robert_ai
Tell me, I will forget; Show me, I may remember; Involve me, I will understand.

昵称：robert_ai
园龄：4年
粉丝：68
关注：2
+加关注

< 2018年9月 >

日	一	二	三	四	五	六
26	27	28	29	30	31	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	1	2	3	4	5	6

搜索
找找看
谷歌搜索

常用链接
我的随笔
我的评论
我的参与
最新评论
我的标签

最新随笔
1. 基线系统需要受到更多关注：基于词向量的简单模型
2. 自然语言处理中的自注意力机制（Self-attention Mechanism）
3. 基于神经网络的实体识别和关系抽取联合学习
4. 神经网络结构在命名实体识别（NER）中的应用
5. 使用维基百科训练简体中文词向量
6. 注意力机制（Attention Mechanism）在自然语言处理中的应用
7. 如何产生好的词向量
8. 谈谈评价指标中的宏平均和微平均
9. 在NLP中深度学习模型何时需要树形结构?
10. Windows下MetaMap工具安装

我的标签
机器学习(9)
NLP(7)
深度学习(6)
Deep Learning(4)
自然语言处理(3)
attention(2)
神经网络(2)
实体识别(1)

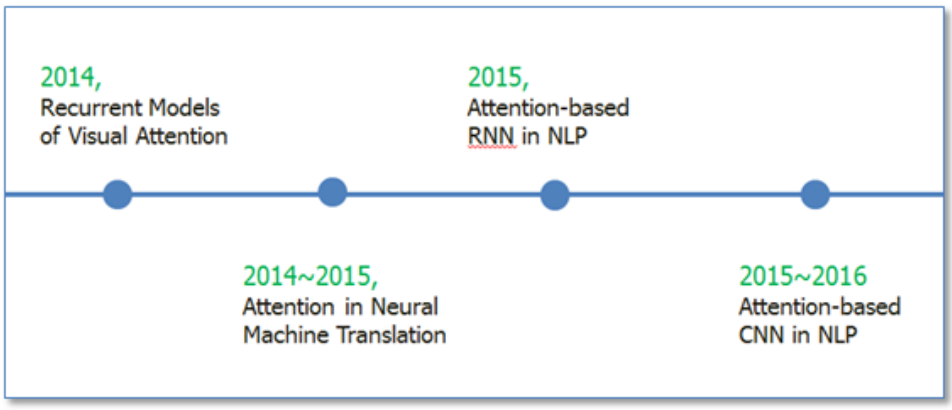
注意力机制（Attention Mechanism）在自然语言处理中的应用

注意力机制（Attention Mechanism）在自然语言处理中的应用

近年来，深度学习的研究越来越深入，在各个领域也都获得了不少突破性的进展。基于注意力（attention）机制的神经网络成为了最近神经网络研究的一个热点，本人最近也学习了一些基于attention机制的神经网络在自然语言处理（NLP）领域的论文，现在来对attention在NLP中的应用进行一个总结，和大家一起分享。

1 Attention研究进展

Attention机制最早是在视觉图像领域提出来的，应该是在九几年思想就提出来了，但是真正火起来应该算是google mind团队的这篇论文《Recurrent Models of Visual Attention》[14]，他们在RNN模型上使用了attention机制来进行图像分类。随后，Bahdanau等人在论文《Neural Machine Translation by Jointly Learning to Align and Translate》[1]中，使用类似attention的机制在机器翻译任务上将翻译和对齐同时进行，他们的工作算是是第一个提出attention机制应用到NLP领域中。接着类似的基于attention机制的RNN模型扩展开始应用到各种NLP任务中。最近，如何在CNN中使用attention机制也成为了大家的研究热点。下图表示了attention研究进展的大概趋势。



2 Recurrent Models of Visual Attention

在介绍NLP中的Attention之前，我想大致说一下图像中使用attention的思想。就具代表性的这篇论文《Recurrent Models of Visual Attention》[14]，他们研究的动机其实也是受到人类注意力机制的启发。人们在进行观察图像的时候，其实并不是一次就把整幅图像的每个位置像素都看过，大多是根据需求将注意力集中到图像的特定部分。而且人类会根据之前观察的图像学习到未来要观察图像注意力应该集中的位置。下图是这篇论文的核心模型示意图。

数学理论(1)
特征选择(1)
更多

随笔分类(34)

