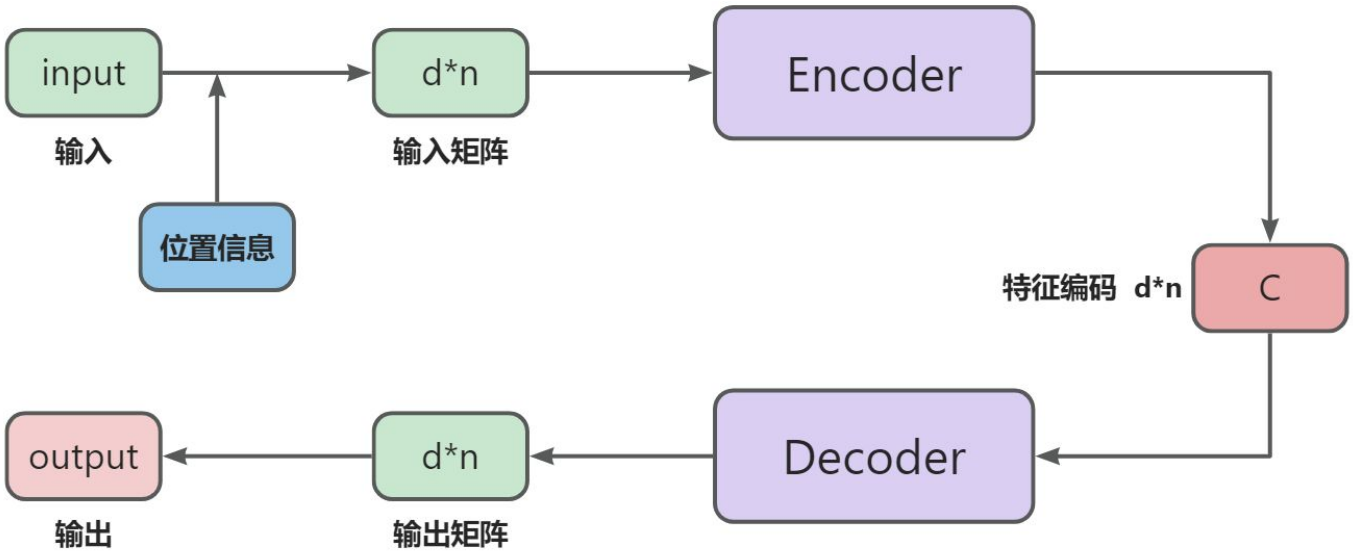
Transformer

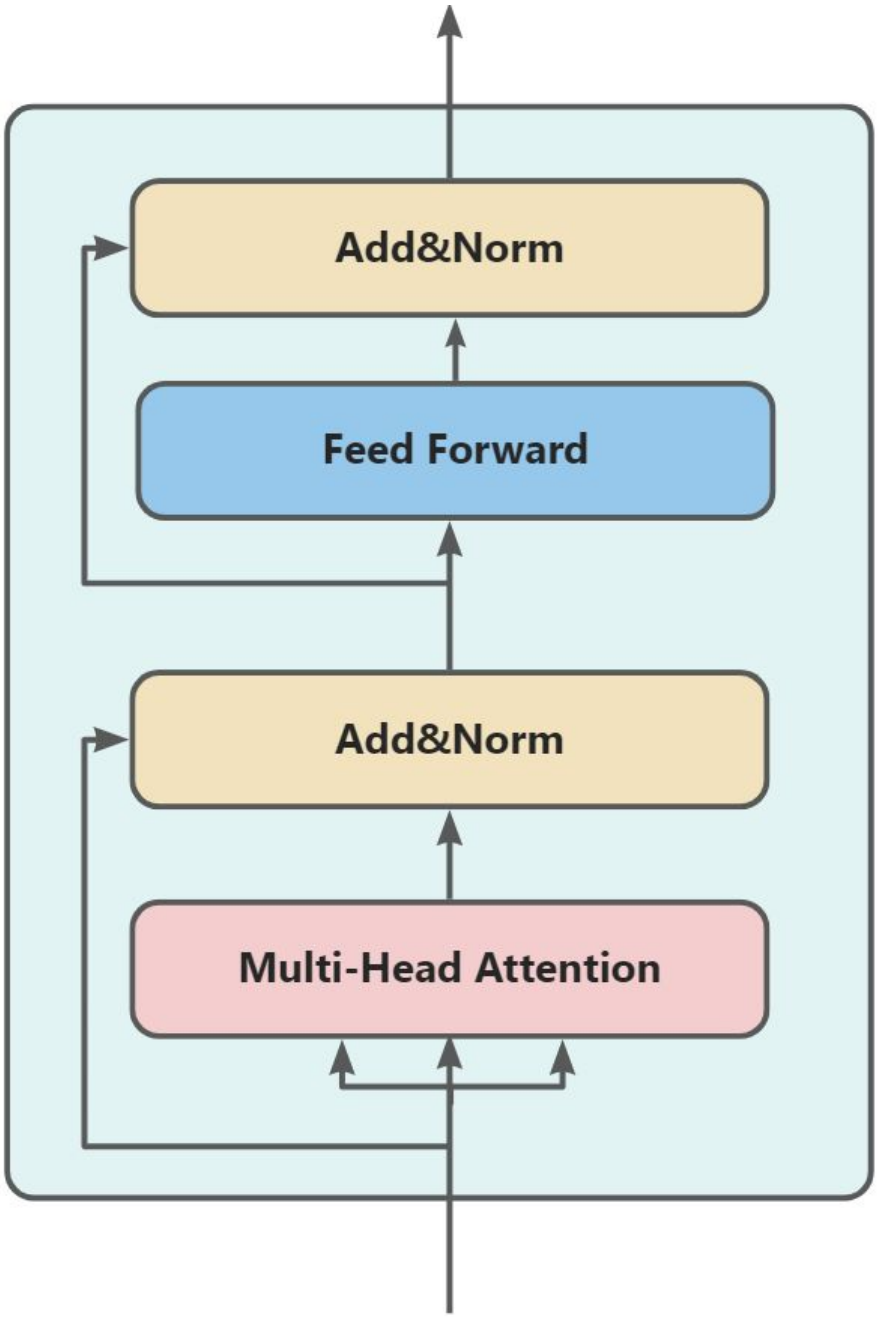
Transformer模型的整体架构为



其输入通常为文本，经过分词成n块后每个词对应到一个d维矩阵，形成一个d*n的输入矩阵（还需要额外加入词的位置信息），n为可以变动过的值，所以其对输入的大小没有限制。

