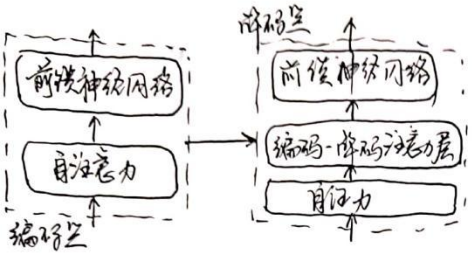
Transformer

① 6个编码器，6个解码器

所有的编码器结构相同，但不共享参数



②每个单词被编码成512维向量

1)词嵌入只在最底层编码器的输入

高层编码器输入是低层编码器输出

向量列表大小是超参，一般是最长句子的长度.

2)位置编码

$$\vec{p}_t^{(i)} = f(t)^{(i)} = \begin{cases} \sin(\omega_k \cdot t), & \text{if } i=2k \\ \cos(\omega_k \cdot t), & \text{if } i=2k+1 \end{cases}$$

$$\omega_k = \frac{1}{10000^{2k/d}}$$

较高维向量维度减小，pt是一个包含每个位置信息的正弦余弦对

$$\vec{p}_t = \begin{bmatrix} \sin(\omega_1 \cdot t) \\ \cos(\omega_1 \cdot t) \\ \sin(\omega_2 \cdot t) \\ \cos(\omega_2 \cdot t) \\ \vdots \\ \sin(\omega_{\frac{d}{2}} \cdot t) \\ \cos(\omega_{\frac{d}{2}} \cdot t) \end{bmatrix}$$

将论文加位置编码加到模型输入之上

$$\psi'(w_t) = \psi(w_t) + \vec{p}_t \quad d_{we} = d_{PE}$$

3)能够更容易关注相对位置信息，相对位置之间具有线性关系

$$e_t = E_{t:} = \begin{bmatrix} \sin(\frac{t}{f_1}) \\ \cos(\frac{t}{f_1}) \\ \sin(\frac{t}{f_2}) \\ \cos(\frac{t}{f_2}) \\ \vdots \\ \sin(\frac{t}{f_d}) \\ \cos(\frac{t}{f_d}) \end{bmatrix} \qquad f_m = \frac{1}{\lambda_m} = 10000^{\frac{2m}{dm}}$$

存在一个线性转置 $T^k \in R^{dm \times dm}$，对 $\forall t \in \{1, \cdots nk\}$

则t位置偏移量 $k \in \{1, \cdots n\}$，下列

$$T^{(k)} E_{t,:} = E_{t+k,:}$$

3)自注意力机制

$Q = XW_Q$，$K = XW_K$，$V = XW_V$

X在第1层是编码，前后面是前一层输出

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{dk}}\right)V$$

(1)使用不同的Q、K，在不同空间投影，增强表达能力，提高泛化能力

(2)"多头"机制

多头扩展了模型专注不同位置的能力

给出了多个"表示子空间"

4)残差网络：

残差+归一化

每个子层都有一个残差连接.

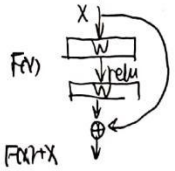
①梯度消失/爆炸

②网络退化：

随着深度增加，网络的表现是先增加至饱和，然后迅速下降.

对神经网络来说，恒等映射并不容易拟合(更深的网络不应该表现更差)

$$z^{(l)} = H(a^{(l-1)}) = a^{(l-1)} + F(a^{(l)})$$



(1)解决网络退化 } 信息前后传播更顺畅

(2)缓解梯度消失

(3)集成学习的角度

(4)缓解梯度破碎