BioNLP(1)
Deep Learning(11)
Machine Learning(7)
NLP(12)
Tool(2)
数学基础(1)

随笔档案(26)

2018年6月 (1)
2018年3月 (1)
2017年10月 (1)
2017年5月 (1)
2017年3月 (1)
2016年10月 (1)
2016年6月 (1)
2016年3月 (1)
2015年11月 (1)
2015年6月 (1)
2015年4月 (2)
2015年3月 (2)
2015年2月 (1)
2015年1月 (1)
2014年11月 (3)
2014年10月 (3)
2014年9月 (4)

文章分类(1)

BioNLP
Deep Learning
Machine Learning(1)
NLP
Tools

文章档案(2)

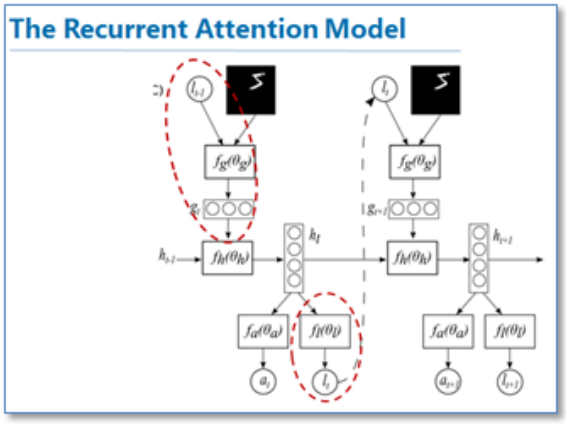
2015年6月 (1)
2015年1月 (1)

积分与排名

积分 - 38649
排名 - 12027

最新评论

1. Re:谈谈评价指标中的宏平均和微平均
我是软院的 互相学习 /抱拳
--shengchaohua
2. Re:谈谈评价指标中的宏平均和微平均
@shengchaohua计算机，互相学习交流~...
--robert_ai
3. Re:谈谈评价指标中的宏平均和微平均

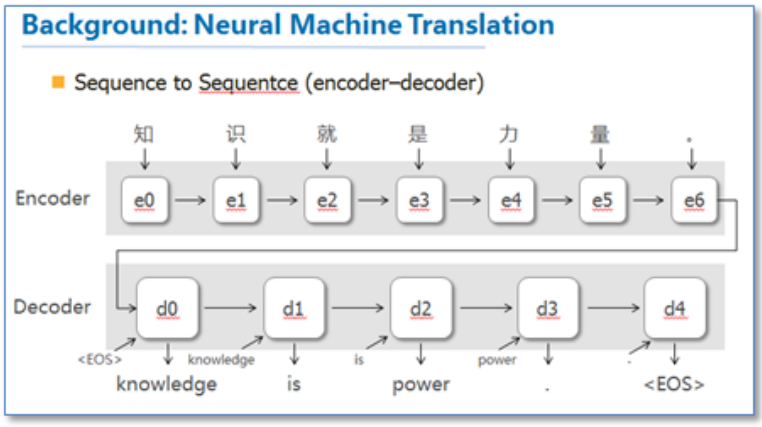


该模型是在传统的RNN上加入了attention机制（即红圈圈出来的部分），通过attention去学习一幅图像要处理的部分，每次当前状态，都会根据前一个状态学习得到的要关注的位置和当前输入的图像，去处理注意力部分像素，而不是图像的全部像素。这样的好处就是更少的像素需要处理，减少了任务的复杂度。可以看到图像中应用attention和人类的注意力机制是很类似的，接下来我们看看在NLP中使用的attention。

