# Attention is all you need

## 我的总结

本文提出一种不用CNN和RNN，只用attention的结构transformer。相比RNN，transformer可以并行，且可以考虑任意两个位置之间的信息。

比较类似seq2seq模型，适合机器翻译，但也可以做其他任务。

## 背景与贡献

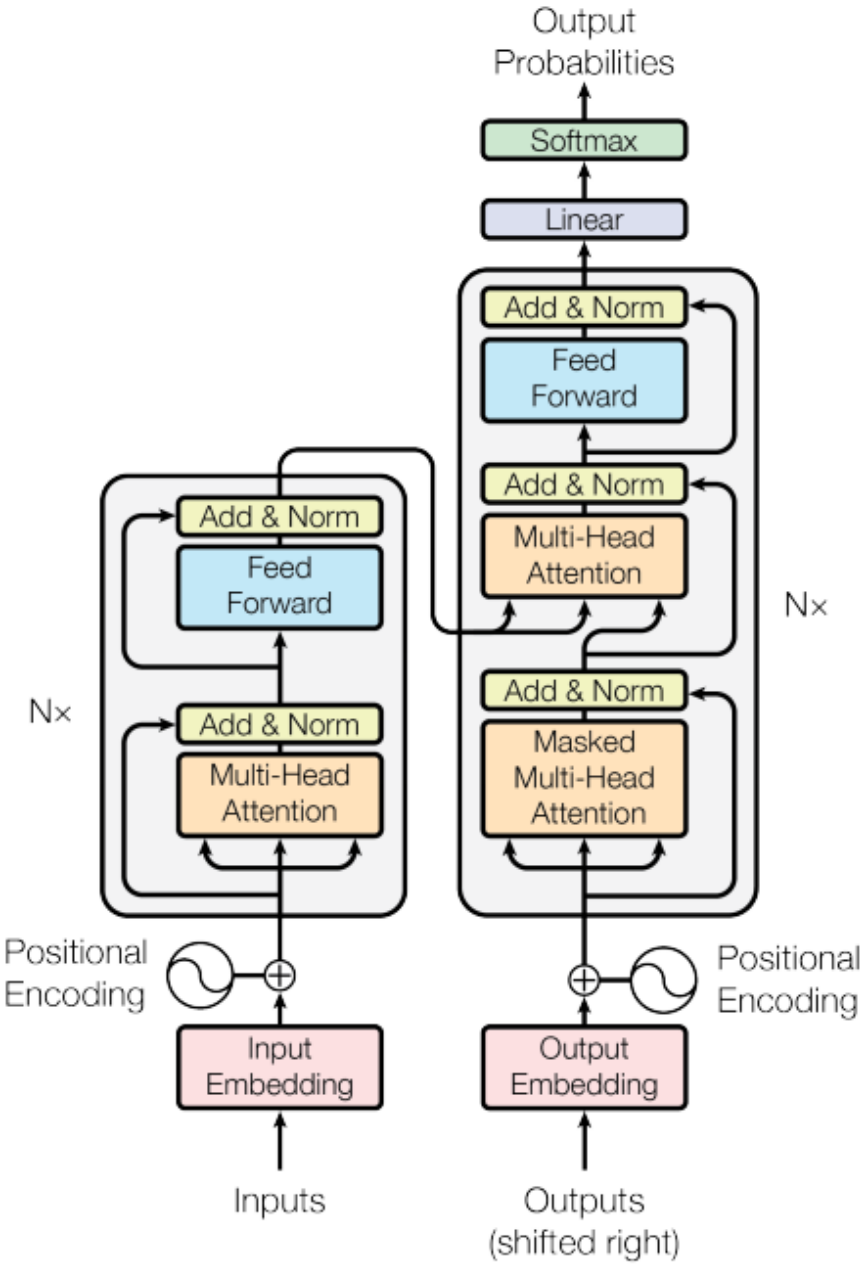
RNN模型**不能并行。**

**贡献**

- 摒弃RNN和CNN结构，采用**全attention**结构，增加了并行度。
- 在各个任务达到SOTA。

## 思路与方法

### 模型结构




## Encoder

$N = 6$层的encoder，每个sublayer(图中feed forward/attention)的输出为$LayerNorm(x + Sublayer(x))$。

