Week4

Transformer 概览

Transformer 的基本框架是 Encoder-Decoder 框架，由 6 个相同的 Encoder(编码器)和 6 个相同的 Decoder(解码器)叠加而成。Encoder 组件是 self-attention(自注意力)和 feed forward(前馈神经网络)，而 Decoder 组件除开这两个，中间多了一个连接 Encoder 和 Decoder 的 Encode-Decoder attention 机制。整体架构图见图一。

1. Encoder 组成

a. 把 input 生成的词向量和位置编码(positional encoding)叠加，作为嵌入向量 X 输入到 encoder 中，假定 X 的维度是 m*512(m 为单词数，512 为向量维度)；

b. 计算嵌入向量的 Query(Q)、Key(K)和 Value(V)矩阵。为了加快运算和扩展关注不同位置，运用多头注意力(Mult-head attention)计算。即可引入 8 个(也可引入其他数目)Q、K、V 的权重矩阵，维度为 512*64，和嵌入向量相乘后，分别得到 8 个 m*64 维的 Q、K、V 向量;

c. 计算注意力矩阵，得到 8 个 z1 向量(注意力矩阵)，公式如下：

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

把此 8 个向量进行拼接，得到 m*512 维的向量。乘以一个权重矩阵(降维及给不同的头赋予权重)，最终得到 self-attentiion 过程的输出向量 Z;

d. 做 Add 和 Normalize 处理，即把 Z 和 X 进行相加后，再进行单一样本归一化。再把此结果输出给 Feed Forwrad 网络，通过非线性的变化来转换特征，再基于神经元做相加和归一化处理，得到 Encoder 的输出；

每一层 Encoder 的输出是下一层 Encoder 的输入,最后一个 Encoder 的输出进入到 Decoder 层中；

2. Decoder 组成

a. 此层 self-attention 计算和 Encoder 层的类似，但因为在预测中不知道下一个序列，只允许关注输出序列中较前的位置，因此这里采用 mask 遮罩；

b. 把 self-attention 层得到的 Query 矩阵，和 Encoder 最终层的 Key 和 Value 矩阵,输入到 Encoder-Decoder attention 中，计算自注意力矩阵；

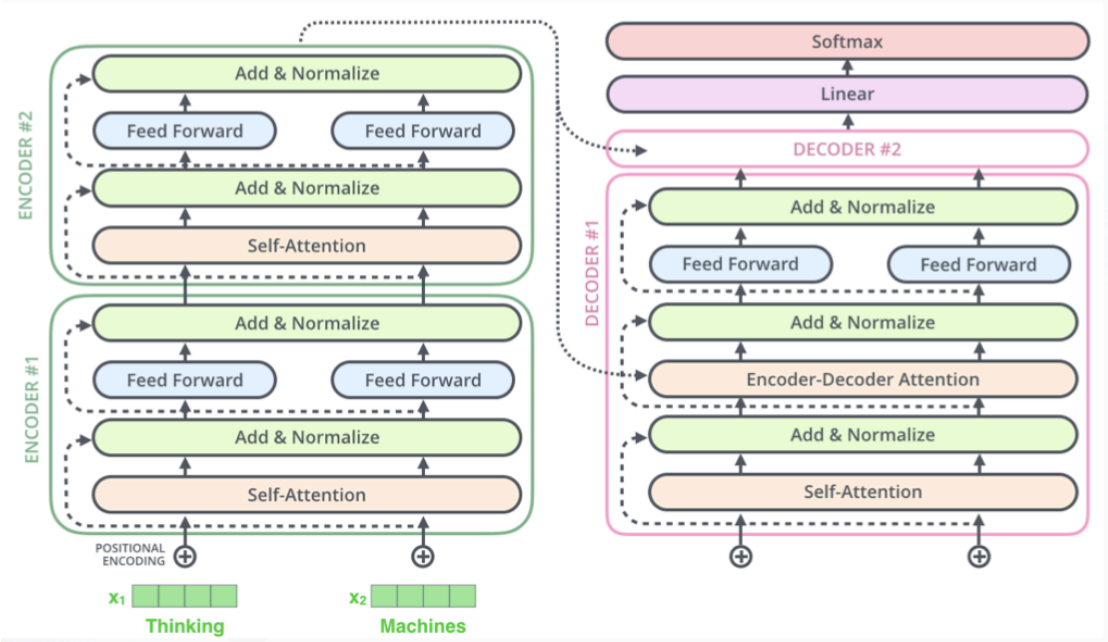
c. 和 Encoder 层一样，把得到的自注意力矩阵输入到 feed forward 中，同样进行 Add 和 Normalize 处理，得到 Decoder 层的输出矩阵；

每一层 Decoder 的输出是下一层 Decoder 的输入,将最后一个 Decoder 的输出映射到 vocab size 维度大小的向量，softmax 将得分转化为对应词概率最大的输出。

3. 优缺点

可做并行计算以及学习长距离依赖，但如果输入数据由时间或空间关系，必须加上位置编码。

图一


Softmax
Linear
DECODER #2
DECODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Encoder-Decoder Attention
Add & Normalize
Self-Attention
ENCODER #2
ENCODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
POSITIONAL ENCODING
x1
x2
Thinking
Machines