Multi-Head Self-Attention（从空间角度解释为什么做多头）

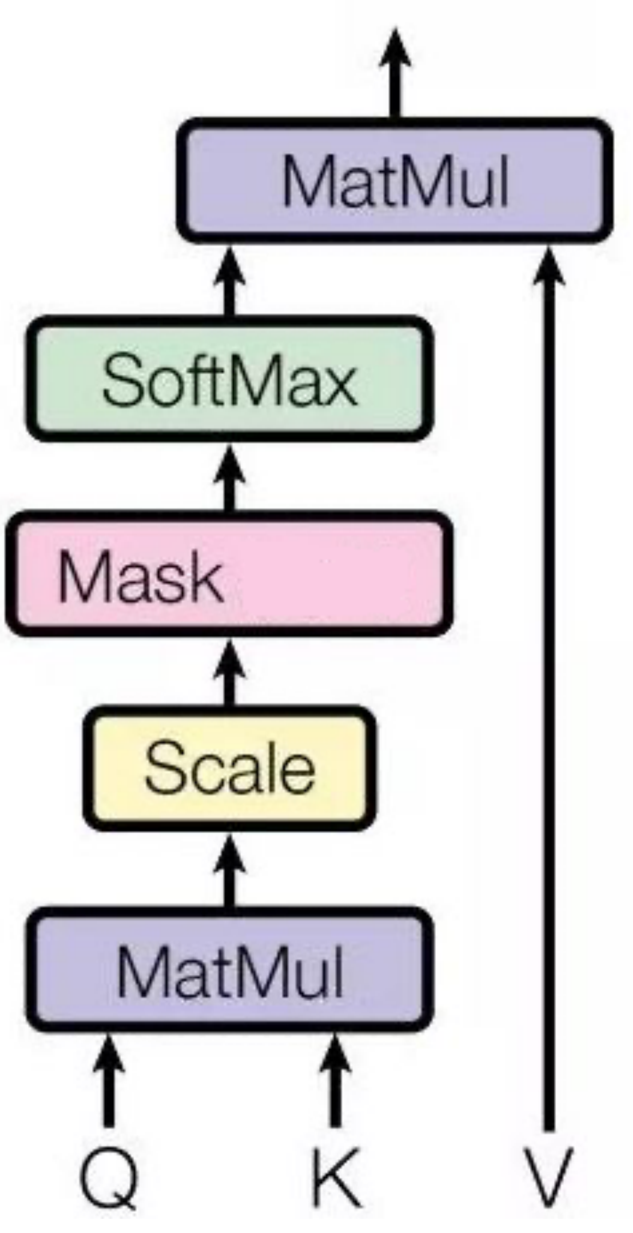
2022年9月3日 星期六 17:30

随笔 - 832 文章 - 0 评论 - 313 阅读 - 151万

