

模型汇总24 - 深度学习中Attention Mechanism详细介绍：原理、分类及应用



lqfarmer

深度学习研究员。欢迎扫描头像二维码，关注微信公众号：深度学习与NLP

关注他

373 人赞了该文章

Attention是一种用于提升基于RNN (LSTM或GRU) 的Encoder + Decoder模型的效果的的机制 (Mechanism)，一般称为Attention Mechanism。Attention Mechanism目前非常流行，广泛应用于机器翻译、语音识别、图像标注 (Image Caption) 等很多领域，之所以它这么受欢迎，是因为Attention给模型赋予了区分辨别的能力，例如，在机器翻译、语音识别应用中，为句子中的每个词赋予不同的权重，使神经网络模型的学习变得更加灵活 (soft)，同时Attention本身可以做作为一种对齐关系，解释翻译输入/输出句子之间的对齐关系，解释模型到底学到了什么知识，为我们打开深度学习的黑箱，提供了一个窗口，如图1所示。

▲ 赞同 373

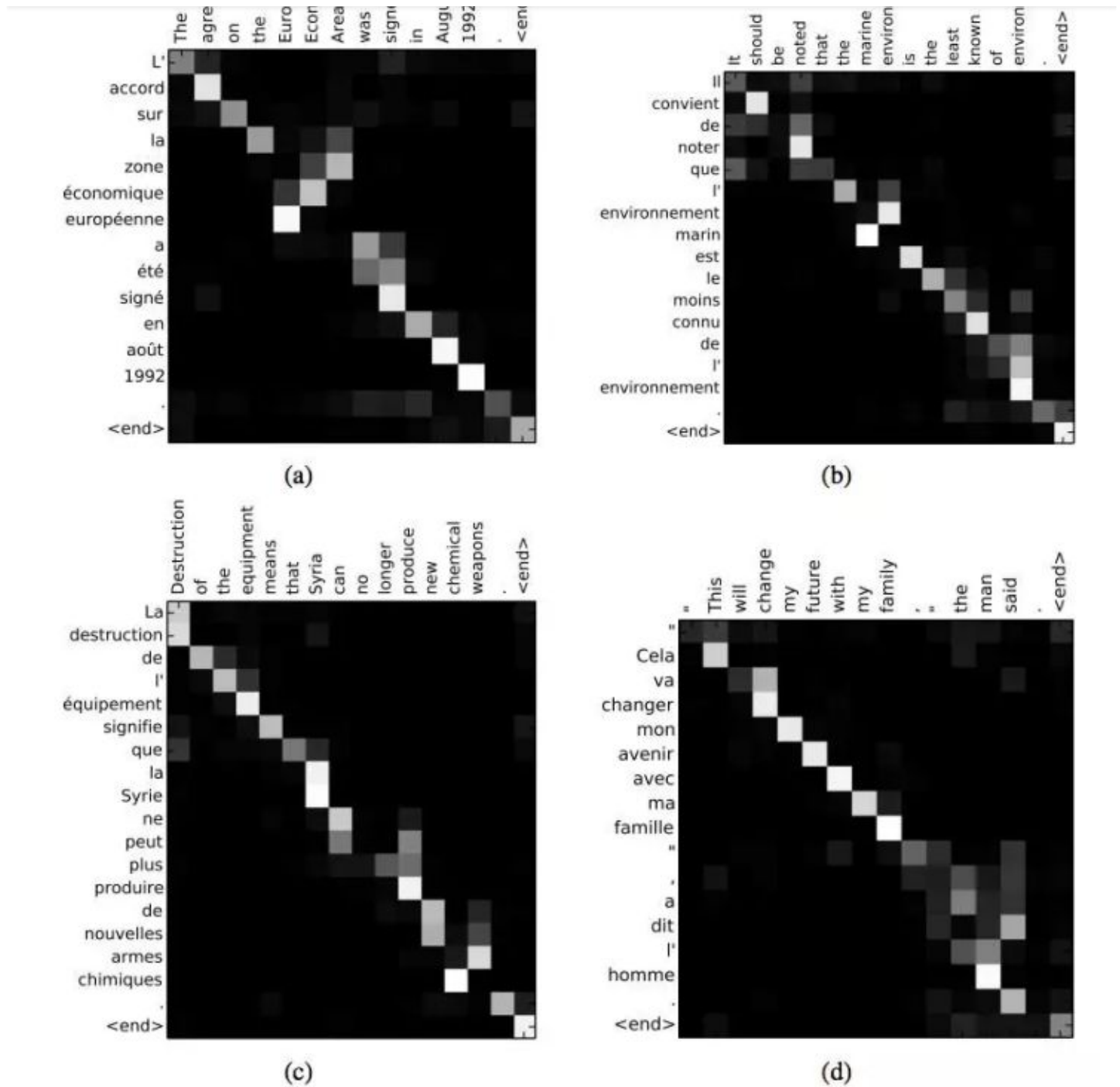


图1 NLP中的attention可视化

又比如在图像标注应用中，可以解释图片不同的区域对于输出Text序列的影响程度。

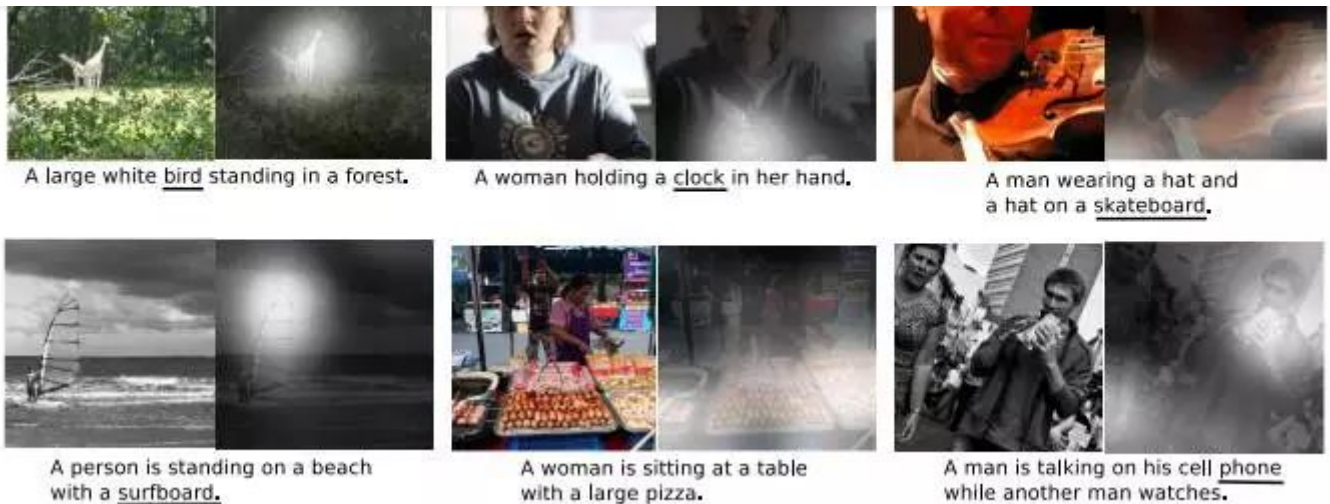


图2 图像标注中的attention可视化

通过上述Attention Mechanism在图像标注应用的case可以发现，Attention Mechanism与人类对外界事物的观察机制很类似，当人类观察外界事物的时候，一般不会对事物当成一个整体去看，往往倾向于根据需要进行选择性的去获取被观察事物的某些重要部分，比如我们看到一个人时，往往先Attention到这个人的脸，然后再把不同区域的信息组合起来，形成一个对被观察事物的整体印象。**因此，Attention Mechanism可以帮助模型对输入的X每个部分赋予不同的权重，抽取出来更加关键及重要的信息，使模型做出更加准确的判断，同时不会对模型的计算和存储带来更大的开销，这也是Attention Mechanism应用如此广泛的原因。**

有了这些背景知识的铺垫，接下来就——介绍下Attention Mechanism其他细节，在接写来的内容里，我会主要介绍以下一些知识：

1. Attention Mechanism原理

1.1 Attention Mechanism主要需要解决的问题

1.2 Attention Mechanism原理

2. Attention Mechanism分类

基本attention结构

2.1 soft Attention 与Hard Attention

2.2 Global Attention 和 Local Attention

2.3 Self Attention





2.4 Hierarchical Attention

2.5 Attention in Attention

2.3 Multi-Step Attention

3. Attention的应用场景

3.1 机器翻译 (Machine Translation)

