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)

自然语言处理中的自注意力机制（Self-attention Mechanism）

自然语言处理中的自注意力机制（Self-attention Mechanism）

近年来，注意力（Attention）机制被广泛应用到基于深度学习的自然语言处理(NLP)各个任务中，之前我对早期注意力机制进行过一些学习总结（可见http://www.cnblogs.com/robert-dlut/p/5952032.html）。随着注意力机制的深入研究，各式各样的attention被研究者们提出。在2017年6月google机器翻译团队在arXiv上放出的《Attention is all you need》论文受到了大家广泛关注，自注意力（self-attention）机制开始成为神经网络attention的研究热点，在各个任务上也取得了不错的效果。本人就这篇论文中的self-attention以及一些相关工作进行了学习总结（其中也参考借鉴了张俊林博士的博客"深度学习中的注意力机制(2017版)"和苏剑林的"《Attention is All You Need》浅读（简介+代码）"），和大家一起分享。

1 背景知识

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2014年google mind团队的这篇论文《Recurrent Models of Visual Attention》，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》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，他们的工作算是第一个将attention机制应用到NLP领域中。接着attention机制被广泛应用在基于RNN/CNN等神经网络模型的各种NLP任务中。2017年，google机器翻译团队发表的《Attention is all you need》中大量使用了自注意力（self-attention）机制来学习文本表示。自注意力机制也成为了大家近期的研究热点，并在各种NLP任务上进行探索。下图维attention研究进展的大概趋势。

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)

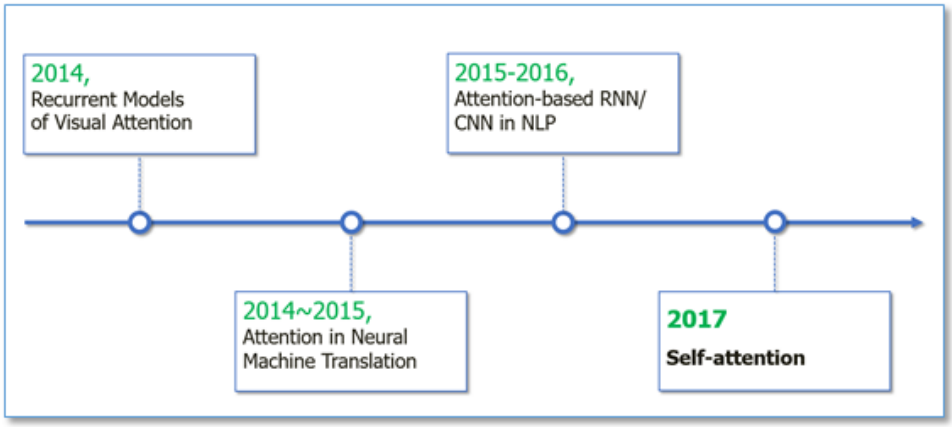
2018年6月 (1)
2018年3月 (1)
2017年10月 (1)
2017年5月 (1)
2017年3月 (1)
2016年10月 (1)
2016年6月 (1)
2016年3月 (1)
2015年11月 (1)
2015年6月 (1)
2015年4月 (2)
2015年3月 (2)
2015年2月 (1)
2015年1月 (1)
2014年11月 (3)
2014年10月 (3)
2014年9月 (4)

文章分类(1)

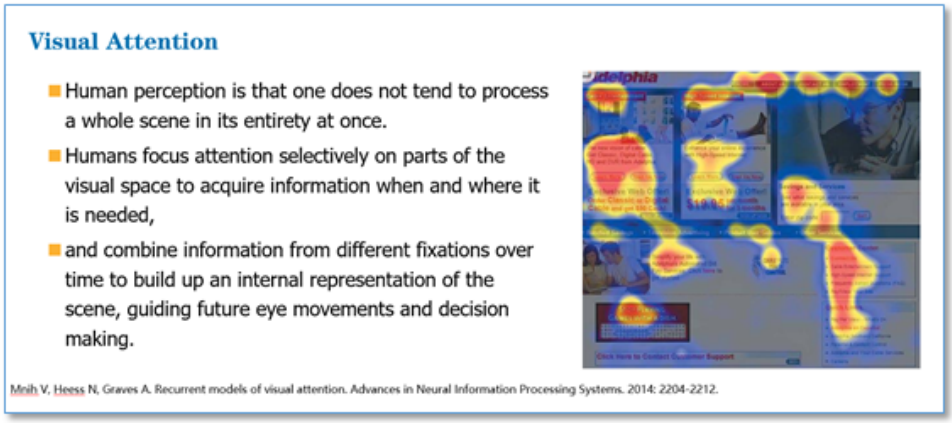
BioNLP
Deep Learning
Machine Learning(1)
NLP
Tools

文章档案(2)

