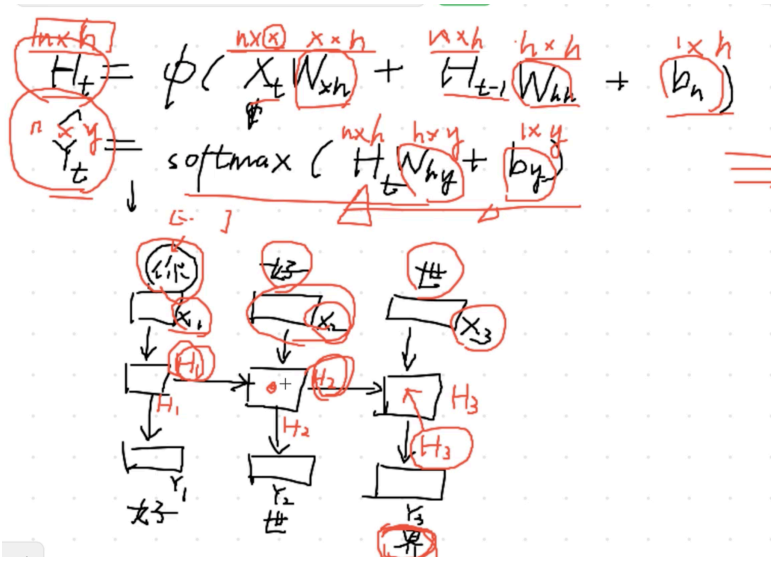
1 自回归解码

- 自回归

  用历史数据的线形加权和+噪声扰动表示当前当前时序数据

RNN

t-1时刻的输出是t时刻的输入



transfomer

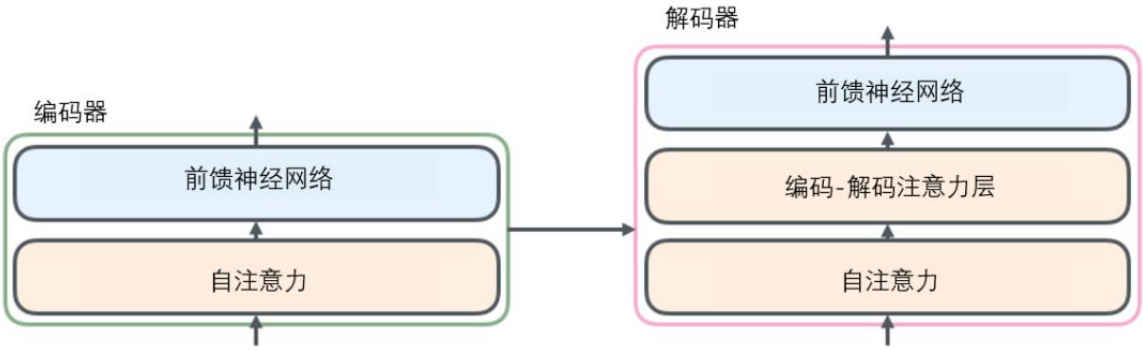
推断时把上一次的输出embedding后作为这一次输入

2 架构

输入序列经过**word embedding**和**positional encoding**相加后，输入到encoder。

输出序列经过**word embedding**和**positional encoding**相加后，输入到decoder。

最后，decoder输出的结果，经过一个线性层，然后计算softmax。



3 Encoder

encoder由6层相同的层组成，每一层分别由两部分组成：

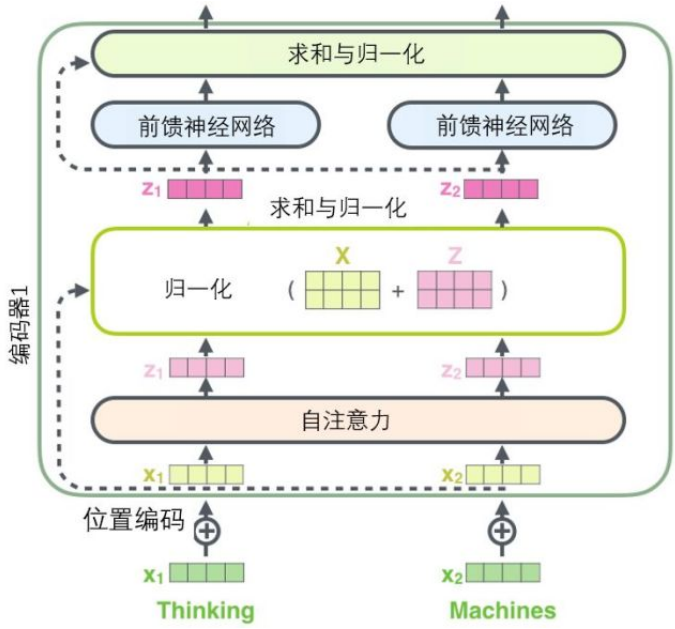
- 第一部分是一个**multi-head self-attention mechanism**