encoder可以整句输入。

## Decoder

比encoder多了第二层，接受encoder的输出，为target在source的维度做attention。

decoder是一个一个word输入的，因为输入取决于上一次的预测结果。

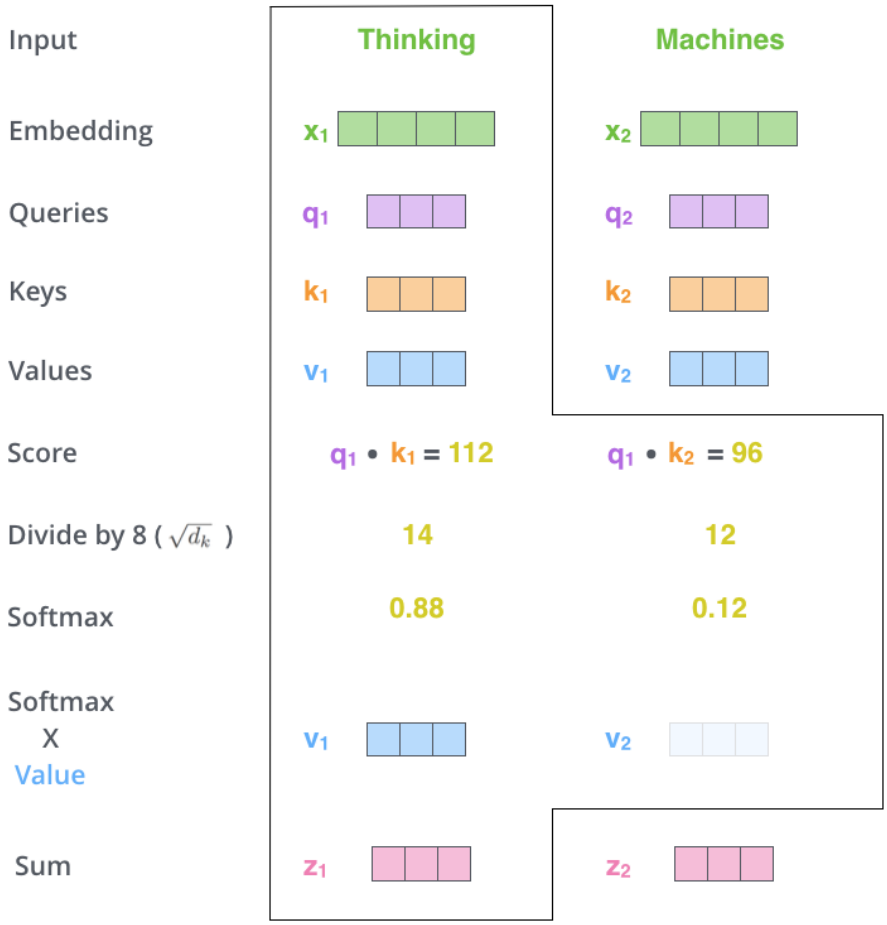
# Attention

## Scaled Dot-Product Attention

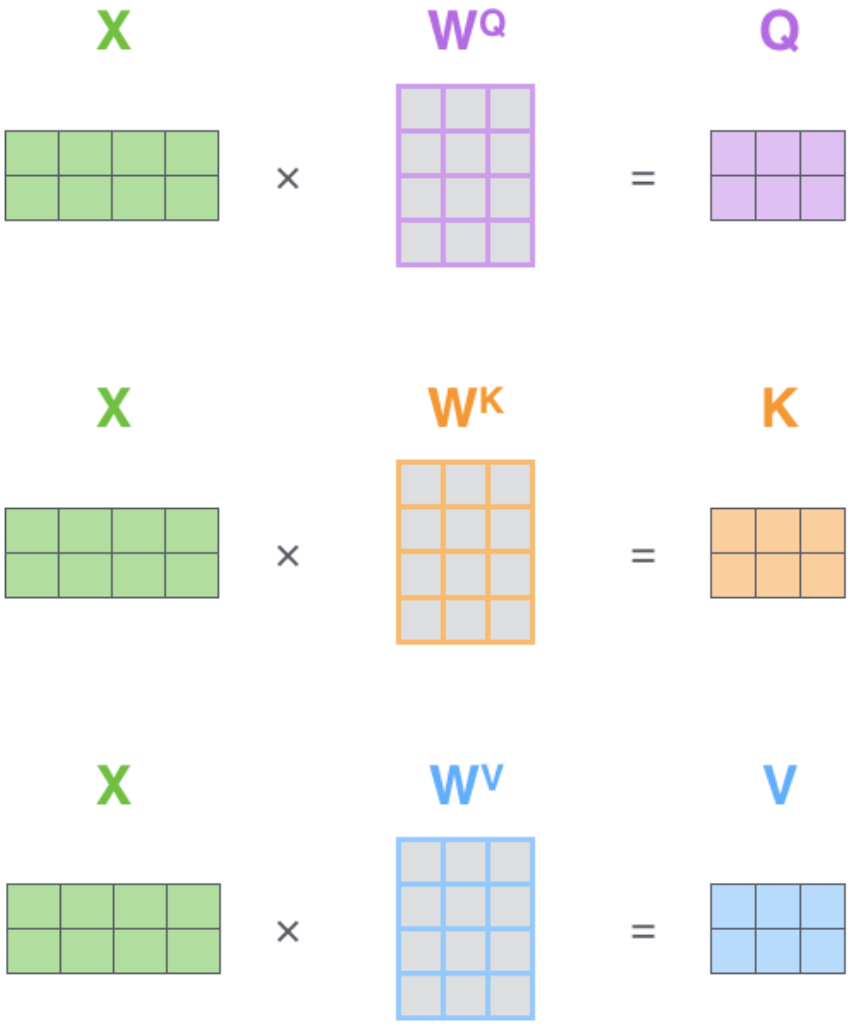
对于每个word，输入$x_i \in R^{d_m}$的input embedding。

