理解LSTM/RNN中的Attention机制

Posted on 2017-07-03 | In Deep Learning | 20116 Views

导读

目前采用编码器-解码器 (Encode-Decode) 结构的模型非常热门，是因为它在许多领域较其他的传统模型方法都取得了更好的结果。这种结构的模型通常将输入序列编码成一个固定长度的向量表示，对于长度较短的输入序列而言，该模型能够学习出对应合理的向量表示。然而，这种模型存在的问题在于：**当输入序列非常长时，模型难以学到合理的向量表示**。

在这篇博文中，我们将探索加入LSTM/RNN模型中的attention机制是如何克服传统编码器-解码器结构存在的问题的。

通过阅读这篇博文，你将会学习到：

- 传统编码器-解码器结构存在的问题及如何将输入序列编码成固定的向量表示；
- Attention机制是如何克服上述问题的，以及在模型输出时是如何考虑输出与输入序列的每一项关系的；
- 基于attention机制的LSTM/RNN模型的5个应用领域：机器翻译、图片描述、语义蕴涵、语音识别和文本摘要。

让我们开始学习吧。

长输入序列带来的问题

使用传统编码器-解码器的RNN模型先用一些LSTM单元来对输入序列进行学习，编码为固定长度的向量表示；然后再用一些LSTM单元来读取这种向量表示并解码为输出序列。

采用这种结构的模型在许多比较难的序列预测问题（如文本翻译）上都取得了最好的结果，因此迅速成为了目前的主流方法。

例如：

- Sequence to Sequence Learning with Neural Networks, 2014
- Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2014

这种结构在很多其他的领域上也取得了不错的结果。然而，它存在一个问题在于：**输入序列不论长短都会被编码成一个固定长度的向量表示，而解码则受限于该固定长度的向量表示。**

这个问题限制了模型的性能，尤其是**当输入序列比较长时，模型的性能会变得很差**（在文本翻译任务上表现为待翻译的原始文本长度过长时翻译质量较差）。

> “一个潜在的问题是，采用编码器-解码器结构的神经网络模型需要将输入序列中的必要信息表示为一个固定长度的向量，而当输入序列很长时则难以保留全部的必要信息（因为太多），尤其是当输入序列的长度比训练数据集中的更长时。”
>
> — Dzmitry Bahdanau, et al., Neural machine translation by jointly learning to align and translate, 2015

使用attention机制

Attention机制的基本思想是，**打破了传统编码器-解码器结构在编解码时都依赖于内部一个固定长度向量的限制。**

Attention机制的实现是**通过保留LSTM编码器对输入序列的中间输出结果，然后训练一个模型来对这些输入进行选择性的学习并且在模型输出时将输出序列与之进行关联。**

换一个角度而言，输出序列中的每一项的生成概率取决于在输入序列中选择了哪些项。

> “在文本翻译任务上，使用attention机制的模型每生成一个词时都会在输入序列中找出一个与之最相关的词集合。之后模型根据当前的上下文向量 (context vectors) 和所有之前生成出的词来预测下一个目标词。
>
> … 它将输入序列转化为一堆向量的序列并自适应地从中选择一个子集来解码出目标翻译文本。这感觉上像是用于文本翻译的神经网络模型需要“压缩”输入文本中的所有信息为一个固定长度的向量，不论输入文本的长短。”

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

虽然模型使用attention机制之后会增加计算量，但是性能水平能够得到提升。另外，使用attention机制便于理解在模型输出过程中输入序列中的信息是如何影响最后生成序列的。这有助于我们更好地理解模型的内部运作机制以及对一些特定的输入-输出进行debug。

> “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对齐关系，这可以通过对标注 (annotations) 权重参数可视化来实现…每个图中矩阵的每一行表示与标注相关联的权重。由此我们可以看出在生成目标词时，源句子中的位置信息会被认为更重要。”
>
> — Dzmitry Bahdanau, et al., Neural machine translation by jointly learning to align and translate, 2015

大型图片带来的问题

被广泛应用于计算机视觉领域的卷积神经网络模型同样存在类似的问题： 对于特别大的图片输入，模型学习起来比较困难。

由此，一种启发式的方法是将在模型做预测之前先对大型图片进行某种近似的表示。

> “人类的感知有一个重要的特性是不会立即处理外界的全部输入，相反的，人类会将注意力专注于所选择的部分来得到所需要的信息，然后结合不同时间段的局部信息来建立一个内部的场景表示，从而引导眼球的移动及做出决策。”
>
> — Recurrent Models of Visual Attention, 2014

