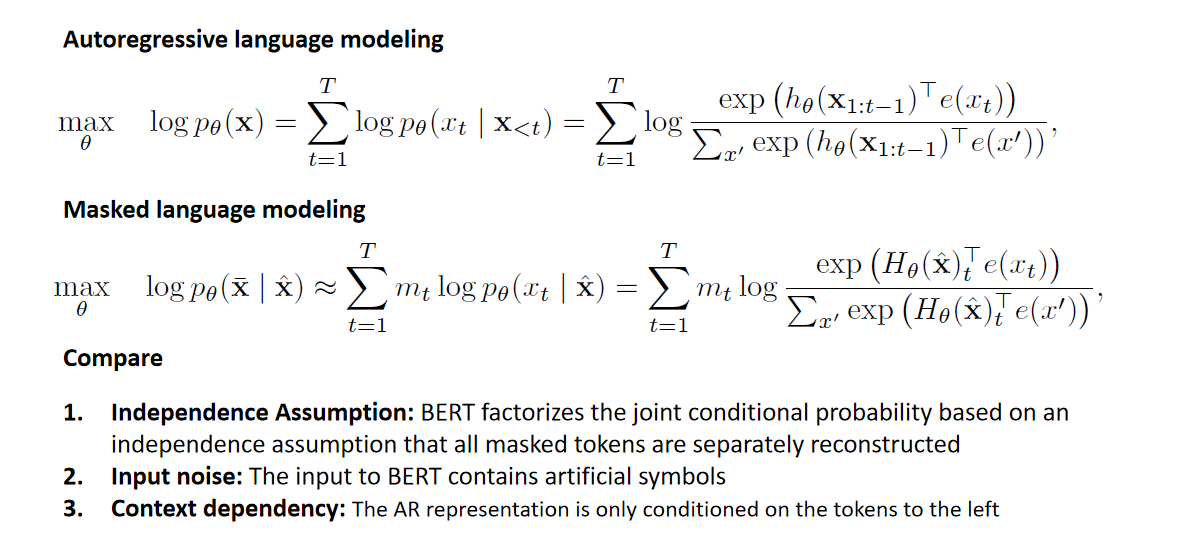
# 语言模型的任务

语言模型的训练目标：

传统语言模型autoregressive language modeling

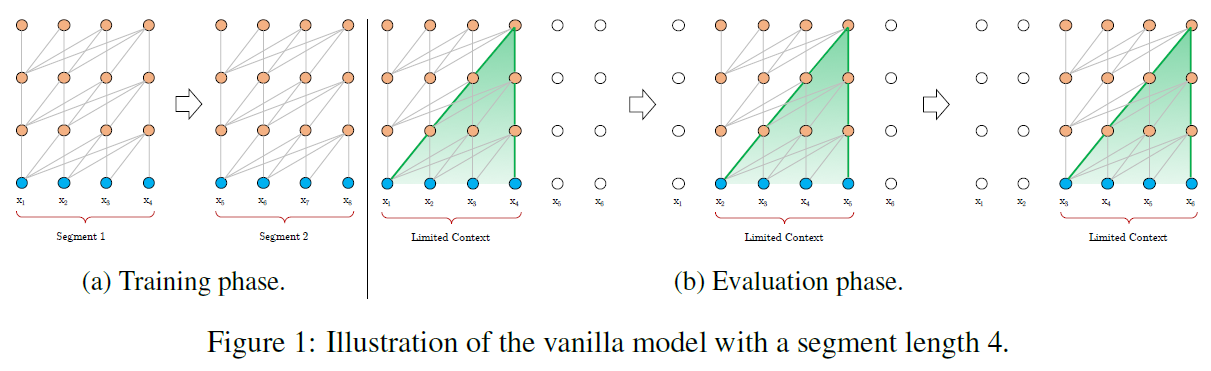
Vs

Bert中提出的 masked language modeling



# XLNet

## Transformer-XL



No long sequence information (hard to predict first few tokens)