$q_i, k_i \in R^{d_k}, v_i \in R^{d_v}$通过$x_i$乘以权重矩阵$W^Q, W^k \in R^{d_m * d_k}, W^v \in R^{d_m * d_v}$得到。

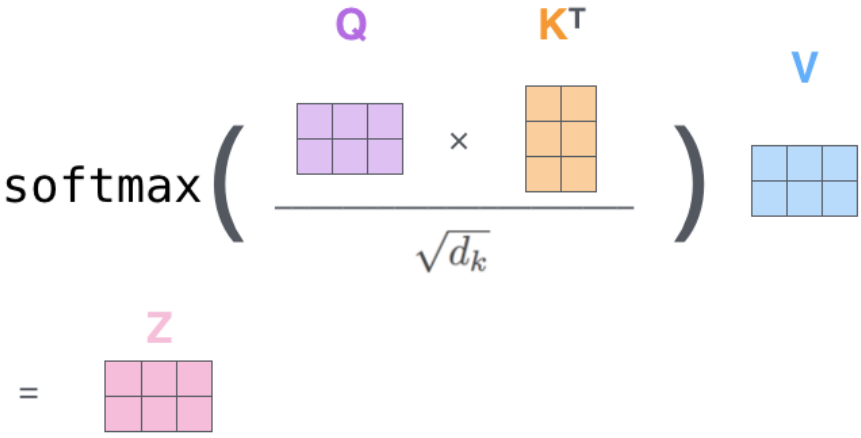
在attention中，对于第i个word，用$q_i^T k_j$计算和其他word的score，再经过softmax计算权值。最后累加加权的$v_j$得到$\sum_j \alpha_j v_j$(j可以等于i)作为第i个word的attention输出。




n个word组成矩阵$Q, K \in R^{n * d_k}, V \in R^{n * d_v}$。下图中$n = 2, d_m = 4, d_k = d_v = 3$。




以矩阵形式计算n个word的attention，其中softmax是作用在每一行，softmax输出维度$(2,)$的矩阵，和$V$进行元素乘。得到的$Z \in R^{n*d_v}$。




## Multi-Head Attention

不同head的attention可以关注到不同信息，比如$head_1$关注词性，$head2$关注entity。因此采用h个attention，文中$h = 8$。

具体来说，第i个head有自己的$W_i^Q, W_i^K, W_i^V$，不同head经过attention后的输出concat起来。

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

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

所以一个multi-head attention的输入和输出维度相同，都是$R^{n*d_m}$。

### Applications of attention in the model

- 在"encoder-decoder attention"中(decoder的第二个attention)，$Q$是上一层的decoder输出，$K, V$是encoder的输出。
- 在encoder的attention中，每个position可以和任意其他position计算attention。
- 在decoder的attention中，每个position只能考虑leftward的信息，需要mask。

## Feed Forward

两层的MLP + 第一层ReLU。

## Positional Encoding

由于没有RNN和CNN，仅有attention不能用到position的信息。打乱句子顺序会得到完全相同的结果，这是不合理的。

因此本文加入position embedding：

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

# 实验

## 训练

在WMT上训练，通过BPE构造词典。

## 结果

- 机器翻译在WMT 2014上BLEU第一
- ablation study

Table 3: Variations on the Transformer architecture. Unlisted values are identical to those of the base model. All metrics are on the English-to-German translation development set, newstest2013. Listed perplexities are per-wordpiece, according to our byte-pair encoding, and should not be compared to per-word perplexities.

| | $N$ | $d_{\text{model}}$ | $d_{\text{ff}}$ | $h$ | $d_k$ | $d_v$ | $P_{drop}$ | $\epsilon_{ls}$ | 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 |

- English consistency parsing在WSJ(Wall Street Journal)上，在没有做task-specific tuning的情况下仍然达到接近SOTA水平。