上节课回顾

0：40

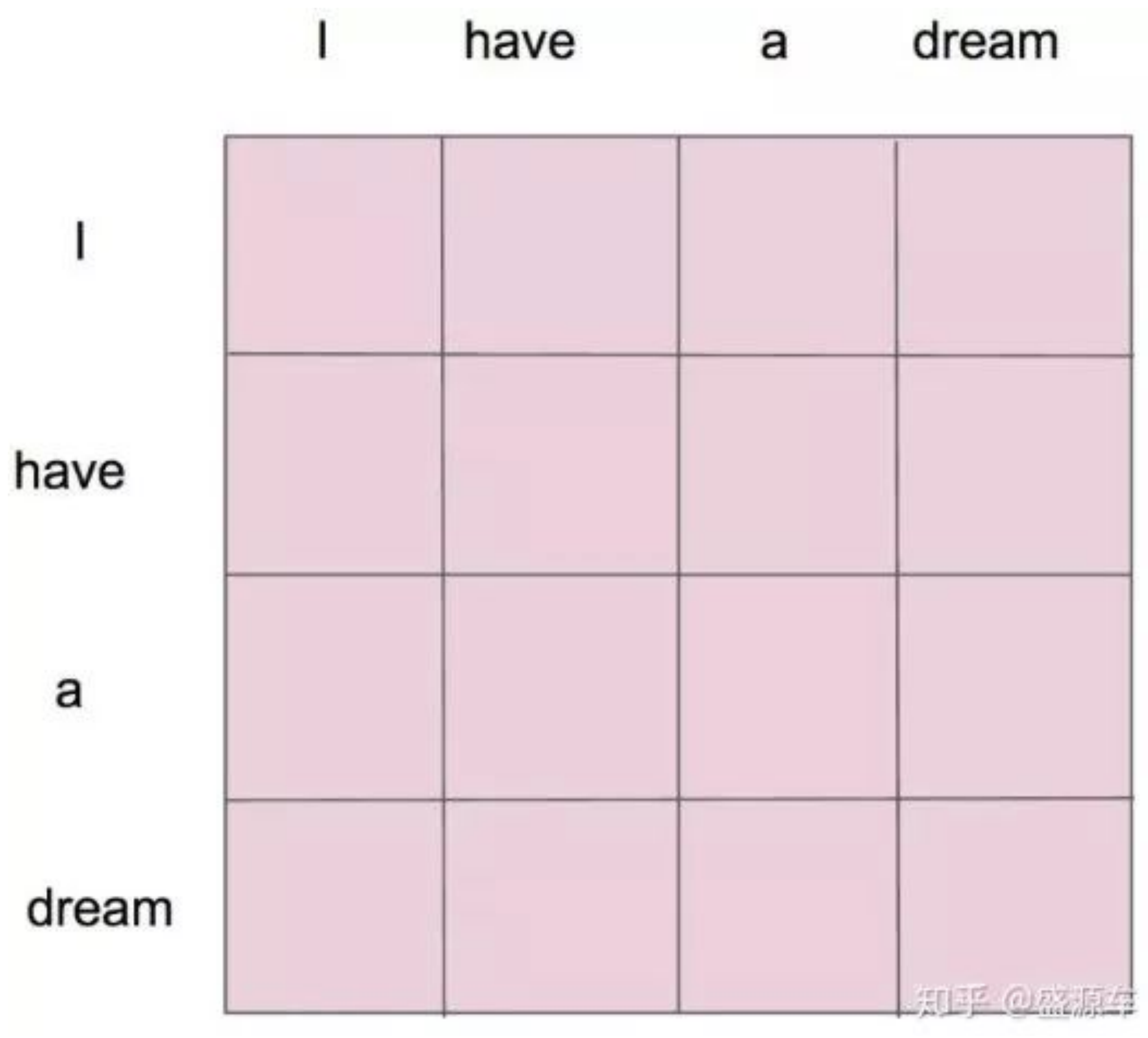
Attention



Self-Attention

Self-Attention 其实是 Attention 的一个具体做法

