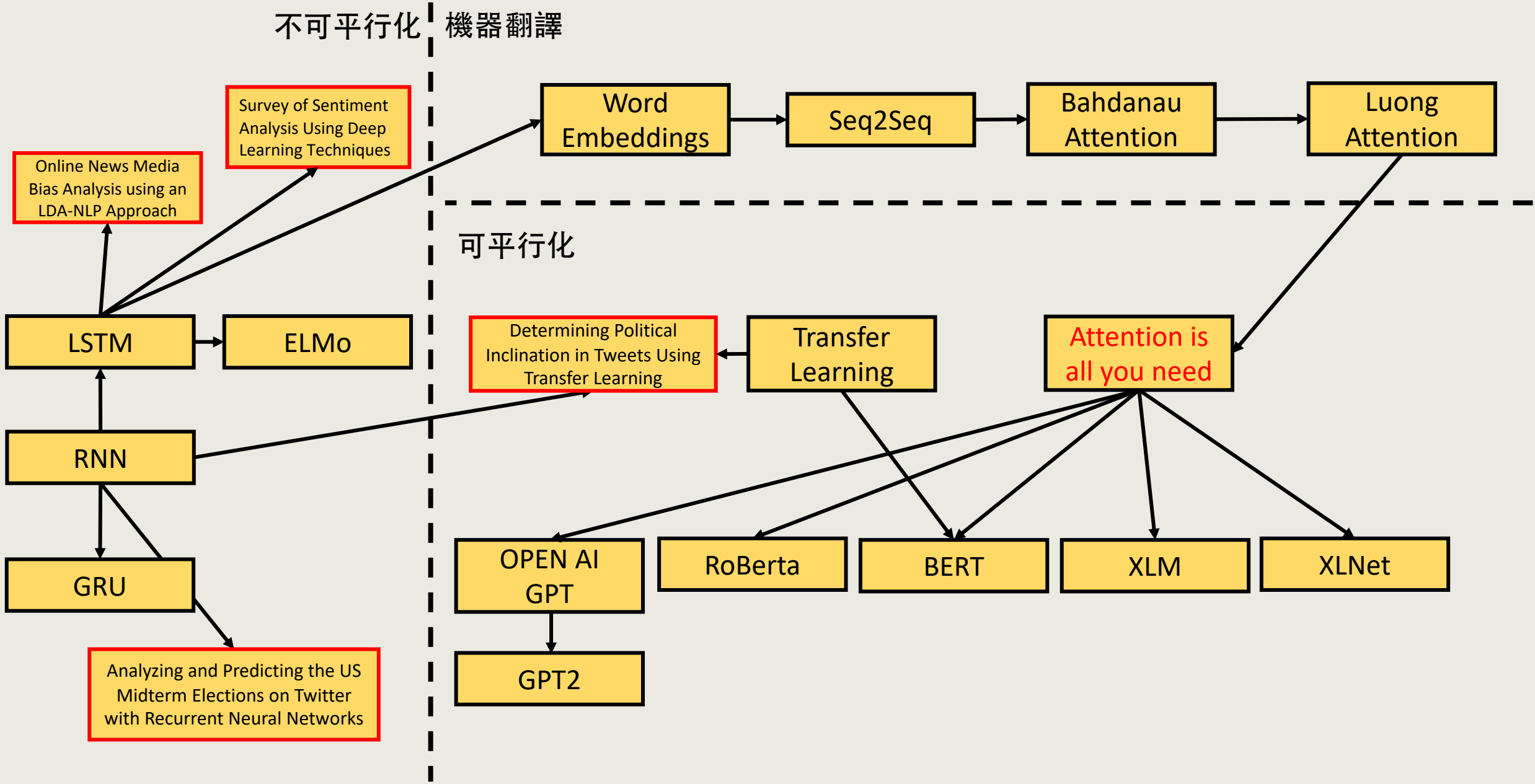
論文報告

圖譜


不可平行化
機器翻譯
Survey of Sentiment Analysis Using Deep Learning Techniques
Online News Media Bias Analysis using an LDA-NLP Approach
Word Embeddings
Seq2Seq
Bahdanau Attention
Luong Attention
可平行化
LSTM
ELMo
Determining Political Inclination in Tweets Using Transfer Learning
Transfer Learning
Attention is all you need
RNN
GRU
OPEN AI GPT
RoBerta
BERT
XLM
XLNet
GPT2
Analyzing and Predicting the US Midterm Elections on Twitter with Recurrent Neural Networks