3.2 图像标注 (Image Caption)

3.3 关系抽取 (Entailment Extraction)

3.4 语音识别 (Speech Recognition)

3.5 自动摘要生成 (Text Summarization)

1. Attention Mechanism原理

1.1 Attention Mechanism主要需要解决的问题

《Sequence to Sequence Learning with Neural Networks》介绍了一种基于RNN的Seq2Seq模型，基于一个Encoder和一个Decoder来构建基于神经网络的End-to-End的机器翻译模型，其中，Encoder把输入X编码成一个固定长度的隐向量Z，Decoder基于隐向量Z解码出目标输出Y。这是一个非常经典的序列到序列的模型，但是却存在**两个明显的问题**：

- 1、把输入X的所有信息有压缩到一个固定长度的隐向量Z，忽略了输入X的长度，当输入句子长度很长，特别是比训练集中最初的句子长度还长时，模型的性能急剧下降。
- 2、把输入X编码成一个固定的长度，对于句子中每个词都赋予相同的权重，这样做是不合理的，比如，在机器翻译里，输入的句子与输出句子之间，往往是输入一个或几个词对应于输出的一个或几个词。因此，对输入的每个词赋予相同权重，这样做没有区分度，往往是模型性能下降。

同样的问题也存在于图像识别领域，卷积神经网络CNN对输入的图像每个区域做相同的处理，这样做没有区分度，特别是当处理的图像尺寸非常大时，问题更明显。因此，2015年，Dzmitry Bahdanau等人在《Neural machine translation by jointly learning to align and translate》提出了Attention Mechanism，用于对输入X的不同部分赋予不同的权重，进而实现软区分的目的。

1.2 Attention Mechanism原理





在论文《Sequence to Sequence Learning with Neural Networks》中使用LSTM来搭建Seq2Seq模型。随后，2015年，Kyunghyun Cho等人在论文《Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation》提出了基于GRU的Seq2Seq模型。两篇文章所提出的Seq2Seq模型，想要解决的主要问题是，如何把机器翻译中，变长的输入 X 映射到一个变长输出 Y 的问题，其主要结构如图3所示。

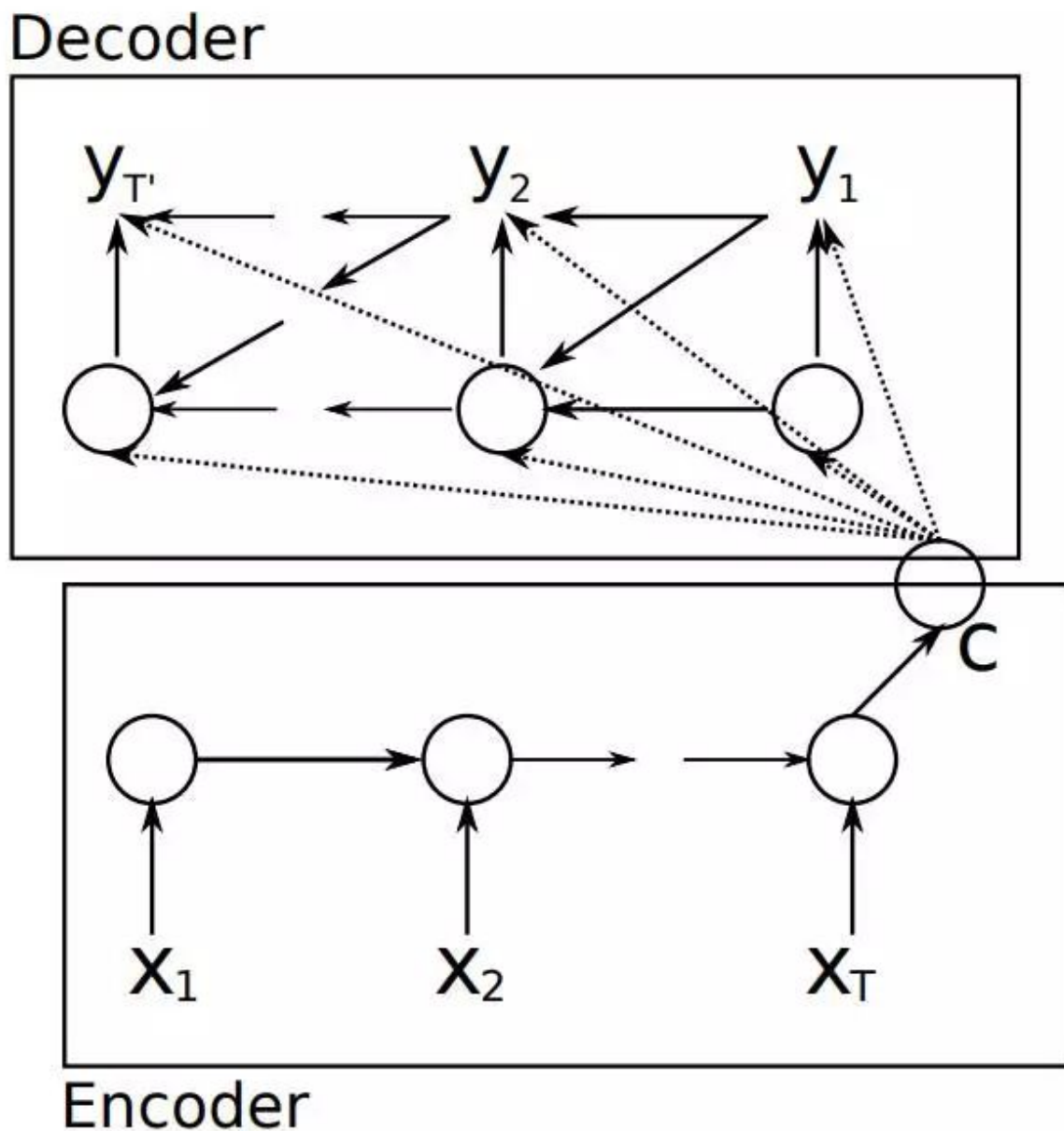


图3 传统的Seq2Seq结构

其中，Encoder把一个变长的输入序列 $x_1, x_2, x_3, \dots, x_t$ 编码成一个固定长度隐向量（背景向量，或上下文向量context） c ， c 有两个作用：1、做为初始向量初始化Decoder的模型，做为decoder模型预测 y_1 的初始向量。2、做为背景向量，指导 y 序列中每一个step的 y 的产出。Decoder主要基于背景向量 c 和上一步的输出 y_{t-1} 解码得到该时刻 t 的输出 y_t ，直到碰到结束标志（ $\langle \text{EOS} \rangle$ ）为止。