给定一个 x，通过自注意力模型，得到一个 z，这个 z 就是对 x 的新的表征（词向量），z 这个词向量相比较 x 拥有了句法特征和语义特征

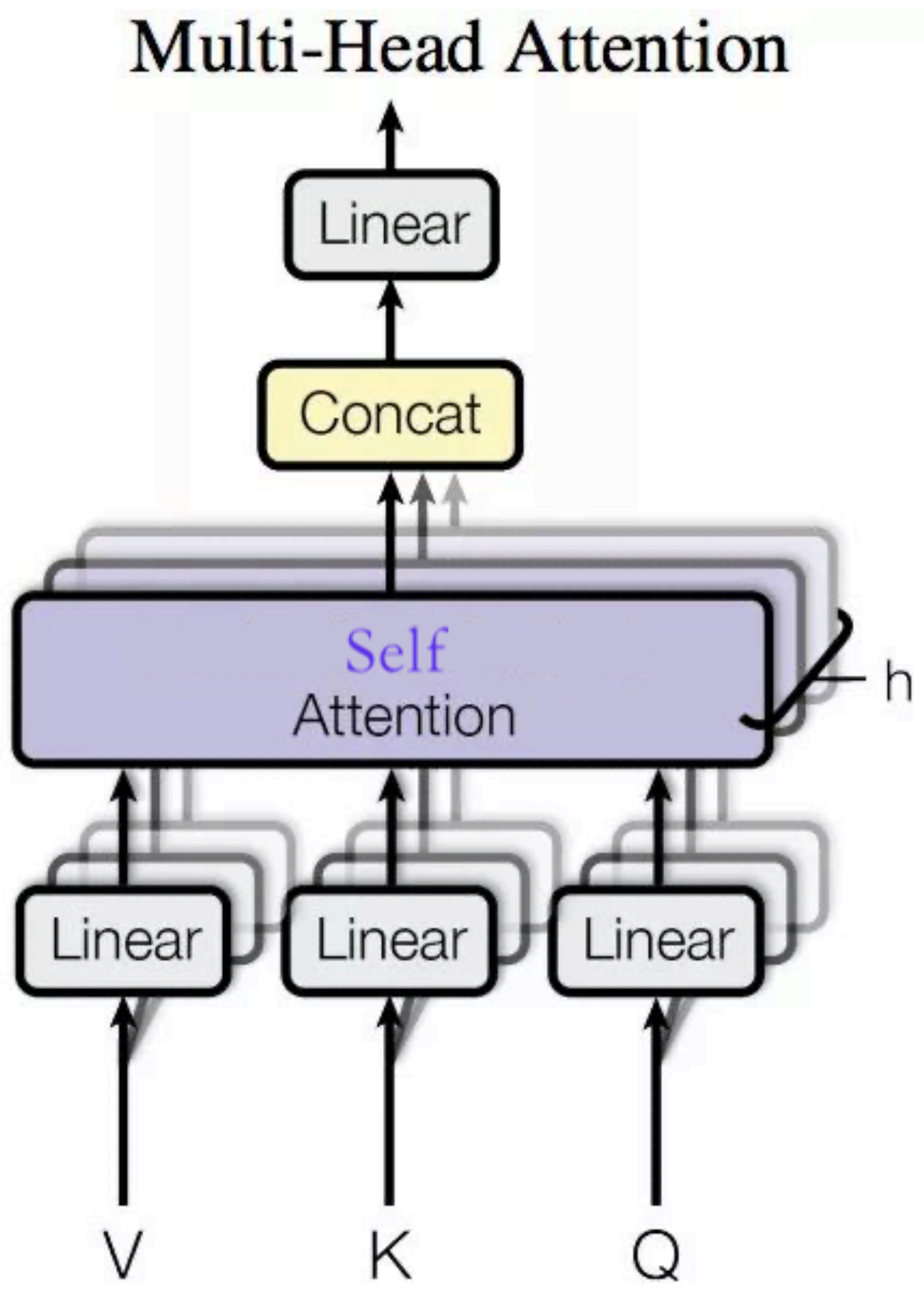


