自然语言处理BERT模型

- 课程安排
  - 通俗讲解知识点，项目实战驱动
  - 当下主流解决框架，一站式搞定NLP任务
  - 环境配置：选一款IDE即可，基于谷歌开源项目
  - 提供所有数据与代码，追随热点持续更新

BERT

- 自然语言处理通用解决方案
  - 需要熟悉word2vec，RNN网络模型，了解词向量如何建模
  - 重点在于Transformer网络架构，BERT训练方法，实际应用
  - 开源项目，都是现成的，套用进去就OK了
  - 提供预训练模型，基本任务拿过来直接用都成

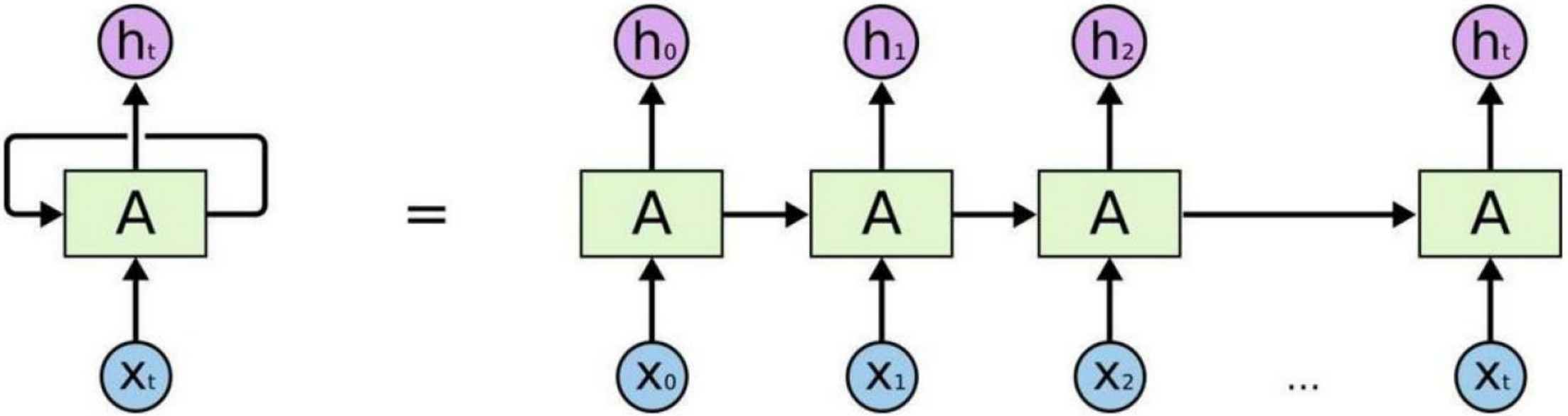
Transformer

- 要做一件什么事呢?
  - 基本组成依旧是机器翻译模型中常见的Seq2Seq网络
  - 输入输出都很直观，其核心架构就是中间的网络设计了

Transformer

- 传统的RNN网络

- 计算时有什么问题?