Machine Translation》中，引入了Attention Mechanism来解决这个问题，他们提出的模型结构如图4所示。

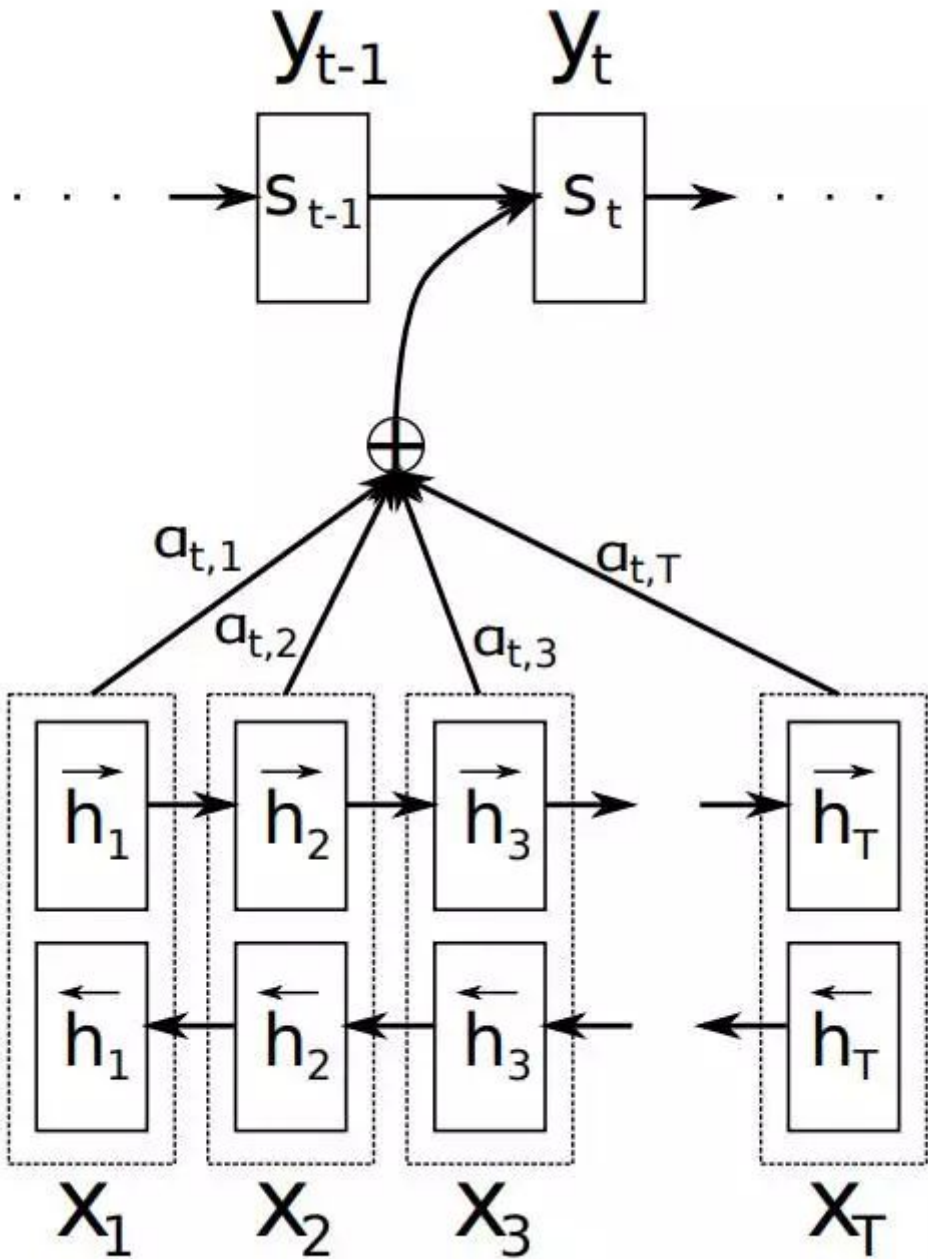


图4 Attention Mechanism模块图解

在该模型中，定义了一个条件概率：

$$p(y_i|y_1, \dots, y_{i-1}, \mathbf{X}) = g(y_{i-1}, s_i, c_i), \tag{4}$$

其中， s_i 是decoder中RNN在在*i*时刻的隐状态，如图4中所示，其计算公式为：

$$s_i = f(s_{i-1}, y_{i-1}, c_i).$$





$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j. \quad (5)$$

其中， i 表示encoder端的第 i 个词， h_j 表示encoder端的第 j 个词的隐向量， α_{ij} 表示encoder端的第 j 个词与decoder端的第 i 个词之间的权值，表示源端第 j 个词对目标端第 i 个词的影响程度， α_{ij} 的计算公式如公式6所示：

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}, \quad (6)$$

$$e_{ij} = a(s_{i-1}, h_j)$$

在公式6中， α_{ij} 是一个softmax模型输出，概率值的和为1。 e_{ij} 表示一个对齐模型，用于衡量encoder端的位置 j 个词，对于decoder端的位置 i 个词的对齐程度（影响程度），换句话说：decoder端生成位置 i 的词时，有多少程度受encoder端的位置 j 的词影响。对齐模型 e_{ij} 的计算方式有很多种，不同的计算方式，代表不同的Attention模型，最简单且最常用的对齐模型是dot product乘积矩阵，即把target端的输出隐状态 h_t 与source端的输出隐状态进行矩阵乘。常见的对齐计算方式如下：

$$\text{score}(h_t, \bar{h}_s) = \begin{cases} h_t^\top \bar{h}_s & \text{dot} \\ h_t^\top W_a \bar{h}_s & \text{general} \\ v_a^\top \tanh(W_a [h_t; \bar{h}_s]) & \text{concat} \end{cases}$$

其中， $\text{Score}(h_t, h_s) = \alpha_{ij}$ 表示源端与目标单词对齐程度。可见，常见的对齐关系计算方式有，点乘（Dot product），权值网络映射（General）和concat映射几种方式。

2. Attention Mechanism分类

2.1 soft Attention 和Hard Attention

Kelvin Xu等人与2015年发表论文《Show, Attend and Tell: Neural Image Caption Generation with Visual Attention》，在Image Caption中引入了Attention，当生成第 i 个关于图片内容描述的词时，用Attention来关联与 i 个词相关的图片的区域。Kelvin Xu等人在论文中使用了两种Attention Mechanism，即Soft Attention和Hard Attention。我们之前所描述的传统Attention Mechanism就是Soft Attention。Soft Attention是参数化的（Parameterization），





相反，Hard Attention是一个随机的过程。Hard Attention不会选择整个encoder的输出做为输入，Hard Attention会依概率 S_i 来采样输入端的隐状态一部分来进行计算，而不是整个encoder的隐状态。为了实现梯度的反向传播，需要采用蒙特卡洛采样的方法来估计模块的梯度。

两种Attention Mechanism都有各自的优势，但目前更多的研究和应用还是更倾向于使用Soft Attention，因为其可以直接求导，进行梯度反向传播。

2.2 Global Attention 和 Local Attention

Global Attention：传统的Attention model一样。所有的hidden state都被用于计算Context vector 的权重，即变长的对齐向量 a_t ，其长度等于encoder端输入句子的长度。结构如图5所示。

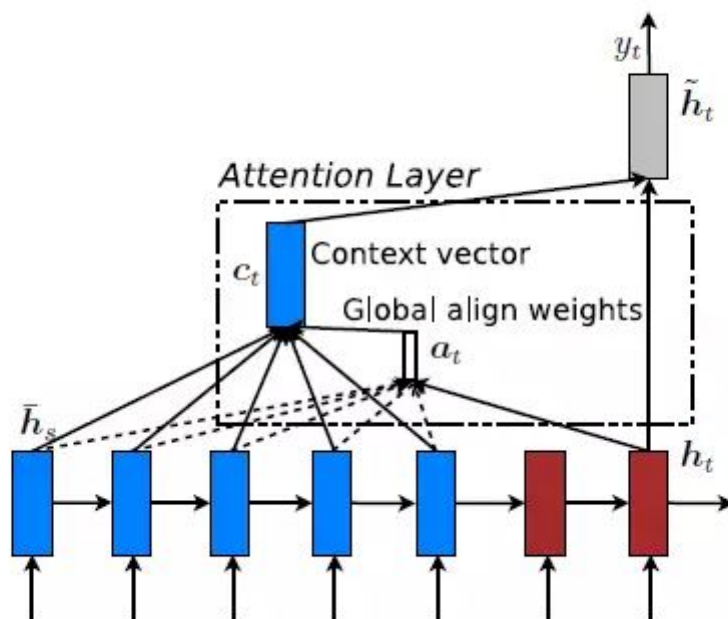


图5 Global Attention模型示意图

在 t 时刻，首先基于decoder的隐状态 h_t 和源端的隐状态 h_s ，计算一个变长的隐对齐权值向量 a_t ，其计算公式如下：

$$\begin{aligned} a_t(s) &= \text{align}(h_t, \bar{h}_s) \\ &= \frac{\exp(\text{score}(h_t, \bar{h}_s))}{\sum_{s'} \exp(\text{score}(h_t, \bar{h}_{s'}))} \end{aligned}$$





得到对齐向量 a_t 之后，就可以通过加权平均的方式，得到上下文向量 c_t 。

Local Attention：Global Attention有一个明显的缺点就是，每一次，encoder端的所有hidden state都要参与计算，这样做计算开销会比较大，特别是当encoder的句子偏长，比如，一段话或者一篇文章，效率偏低。因此，为了提高效率，Local Attention应运而生。

Local Attention是一种介于Kelvin Xu所提出的Soft Attention和Hard Attention之间的一种Attention方式，即把两种方式结合起来。其结构如图6所示。