Multi-Head Self-Attention（多头自注意力）

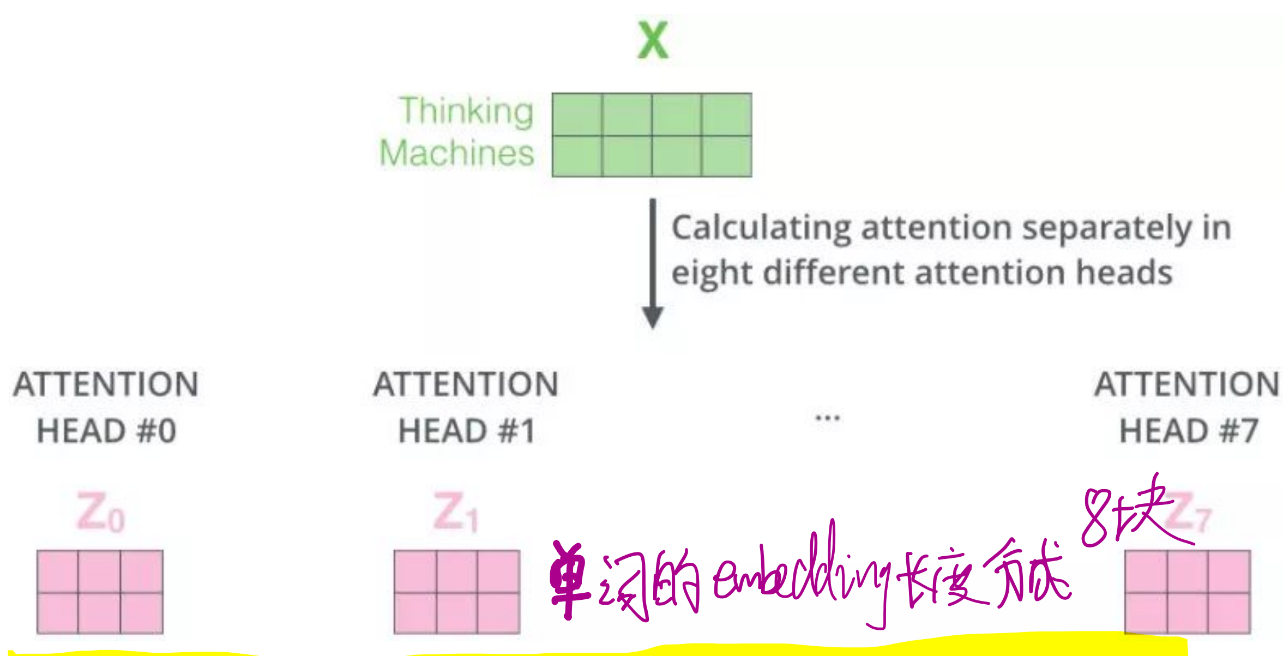
z 相比较 x 有了提升，通过 Multi-Head Self-Attention，得到的 Z′Z′相比较 z 又有了进一步提升

