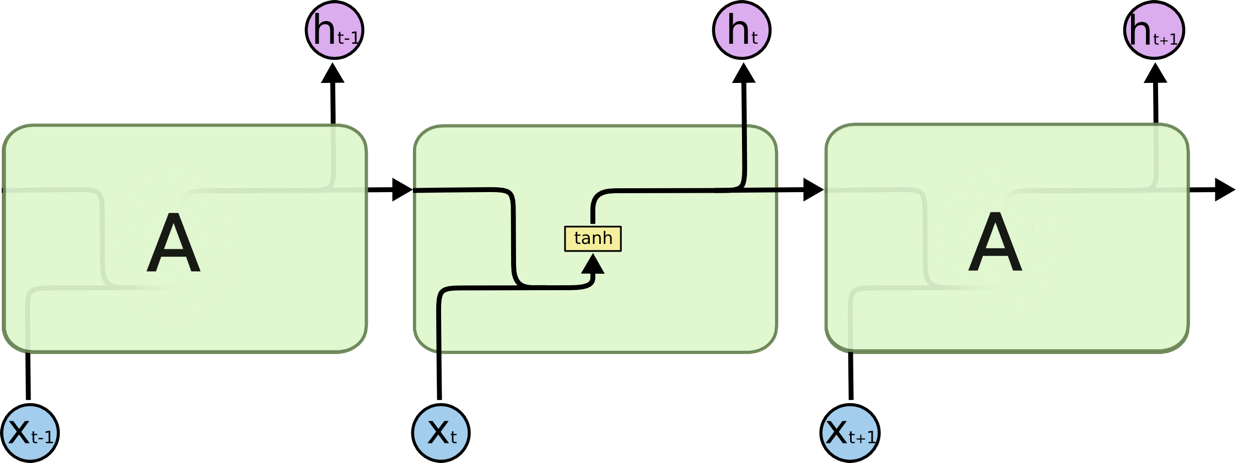
# 3 长短期记忆网络

## 3.1 基本介绍

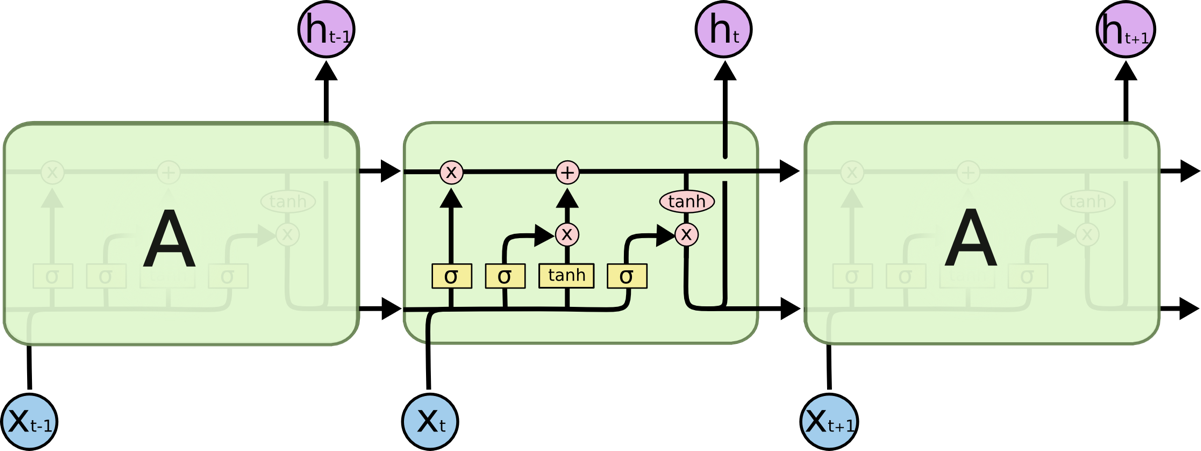
长短期内存网络 - 通常只称为“LSTM” - 是一种特殊的RNN，能够学习长期依赖性。 它们是由Hochreiter＆Schmidhuber（1997）介绍的，并且在以下工作中被许多人提炼和推广，他们在各种各样的问题上工作得非常好，现在被广泛使用。