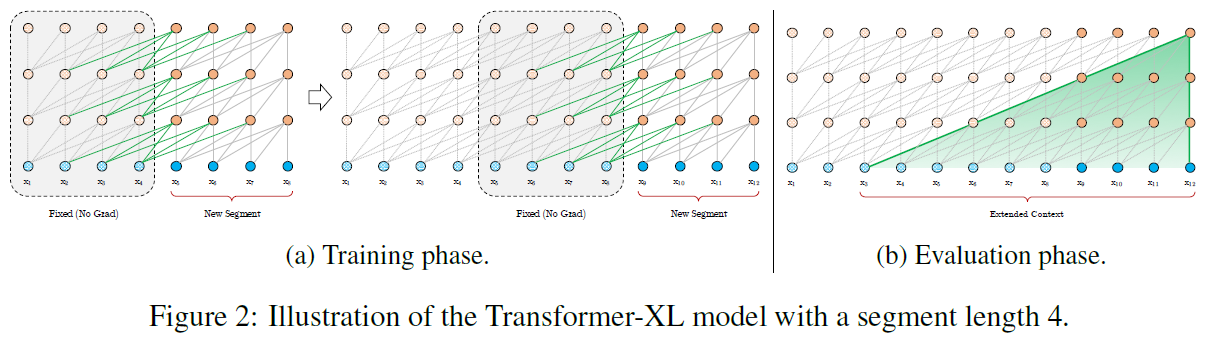
测试的时候，High evaluation cost ，需要重新计算每一个的隐状态