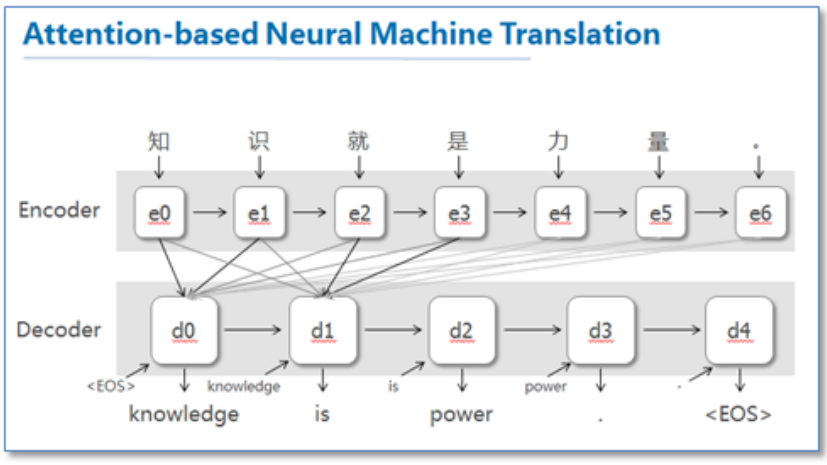
3 Attention-based RNN in NLP

3.1 Neural Machine Translation by Jointly Learning to Align and Translate [1]

这篇论文算是在NLP中第一个使用attention机制的工作。他们把attention机制用到了神经网络机器翻译（NMT）上，NMT其实就是一个典型的sequence to sequence模型，也就是一个encoder to decoder模型，传统的NMT使用两个RNN，一个RNN对源语言进行编码，将源语言编码到一个固定维度的中间向量，然后在使用一个RNN进行解码翻译到目标语言，传统的模型如下图：

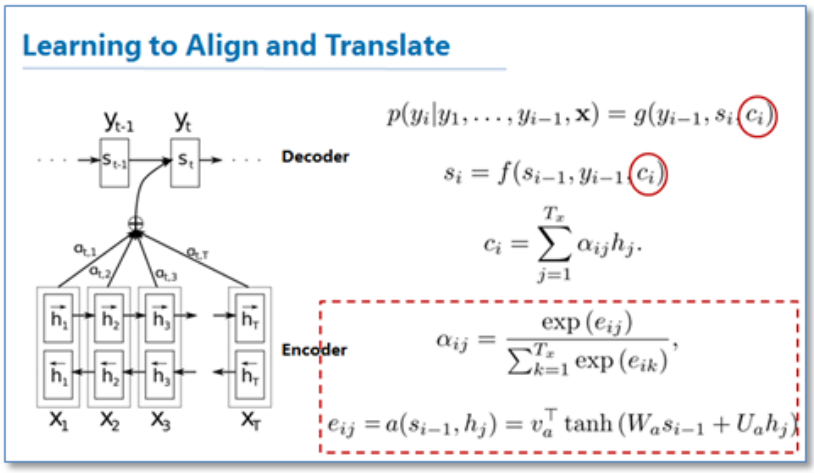


这篇论文提出了基于attention机制的NMT，模型大致如下图：

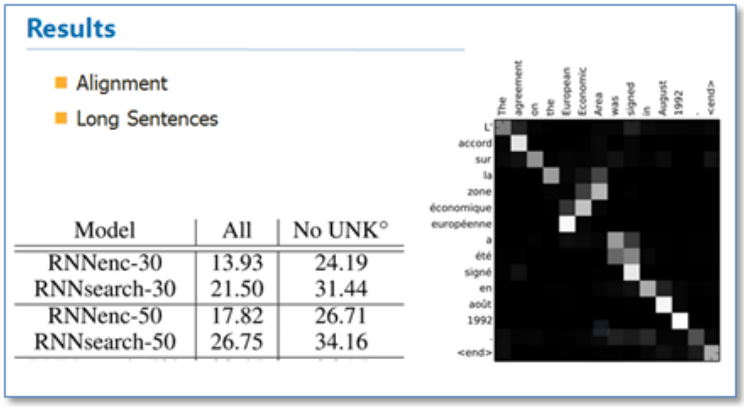


图中我并没有把解码器中的所有连线画玩，只画了前两个词，后面的词其实都一样。可以看到基于attention的NMT在传统的基础上，它把源语言端的每个词学到的表达（传统的只有

最后一个词后学到的表达）和当前要预测翻译的词联系了起来，这样的联系就是通过他们设计的attention进行的，在模型训练好后，根据attention矩阵，我们就可以得到源语言和目标语言的对齐矩阵了。具体论文的attention设计部分如下：