输入矩阵经过Encoder后得到相同大小的特征编码，然后经过Decoder得到最终的输出矩阵，并转化为文本输出。

Encoder

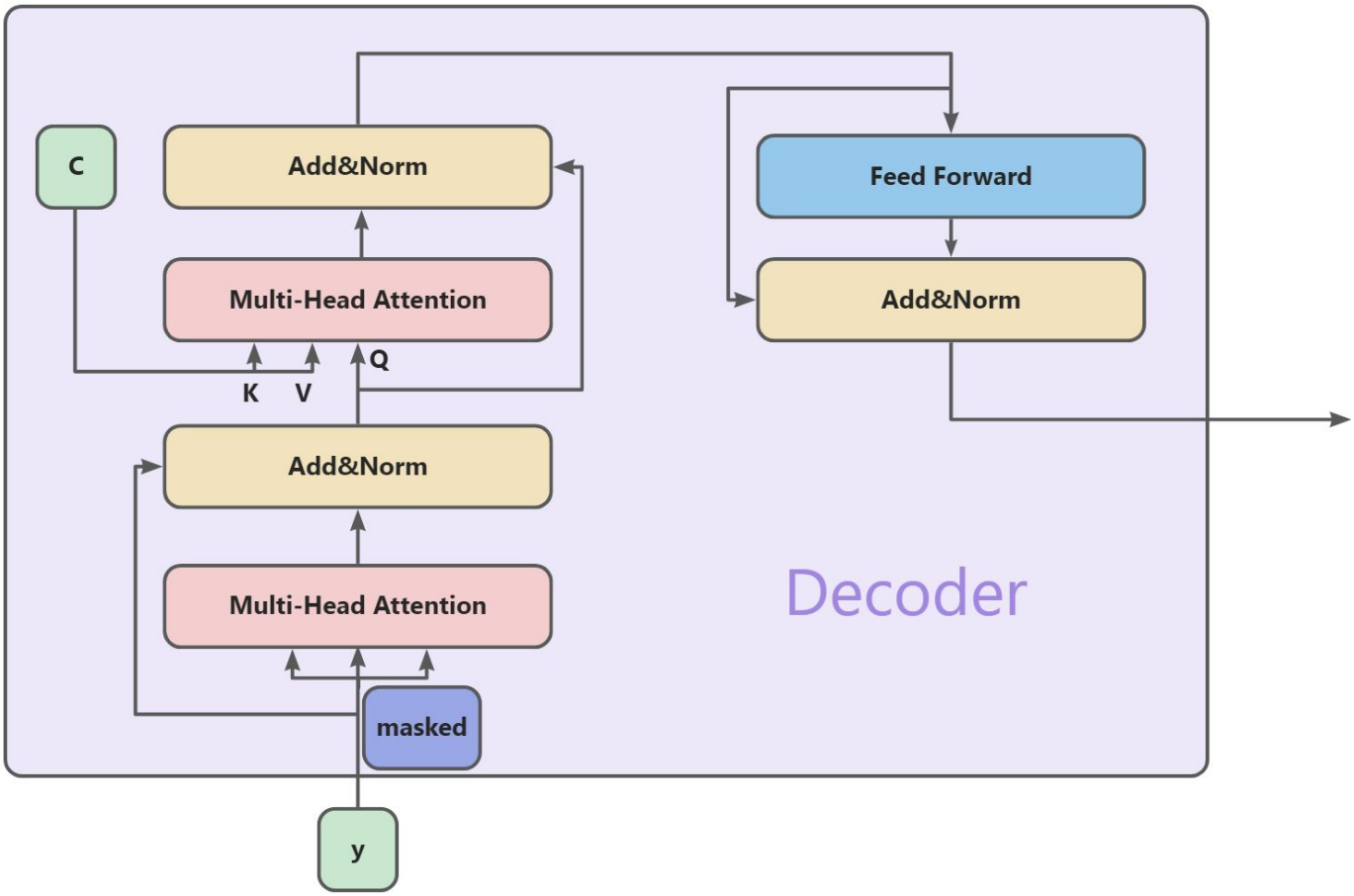


Encoder

Encoder的具体结构如图所示，其中FFN层为一个简单的前馈网络，而ADD&Norm层为残差链接层。最为关键的为其中所使用到的多头注意力层Multi-Head Attention。且实际运行的网络中，多个基本Encoder叠加而成一个更大的Encoder。

残差层通常指的是在经过某个函数 $f(x)$ 运算后，将 $f(x)+x$ 作为输出。在网络中表现就是一条绕过某一层的旁路。用于防止在网络逐步传递中，有效信息逐步减少。同时往往需要做归一化。