Attention is all you need

作者： Ashish Vaswani
Noam Shazeer
Niki Parmar
Jakob Uszkoreit
Llion Jones
Aidan N. Gomez
Łukasz Kaiser
Illia Polosukhin

目　錄

簡　　介

簡介

論文名稱 : Attention is all you need

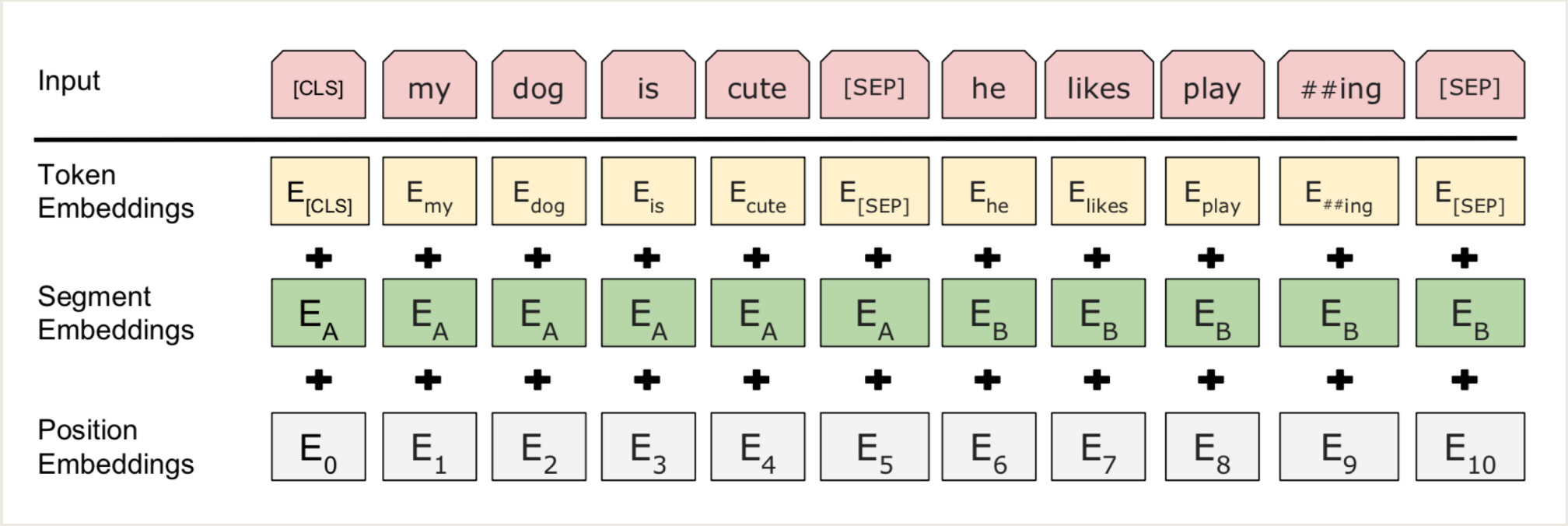
發表時間： 2017

研討會 : NIPS (Conference on Neural Information Processing Systems)

被引用次數 : 5202 (截至2018.12.19)

