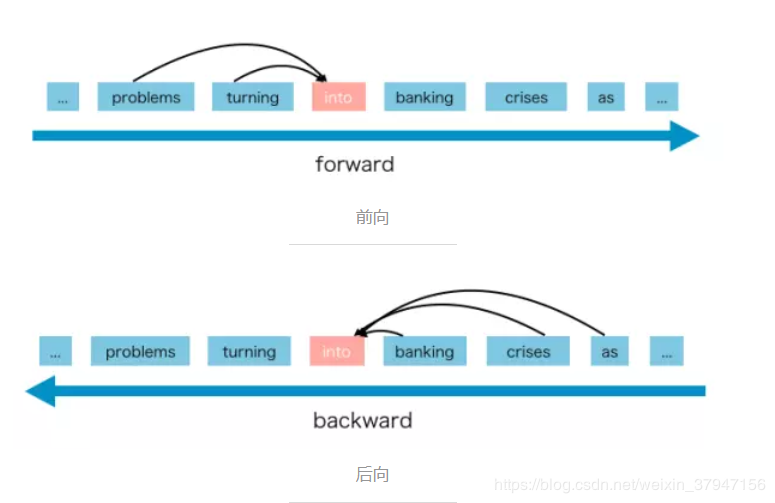
# 自回归语言模型与自编码语言模型

## 自回归语言模型

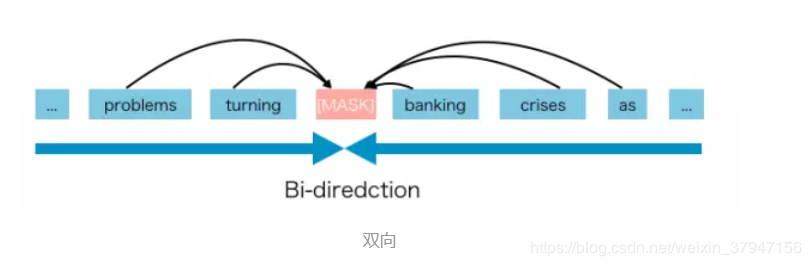
1. 根据上文预测下个单词这种从左到右的语言模型或者根据下文预测上个单词这种从右到左的语言模型



1. 自回归比较适合从左到右进行计算的NLP任务，比如生成任务：机器翻译等

## 自编码语言模型

1. 对输入序列进行Mark操作，用未Mark的数据预测被Mark的数据，类似于降噪自编码(DAE)



1. 自编码能够在训练时看到单词的上下文，更好的学习上下文信息。
2. 自编码面临着训练与应用不一致的问题，因为在训练时有完整的序列，但测试时是没有后续序列的。这导致其对生成类任务表现不太良好。