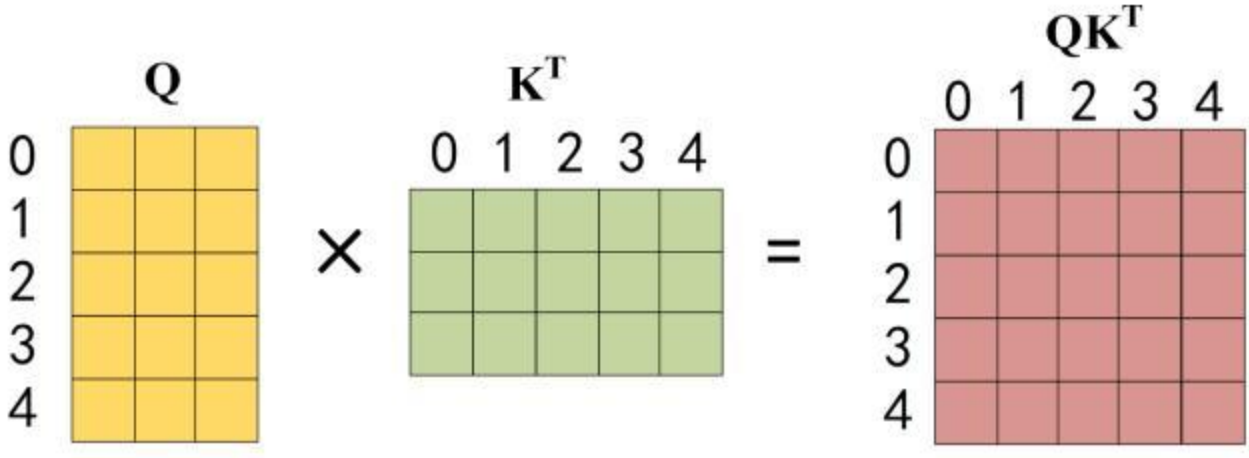
Decoder

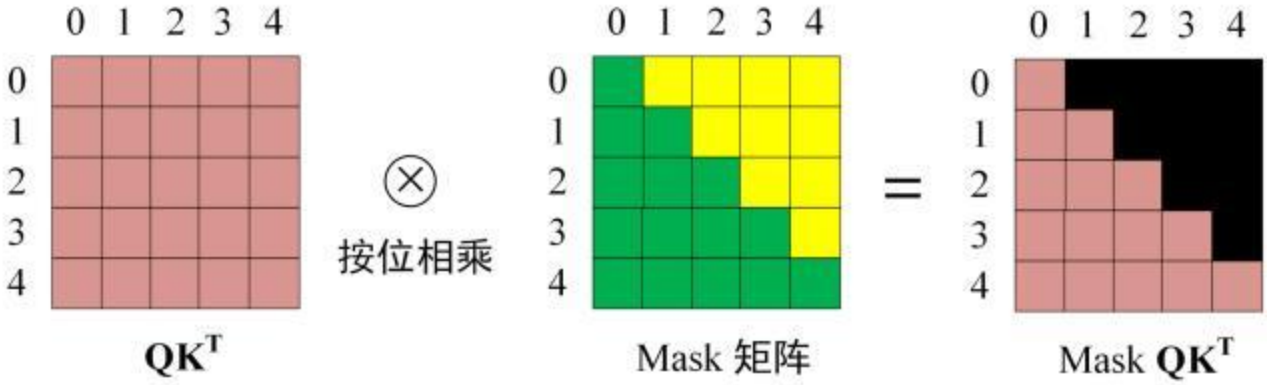


Decoder的结构如图所示，同样实际由多个Decoder叠加而成。不同于Encoder的地方在于，该模型有两次Attention过程，且第一次进行了masked，而第二次则是由encoder的输出c提供K，V，上一个Decoder的输出y提供Q进行。

masked

宏观的理解masked即为，保证每个词都只被其之前的词影响。而微观的实现在于，attention中，QK相乘得到的矩阵，每行都代表了一个分词，而每列则代表该标号的分词对于本词的影响程度。





如果加上如图所示的遮罩，那么可以使得排序之后的词对于该词的影响程度都降为0，因此，达成只被之前词所影响的效果。