多头自注意力，问题来了，多头是什么，多头的个数用 h 表示，一般h=8h=8，我们通常使用的是 8 头自注意力

什么是多头



如何多头 1

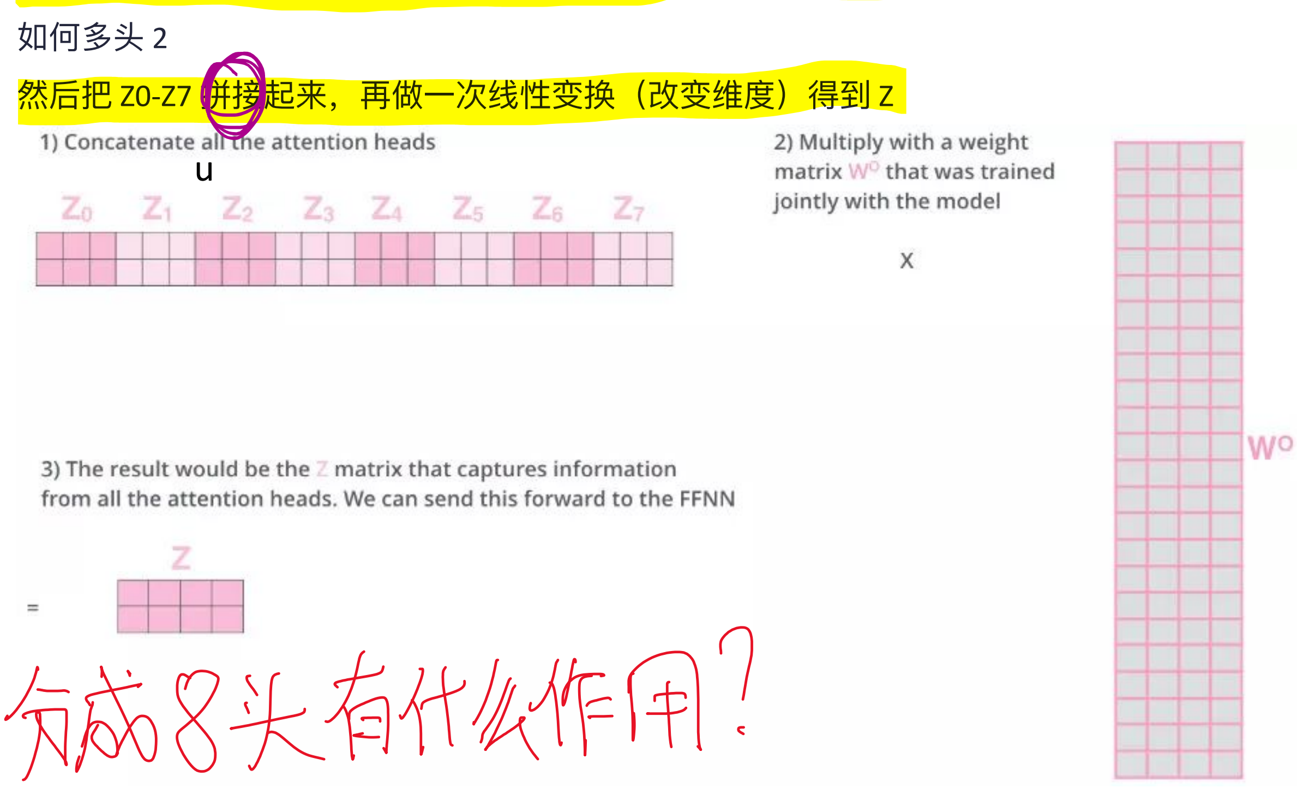


单词的embedding长度分成 8块

对于 x，我们不是说，直接拿 x 去得到 z，而是把 x 分成了 8 块（8 头），得到 Z0-Z7

如何多头 2

然后把 Z0-Z7 拼接起来，再做一次线性变换（改变维度）得到 z



分成8头有什么作用？

有什么作用?

机器学习的本质是什么：y=σ(wx+b)，在做一件什么事情，非线性变换（把一个看起来不合理的东西，通过某个手段（训练模型），让这个东西变得合理）

非线性变换的本质又是什么？改变空间上的位置坐标，任何一个点都可以在维度空间上找到，通过某个手段，让一个不合理的点（位置不合理），变得合理

这就是词向量的本质 词向量的本质

one-hot 编码（0101010）

word2vec（11，222，33）

emlo（15，3，2）

attention（124，2，32）

multi-head attention（1231，23，3），把 X 切分成 8 块（8 个子空间），这样一个原先在一个位置上的 X，去了空间上 8 个位置，通过对 8 个点进行寻找，找到更合适的位置