可以看到他们是使用一个感知机公式来将目标语言和源语言的每个词联系了起来，然后通过soft函数将其归一化得到一个概率分布，就是attention矩阵。

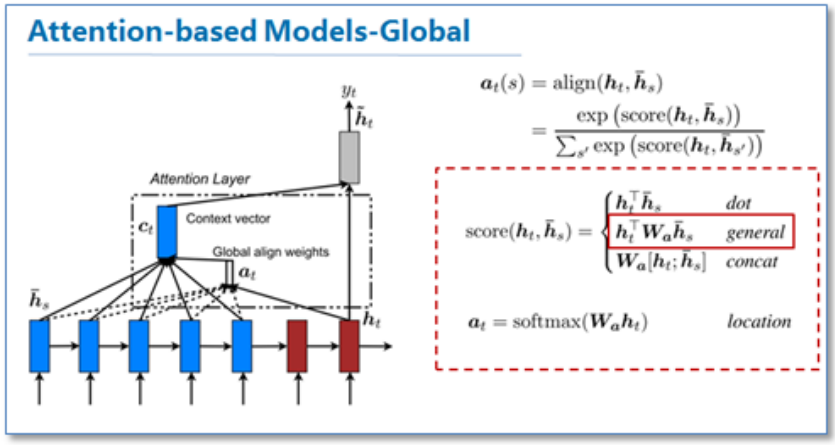


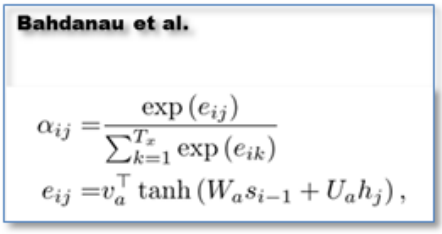
从结果来看相比传统的NMT（RNNsearch是attention NMT，RNNenc是传统NMT）效果提升了不少，最大的特点还在于它可以可视化对齐，并且在长句的处理上更有优势。

3.2 Effective Approaches to Attention-based Neural Machine Translation [2]

这篇论文是继上一篇论文后，一篇很具代表性的论文，他们的工作告诉了大家attention在RNN中可以如何进行扩展，这篇论文对后续各种基于attention的模型在NLP应用起到了很大的促进作用。在论文中他们提出了两种attention机制，一种是全局（global）机制，一种是局部（local）机制。

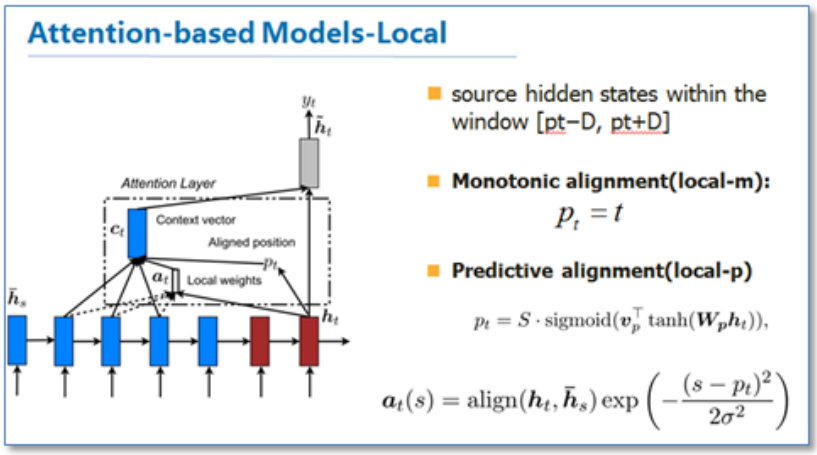
首先我们来看看global机制的attention，其实这和上一篇论文提出的attention的思路是一样的，它都是对源语言对所有词进行处理，不同的是在计算attention矩阵值的时候，他提出了几种简单的扩展版本。





在他们最后的实验中general的计算方法效果是最好的。

我们再来看一下他们提出的local版本。主要思路是为了减少attention计算时的耗费，作者在计算attention时并不是去考虑源语言端的所有词，而是根据一个预测函数，先预测当前解码时要对齐的源语言端的位置Pt，然后通过上下文窗口，仅考虑窗口内的词。