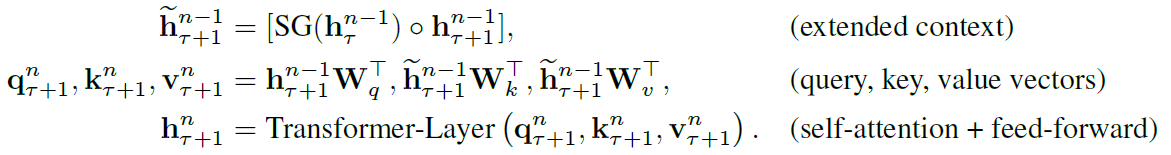
RNN时代：Truncated BPTT

* 1. Segment-level Recurrence

Transformer-XL尝试去**解决长序的依赖问题**

在计算上一个segment进行缓存，后面的segment的时候直接**使用缓存，不用进行反向传播。**



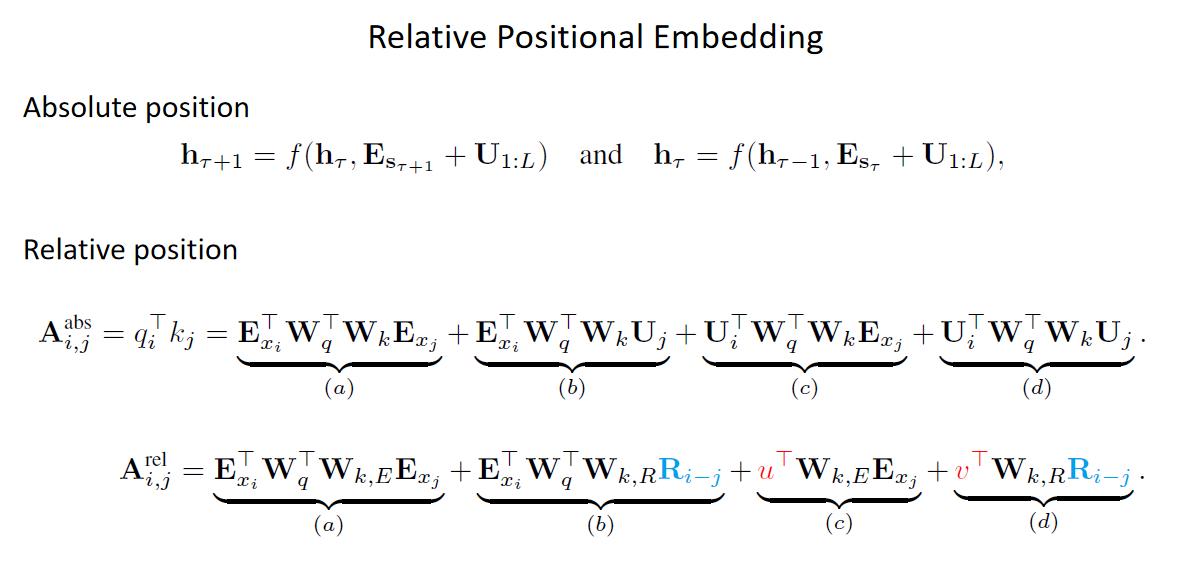


* 1. Relative Positional Embedding

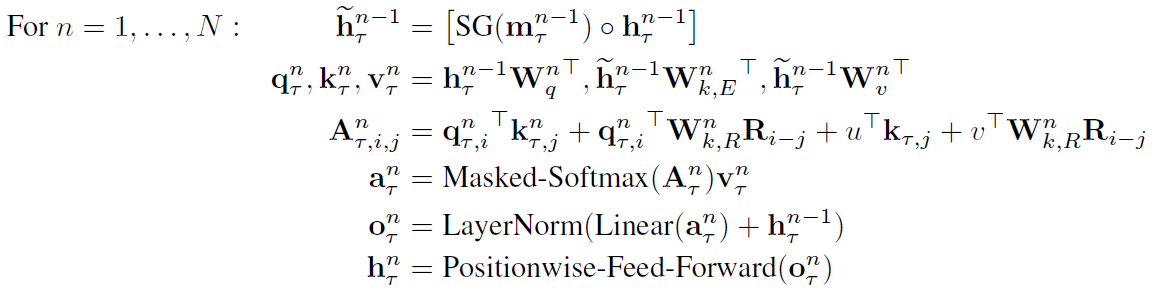
**Absolute position：**存在的问题，在segment中，间隔为4的时候，输入1和输入5的位置是一样的，这样会误导模型。