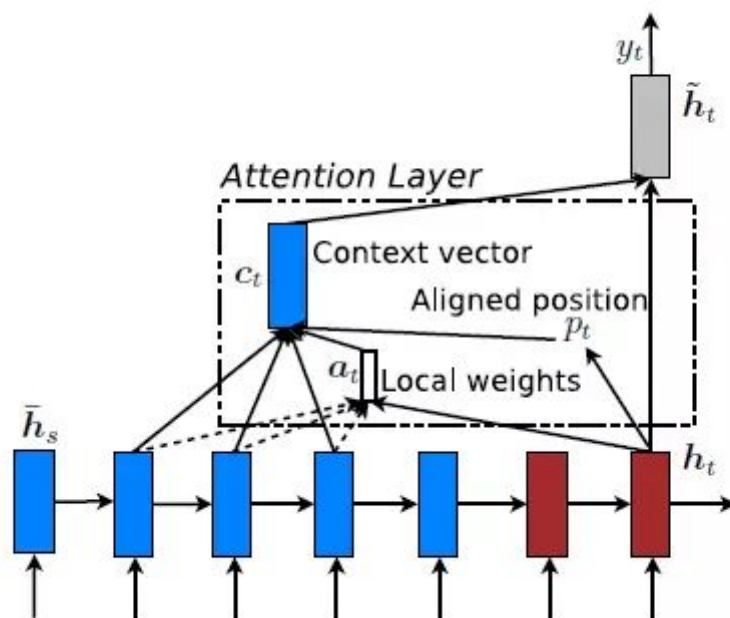


图6 Local Attention模型示意图

Local Attention首先会为decoder端当前的词，预测一个source端对齐位置（aligned position） p_t ，然后基于 p_t 选择一个窗口，用于计算背景向量 c_t 。Position p_t 的计算公式如下：

$$p_t = S \cdot \text{sigmoid}(v_p^\top \tanh(W_p h_t)),$$

其中， S 是encoder端句子长度， v_p 和 w_p 是模型参数。此时，对齐向量 a_t 的计算公式如下：

$$a_t(s) = \text{align}(h_t, \bar{h}_s) \exp\left(-\frac{(s - p_t)^2}{2\sigma^2}\right)$$



是很长时，相对Global Attention，计算量并没有明显减小。2、位置向量pt的预测并不非常准确，这就直接计算的到的local Attention的准确率。

2.3 Self Attention

Self Attention与传统的Attention机制非常的不同：传统的Attention是基于source端和target端的隐变量 (hidden state) 计算Attention的，得到的结果是源端的每个词与目标端每个词之间的依赖关系。但Self Attention不同，它分别在source端和target端进行，仅与source input或者target input自身相关的Self Attention，捕捉source端或target端自身的词与词之间的依赖关系；然后再把source端的得到的self Attention加入到target端得到的Attention中，捕捉source端和target端词与词之间的依赖关系。因此，self Attention Attention比传统的Attention mechanism效果要好，主要原因之一是，传统的Attention机制忽略了源端或目标端句子中词与词之间的依赖关系，相对比，self Attention可以不仅可以得到源端与目标端词与词之间的依赖关系，同时还可以有效获取源端或目标端自身词与词之间的依赖关系，如图7所示。

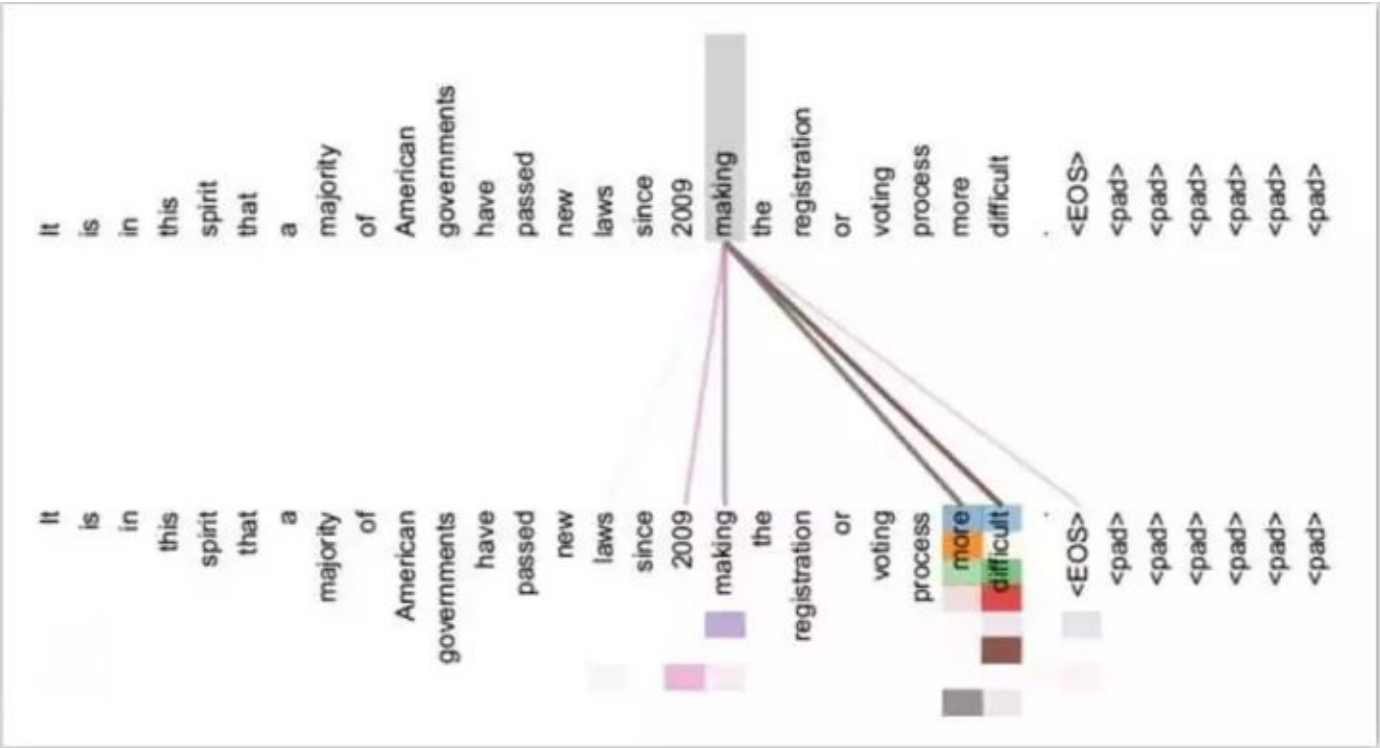


图7 self attention可视化实例，图片摘自《深度学习中的注意力机制》，张俊林

Self Attention的具体计算方式如图8所示：



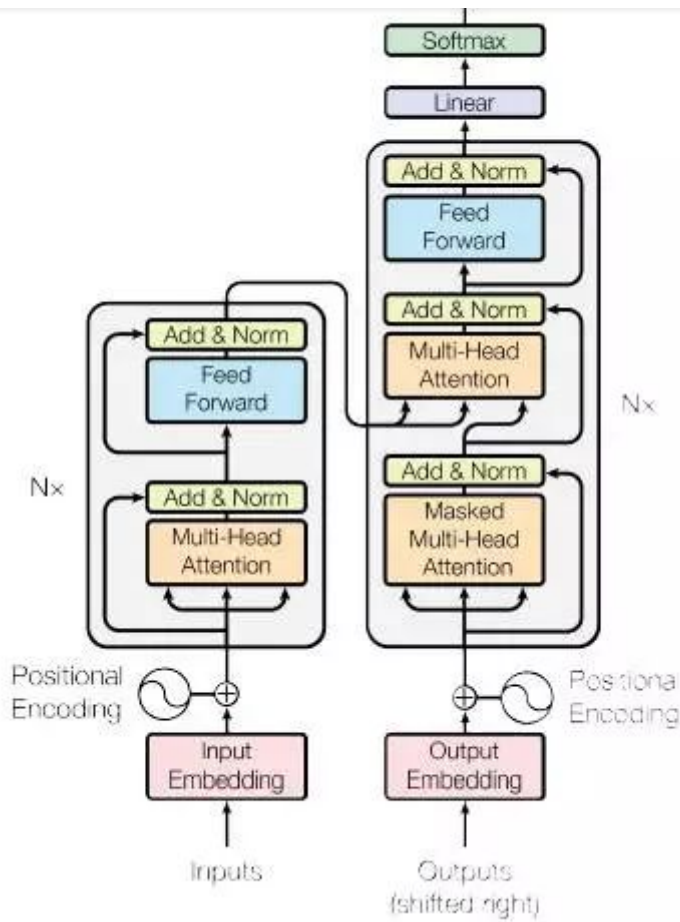


图8 All Attention模型结构示意图

Encoder的输入inputs和decoder的输入outputs，加上position embedding，做为各自的最初的输入，那么问题来了，self Attention具体是怎么实现的呢？从All Attention的结构示意图可以发现，Encoder和decoder是层叠多了类似的Multi-Head Attention单元构成，而每一个Multi-Head Attention单元由多个结构相似的Scaled Dot-Product Attention单元组成，结构如图9所示。