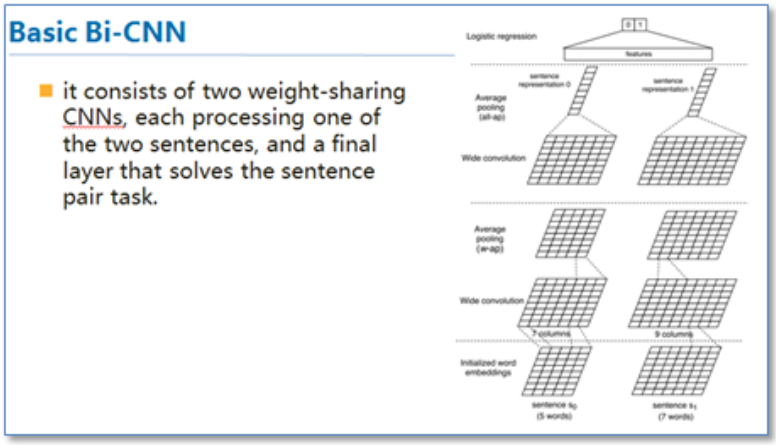


里面给出了两种预测方法，local-m和local-p，再计算最后的attention矩阵时，在原来的基础上去乘了一个pt位置相关的高斯分布。作者的实验结果是局部的比全局的attention效果好。

这篇论文最大的贡献我觉得是首先告诉了我们可以如何扩展attention的计算方式，还有就是局部的attention方法。

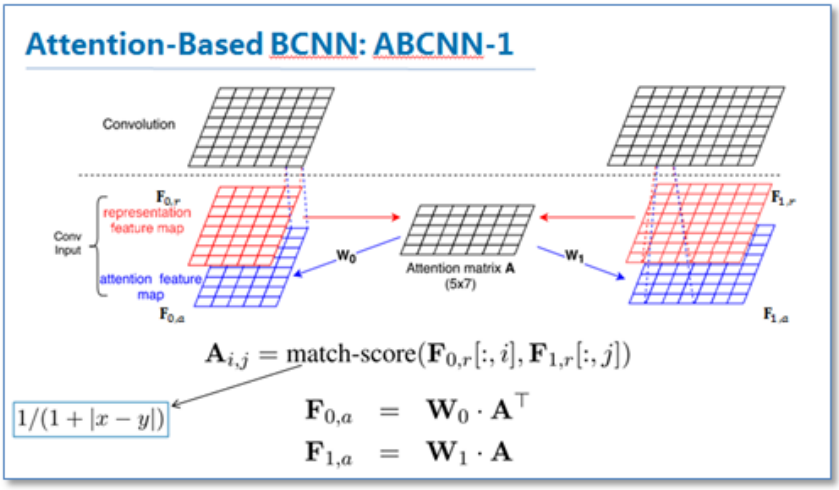
4 Attention-based CNN in NLP

随后基于Attention的RNN模型开始在NLP中广泛应用，不仅仅是序列到序列模型，各种分类问题都可以使用这样的模型。那么在深度学习中与RNN同样流行的卷积神经网络CNN是否也可以使用attention机制呢？《ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs》 [13]这篇论文就提出了3中在CNN中使用attention的方法，是attention在CNN中较早的探索性工作。

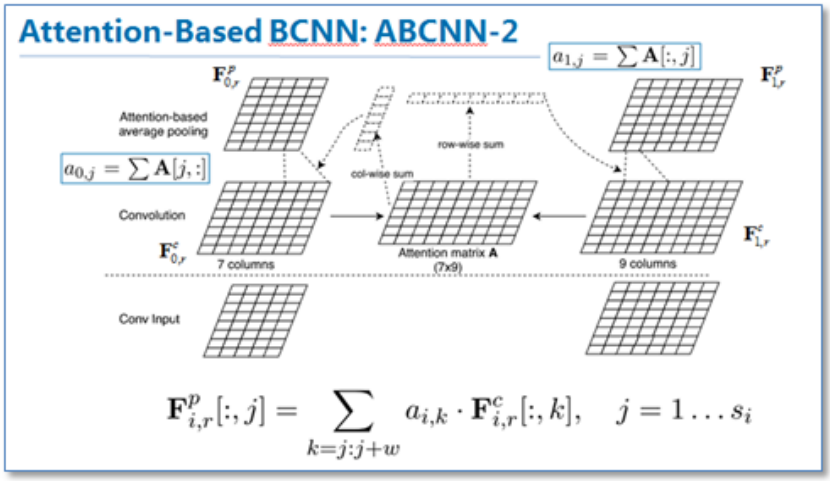


传统的CNN在构建句对模型时如上图，通过每个单通道处理一个句子，然后学习句子表达，最后一起输入到分类器中。这样的模型在输入分类器前句对间是没有相互联系的，作者们就想通过设计attention机制将不同cnn通道的句对联系起来。