（segment size这个超参数会对位置选择有影响）  
 在划分了segment的时候，就使用absolute position 会导致参数量变大

**Relative position：**影响attention的值，a\b\c\d的分工明确，a表示与内容相关，b、c表示位置和内容相关，d表示位置相关

V、U已经是新的超参数了，与之前的W等无关



### 怎么做的self-attention的



## 二、XLNet

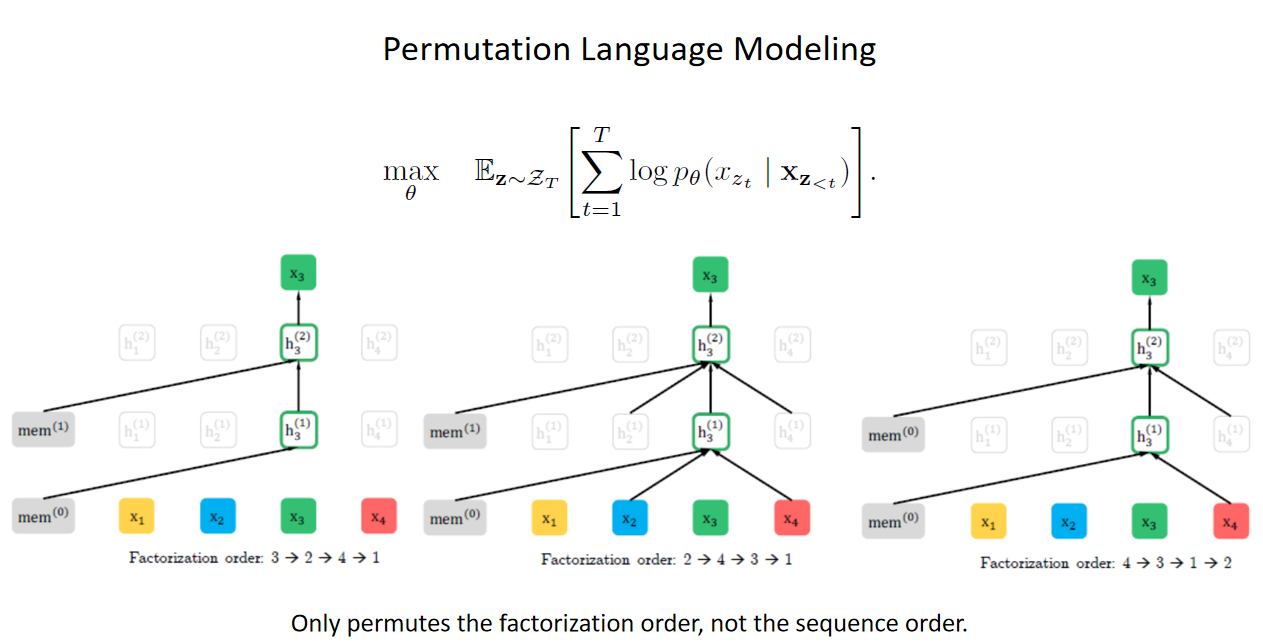
继承了很多transformer的很多工作，让autoregressive language modeling 和 Masked languege modeling之间的进行取长补短。