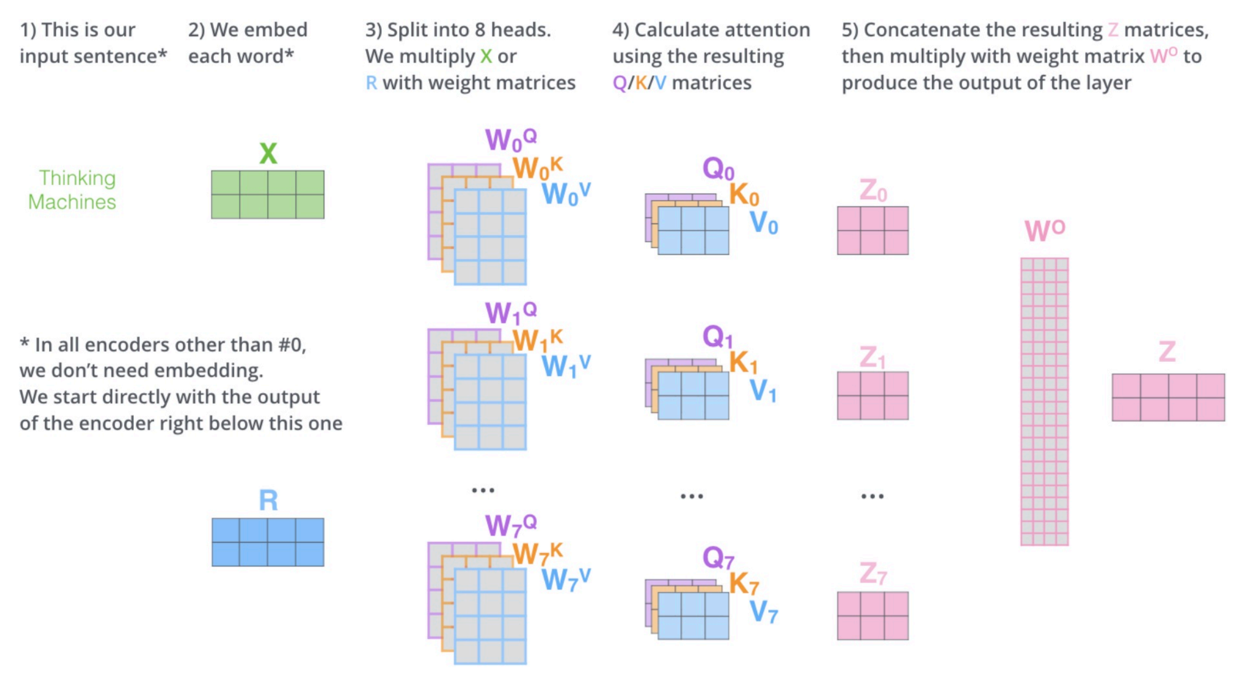
找更合适的位置.

词向量的大小是 512

假设你的任务，视频向量是 5120，80

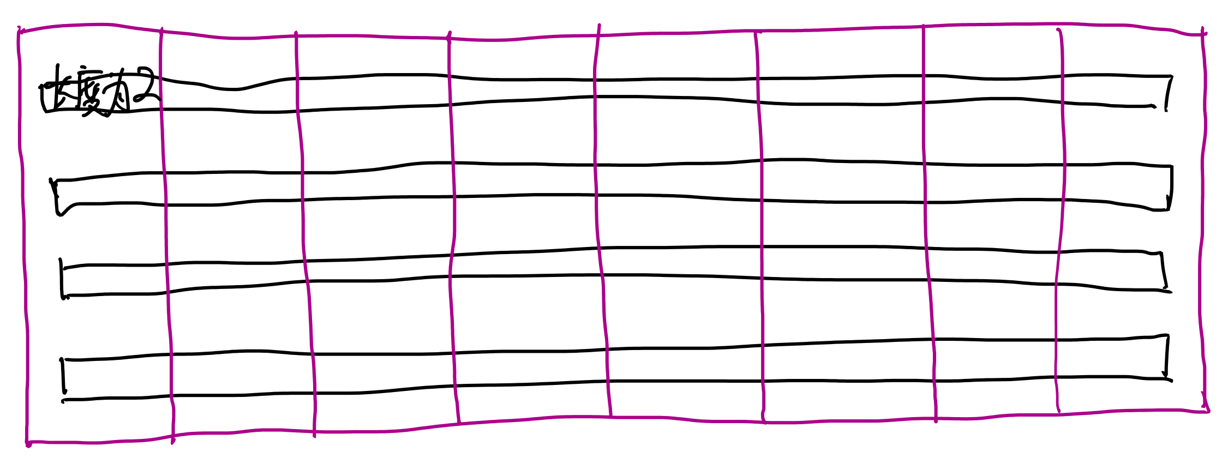
对计算机的性能提出了要求

多头流程图



分割（下面是针对本例子）（

假设，batch=3，词嵌入长度为16。



[3, 4, 16]

↓ view(batch, -1, head, dim_v)

[3, 4, 8, 2]

↓ transpose(-2, -1)

[3, 8, 4, 2]

↓ 多头，self-attention

[3, 8, 4, 2]

↓ （拼接各头结果） view(batch, -1, head x dim_v)

[3, 4, 16]. MultiHead-self Attention 输出。