報告動機

- 一般的BERT跟這句話是誰說的沒有關係，但政治立場需要有



概　　述

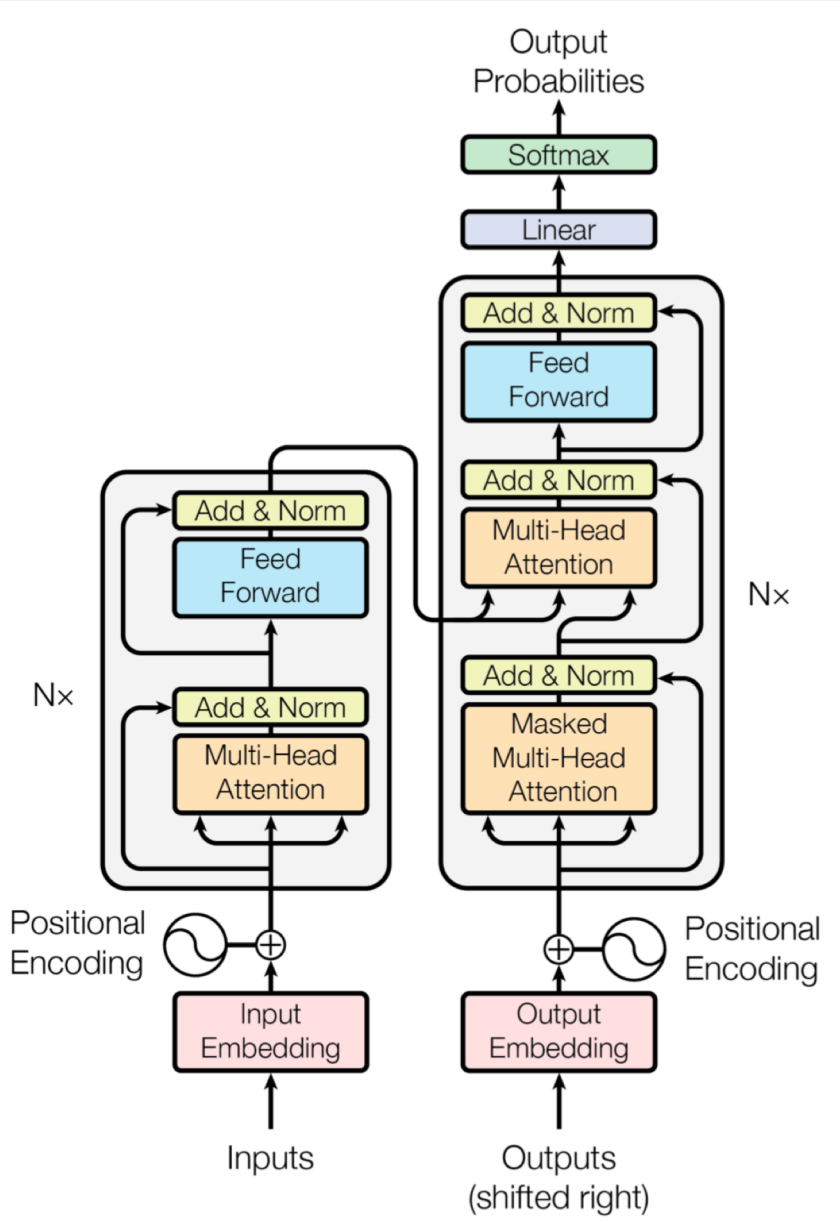
概述

欲解決問題：傳統機器翻譯使用遞歸神經網路重複 Attention 和卷積效率差且無法平行化

解決方法 ：提出 Transformer 模型，透過向量點乘 (self-Attention) 取代傳統 Attention、RNN 需要不斷的考慮輸出順序中的位置及時間點的hidden State花費大量算力又不能並行的問題。

模型及各層介紹

架構



- 跟傳統的機器翻譯一樣，擁有 Encoder，Decoder 結構
- 用Multi-Head Attention取代原本的Attention Hidden Layer