提出新的语言模型的训练目标Permutation Language Modeling

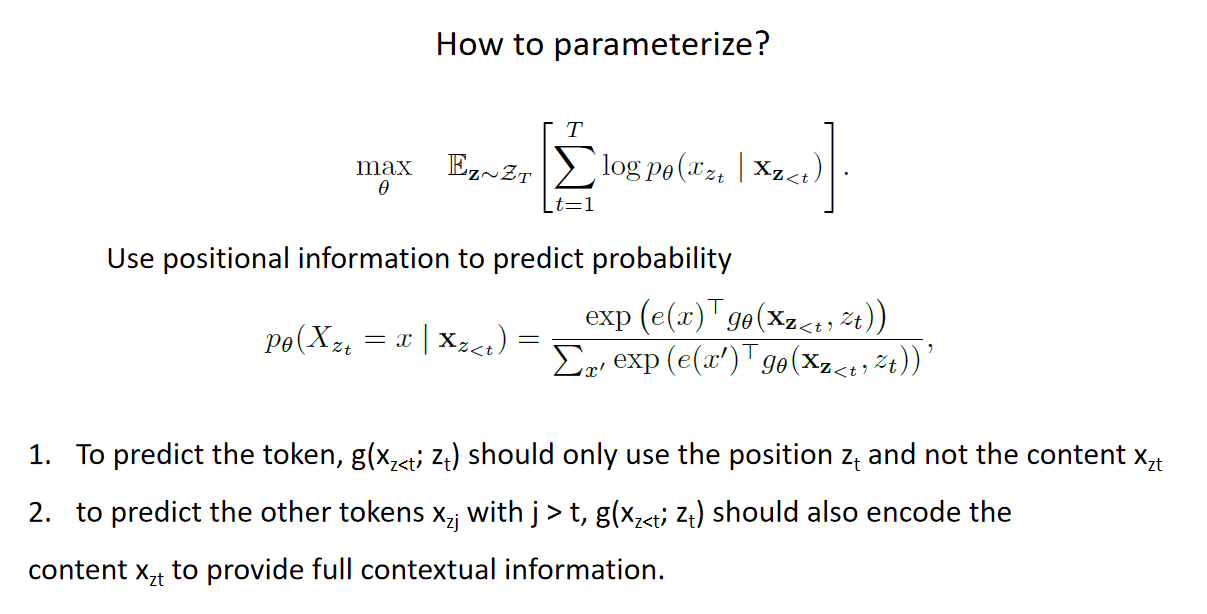
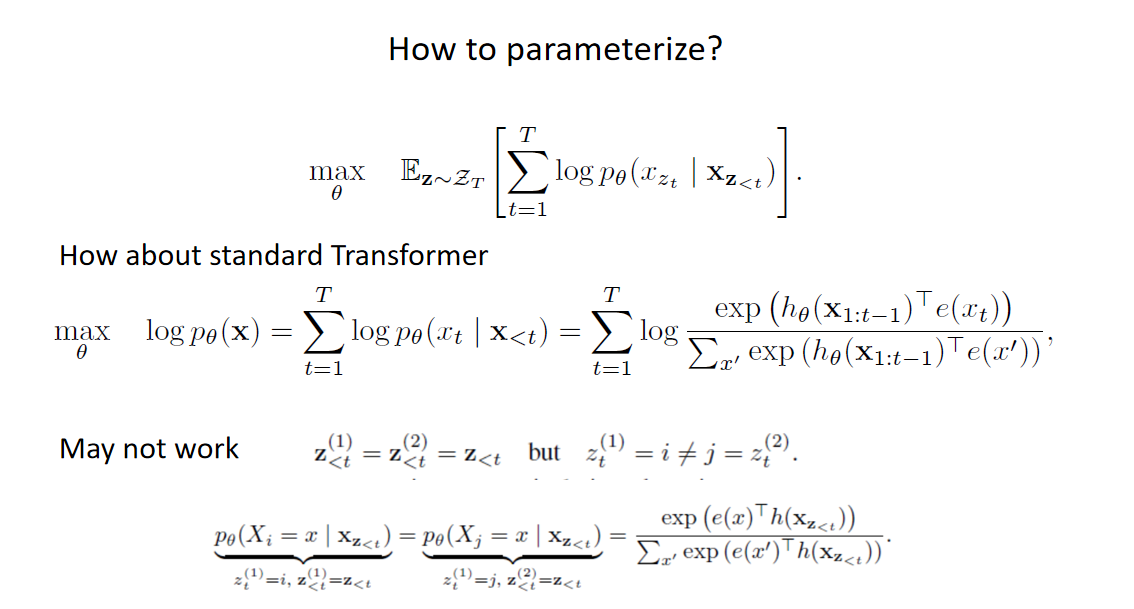
参数化预测： Two-Stream Self-Attention