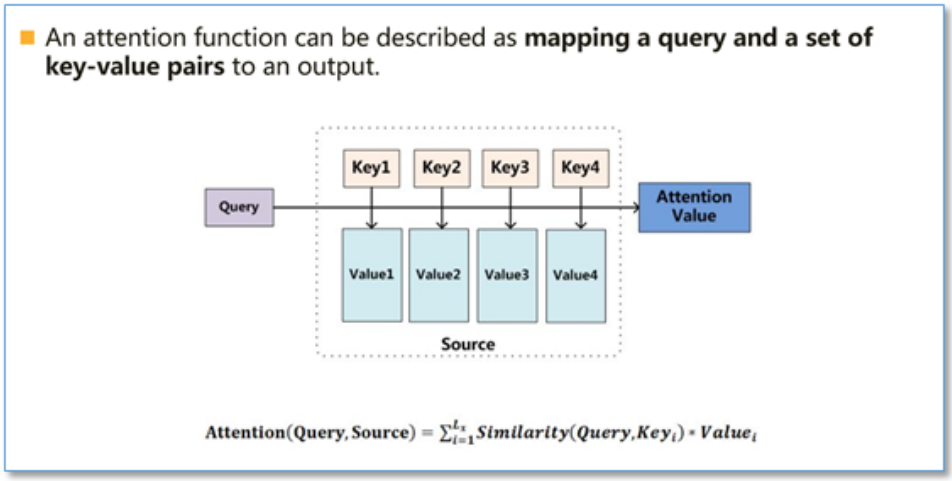
2015年6月 (1)
2015年1月 (1)



Attention机制的本质来自于人类视觉注意力机制。人们视觉在感知东西的时候一般不会是一个场景从到头看到尾每次全部都看，而往往是根据需求观察注意特定的一部分。而且当人们发现一个场景经常在某部分出现自己想观察的东西时，人们会进行学习在将来再出现类似场景时把注意力放到该部分上。



下面我先介绍一下在NLP中常用attention的计算方法（里面借鉴了张俊林博士"深度学习中的注意力机制(2017版)"里的一些图）。Attention函数的本质可以被描述为一个查询（query）到一系列（键key-值value）对的映射，如下图。



在计算attention时主要分为三步，第一步是将query和每个key进行相似度计算得到权重，常用的相似度函数有点积，拼接，感知机等；然后第二步一般是使用一个softmax函数对这些权重进行归一化；最后将权重和相应的键

@robert_ai引用@shengchaohua是的。那是软院还是计算机的啊 你的这篇博客帮我解决了疑惑 抱拳...
--shengchaohua

4. Re:谈谈评价指标中的宏平均和微平均
@shengchaohua是的。...
--robert_ai

5. Re:谈谈评价指标中的宏平均和微平均
博主是大工的吗?
--shengchaohua

阅读排行榜

1. 注意力机制（Attention Mechanism）在自然语言处理中的应用(23799)
2. 神经网络结构在命名实体识别（NER）中的应用(19762)
3. 自然语言处理中的自注意力机制（Self-attention Mechanism）(10387)
4. 基于神经网络的实体识别和关系抽取联合学习(5484)
5. DL一（ML基础知识）(5122)

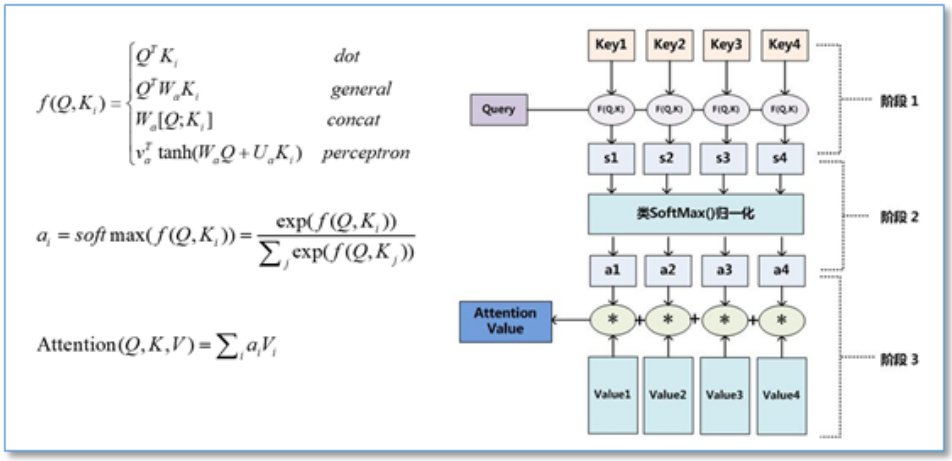
评论排行榜

1. 神经网络结构在命名实体识别（NER）中的应用(24)
2. 注意力机制（Attention Mechanism）在自然语言处理中的应用(8)
3. 使用维基百科训练简体中文词向量(7)
4. 基于神经网络的实体识别和关系抽取联合学习(7)
5. 谈谈评价指标中的宏平均和微平均(6)

推荐排行榜

1. 神经网络结构在命名实体识别（NER）中的应用(6)
2. 注意力机制（Attention Mechanism）在自然语言处理中的应用(6)
3. 自然语言处理中的自注意力机制（Self-attention Mechanism）(5)
4. 谈谈评价指标中的宏平均和微平均(3)
5. 在NLP中深度学习模型何时需要树形结构？(2)