这种启发式方法某种程度上也可以认为是考虑了attention，但在这篇博文中，这种方法并不认为是基于attention机制的。

基于attention机制的相关论文如下：

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

基于attention模型的应用实例

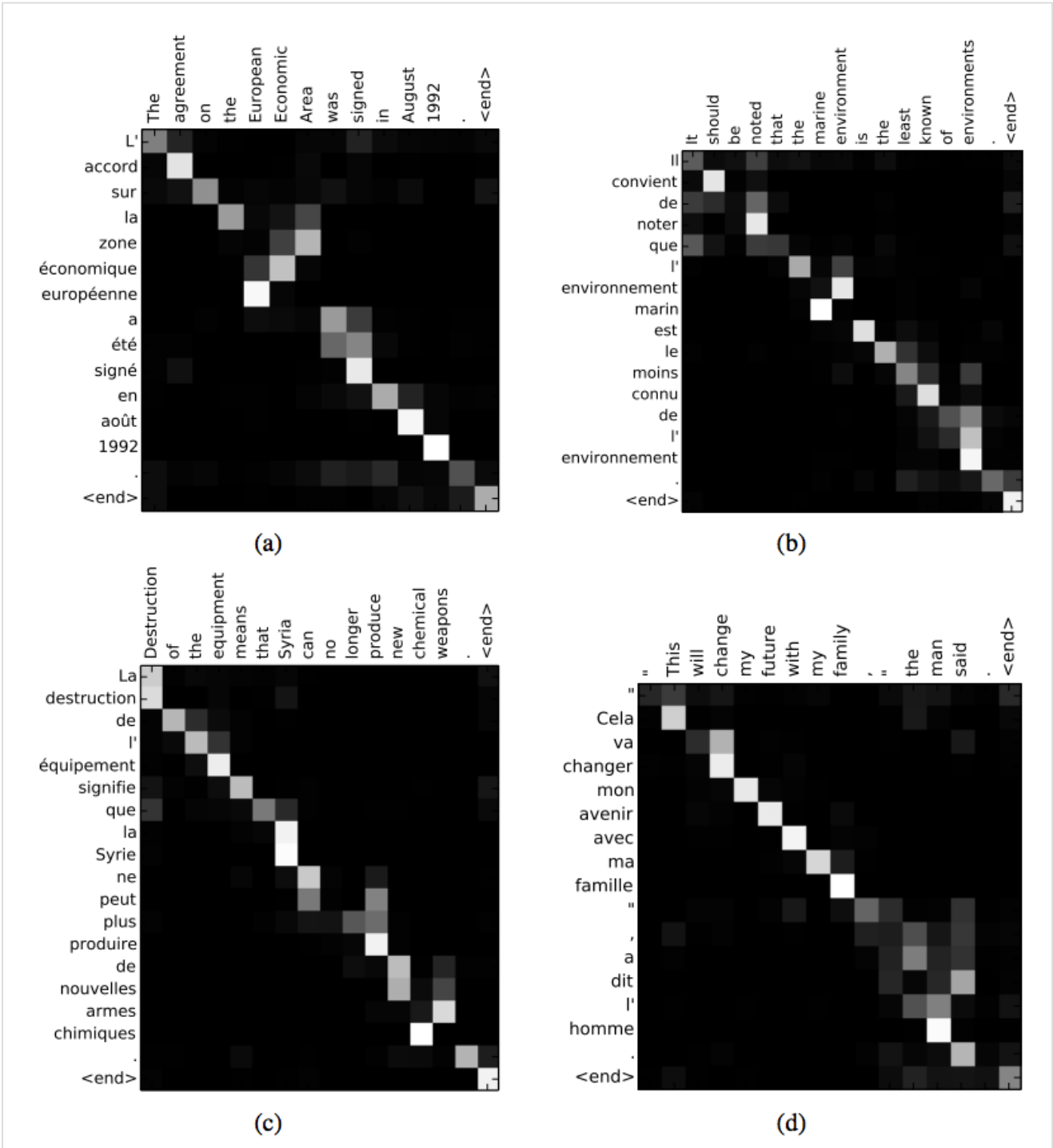
这部分将列举几个具体的应用实例，介绍attention机制是如何用在LSTM/RNN模型来进行序列预测的。