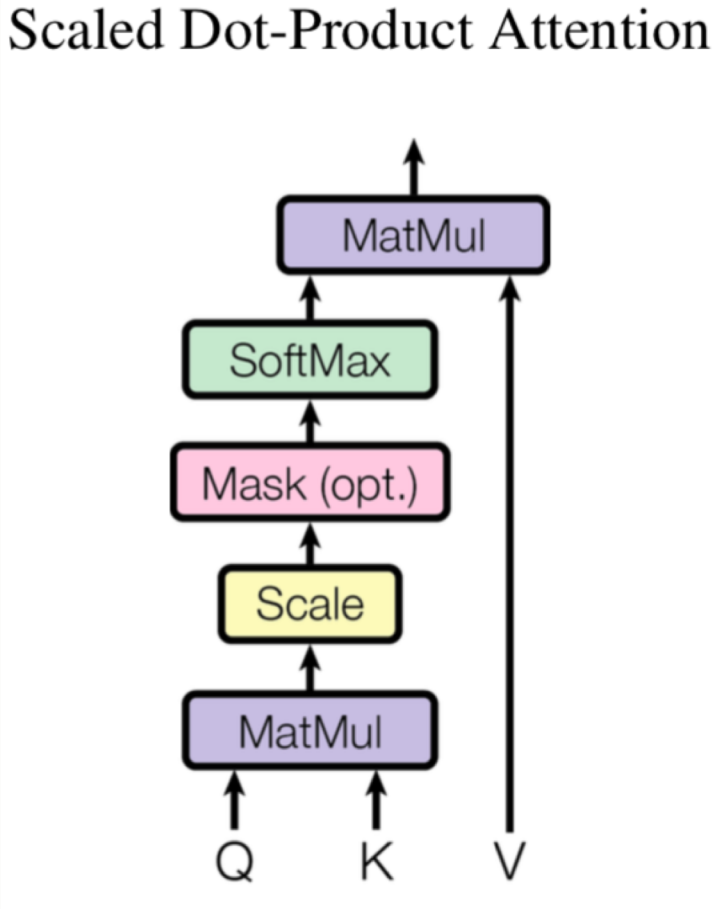
Encoder

- 由六個Encoder堆疊而成
- 每層又分為Multi-Head Attention跟前饋網路
- 每次輸出都要經過歸一化
- 所有子層輸出d = 512

Decoder

- 由六個Decoder堆疊而成
- 第一層需遮蓋後續位置以防止Decoder關注到後面位置的訊息，確保預測只能依賴小於當前時間點之前的訊息。

架構



$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

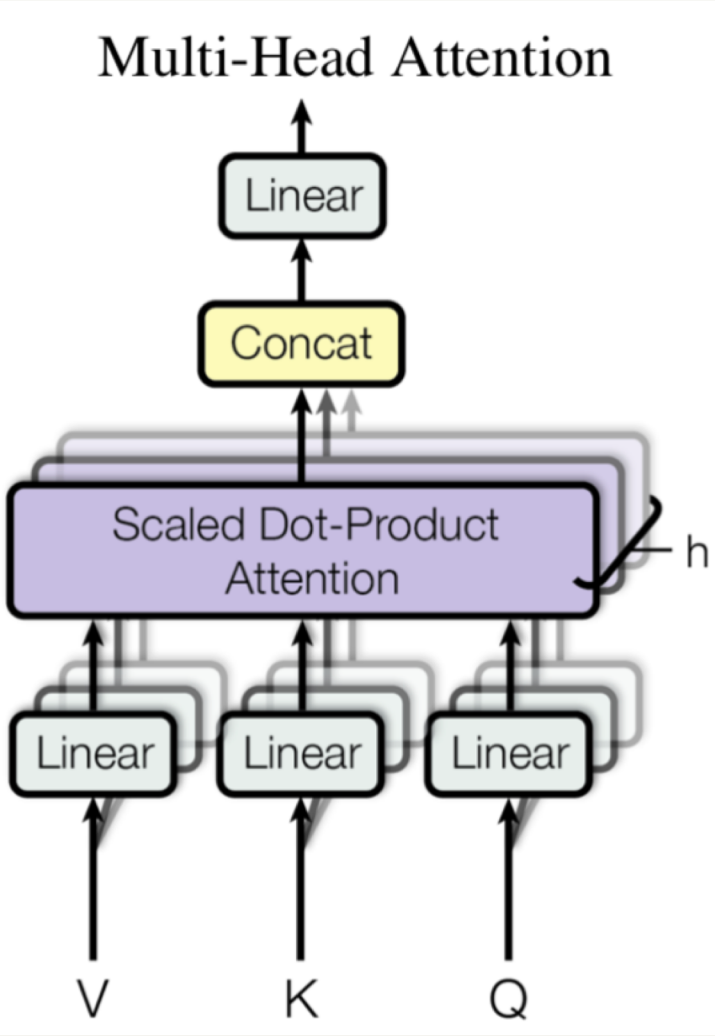
- Q x K^T 目的是為了取得每個字之間的關聯性
- d_k 是指字數轉為向量時的維度數量，相除目的是為了在 softmax 之前先將文字向量平滑化
- 取得 softmax 的值後代表每個字的注意力機率，最後乘上V即可得到每個字的重要程度