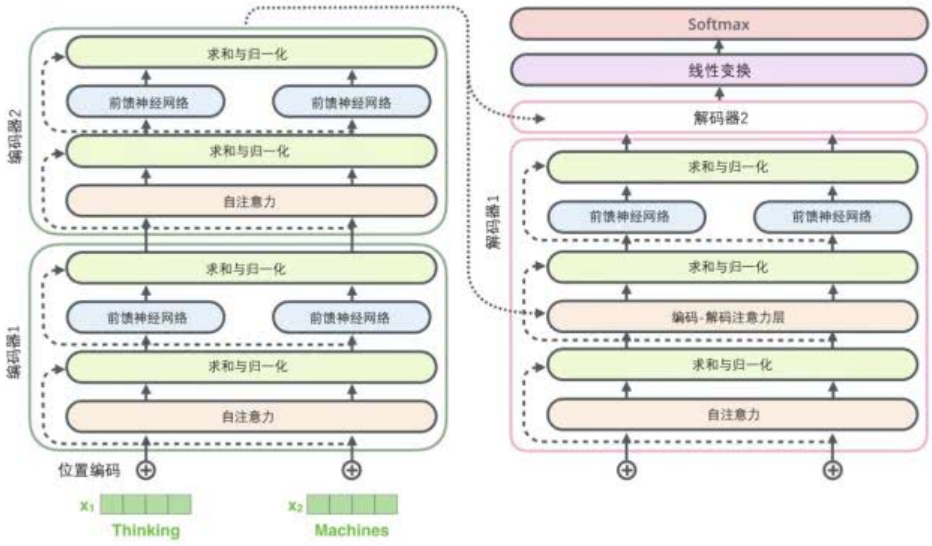
  有依赖关系，无法并行

- 第二部分是一个**position-wise feed-forward network**，是一个全连接层

  无依赖关系，可以并行

两个部分，都有一个**残差连接(residual connection)**，然后接着一个**Layer Normalization**。





- 输入

  word----embedding---->word vector ----自注意层--->得分矩阵z

  embedding可以加入位置编码：词嵌入+ 位置编码矩阵 ->encoder的输入

- multi-head self-attention

$$MultiHead(Q,K,V) = Concat(head_1,\dots,head_h)W^0$$

其中，

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

- 残差连接

  学习目标从f x变为fx+x

  作用：在网络很深时防止梯度消失

  实现

  残差并非一定需要是x

Q

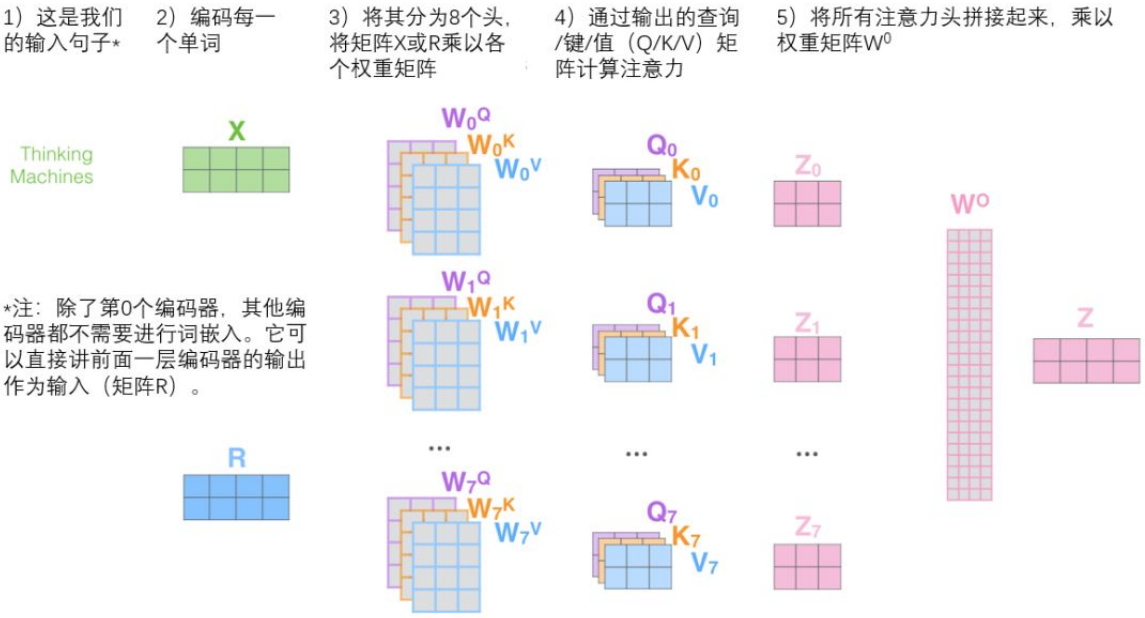
- 多头注意力机制

  为什么能更好地知道it指代的是什么？比如it的不同注意力头部更好地集中在不同单词上，简单地求和会平均掉局部

  怎么work的?

  这里对x的编码怎么编出8*512

  线形分割。假设分成 h 份，在 d_Q 、d_K 和 d_V的维度上进行切分。进入到scaled dot-product attention 的 d_K 实际上等于未进入之前的 $\frac{D_K}{h}$ 。



Decoder

- 第一个部分是 multi-head self-attention mechanism
- 第二部分是 multi-head context-attention mechanism
- 第三部分是一个 position-wise feed-forward network

和 encoder 一样，上面三个部分的每一个部分，都有一个残差连接，后接一个 **Layer Normalization**。

decoder 和 encoder 不同的地方在 multi-head context-attention mechanism

训练

- **输入** 上次的输入+embedding（标记中下一个单词）(实操中一般一次把标记全输进去加一个sequencing mask);第一次输入是一个特殊的token （bos/eos/等等）