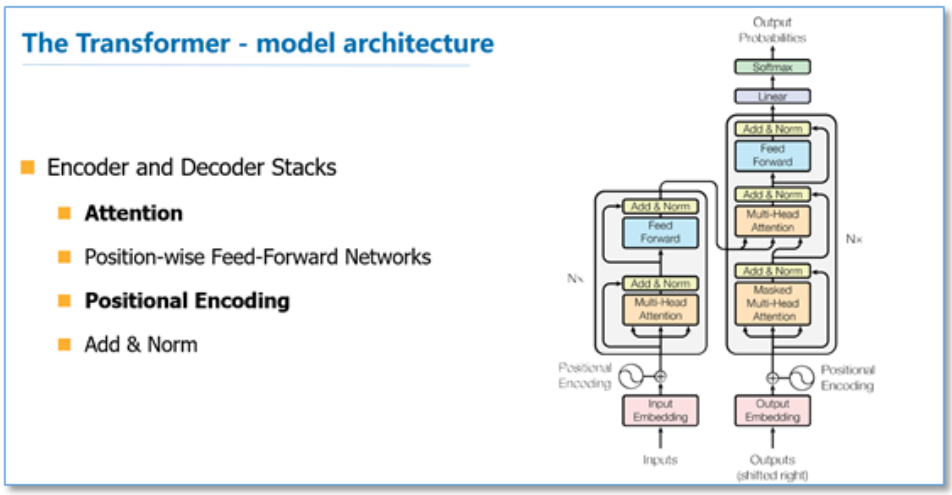
值value进行加权求和得到最后的attention。目前在NLP研究中，key和value常常都是同一个，即key=value。



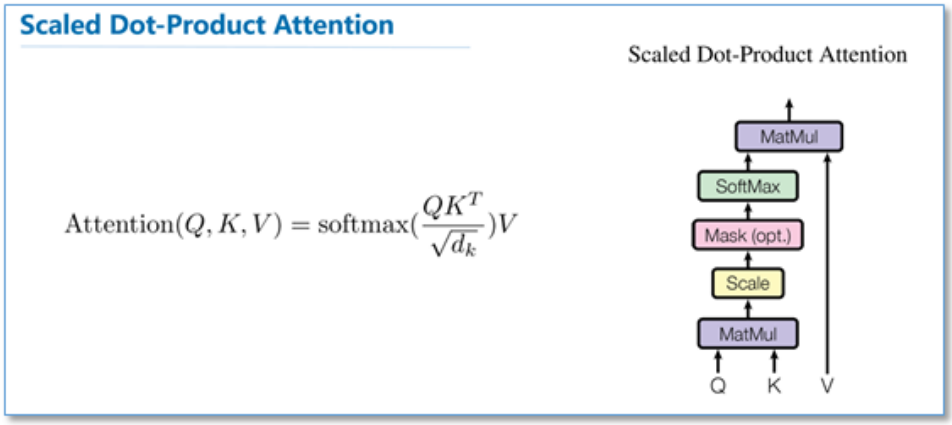
2 Attention is all you need[1]

接下来我将介绍《Attention is all you need》这篇论文。这篇论文是google机器翻译团队在2017年6月放在arXiv上，最后发表在2017年nips上，到目前为止google学术显示引用量为119，可见也是受到了大家广泛关注和应用。这篇论文主要亮点在于1）不同于以往主流机器翻译使用基于RNN的seq2seq模型框架，该论文用attention机制代替了RNN搭建了整个模型框架。2）提出了多头注意力（Multi-headed attention）机制方法，在编码器和解码器中大量的使用了多头自注意力机制（Multi-headed self-attention）。3）在WMT2014语料中的英德和英法任务上取得了先进结果，并且训练速度比主流模型更快。

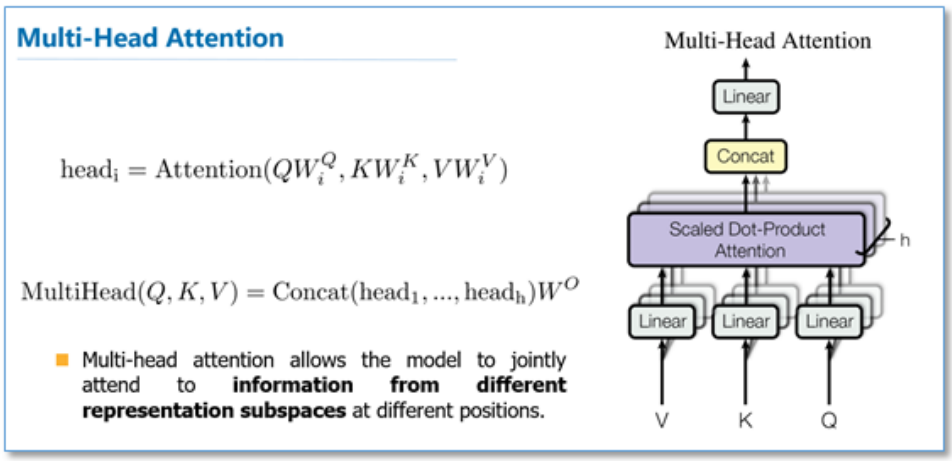
该论文模型的整体结构如下图，还是由编码器和解码器组成，在编码器的一个网络块中，由一个多头attention子层和一个前馈神经网络子层组成，整个编码器栈式搭建了N个块。类似于编码器，只是解码器的一个网络块中多了一个多头attention层。为了更好的优化深度网络，整个网络使用了残差连接和对层进行了规范化（Add&Norm）。