1. Attention在文本翻译任务上的应用

文本翻译这个实例在前面已经提过了。

给定一个法语的句子作为输入序列，需要输出翻译为英语的句子。Attention机制被用在输出输出序列中的每个词时会专注考虑输入序列中的一些被认为比较重要的词。

> 我们对原始的编码器-解码器模型进行了改进，使其有一个模型来对输入内容进行搜索，也就是说在生成目标词时会有一个编码器来做这个事情。这打破了之前的模型是基于将整个输入序列强行编码为一个固定长度向量的限制，同时也让模型在生成下一个目标词时重点考虑输入中相关的信息。
>
> — Dzmitry Bahdanau, et al., Neural machine translation by jointly learning to align and translate, 2015



Attention在文本翻译任务（输入为法语文本序列，输出为英语文本序列）上的可视化（图片来源于Dzmitry Bahdanau, et al., Neural machine translation by jointly learning to align and translate, 2015）

2. Attention在图片描述上的应用

与之前启发式方法不同的是，基于序列生成的attention机制可以应用在计算机视觉相关的任务上，帮助卷积神经网络重点关注图片的一些局部信息来生成相应的序列，典型的任务就是对一张图片进行文本描述。

给定一张图片作为输入，输出对应的英文文本描述。Attention机制被用在输出输出序列的每个词时会专注考虑图片中不同的局部信息。

> 我们提出了一种基于attention的方法，该方法在3个标准数据集上都取得了最佳的结果……同时展现了attention机制能够更好地帮助我们理解模型地生成过程，模型学习到的对齐关系与人类的直观认知非常的接近（如下图）。
>
> — Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, 2016

Attention在图片描述任务（输入为图片，输出为描述的文本）上的可视化（图片来源于Attend and Tell: Neural Image Caption Generation with Visual Attention, 2016）

