Transformer:2017

Attention Is All You Need

论文背景

循环神经网络，特别是其中的GRU和LSTM已经在序列建模问题中得到了很好的效果。

注意力机制作为网络的一部分引入，主要注重输入的依赖关系，而不是在序列中的距离，一般情况下都是注意力和RNN联合使用。

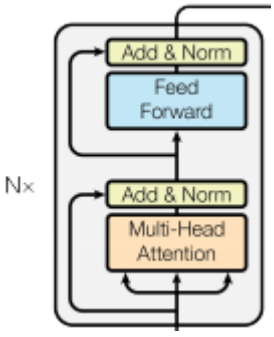
依赖卷积进行序列建模的模型受限于输入的数量，导致难以学习长距离输入的依赖关系

自注意力机制，将单个序列的不同位置联系起来，计算序列的表示

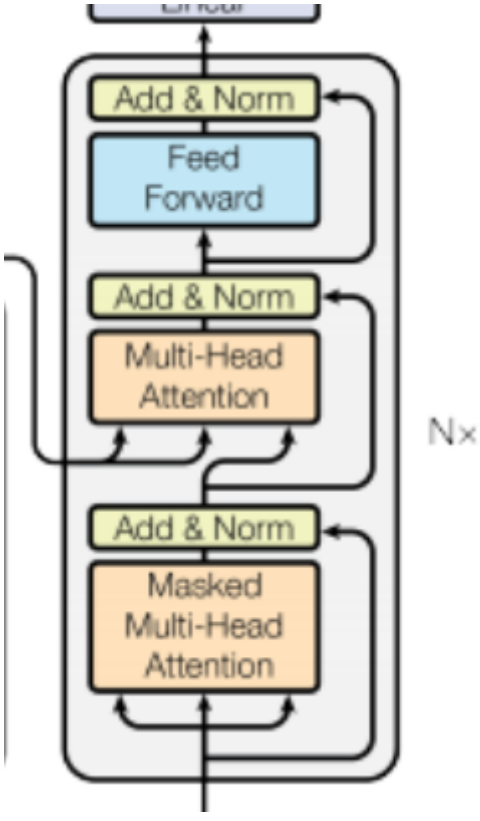
论文创新点

提出了一个完全依赖于注意力机制的模型transformer。

Encoder模块：多头自注意力机制、全连接层，LayerNormalization,残差结构



Decoder模块：mask模块(对位置i的预测只能依赖于小于i位置的token)，对encoder的输出做MHSA



自注意力机制：

点积注意力机制在实现上要易于加性注意力机制的实现

对于d_k比较大的值，其点积得到的值也会很大,进而会将softmax推到梯度很小的区域，所以考虑引入一个缩放因子$\sqrt{d_k}$.

Value，Query，Key——> $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$

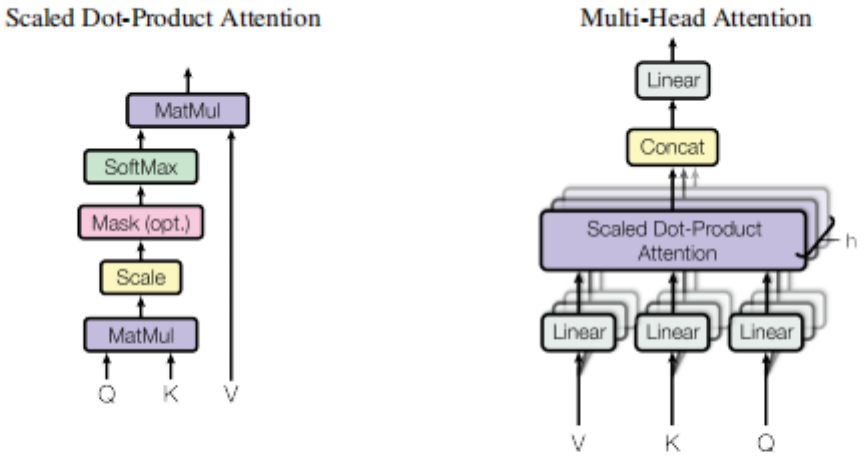


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

多头注意力机制

使得模型关注到不同位置的信息，然后通过W^o矩阵给汇总一下

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

FFN模块：

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

位置编码：为了让模型利用序列的顺序，同时又不具有递归性和卷积，采用嵌入位置编码的方式引入位置信息

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{\text{mak 1}}}\right)$$
$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{\text{makl}}}\right)$$

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

论文方法

左边为编码器部分，右边为解码器部分，通过使用堆叠的自注意力层和全连接层实现

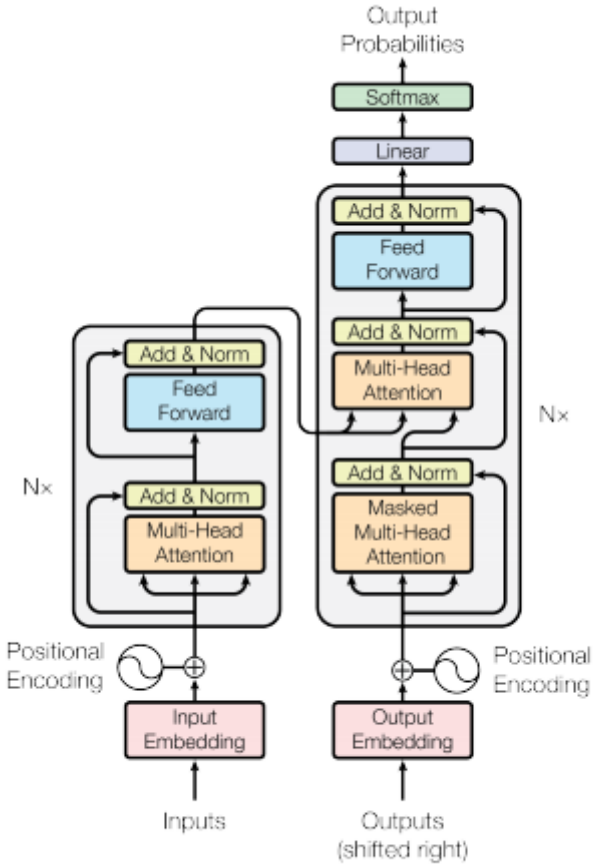


Figure 1: The Transformer - model architecture.

具体效果

	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

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 [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	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**	**41.0**	$2.3 \cdot 10^{19}$	

个人看法

论文的想法就是通过设计一个纯注意力机制来取缔NLP领域中卷积神经网络和循环神经网络的使用，主要是由于循环神经网络在对序列模型建模的时候需要进行递归以及容易发生梯度消失或者梯度爆炸的现象，transformer可以并行计算，进而减小了训练实践。Tramsformer的优越性在于其自注意力机制，是一种全局的注意力，虽然其SA模块不需要学习参数，但是其计算复杂度与序列的输入长度成二次方，因而在输入维度较大是很吃显存。