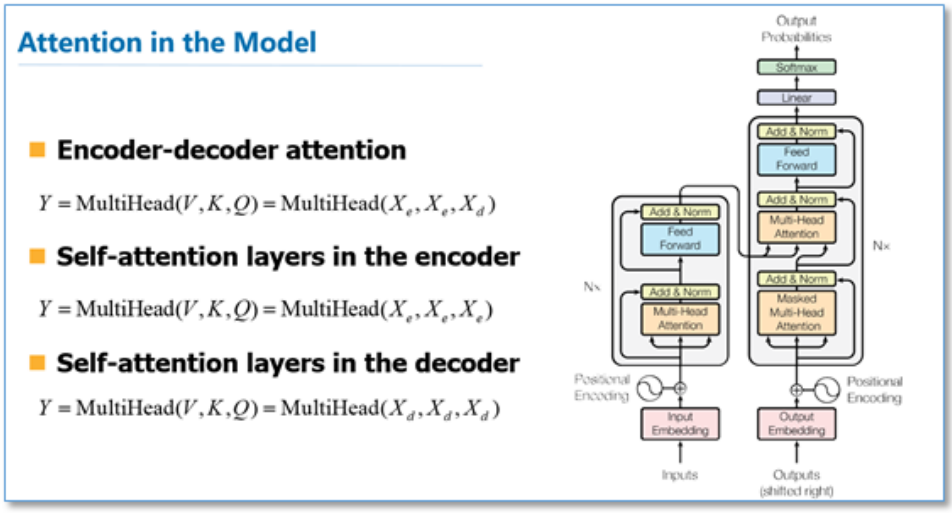
下面我们重点关注一下这篇论文中的attention。在介绍多头attention之前，我们先看一下论文中提到的放缩点积attention（scaled dot-Product attention）。对比我在前面背景知识里提到的attention的一般形式，其实scaled dot-Product attention就是我们常用的使用点积进行相似度计算的attention，只是多除了一个（为K的维度）起到调节作用，使得内积不至于太大。



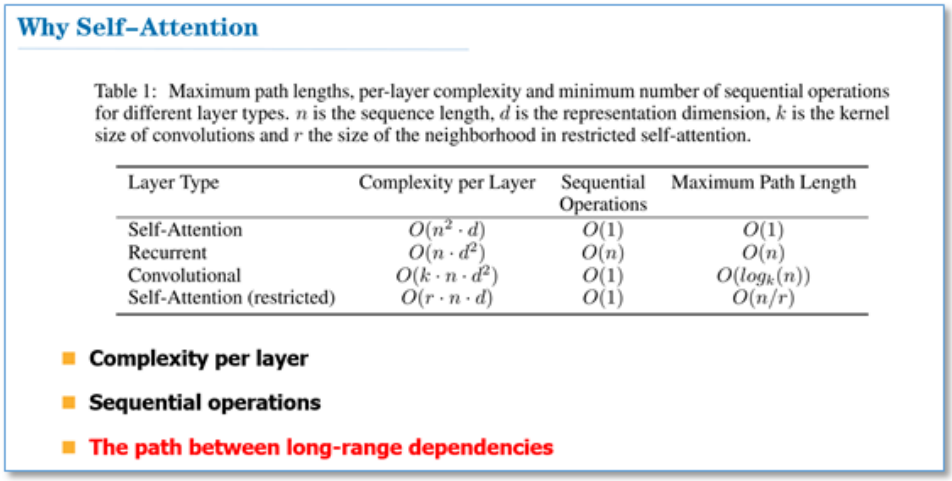
多头attention（Multi-head attention）结构如下图，Query，Key，Value首先进过一个线性变换，然后输入到放缩点积attention，注意这里要做h次，其实也就是所谓的多头，每一次算一个头。而且每次Q，K，V进行线性变换的参数W是不一样的。然后将h次的放缩点积attention结果进行拼接，再进行一次线性变换得到的值作为多头attention的结果。可以看到，google提出来的多头attention的不同之处在于进行了h次计算而不仅仅算一次，论文中说到这样的好处是可以允许模型在不同的表示子空间里学习到相关的信息，后面还会根据attention可视化来验证。



那么在整个模型中，是如何使用attention的呢？如下图，首先在编码器到解码器的地方使用了多头attention进行连接，K，V，Q分别是编码器的层输出（这里K=V）和解码器中都头attention的输入。其实就和主流的机器翻译模型中的attention一样，利用解码器和编码器attention来进行翻译对齐。然后在编码器和解码器中都使用了多头自注意力self-attention来学习文本的表示。Self-attention即K=V=Q，例如输入一个句子，那么里面的每个词都要和该句子中的所有词进行attention计算。目的是学习句子内部的词依赖关系，捕获句子的内部结构。



对于使用自注意力机制的原因，论文中提到主要从三个方面考虑（每一层的复杂度，是否可以并行，长距离依赖学习），并给出了和RNN，CNN计算复杂度的比较。可以看到，如果输入序列n小于表示维度d的话，每一层的时间复杂度self-attention是比较有优势的。当n比较大时，作者也给出了一种解决方案self-attention（restricted）即每个词不是和所有词计算attention，而是只与限制的r个词去计算attention。在并行方面，多头attention和CNN一样不依赖于前一时刻的计算，可以很好的并行，优于RNN。在长距离依赖上，由于self-attention是每个词和所有词都要计算attention，所以不管他们中间有多长距离，最大的路径长度也都只是1。可以捕获长距离依赖关系。



