 将注意力机制引入RNN，解决5大应用领域的序列预测问题

[机器之心](javascript:void(0);) 2017-08-02

选自MachineLearningMastery

**作者：Jason Brownlee**

**机器之心编译**

**参与：Nurhachu Null、路雪**

编码器-解码器结构在多个领域展现出先进水平，但这种结构会将输入序列编码为固定长度的内部表征。这限制了输入序列的长度，也导致模型对特别长的输入序列的性能变差。将注意力机制引入循环神经网络帮助解决这一局限性。该方法可用于多个领域的序列预测中，包括文本翻译、语音识别等。

编码器-解码器结构流行的原因是它在众多领域展现出了当前最高水平。这种结构的局限性在于它会将输入序列编码为固定长度的内部表征。这限制了输入序列的长度，并且导致模型对特别长的输入序列的性能变差。

在这篇博客中，我们将发现可以在循环神经网络中采用注意力机制（attention mechanism）来克服这种局限性。

读完这篇博客，你会知道：

* 编码器-解码器结构和固定长度内部表征的局限性
* 让网络学会针对输出序列的每一项去注意输入序列中对应位置的注意力机制
* 带有注意力机制的循环神经网络在文本翻译、语音识别等 5 个领域中的应用。

**长序列的问题**

在编码器-解码器循环神经网络中，一系列长短期记忆网络（LSTM）学习将输入序列编码为固定长度的内部表征，另一部分长短期记忆网络读取内部表征并将它解码到输出序列中。这种结构在困难的序列预测问题（例如文本翻译）中展示出了当前最先进的水平，并迅速成为了主导方法。例如，下面这两篇论文：

* 使用神经网络的序列到序列学习（Sequence to Sequence Learning with Neural Networks, 2014）
* 使用 RNN 编码器-解码器学习短语表征，用于统计机器翻译（Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2014）

编码器-解码器结构仍然能够在很多问题上实现优秀的结果。然而，它受到了一个限制，即所有的输入序列都被强制编码成固定长度的内部向量。这一局限性限制了这些网络的性能，尤其是考虑到比较长的输入序列时，例如文本翻译中的长句子。

「这种编码器-解码器方法的一个潜在问题是神经网络需要把源句子中的所有必要信息压缩成一个固定长度的向量。这使得神经网络在处理长句子，尤其是比训练语料长的句子时会比较困难。」

——Dzmitry Bahdanau, et al.,Neural machine translation by jointly learning to align and translate, 2015

**序列中的注意力机制**

注意力机制是一个将编码器-解码器结构从固定长度的内部表征中解放出来的方法。它通过保持 LSTM 编码器对输入序列每一步的中间输出结果，然后训练模型学习如何选择性地关注输入，并将它们与输出序列中的项联系起来。换句话说，输出序列中的每一项都取决于输入序列中被选中的项。

「论文中提出的模型在翻译的时候每生成一个词，就会在源句子中的一系列位置中搜索最相关信息集中的地方。然后它会基于上下文向量以及这些源文本中的位置和之前生成的目标词来预测下一个目标词。」「……该模型将输入语句编码成向量序列，并在解码翻译的时候适应性地选择这些向量的子集。这使得神经翻译模型不必再将各种长度的源句子中的所有信息压缩成一个固定长度的向量。」

——Dzmitry Bahdanau, et al., Neural machine translation by jointly learning to align and translate (https://arxiv.org/abs/1409.0473), 2015

虽然这样做会增加模型的计算负担，但是会形成目标性更强、性能更好的模型。此外，模型还能够展示在预测输出序列的时候，如何将注意力放在输入序列上。这会帮助我们理解和分析模型到底在关注什么，以及它在多大程度上关注特定的输入-输出对。

「论文提出的方法能够直观地观察到生成序列中的每个词与输入序列中一些词的（软）对齐关系，这可以通过对标注权重的可视化来实现……每个图中矩阵的每一行代表与标注相关联的权重。由此我们可以看出在生成目标词时，源句子中的哪一个位置受到了重视。」

——Dzmitry Bahdanau, et al., Neural machine translation by jointly learning to align and translate (https://arxiv.org/abs/1409.0473), 2015

**使用大幅图片时的问题**

应用在计算机视觉问题中的卷积神经网络也面临类似问题，用特别大的图片训练模型会很困难。由此引发的结果就是在做出预测之前，图像会被大量观察，得到其近似表示（approximate impression）。

「人类感知的一个重要特点就是不会倾向于一次性处理场景的全貌，而是选择性地将注意力聚焦于视觉空间中的某些部分来获取所需信息，并且结合不同时间点的局部信息来构建整个场景的内部表征，以此指导随后的眼动和决策。」

——Recurrent Models of Visual Attention (https://arxiv.org/abs/1406.6247), 2014

这些基于 glimpse 的修正也可以被认为是注意力机制，但是并不是本文要说的注意力机制。

相关论文：

* Recurrent Models of Visual Attention, 2014
* DRAW: A Recurrent Neural Network For Image Generation, 2014
* Multiple Object Recognition with Visual Attention, 2014

**使用注意力机制进行序列预测的 5 个例子**

这一部分给出了一些将注意力机制与循环神经网络结合起来用于序列预测的具体例子。

1. 文本翻译中的注意力机制

文本翻译这个例子我们在前面已经提过了。给定一个法语句子的输入序列，将它翻译并输出英文句子。注意力机制用于观察输入序列中与输出序列每一个词相对应的具体单词。

「生成每个目标词时，我们让模型搜索一些输入单词或由编码器计算得到的单词标注，进而扩展基本的编码器-解码器结构。这让模型不再必须将整个源句子编码成一个固定长度的向量，还能让模型仅聚焦于和下一个目标词相关的信息。」