- **输出** 一个代表下一个预测单词的实数向量

inference

- **输入** 上次的输入+embedding（上次的输出）
- **输出** 一个代表下一个预测单词的实数向量

一些细节

Attention层

k q v的含义

- encoder： self-attention

  Q、K、V 是上一层 encoder 的输出

  （第一层 encoder，是 word embedding 和 positional encoding 相加得到的输入）
- decoder： self-attention 中

  Q、K、V 是上一层 decoder 的输出

  （第一层 decoder是 word embedding 和 positional encoding 相加得到的输入）

  但是对于 decoder，不希望它能获得将来的信息，因此还需要需要进行 sequence masking。
- 在 encoder-decoder attention 中，Q 来自于 decoder 的上一层的输出，K 和 V 是encoder 的输出，**k=v**
- Q、K、V 的维度都是一样的

层归一化

区别于批量归一化(对每个batch求均值),层归一化是对每一个样本取均值

Mask

- 为保证超过样本长度的部分不处理的paddling mask
- decoder self attention 的sequential mask：上三角矩阵，保证读不到未来数据

位置编码

优点：能扩展到未知序列长度（类似变长数组）

$$PE(pos, 2i) = sin(pos/10000^{2i/d_{model}})$$

$$PE(pos, 2i+1) = cos(pos/10000^{2i/d_{model}})$$

pos 是指词语在序列中的位置。偶数位置使用正弦编码，奇数位置使用余弦编码

note 上述是绝对位置编码。但是使用三角函数保证了相对位置信息也被利用了

reference

1.动手学深度学习第10章，建议跟着视频一起看
2.Transformer中文图解：https://blog.csdn.net/longxinchen_ml/article/details/86533005
3.Transformer代码+注释：http://nlp.seas.harvard.edu/2018/04/03/attention.html?tdsourcetag=s_pcqq_aiomsg
4.https://juejin.im/post/6844903680487981069#comment

5.https://www.nowcoder.com/discuss/258321