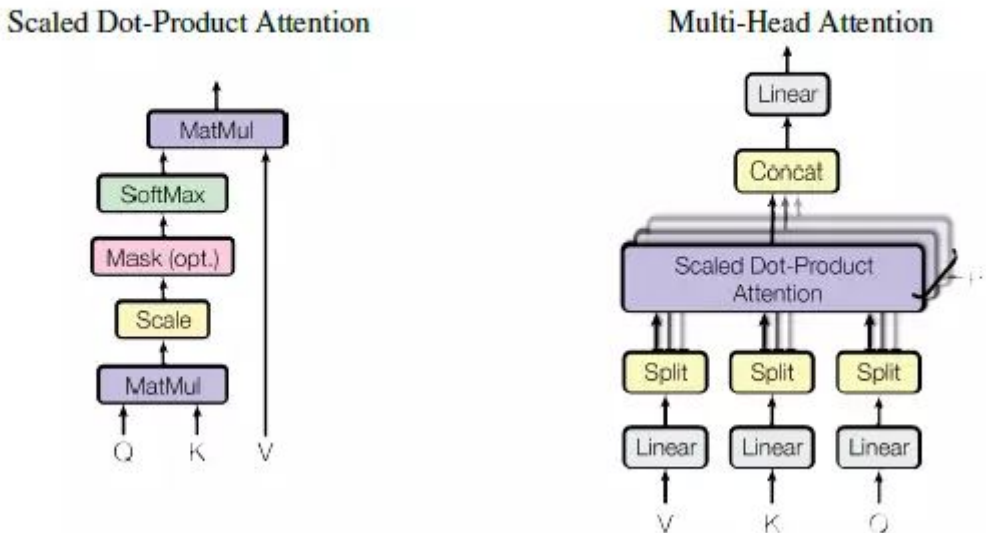


图9 Multi-Head Attention结构示意图

Self Attention也是在Scaled Dot-Product Attention单元里面实现的，如上图左图所示，首先把输入Input经过线性变换分别得到Q、K、V，注意，Q、K、V都来自于Input，只不过是线性变换的矩阵的权值不同而已。然后把Q和K做dot Product相乘，得到输入Input词与词之间的依赖关系，然后经过尺度变换（scale）、掩码（mask）和softmax操作，得到最终的Self Attention矩阵。尺度变换是为了防止输入值过大导致训练不稳定，mask则是为了保证时间的先后关系。

最后，把encoder端self Attention计算的结果加入到decoder做为k和V，结合decoder自身的输出做为q，得到encoder端的attention与decoder端attention之间的依赖关系。

Attention其他一些组合使用

2.4 Hierarchical Attention

Zichao Yang等人在论文《Hierarchical Attention Networks for Document Classification》提出了Hierarchical Attention用于文档分类。Hierarchical Attention构建了两个层次的Attention Mechanism，第一个层次是对句子中每个词的attention，即word attention；第二个层次是针对文档中每个句子的attention，即sentence attention。网络结构如图10所示。



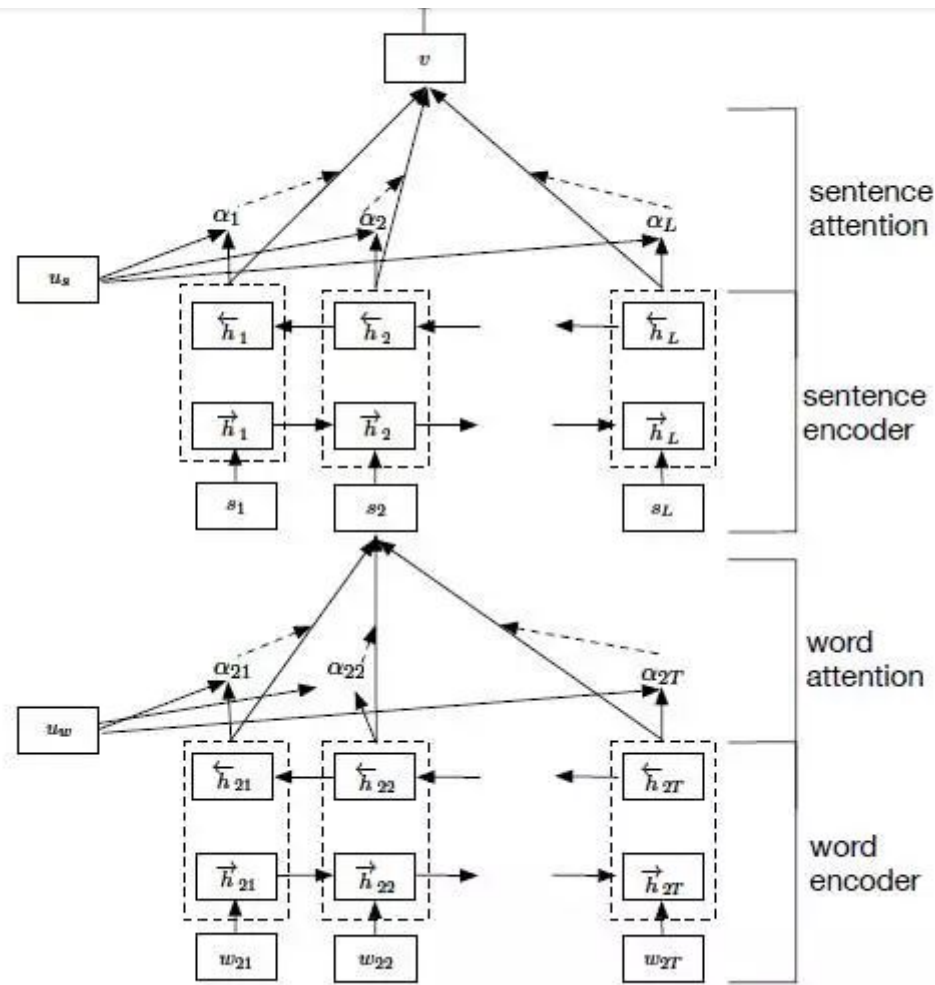


图10 Hierarchical Attention结构示意图

整个网络结构由四个部分组成：一个由双向RNN（GRU）构成的word sequence encoder，然后是一个关于词的word-level的attention layer；基于word attention layer之上，是一个由双向RNN构成的sentence encoder，最后的输出层是一个sentence-level的attention layer。

2.5 Attention over Attention

Yiming Cui与2017年在论文《Attention-over-Attention Neural Networks for Reading Comprehension》中提出了Attention Over Attention的Attention机制，结构如图11所示。