# XLNet

## 主要改进

1. 使用Transformer-XL代替Transformer，更适合处理长序列
2. 特殊的自回归预训练语言模型
3. 双流自注意力机制

## Transformer-XL特点

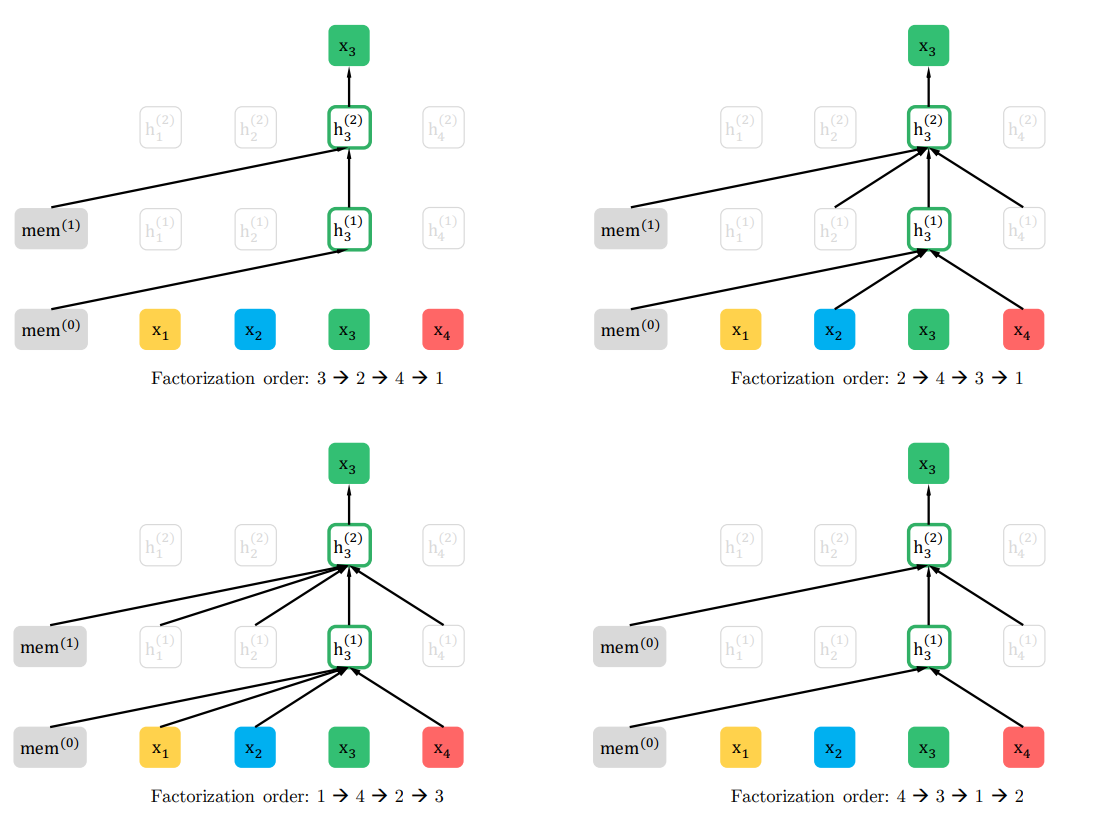
相对位置编码，分段运算机制

## 特殊的自回归预训练语言模型

在预训练语言模型方面，XLNet采用的是一种特殊的自回归语言模型训练方式，整体思路依旧是在循环中根据上文，也就是前面的单词预测后面的单词。但是XLNet又希望像自编码那样能够看到下文的信息，因此做了一些特殊的改变。

比如对于序列长度为4的序列，其值分别为，在预测位置的词时，XLNet会进行如下操作：

1. 将的位置随机打乱，从所有可能中随机抽取，产生新的序列，但是的Position Embedding依旧是原始位置对应的相对位置编码。
2. 对于所有可能的新序列，使用正常的自回归预训练模型的训练方式进行训练，也就是使用新序列中之前的单词来预测。如下图，第一个序列中排在最前面，所以只用mem进行预测，第二个序列中前有和，所以使用来进行预测。



1. 在这种方式下，预测时，模型训练时就不仅仅只能获取到上文信息，也会获取到下午信息，就能够进行更好的参数选择

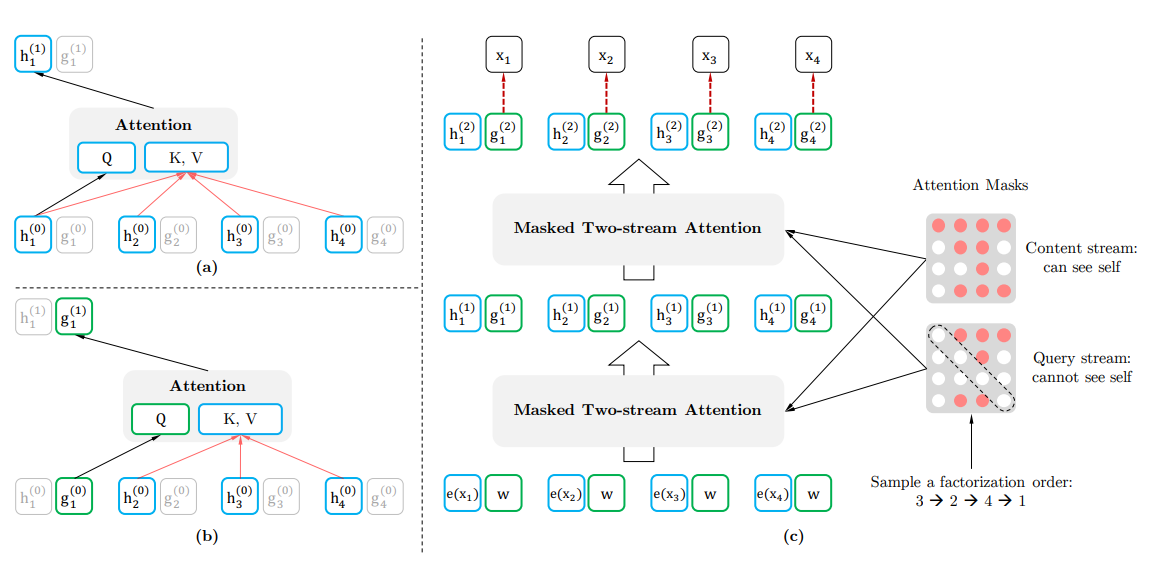
## 双流自注意力机制

### 内容介绍

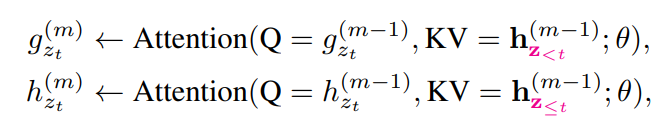
双流指的是内容表征通道Content Stream和语境表征通道Query Stream，对于单词，其在第m层Attention层计算出的内容表征向量记为，语境表征向量记为。

所以第m层Attention会产生两组表征向量

XLNet的想法是在计算语境表征向量时，模型只能看到单词重排序后前面的单词，而**看不到单词本身**，只能将**的位置信息**加入学习。而在计算内容表征向量时就和一般的Transformer一样，可以看到本身和其前面的单词信息。



### 公式计算



计算语境表征向量所需的Q、K、V

1. Q由在上一层Attention层产生的语境表征向量产生
2. K、V则由位于前面的单词在上一层Attention层产生的内容表征向量计算得到
3. 这里的位置序号t是指的重排序后的序号

计算内容表征向量所需的Q、K、V就比较简单，它可以学习单词本身的信息

1. Q由在上一层Attention层产生的内容表征向量产生
2. K、V则由位于前面的单词以及本身在上一层Attention层产生的内容表征向量计算得到