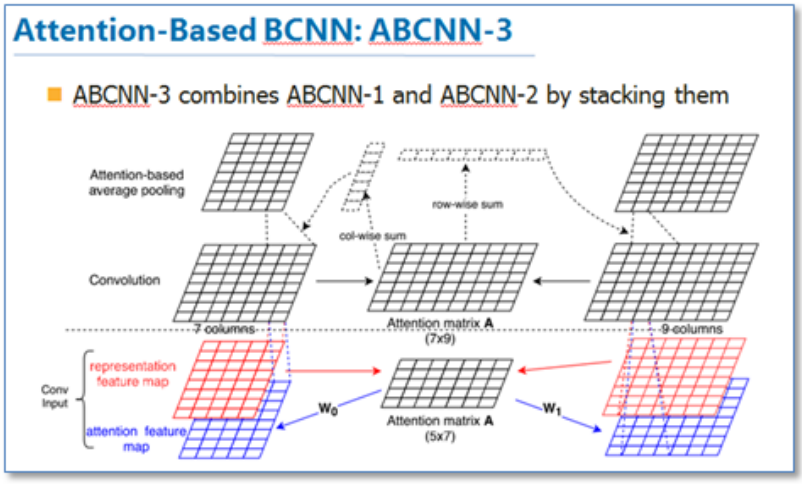
第一种方法ABCNN0-1是在卷积前进行attention，通过attention矩阵计算出相应句对的attention feature map，然后连同原来的feature map一起输入到卷积层。具体的计算方法如下。



第二种方法ABCNN-2是在池化时进行attention，通过attention对卷积后的表达重新加权，然后再进行池化，原理如下图。



第三种就是把前两种方法一起用到CNN中，如下图



这篇论文提供了我们在CNN中使用attention的思路。现在也有不少使用基于attention的CNN工作，并取得了不错的效果。

5 总结

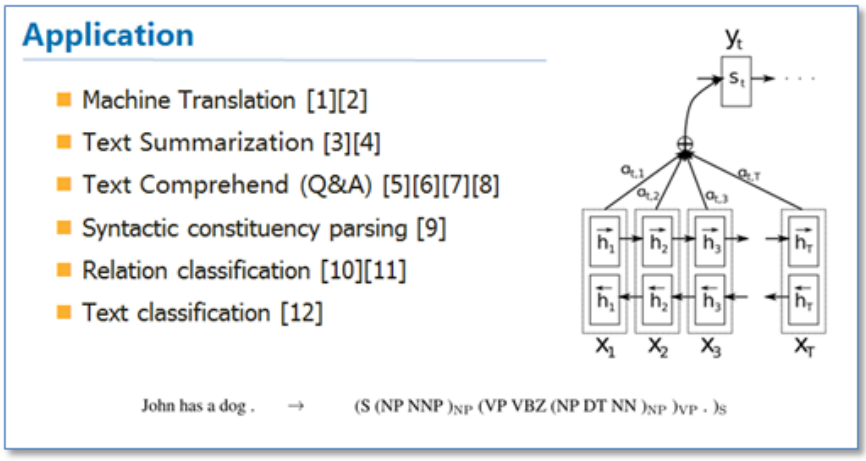
最后进行一下总结。Attention在NLP中其实我觉得可以看成是一种自动加权，它可以把两个你想要联系起来的不同模块，通过加权的形式进行联系。目前主流的计算公式有以下几种：

$$a_t = align(m_t, m_s) = \frac{\exp(f(m_t, m_s))}{\sum_{s'} \exp(f(m_t, m_{s'}))}$$

$$f(m_t, m_s) = \begin{cases} m_t^{\mathrm{T}} m_s & \text{dot} \\ m_t^{\mathrm{T}} W_a m_s & \text{general} \\ W_a [m_t; m_s] & \text{concat} \\ v_a^{\mathrm{T}} \tanh(W_a m_t + U_a m_s) & \text{perceptron} \end{cases}$$

通过设计一个函数将目标模块mt和源模块ms联系起来，然后通过一个soft函数将其归一化得到概率分布。

目前Attention在NLP中已经有广泛的应用。它有一个很大的优点就是可以可视化attention矩阵来告诉大家神经网络在进行任务时关注了哪些部分。



不过在NLP中的attention机制和人类的attention机制还是有所区别，它基本还是需要计算所有要处理的对象，并额外用一个矩阵去存储其权重，其实增加了开销。而不是像人类一样可以忽略不想关注的部分，只去处理关注的部分。

参考文献

[1] Bahdanau, D., Cho, K. & Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. Iclr 2015 1–15 (2014).