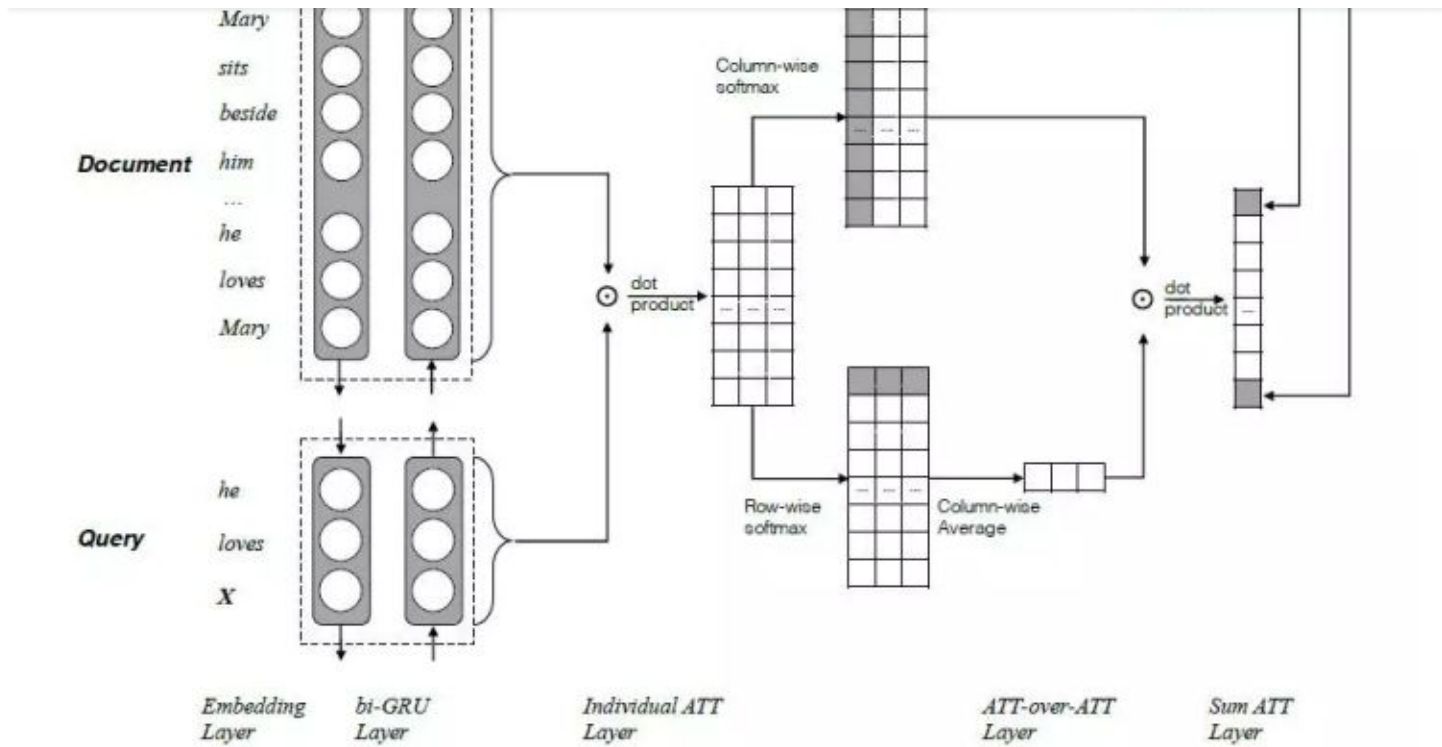


图11 Attention over Attention结构示意图

两个输入，一个Document和一个Query，分别用一个双向的RNN进行特征抽取，得到各自的隐状态 $h(doc)$ 和 $h(query)$ ，然后基于query和doc的隐状态进行dot product，得到query和doc的attention关联矩阵。然后按列（column）方向进行softmax操作，得到query-to-document的attention值 $a(t)$ ；按照行（row）方向进行softmax操作，得到document-to-query的attention值 $b(t)$ ，再按照列方向进行累加求平均得到平均后的attention值 $b(t)$ 。最后再基于上一步attention操作得到 $a(t)$ 和 $b(t)$ ，再进行attention操作，即attention over attention得到最终query与document的关联矩阵。

2.6 Multi-step Attention

2017年，FaceBook 人工智能实验室的Jonas Gehring等人在论文《Convolutional Sequence to Sequence Learning》提出了完全基于CNN来构建Seq2Seq模型，除了这一最大的特色之外，论文中还采用了多层Attention Mechanism，来获取encoder和decoder中输入句子之间的关系，结构如图12所示。



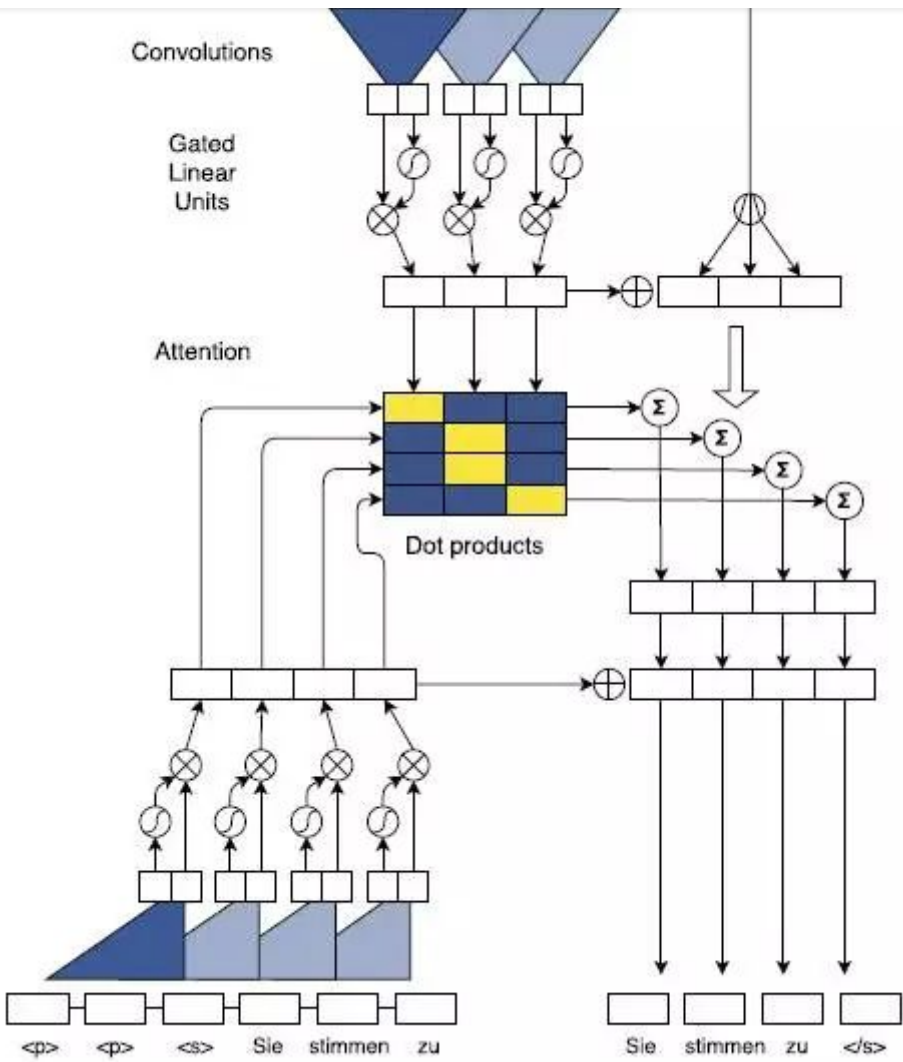


图12 Multi-step Attention结构示意图

完全基于CNN的Seq2Seq模型需要通过层叠多层来获取输入句子中词与词之间的依赖关系，特别是当句子非常长的时候，我曾经实验证明，层叠的层数往往达到10层以上才能取得比较理想的结果。针对每一个卷记得step（输入一个词）都对encoder的hidden state和decoder的hidden state进行dot product计算得到最终的Attention 矩阵，并且基于最终的attention矩阵去指导decoder的解码操作。

3. Attention的应用场景

本节主要给出一些基于Attention去处理序列预测问题的例子，以下内容整理翻译自：
machinelearningmastery.com...

1.机器翻译

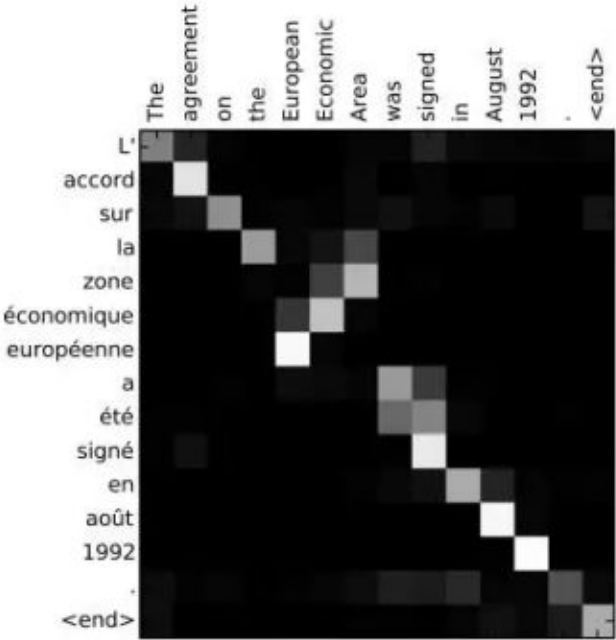
给定一个法语句子做为输入序列，翻译并输出一个英文句子做为输出序列。Attention用于关联输出序列中每个单词与输入序列中的某个特定单词的关联程度。