架構

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

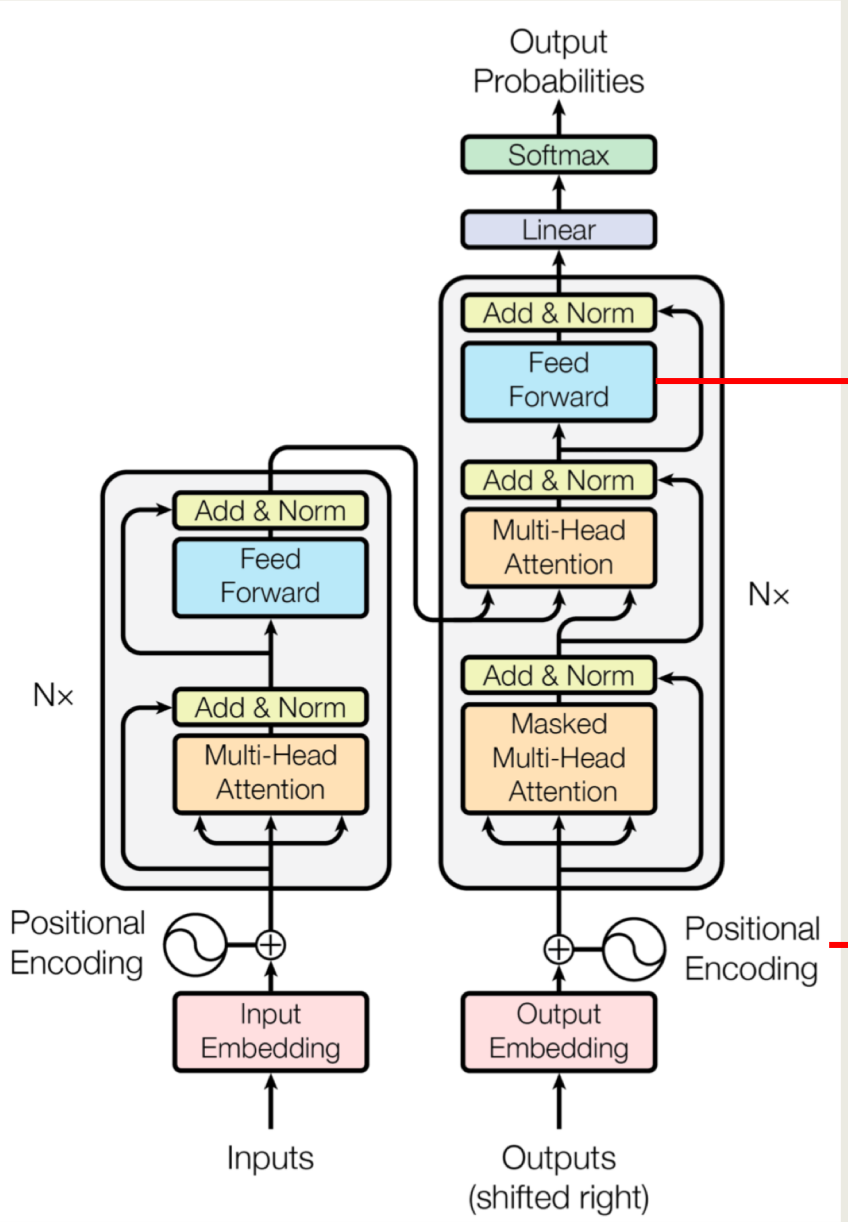
	機器	翻譯
Embedding	X1	X2
Q	X1 x Wq	X2 x Wq
K	X1 x Wk	X2 x Wk
V	X1 x Wv	X2 x Wv
QK^T	Q1 x $K1^T$ =120	Q1 x $K2^T$= 80
Divide by 8 $d_2^{1/2}$	15	10
Softmax	0.6	0.4
Softmax x Value	0.6 x V1	0.4 x V2
Sum	Z	

架構



- 用很多不同的 W_{qkv} 矩陣做運算之後再將結果Concat起來
- 論文為head數=8

架構



- 加入Relu的前饋網路

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

- 提出兩種PE,一種是用現在時間點另一種是用下一個

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

架構

- 比較其他種類的層

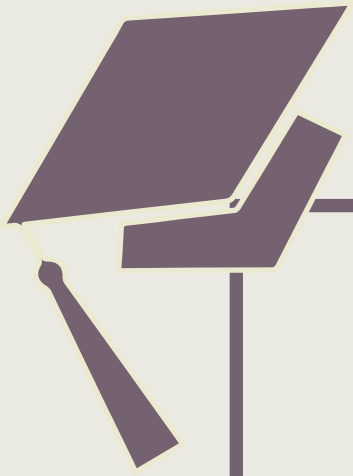
Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

- 比較其他模型

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**	$2.3 \cdot 10^{19}$	

架構

	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



结　　論

結論

本文介紹了Transformer，這是完全基於注意力的第一個序列轉導模型，用Multi-Head Self-Attention代替了Encoder-Decoder體系結構中最常用的循環層。對於翻譯任務，與基於循環層或卷積層的體系結構相比，可以大大加快Transformer的訓練速度。在WMT 2014英語轉德語和WMT 2014英語轉法語翻譯任務上，都達到了最新水平。在前一項任務中，我們最好的模型甚至勝過所有先前模型的Ensembles。計畫擴展到涉及文本以外的涉及輸入和輸出方式的問題，並研究局部受限的注意機制，以有效處理大型輸入和輸出，例如圖像，音頻和視頻。