LSTM主要目标解决长期依赖性问题。长时间记住信息实际上是他们的默认行为，就是他们自身的一种特性。

所有递归神经网络都具有神经网络重复模块链的形式。 在标准RNN中，重复模块将具有非常简单的结构，例如单个Tanh层。



LSTM也具有这种类似链的结构，但重复模块具有不同的结构。 与单个神经网络层不同，他们有四个神经网络层，以一种特殊的方式进行交互。



通过对这些图中相关组建的认识，让我们更加深入的去了解整个运行机制是怎么一回事。



在上图中，每个组件都有一个数据，从一个节点的输出到其他节点的输入。 粉色圆圈表示逐点运算，如矢量加法，而黄色框表示神经网络层。 行合并表示连接，而行分叉表示其内容被复制，副本将转移到不同的位置。

## 3.2 LSTM背后的核心理念

LSTM的关键是单元状态，水平线贯穿图的顶部。

单元状态有点像传送带，它直接沿着整个链运行，只有一些微小的线性相互作用，信息很容易沿着它不变地流动下去。



LSTM确实能够移除或添加信息到单元状态，由称为门的结构精心调节。

门是一种可选择通过信息的方式。 它们由S形神经网络层和逐点乘法运算组成。



sigmoid层输出0到1之间的数字，描述每个组件应该通过多少。 值为零意味着“不让任何东西通过”，而值为1则意味着“让一切都通过！”

LSTM具有三个这样的门，用于保护和控制电池状态。

## 3.3 LSTM计算过程

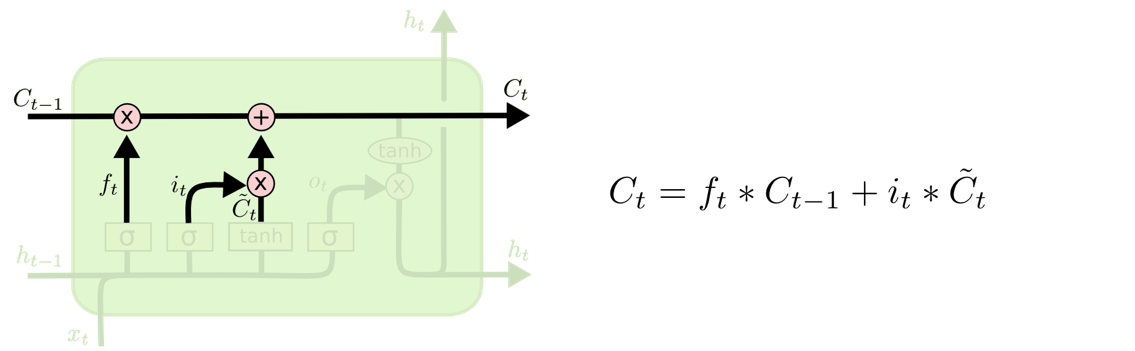
LSTM的第一步是确定我们将从细胞状态中丢弃的信息。 这个决定是由一个叫做“遗忘门”的sigmoid层决定的。输入和，对每一个单元胞输出一个处于0-1的的数字，1表示完全保留这个单元胞的信息，0表示完全遗弃这个单元胞的信息。

以一个语言模型示例，试图根据以前的所有单词预测下一个单词。在这样的问题中，单元胞状态可能包括当前受试者的性别，因此可以使用正确的代词。 当我们看到一个新主体时，我们想要忘记旧主体的性别。



下一步是确定我们将在单元状态中存储哪些新信息。 有两个部分；首先，称为“输入门层”的sigmoid层决定我们将更新哪些值。接下来，tanh层创建新候选值的向量，这可以添加到状态。在下一步中，我们将结合这两个来对状态进行更新。