2.1 Permutation Language Modeling

随机取出一个x，进行全排列，排列只影响预测的顺序，对token本身的顺序是不变的

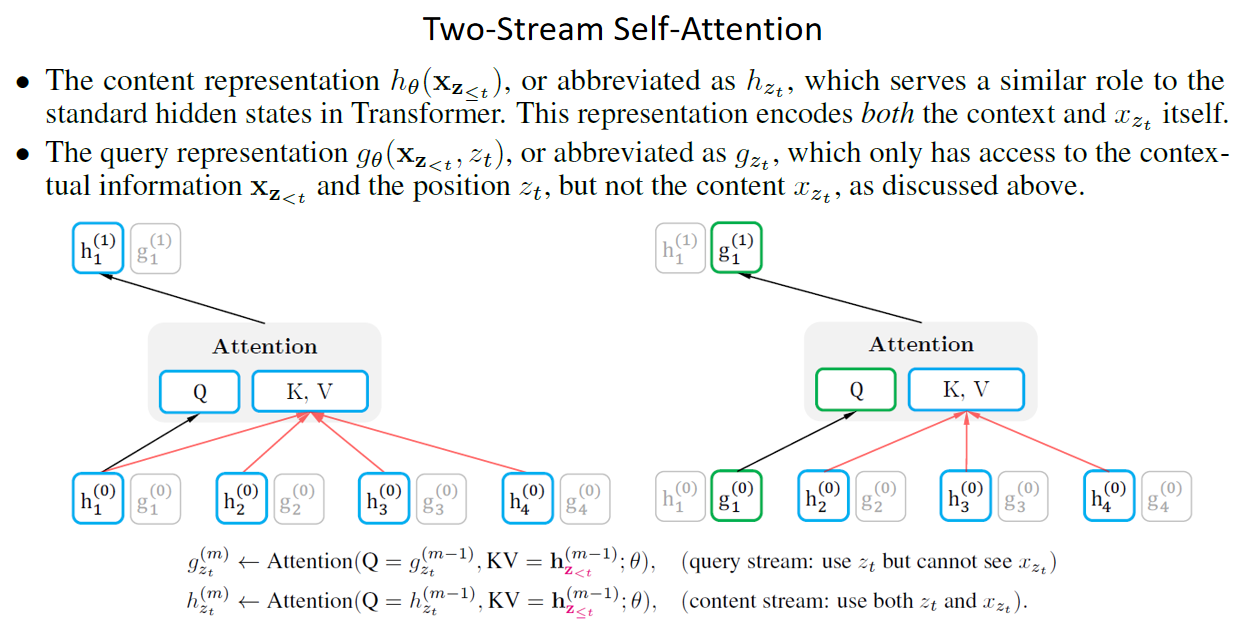


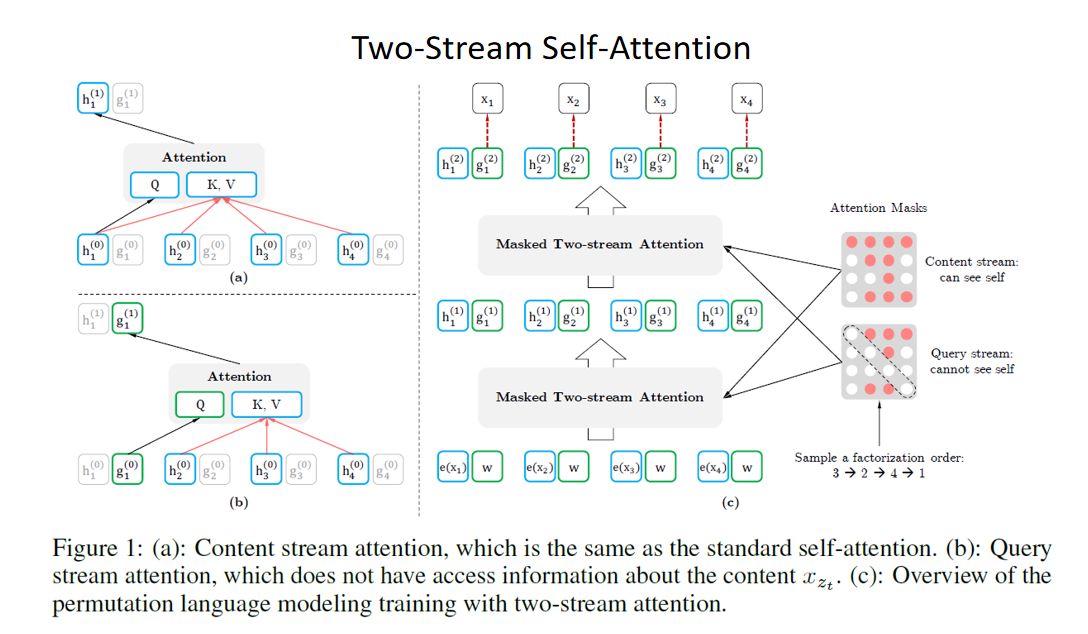
May not work:1243和1234是一样的