最后我们看一下实验结果，在WMT2014的英德和英法机器翻译任务上，都取得了先进的结果，且训练速度优于其他模型。

Machine Translation

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

在模型的超参实验中可以看到，多头attention的超参h太小也不好，太大也会下降。整体更大的模型比小模型要好，使用dropout可以帮助过拟合。

Model Variations

- While **single-head** attention is 0.9 BLEU **worse** than the best setting, quality also **drops off with too many heads.**
- **Bigger models are better,** and dropout is very helpful in avoiding over-fitting.
- We replace our sinusoidal positional encoding with learned positional embeddings, and observe **nearly identical results** to the base model.

heads

Unlisted values are identical to those of the base model.

	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train steps	PPL (dev)	BLEU (dev)	params $\times 10^6$
base	6	512	2048	8	64	64	0.1	0.1	100K	4.92	25.8	65
(A)				1	512	512				5.29	24.9	
				4	128	128				5.00	25.5	
				16	32	32				4.91	25.8	
				32	16	16				5.01	25.4	
(B)					16					5.16	25.1	58
					32					5.01	25.4	60
(C)	2									6.11	23.7	36
	4									5.19	25.3	50
	8									4.88	25.5	80
		256			32	32				5.75	24.5	28
		1024			128	128				4.66	26.0	168
			1024							5.12	25.4	53
			4096							4.75	26.2	90
(D)							0.0			5.77	24.6	
							0.2			4.95	25.5	
								0.0		4.67	25.3	
								0.2		5.47	25.7	
(E)	positional embedding instead of sinusoids									4.92	25.7	
big	6	1024	4096	16			0.3		300K	**4.33**	**26.4**	213

作者还将这个模型应用到了句法分析任务上也取得了不错的结果。

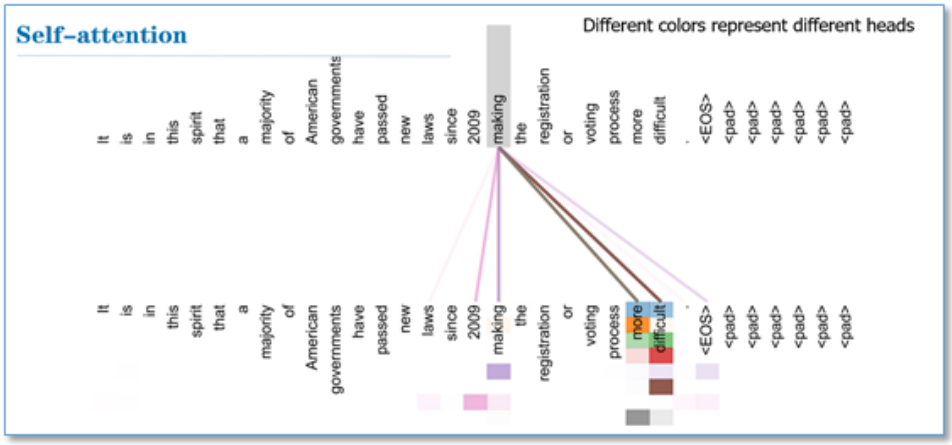
English Constituency Parsing

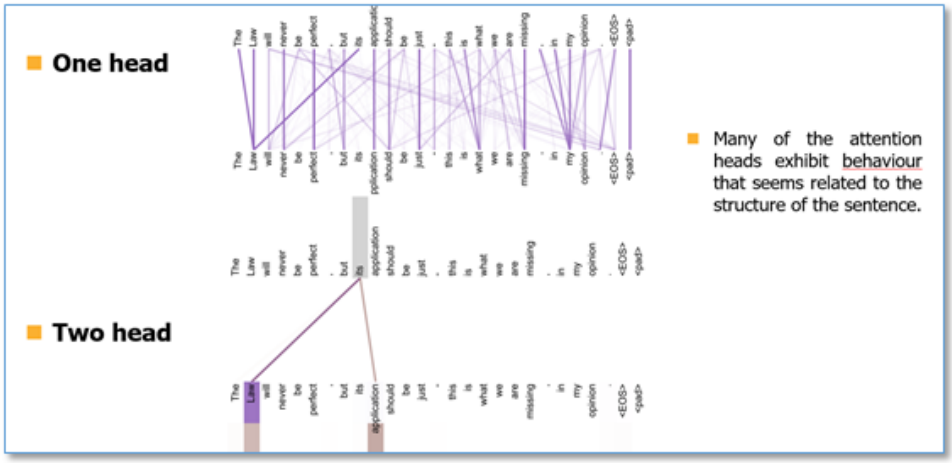
Table 4: The Transformer generalizes well to English constituency parsing (Results are on Section 23 of WSJ)

Parser	Training	WSJ 23 F1
Vinyals & Kaiser el al. (2014) [37]	WSJ only, discriminative	88.3
Petrov et al. (2006) [29]	WSJ only, discriminative	90.4
Zhu et al. (2013) [40]	WSJ only, discriminative	90.4
Dyer et al. (2016) [8]	WSJ only, discriminative	91.7
Transformer (4 layers)	WSJ only, discriminative	91.3
Zhu et al. (2013) [40]	semi-supervised	91.3
Huang & Harper (2009) [14]	semi-supervised	91.3
McClosky et al. (2006) [26]	semi-supervised	92.1
Vinyals & Kaiser el al. (2014) [37]	semi-supervised	92.1
Transformer (4 layers)	semi-supervised	92.7
Luong et al. (2015) [23]	multi-task	93.0
Dyer et al. (2016) [8]	generative	93.3

The Transformer can generalize to other tasks.