在我们的语言模型的例子中，我们想要将新主体的性别添加到单元胞状态，以取代我们已经遗忘的旧的。



现在是时候更新旧的单元状态了，进入新的细胞状态。前面的步骤已经决定要做什么，我们只需要实际做到这一点。

我们将旧状态乘以,忘记了我们之前决定忘记的事情。 然后我们添加。这是新的候选值，根据我们决定更新每个状态的值来缩放。

在语言模型的情况下，我们实际上放弃了关于旧主体的性别的信息并添加新信息，正如我们在前面的步骤中所做的那样。



最后，我们需要决定我们要输出的内容。 此输出将基于我们的单元状态，但将是过滤版本。首先，我们运行一个sigmoid层，它决定我们要输出的单元状态的哪些部分。 然后，我们把细胞状态通过tanh（将值变换到-1和1之间)并将它乘以sigmoid门的输出，这样我们只输出我们决定的部分。

对于语言模型示例，由于它只是看到一个主体，它可能想要输出与动词相关的信息，以防接下来会发生什么。 例如，它可能输出主语是单数还是复数，以便我们知道动词应该与什么形式共轭。



## 3.4 长期短期记忆的变种

到目前为止我所描述的是一个正常的LSTM。但并非所有LSTM都与上述相同。 事实上，似乎几乎所有涉及LSTM的论文都使用略有不同的版本。 差异很小，但值得一提的是其中一些。

由Gers＆Schmidhuber（2000）引入的一种流行的LSTM变体是添加“窥视孔连接”。这意味着我们让栅极层看到单元状态。

上面的图表为所有门增加了窥视孔，但是许多论文会给一些窥视孔而不是其他的。

另一种变化是使用耦合的遗忘和输入门。 我们不是单独决定忘记什么以及应该添加新信息，而是一起做出这些决定。 我们只会忘记什么时候输入一些东西。 当我们忘记更旧的东西时，我们只向州输入新值。

LSTM稍微有点戏剧性的变化是由Cho等人引入的门控循环单元（GRU）。它将遗忘和输入门组合成一个“更新门”。它还合并了单元状态和隐藏状态，并进行了一些其他更改。 由此产生的模型比标准LSTM模型简单，并且越来越受欢迎。

这些只是最着名的LSTM变种中的一小部分。 还有很多其他的东西，比如Yao等人的Depth Gated RNNs。还有一些完全不同的解决长期依赖关系的方法，如Koutnik等人的Clockwork RNNs。

哪种变体最好？差异是否重要？ 格雷夫等人。对流行变体进行了很好的比较，发现它们几乎完全相同。Jozefowicz测试了超过一万个RNN架构，找到了一些在某些任务上比LSTM更好的架构。

# 结论

早些时候，人们用RNN在各个领域取得的显着成果。基本上所有这些都是使用LSTM实现的。对于大多数任务来说，它们确实有很好的效果。

作为一组方程写下来，LSTM看起来非常令人生畏。在这篇文章中希望阅读者能够明白长短期记忆网络的基本原理。

深入的了解长短期记忆网络的核心原理，在接下来，在我们对时序数据的处理过程中进行模型训练以及检测将会起到很大的作用，将会在模型调整中更加明确那一方面的问题。

# 参考文献

1. Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//Eleventh Annual Conference of the International Speech Communication Association. 2010.
2. Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//Acoustics, speech and signal processing (icassp), 2013 ieee international conference on. IEEE, 2013: 6645-6649.
3. Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling[C]//Fifteenth annual conference of the international speech communication association. 2014.
4. Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling[C]//Thirteenth annual conference of the international speech communication association. 2012.
5. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.
6. Malhotra P, Vig L, Shroff G, et al. Long short term memory networks for anomaly detection in time series[C]//Proceedings. Presses universitaires de Louvain, 2015: 89.