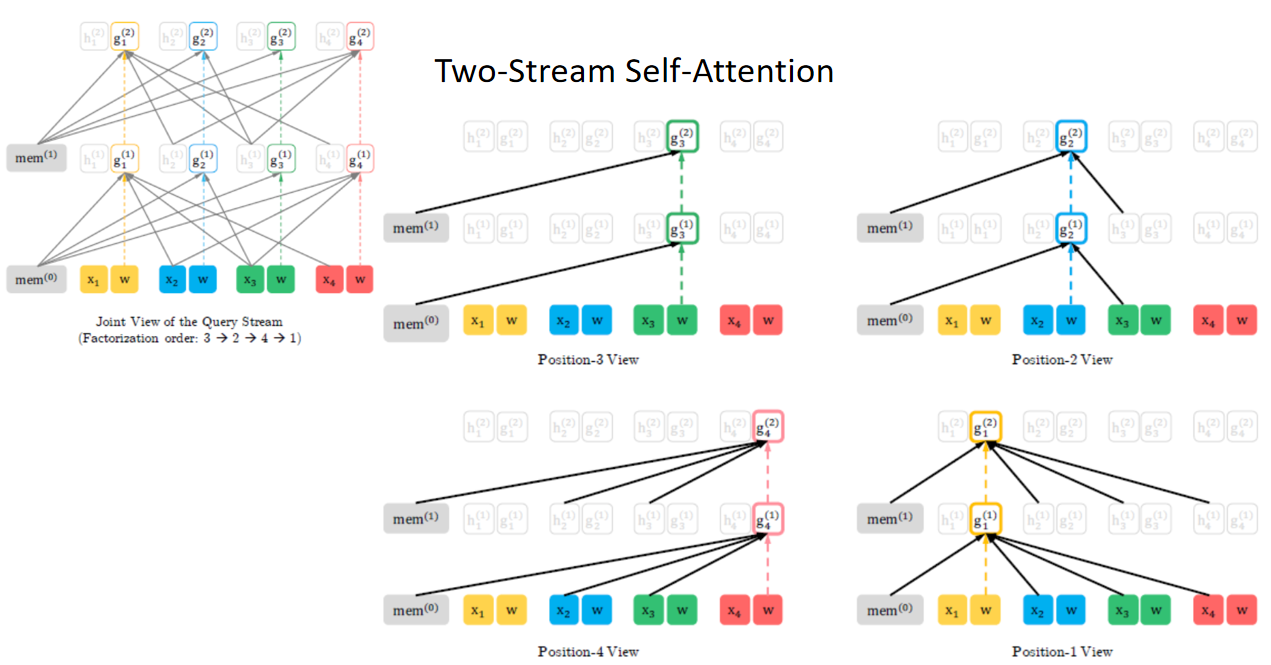
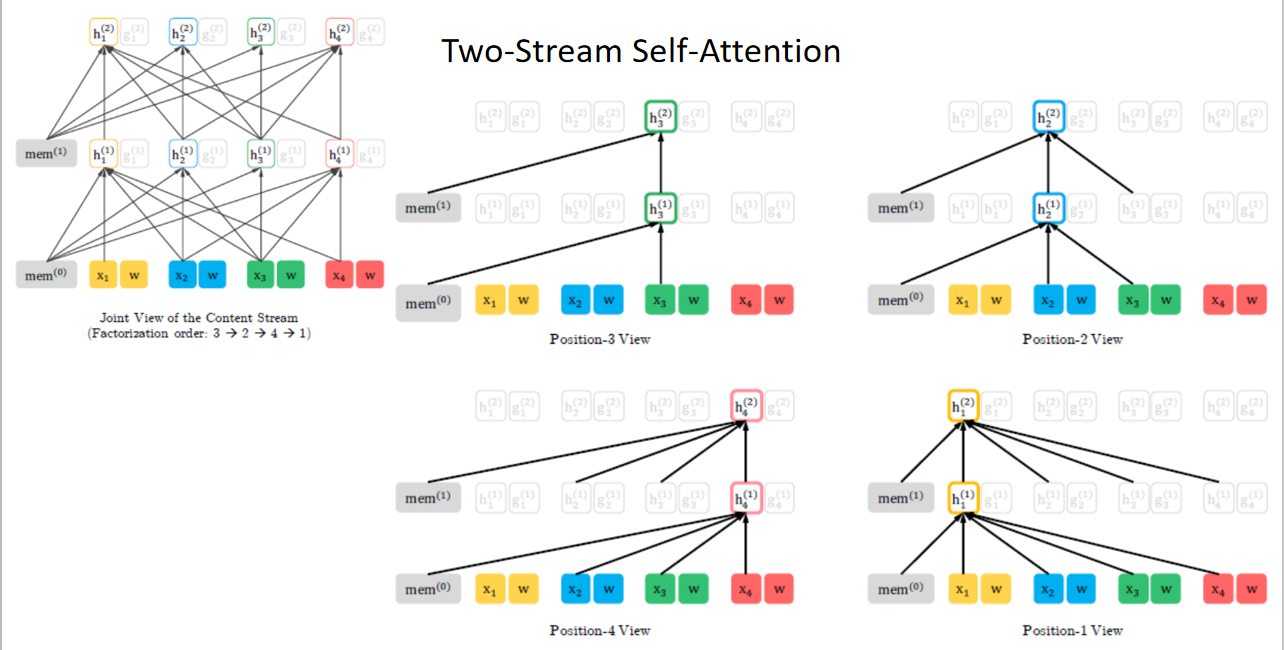
2.2 Two-Stream Self-Attention

计算h的时候其实就是multi self-attention





用3去预测2，再用3->2去预测4。。。



2.3 More trick