——Dzmitry Bahdanau, et al., Neural machine translation by jointly learning to align and translate (https://arxiv.org/abs/1409.0473, 2015

法语-英语翻译中的注意力编译

图注：列为输入序列，行为输出序列，高亮块代表二者的关联，颜色越浅，则关联性越强。

图片来自论文：Dzmitry Bahdanau, et al., Neural machine translation by jointly learning to align and translate, 2015

2. 图像描述中的注意力机制

与 glimpse 方法不同，基于序列的注意力机制可以应用在计算机视觉问题上，来帮助找出方法，使输出序列时更好地利用卷积神经网络来关注输入的图片，例如在典型的图像描述任务中。给定一幅输入图像，输出对该图像的英文描述。注意力机制用于关注与输出序列中的每一个词相关的局部图像。

「我们提出了一个基于注意力机制的方法，该方法在三个基准数据集上都达到了最先进的性能表现……我们还展示了如何使用学得的注意力机制为模型生成过程提供更多可解释性，演示了学得的对齐与人类直觉一致性很高。」

—— Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, 2016

输出单词与输入图像特定区域的注意力编译

和上图类似，输出文本中下划线处单词对应右侧图片中的泛光区域。图片来自论文：Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, 2016

3. 语义蕴涵中的注意力机制

给定一个前提场景，并且用英文给出关于该场景的假设，输出内容是前提和假设是否矛盾、二者是否相互关联，或者前提是否蕴涵假设。

例如：

* 前提:「婚礼上的照片」
* 假设:「某人在结婚」

注意力机制用于将假设中的每一个词与前提中的词关联起来，反之亦然。

我们提出了一个基于 LSTM 的神经模型，它能够将两个句子读取成一个句子来进行语义蕴涵分析，而不是把每个句子都独立地编码成一个语义向量。然后，我们用逐词注意力机制（neural word-by-word attention mechanism）来扩展该模型，以鼓励对成对单词和词组是否存在蕴涵关系作出推理……该扩展模型的基准测试分数比 LSTM 高了 2.6%，创造了一项新的准确率记录……

——Reasoning about Entailment with Neural Attention (https://arxiv.org/abs/1509.06664), 2016

前提中的单词到假设中单词的注意力编译。图片来自论文：Reasoning about Entailment with Neural Attention, 2016

4. 语音识别中的注意力机制

给定一个英文语音片段作为输入，输出一个音素序列。注意力机制被用来关联输出序列中的每一个音素和输入序列中特定的语音帧。

「……提出一个新型的基于混合注意力机制的端到端可训练的语音识别结构，为了在解码的时候在输入序列中选择下一个位置，它同时结合了内容信息和位置信息。该模型令人满意的一点是它能够识别比训练所用的语料更长的语音。」

——Attention-Based Models for Speech Recognition (https://arxiv.org/abs/1506.07503), 2015.

输出音素到输入语音帧的注意力编译。图片来自论文：Attention-Based Models for Speech Recognition, 2015.

5. 文本摘要中的注意力机制

给定一段英文文章作为输入序列，输出一段英文文本来总结输入序列。注意力机制被用来关联摘要文本中的每一个词语与源文本中的对应单词。

「……提出一个基于 neutral attention 机制的模型用于抽象摘要，该模型在最近神经机器翻译的进展基础上发展而来。我们将该概率模型与一个能够生成准确抽象摘要的生成算法结合起来。」

——A Neural Attention Model for Abstractive Sentence Summarization (https://arxiv.org/abs/1509.00685), 2015

输入文本中的单词与输出摘要的注意力编译。图片来自论文：A Neural Attention Model for Abstractive Sentence Summarization, 2015.

**延伸阅读**

如果你对在 LSTM 中添加注意力机制感兴趣，可以阅读下面的内容：

* 深度学习和自然语言处理中的注意力和记忆（Attention and memory in deep learning and NLP，http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp/）
* 注意力机制（Attention Mechanism，https://blog.heuritech.com/2016/01/20/attention-mechanism/）
* 基于注意力机制的模型在自然语言处理中的应用（Survey on Attention-based Models Applied in NLP，http://yanran.li/peppypapers/2015/10/07/survey-attention-model-1.html）
* [Quora 问答] 将注意力机制引入 RNN 究竟是什么？（What is exactly the attention mechanism introduced to RNN? https://www.quora.com/What-is-exactly-the-attention-mechanism-introduced-to-RNN-recurrent-neural-network-It-would-be-nice-if-you-could-make-it-easy-to-understand）
* 神经网络中的注意力机制是什么？（What is Attention Mechanism in Neural Networks? https://www.quora.com/What-is-Attention-Mechanism-in-Neural-Networks）

**总结**

这篇博文介绍了在 LSTM 循环神经网络中使用注意力机制用于序列预测。

具体而言：

* 循环神经网络中的编码器-解码器结构使用固定长度的内部表征，给超长序列的学习带来了限制。
* 通过允许网络学习将输出序列中的每一项与输入序列中的相关项相对应，注意力机制克服了编码器-解码器结构的这种局限性。
* 这种方法在多种序列预测问题中得到应用，包括文本翻译、语音识别等。

*原文链接：http://machinelearningmastery.com/attention-long-short-term-memory-recurrent-neural-networks/*