最后我们看一下attention可视化的效果（这里不同颜色代表attention不同头的结果，颜色越深attention值越大）。可以看到self-attention在这里可以学习到句子内部长距离依赖"making……more difficult"这个短语。在两个头和单头的比较中，可以看到单头"its"这个词只能学习到"law"的依赖关系，而两个头"its"不仅学习到了"law"还学习到了"application"依赖关系。多头能够从不同的表示子空间里学习相关信息。

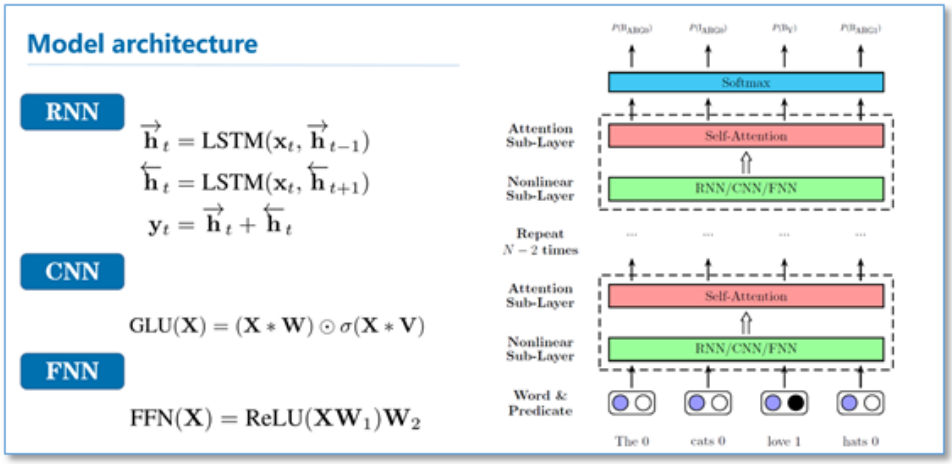




3 Self-attention in NLP

3.1 Deep Semantic Role Labeling with Self-Attention[8]

这篇论文来自AAAI2018，厦门大学Tan等人的工作。他们将self-attention应用到了语义角色标注任务（SRL）上，并取得了先进的结果。这篇论文中，作者将SRL作为一个序列标注问题，使用BIO标签进行标注。然后提出使用深度注意力网络（Deep Attentional Neural Network）进行标注，网络结构如下。在每一个网络块中，有一个RNN/CNN/FNN子层和一个self-attention子层组成。最后直接利用softmax当成标签分类进行序列标注。



该模型在CoNLL-2005和CoNll-2012的SRL数据集上都取得了先进结果。我们知道序列标注问题中，标签之间是有依赖关系的，比如标签I，应该是出现在标签B之后，而不应该出现在O之后。目前主流的序列标注模型是BiLSTM-CRF模型，利用CRF进行全局标签优化。在对比实验中，He et al和Zhou and Xu的模型分别使用了CRF和constrained decoding来处理这个问题。

可以看到本论文仅使用self-attention，作者认为在模型的顶层的attention层能够学习到标签潜在的依赖信息。

CoNLL-2005

Zhou and Xu: BiLSTM+CRF
He et al.: Highway BiLSTM+constrained decoding
DEEPATT: self-attention

Model	Development				WSJ Test				Brown Test				Combined
	P	R	F1	Comp.	P	R	F1	Comp.	P	R	F1	Comp.	F1
He et al. (Ensemble) (2017)	83.1	82.4	82.7	64.1	85.0	84.3	84.6	66.5	**74.9**	72.4	73.6	46.5	83.2
He et al. (Single) (2017)	81.6	81.6	81.6	62.3	83.1	83.0	83.1	64.3	72.8	71.4	72.1	44.8	81.6
Zhou and Xu (2015)	79.7	79.4	79.6	-	82.9	82.8	82.8	-	70.7	68.2	69.4	-	81.1
FitzGerald et al. (Struct., Ensemble) (2015)	81.2	76.7	78.9	55.1	82.5	78.2	80.3	57.3	74.5	70.0	72.2	41.3	-
Täckström et al. (Struct.) (2015)	81.2	76.2	78.6	54.4	82.3	77.6	79.9	56.0	74.3	68.6	71.3	39.8	-
Toutanova et al. (Ensemble) (2008)	-	-	78.6	58.7	81.9	78.8	80.3	60.1	-	-	68.8	40.8	-
Punyakanok et al. (Ensemble) (2008)	80.1	74.8	77.4	50.7	82.3	76.8	79.4	53.8	73.4	62.9	67.8	32.3	77.9
DEEPATT (RNN)	81.2	82.3	81.8	62.4	83.5	84.0	83.7	65.2	72.5	73.4	72.9	44.7	82.3
DEEPATT (CNN)	82.1	82.8	82.4	63.6	83.6	83.9	83.8	65.4	72.8	72.7	72.7	45.9	82.3
DEEPATT (FFN)	82.6	83.6	83.1	65.2	84.5	85.2	84.8	66.4	73.5	74.6	74.1	48.4	83.4
DEEPATT (FFN, Ensemble)	**84.3**	**84.9**	**84.6**	**67.3**	**85.9**	**86.3**	**86.1**	**69.0**	74.6	**75.0**	**74.8**	**48.6**	**84.6**