3. Attention在语义蕴涵 (Entailment) 中的应用

给定一个用英文描述的前提和假设作为输入，输出假设与前提是否矛盾、是否相关或者是否成立。

举个例子：

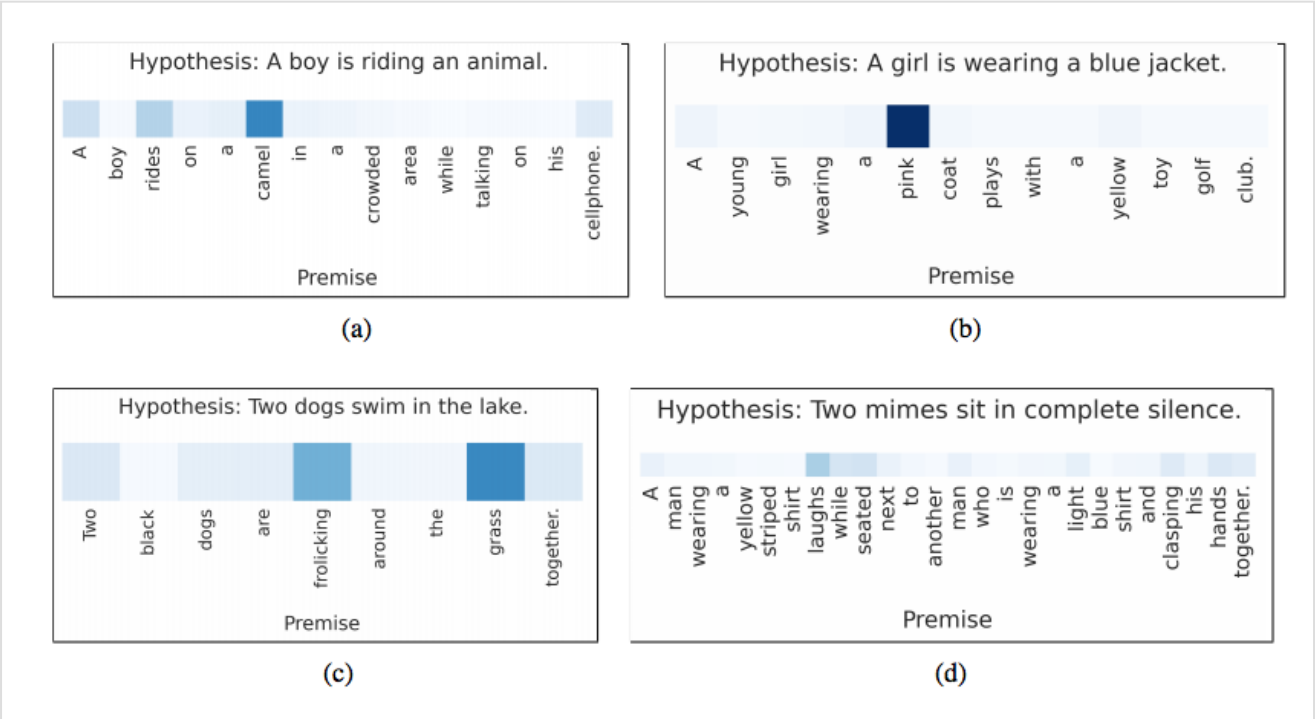
前提：在一个婚礼派对上拍照

假设：有人结婚了

该例子中的假设是成立的。

Attention机制被用于关联假设和前提描述文本之间词与词的关系。

> 我们提出了一种基于LSTM的神经网络模型，和把每个输入文本都独立编码为一个语义向量的模型不同的是，该模型同时读取前提和假设两个描述的文本序列并判断假设是否成立。我们在模型中加入了attention机制来找出假设和前提文本中词/短语之间的对齐关系。……加入attention机制能够使模型在实验结果上有2.6个点的提升，这是目前数据集上取得的最好结果…
>
> — Reasoning about Entailment with Neural Attention, 2016



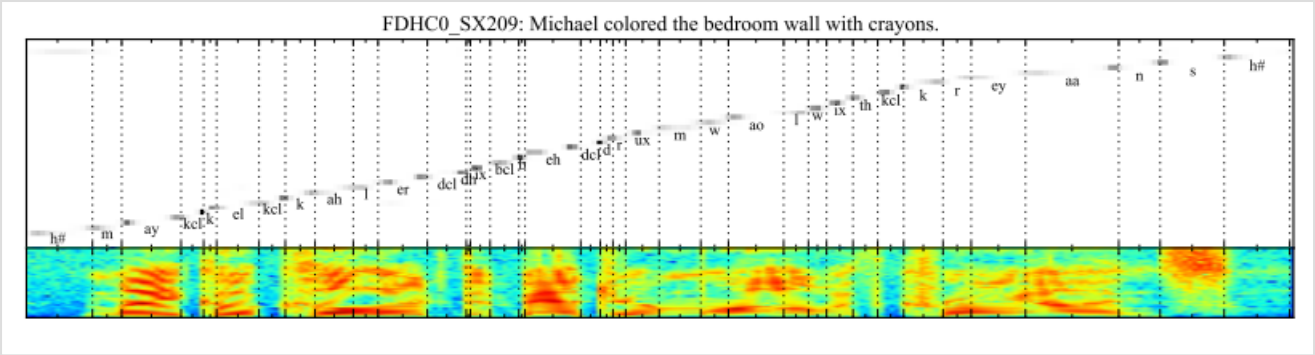
Attention在语义蕴涵任务（输入是前提文本，输出是假设文本）上的可视化（图片来源于Reasoning about Entailment with Neural Attention, 2016）

4. Attention在语音识别上的应用

给定一个英文的语音片段作为输入，输出对应的音素序列。

Attention机制被用于对输出序列的每个音素和输入语音序列中一些特定帧进行关联。

> …一种基于attention机制的端到端可训练的语音识别模型，能够结合文本内容和位置信息来选择输入序列中下一个进行编码的位置。该模型有一个优点是能够识别长度比训练数据长得多的语音输入。
>
> — Attention-Based Models for Speech Recognition, 2015.