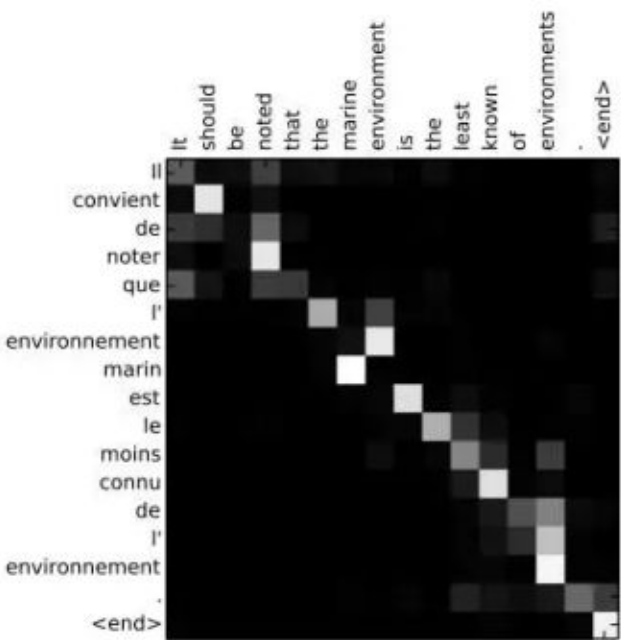


度的向量，并且还使模型只关注源端与下一个目标词的生成有关的信息。”

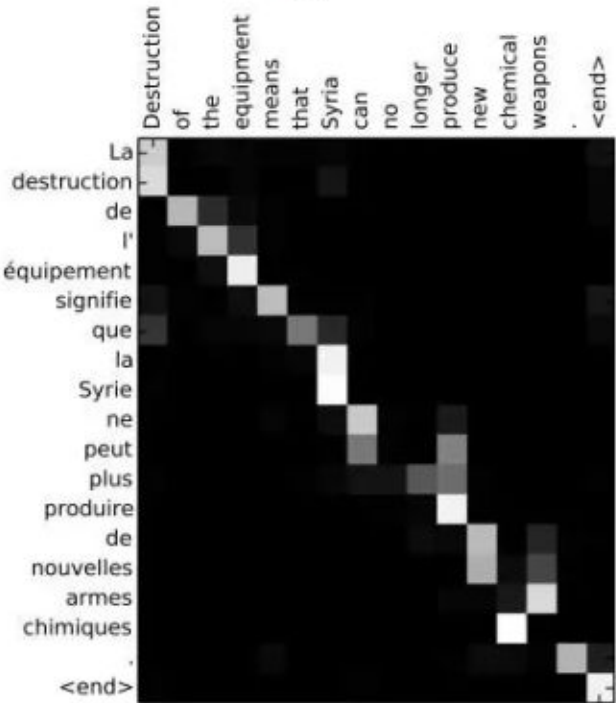
- Dzmitry Bahdanau等人，《Neural machine translation by jointly learning to align and translate》，2015。



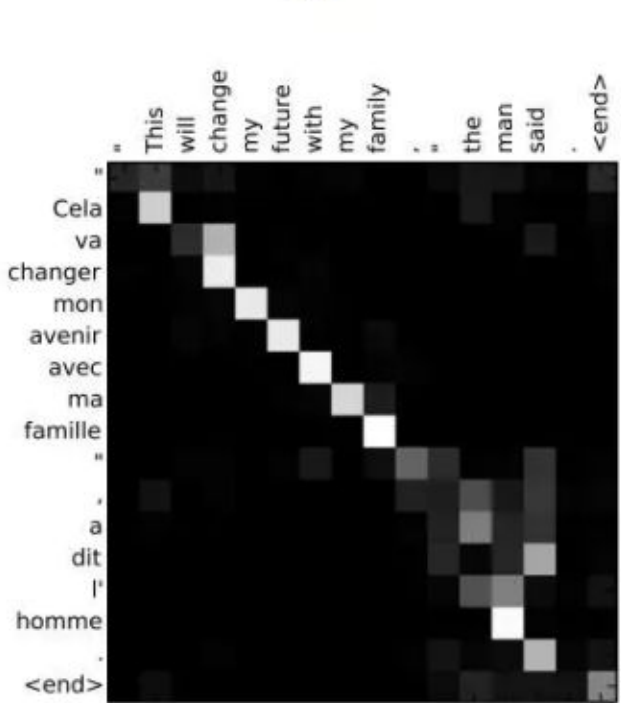
(a)



(b)



(c)



(d)

通过Attention来解释法语到英语单词之间的对应关系。摘自Dzmitry Bahdanau的论文

2.图像标注 (Image Caption)

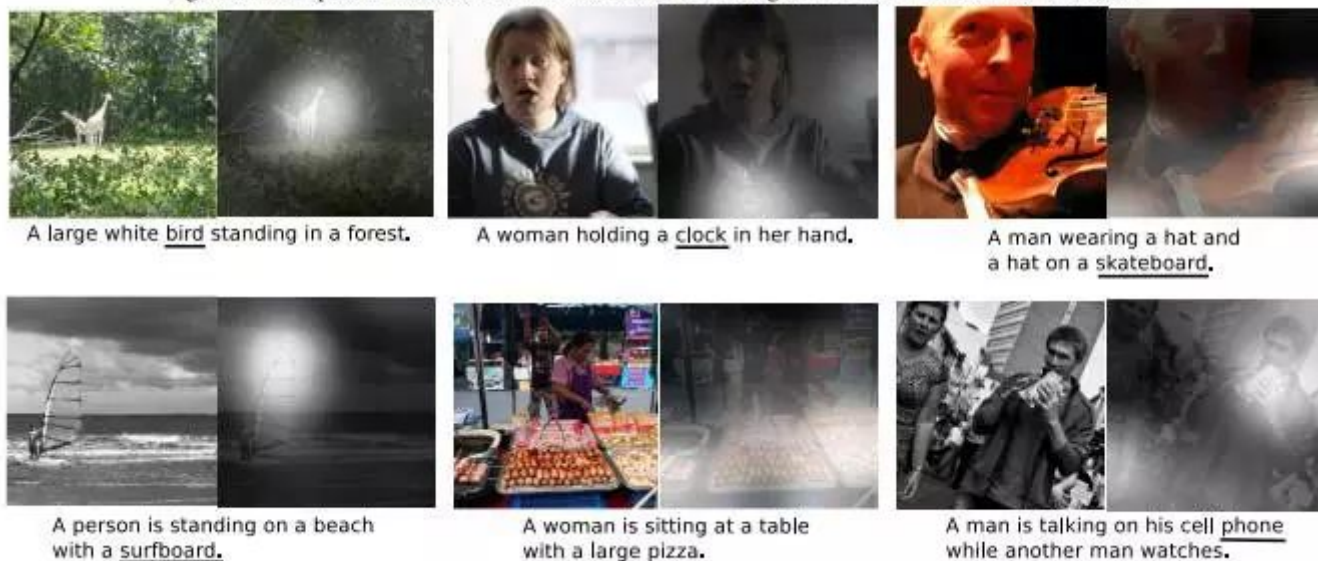


给定输入图像，输出图像的英文描述。使用Attention是为输出序列中的每个单词关注图像中不同部分。

“我们提出了一种基于Attention mechanism的方法，并在三个标准数据集上都取得了最好的成绩...我们还展示了如何利用学到的Attention来提供更多对模型生成过程的解释，并且证明Attention学习到的对齐与人类视觉感知非常一致。”

Kelvin Xu等人，《Attend and Tell: Neural Image Caption Generation with Visual Attention》，2016

Figure 5. Examples of mistakes where we can use attention to gain intuition into what the model saw.