Table 1: Comparison with previous methods on the CoNLL-2005 dataset. We report the results in terms of precision (P), recall (R), F_1 and percentage of completely correct predicates (Comp.). Our single and ensemble model lead to substantial improvements over the previous state-of-the-art results.

CoNLL-2012

Latent dependency information is embedded in the topmost attention sub-layer learned by DEEPATT.

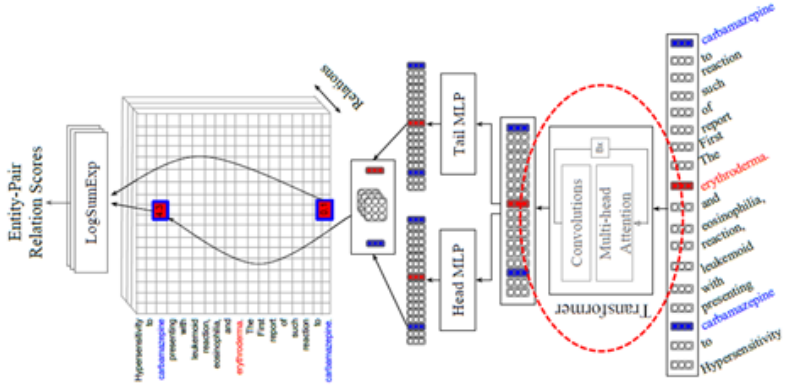
Model	Development				Test			
	P	R	F1	Comp.	P	R	F1	Comp.
He et al. (Ensemble) (2017)	83.5	83.2	83.4	67.5	**83.5**	83.3	83.4	68.5
He et al. (Single) (2017)	81.7	81.4	81.5	64.6	81.8	81.6	81.7	66.0
Zhou and Xu (2015)	-	-	81.1	-	-	-	81.3	-
FitzGerald et al. (Struct., Ensemble) (2015)	81.0	78.5	79.7	60.9	81.2	79.0	80.1	62.6
Täckström et al. (Struct., Ensemble) (2015)	80.5	77.8	79.1	60.1	80.6	78.2	79.4	61.8
Pradhan et al.(Revised) (2013)	-	-	-	-	78.5	76.6	77.5	55.8
DEEPATT (RNN)	81.0	82.3	81.6	64.6	80.9	82.2	81.5	65.7
DEEPATT (CNN)	80.1	82.5	81.3	65.0	79.8	82.6	81.2	66.1
DEEPATT (FFN)	82.2	83.6	82.9	66.7	81.9	83.6	82.7	67.5
DEEPATT (FFN, Ensemble)	**83.6**	**84.7**	**84.1**	**68.7**	83.3	**84.5**	**83.9**	**69.3**

Table 2: Experimental results on the CoNLL-2012 dataset. The metrics are the same as above. Again, our model achieves the state-of-the-art performance.

3.2 Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction[7]

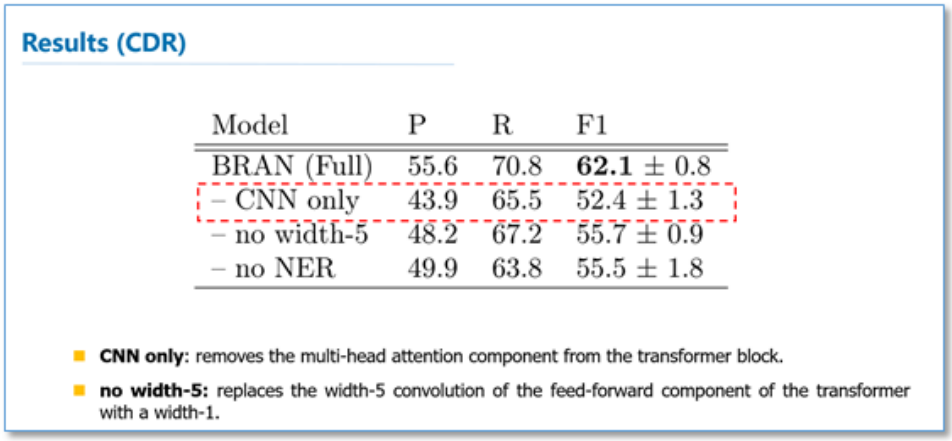
这篇论文是Andrew McCallum团队应用self-attention在生物医学关系抽取任务上的一个工作，应该是已经被NAACL2018接收。这篇论文作者提出了一个文档级别的生物关系抽取模型，里面做了不少工作，感兴趣的读者可以更深入阅读原文。我们这里只简单提一下他们self-attention的应用部分。论文模型的整体结构如下图，他们也是使用google提出包含self-attention的transformer来对输入文本进行表示学习，和原始的transformer略有不同在于他们使用了窗口大小为5的CNN代替了原始FNN。