[2] Luong, M. & Manning, C. D. Effective Approaches to Attention-based Neural Machine Translation. 1412–1421 (2015).

[3] Rush, A. M. & Weston, J. A Neural Attention Model for Abstractive Sentence Summarization. EMNLP (2015).

[4] Allamanis, M., Peng, H. & Sutton, C. A Convolutional Attention Network for Extreme Summarization of Source Code. Arxiv (2016).

[5] Hermann, K. M. et al. Teaching Machines to Read and Comprehend. arXiv 1–13 (2015).

[6] Yin, W., Ebert, S. & Schütze, H. Attention-Based Convolutional Neural Network for Machine Comprehension. 7 (2016).

[7] Kadlec, R., Schmid, M., Bajgar, O. & Kleindienst, J. Text Understanding with the Attention Sum Reader Network. arXiv:1603.01547v1 [cs.CL] (2016).

[8] Dhingra, B., Liu, H., Cohen, W. W. & Salakhutdinov, R. Gated-Attention Readers for Text Comprehension. (2016).

[9] Vinyals, O. et al. Grammar as a Foreign Language. arXiv 1–10 (2015).

[10] Wang, L., Cao, Z., De Melo, G. & Liu, Z. Relation Classification via Multi-Level Attention CNNs. Acl 1298–1307 (2016).

[11] Zhou, P. et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. Proc. 54th Annu. Meet. Assoc. Comput. Linguist. (Volume 2 Short Pap. 207–212 (2016).

[12] Yang, Z. et al. Hierarchical Attention Networks for Document Classification. Naacl (2016).

[13] Yin W, Schütze H, Xiang B, et al. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. arXiv preprint arXiv:1512.05193, 2015.

[14] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]//Advances in Neural Information Processing Systems. 2014: 2204-2212.

分类: Deep Learning,NLP

标签: attention, Deep Learning, 深度学习, 注意力机制, NLP, 自然语言处理

robert_ai
关注 - 2
粉丝 - 68
+加关注

6 0

« 上一篇: 如何产生好的词向量
» 下一篇: 使用维基百科训练简体中文词向量

posted on 2016-10-12 11:15 robert_ai 阅读(23802) 评论(8) 编辑 收藏

评论:

#1楼 2016-10-18 12:41 | 分析挖掘机
申请转载至微信公众号大数据分析挖掘

支持(0) 反对(0)

#2楼[楼主] 2016-10-18 13:33 | robert_ai
@ 分析挖掘机
好的，谢谢。

支持(0) 反对(0)

#3楼 2016-12-13 13:54 | 洞明
博主的博客中，作图风格很统一，赞!
请问，博主是用什么作图的?
支持(0) 反对(0)

#4楼[楼主] 2016-12-13 16:01 | robert_ai
@ 洞明
我用PPT做的~
支持(0) 反对(0)

#5楼 2017-06-29 17:11 | 职涯有乐
毛遂自荐，我是猎头，正好有个相关的职位招聘，欢迎咨询：qq 3091309630 邮箱：davi@andxy.cn。大家有兴趣也欢迎关注公众号：职涯有乐（topjob100）。平台提供值招聘、名企业、乐文章优质内容，适时发布IT/互联网金融行业topjob100岗位招聘信息，希望通过我的努力给大家的工作带来方便。
支持(0) 反对(0)

#6楼 2018-04-30 11:39 | 千里缘
大神请教个问题，在计算attention矩阵的aij时用到了eij，这个是什么值呢?
支持(1) 反对(0)

#7楼[楼主] 2018-05-01 08:45 | robert_ai
@ 千里缘
你好，不同的论文实现的eij不同，一般点积，感知机等是常用的方式，主要就是计算出目标端和源端的对齐函数得分。然后一般再经过一个softmax进行归一化就得到attention权重了。
支持(1) 反对(0)

#8楼 2018-07-11 14:16 | 我乐飞
很系统，很好的介绍attention的文章，感谢博主
支持(0) 反对(0)

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册，访问网站首页。