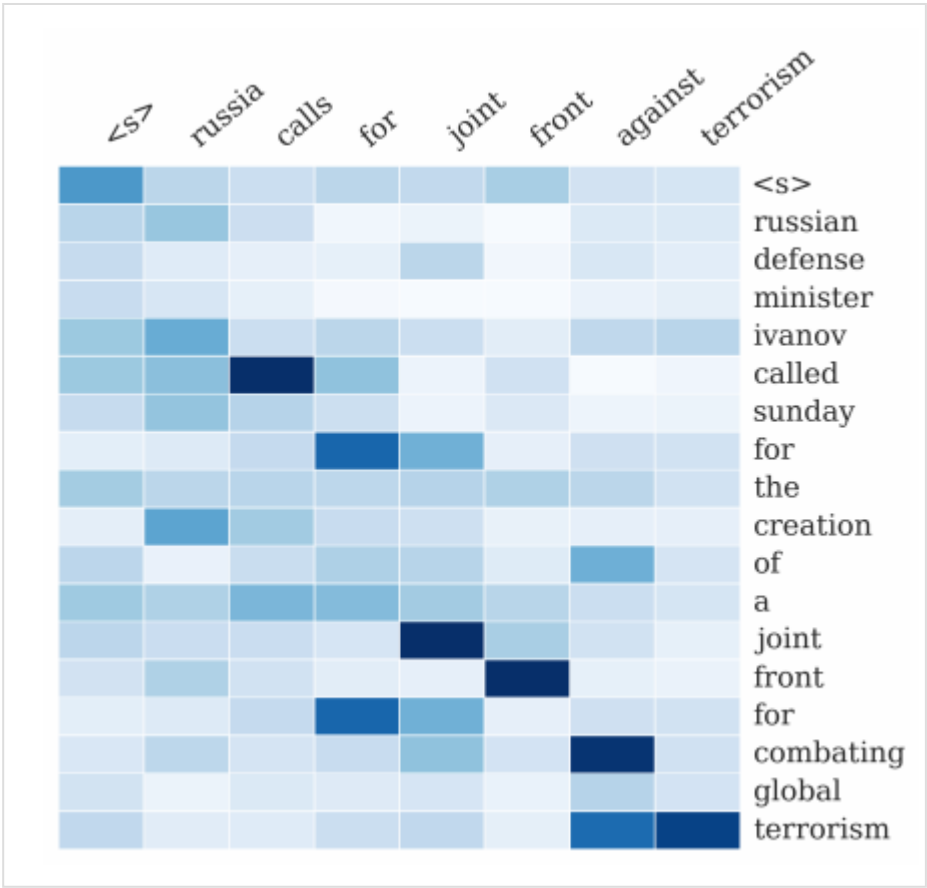
Attention在语音识别任务（输入是音帧，输出是音素的位置）上的可视化（图片来源于Attention-Based Models for Speech Recognition, 2015）

5. Attention在文本摘要上的应用

给定一篇英文文章作为输入序列，输出一个对应的摘要序列。

Attention机制被用于关联输出摘要中的每个词和输入中的一些特定词。

> … 在最近神经网络翻译模型的发展基础之上，提出了一个用于生成摘要任务的基于attention的神经网络模型。通过将这个概率模型与一个生成式方法相结合来生成出准确的摘要。
>
> — A Neural Attention Model for Abstractive Sentence Summarization, 2015



Attention在文本摘要任务（输入为文章，输出为文本摘要）上的可视化（图片来源于A Neural Attention Model for Abstractive Sentence Summarization, 2015）

进一步的阅读

如果你想进一步地学习如何在LSTM/RNN模型中加入attention机制，可阅读以下论文：

- Attention and memory in deep learning and NLP
- Attention Mechanism
- Survey on Attention-based Models Applied in NLP
- What is exactly the attention mechanism introduced to RNN?（来自Quora）
- What is Attention Mechanism in Neural Networks?

目前Keras官方还没有单独将attention模型的代码开源，下面有一些第三方的实现：

- Deep Language Modeling for Question Answering using Keras
- Attention Model Available!
- Keras Attention Mechanism
- Attention and Augmented Recurrent Neural Networks
- How to add Attention on top of a Recurrent Layer (Text Classification)
- Attention Mechanism Implementation Issue
- Implementing simple neural attention model (for padded inputs)
- Attention layer requires another PR
- seq2seq library

总结

通过这篇博文，你应该学习到了attention机制是如何应用在LSTM/RNN模型中来解决序列预测存在的问题。

具体而言，采用传统编码器-解码器结构的LSTM/RNN模型存在一个问题：不论输入长短都将其编码成一个固定长度的向量表示，这使模型对于长输入序列的学习效果很差（解码效果很差）。而attention机制则克服了上述问题，原理是在模型输出时会选择性地专注考虑输入中的对应相关的信息。使用attention机制的方法被广泛应用在各种序列预测任务上，包括文本翻译、语音识别等。

本文结束，感谢欣赏。

感谢原作者**Jason Brownlee**。原文链接见：**Attention in Long Short-Term Memory Recurrent Neural Networks**