基于Attention来解释，生成英文描述中某一个词时，与图片中某一区域的高度依赖关系。

3. 蕴含关系推理 (Entailment Reasoning)


给定一个用英语描述前景描述 (premise scenario) 和假设 (hypothesis)，判读假设 (premise) 与假设 (hypothesis) 的关系：矛盾，相关或包含。

例如：

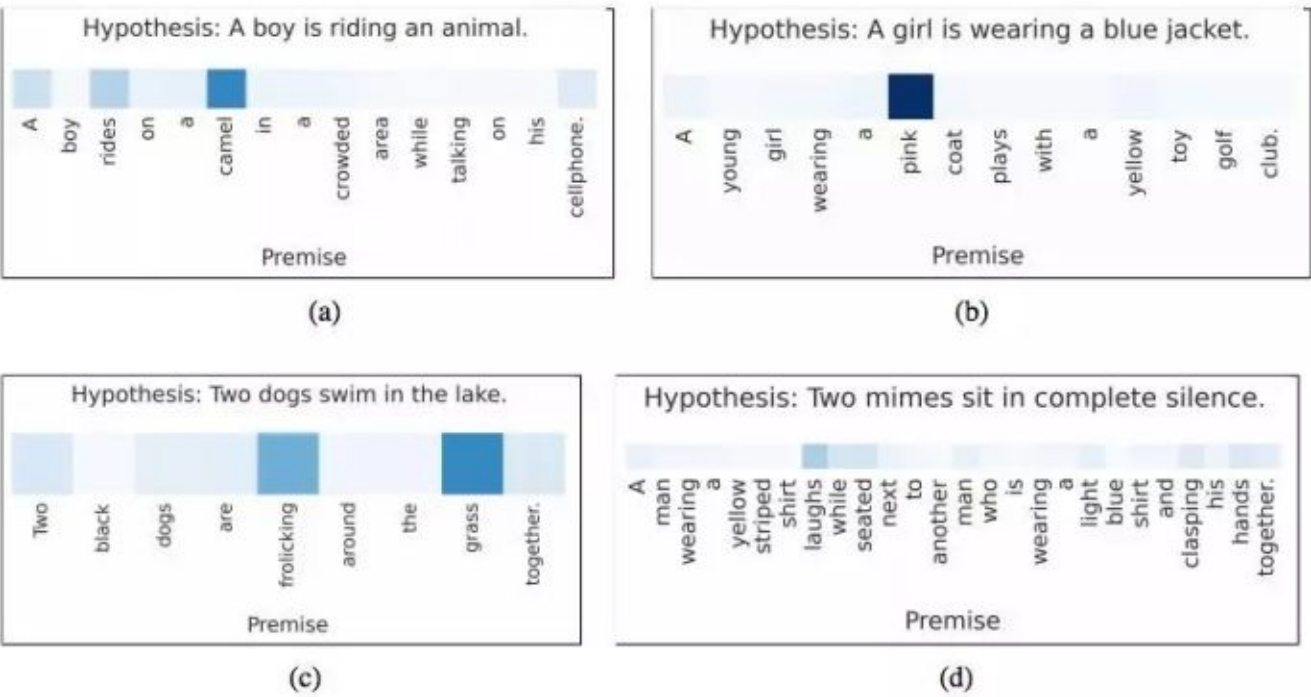
前提：“一场婚礼中拍照”

假设：“有人结婚”

Attention被用来把假设中的每个单词与前提中的单词联系起来，反之亦然。

“我们提出了一个基于LSTM的神经模型，它一次读取两个句子来确定两个句子之间的蕴含关系，而不是将每个句子独立映射到一个语义空间。我们引入逐字的 (word-by-word) Attention 

-Tim Rocktäschel , 《Reasoning about Entailment with Neural Attention》 , 2016



基于Attention来解释前提和假设中词与词之间的对应关系

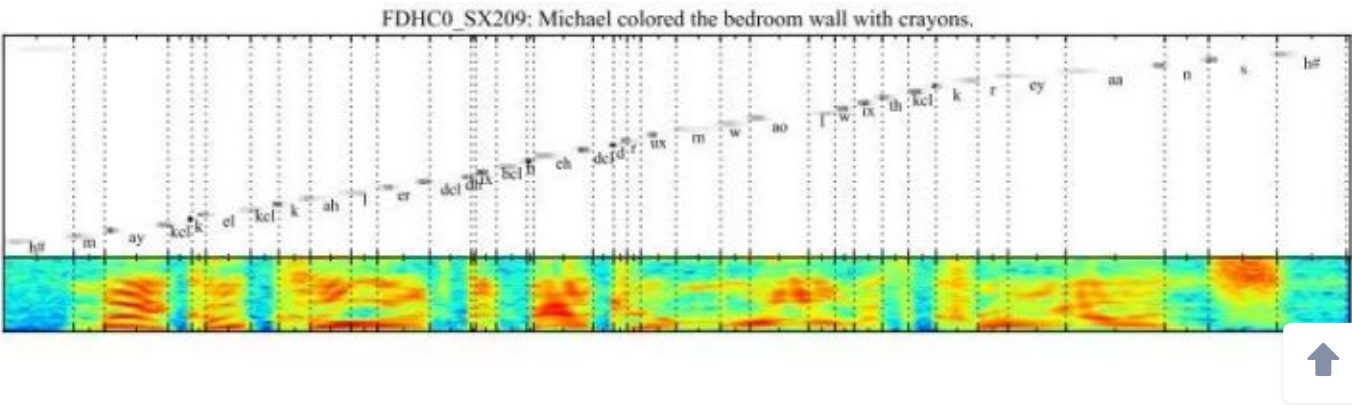
4. 语音识别

给定一段英语语音片段做为输入序列，输出对应的音素序列。

Attention被用联将输出序列中的每个音素与输入序列中的特定音频帧相关联。

“基于混合Attention机制的新型端到端可训练语音识别体系结构，其结合内容和位置信息帮助选择输入序列中的下一个位置用于解码。所提出的模型的一个理想特性就是它可以识别比训练集中句子的更长的句子。”

-Jan Chorowski , 《Attention-Based Models for Speech Recognition》 , 2015.。





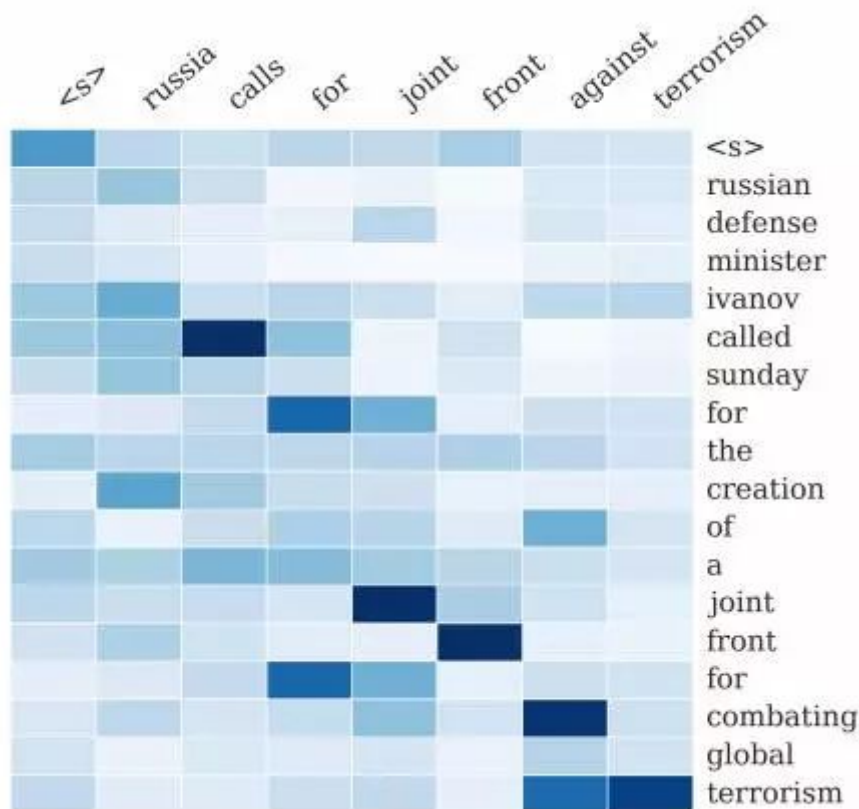
5.文字摘要生成

给定一篇英文文章做为输入顺序，输出一个总结英文文章注意内容的摘要句子。

Attention用于将输出摘要中的每个单词与输入文档中的特定单词相关联。

“将基于Attention的神经网络用语摘要抽取。我们将这个概率模型与可以产生准确的摘要的生成算法相结合。”

-Alexander M. Rush , 《A Neural Attention Model for Abstractive Sentence Summarization》, 2015



基于Attention来解释输入Sentence与输出Summary之间单词的对应关系

Attention Mechanism现在应用非常广泛，这里就列出这几个case供大家参考。

往期精彩内容推荐：

[纯干货-17 分布式深度学习原理、算法详细介绍](#)

[老铁，邀请你来免费学习人工智能！！](#)

[模型汇总23 - 卷积神经网络中不同类型的卷积方式介绍](#)