Relation extraction architecture



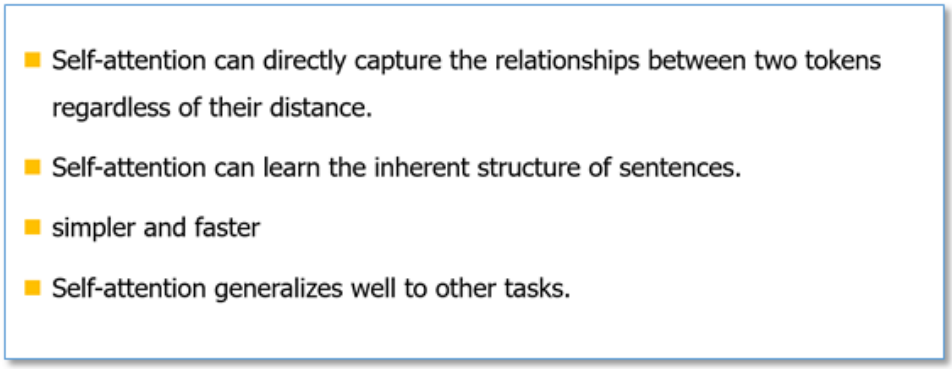
我们关注一下attention这部分的实验结果。他们在生物医学药物致病数据集上（Chemical Disease Relations，CDR）取得了先进结果。去掉self-attention

这层以后可以看到结果大幅度下降，而且使用窗口大小为5的CNN比原始的FNN在这个数据集上有更突出的表现。

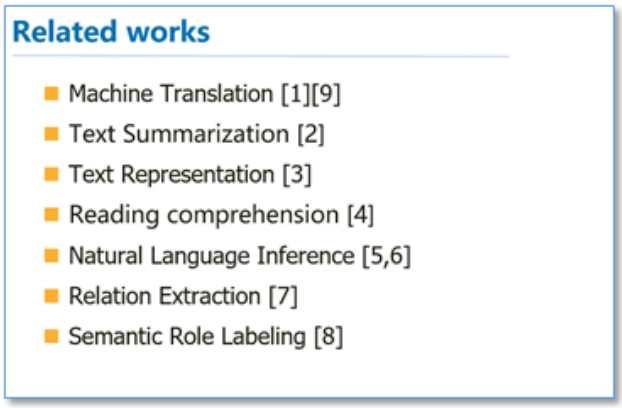


4 总结

最后进行一下总结，self-attention可以是一般attention的一种特殊情况，在self-attention中，Q=K=V每个序列中的单元和该序列中所有单元进行attention计算。Google提出的多头attention通过计算多次来捕获不同子空间上的相关信息。self-attention的特点在于无视词之间的距离直接计算依赖关系，能够学习一个句子的内部结构，实现也较为简单并行可以并行计算。从一些论文中看到，self-attention可以当成一个层和RNN，CNN，FNN等配合使用，成功应用于其他NLP任务。



除了Google提出的自注意力机制，目前也有不少其他相关工作，感兴趣的读者可以继续阅读。



参考文献

[1] Vaswani, Ashish, et al. **Attention is all you need**. Advances in Neural Information Processing Systems. 2017.

[2] Romain Paulus, Caiming Xiong, and Richard Socher**. A deep reinforced model for abstractive summarization**. arXiv preprint arXiv:1705.04304, 2017.

[3] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. **A structured self-attentive sentence embedding**. arXiv preprint arXiv:1703.03130, 2017.

[4] Jianpeng Cheng, Li Dong, and Mirella Lapata. **Long short-term memory-networks for machine reading**. arXiv preprint arXiv:1601.06733, 2016.

[5] Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Pan, S.; and Zhang, C. **Disan: Directional self-attention network for rnn/cnn-free language understanding.** arXiv preprint arXiv:1709.04696, 2017.

[6] Im, Jinbae, and Sungzoon Cho. **Distance-based Self-Attention Network for Natural Language Inference**. *arXiv preprint arXiv:1712.02047*, 2017.

[7] Verga P, Strubell E, McCallum A. **Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction**. arXiv preprint arXiv:1802.10569, 2018.

[8] Tan Z, Wang M, Xie J, et al. **Deep Semantic Role Labeling with Self-Attention**. AAAI 2018.

[9] Shaw, Peter, Jakob Uszkoreit, and Ashish Vaswani. **Self-Attention with Relative Position Representations**. *arXiv preprint arXiv:1803.02155* ,2018.

参考博客

张俊林，深度学习中的注意力机制(2017版)，
https://blog.csdn.net/malefactor/article/details/78767781

苏剑林，《Attention is All You Need》浅读（简介+代码），https://kexue.fm/archives/4765

分类: Deep Learning,NLP

标签: 自注意力机制, attention, self-attention, NLP, 深度学习

robert_ai
关注 - 2
粉丝 - 68
+加关注

5 0

« 上一篇: 基于神经网络的实体识别和关系抽取联合学习
» 下一篇: 基线系统需要受到更多关注：基于词向量的简单模型

posted on 2018-03-24 11:46 robert_ai 阅读(10391) 评论(0) 编辑 收藏