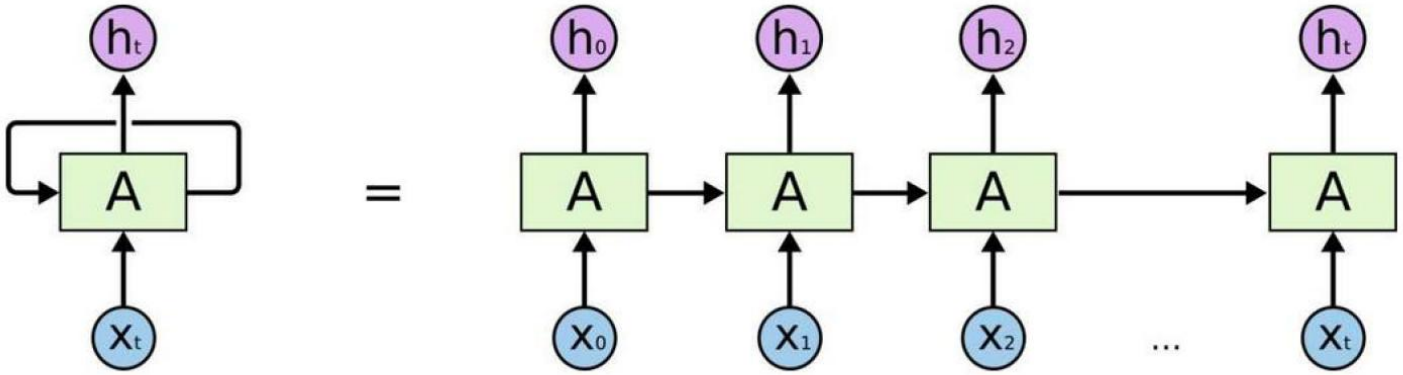


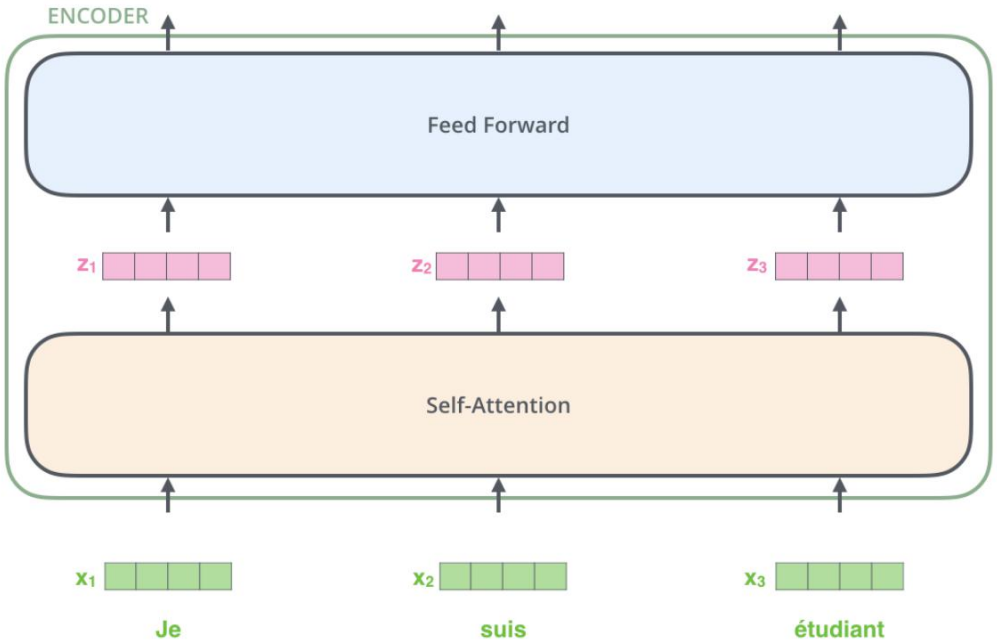
Transformer

- 传统的RNN网络

- Self-Attention机制来进行并行计算，在输入和输出都相同

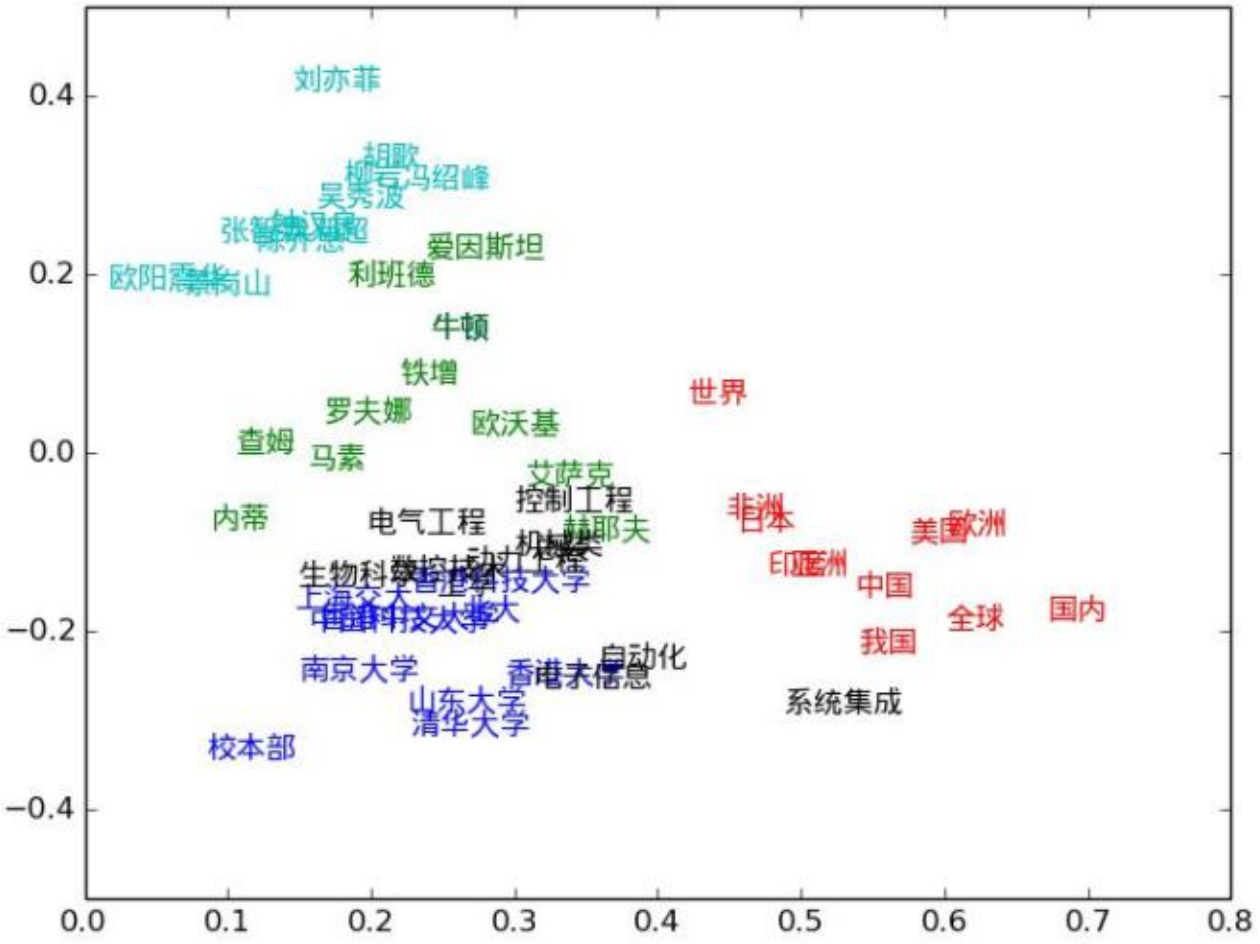
- 输出结果是同时被计算出来的，现在基本已经取代RNN了





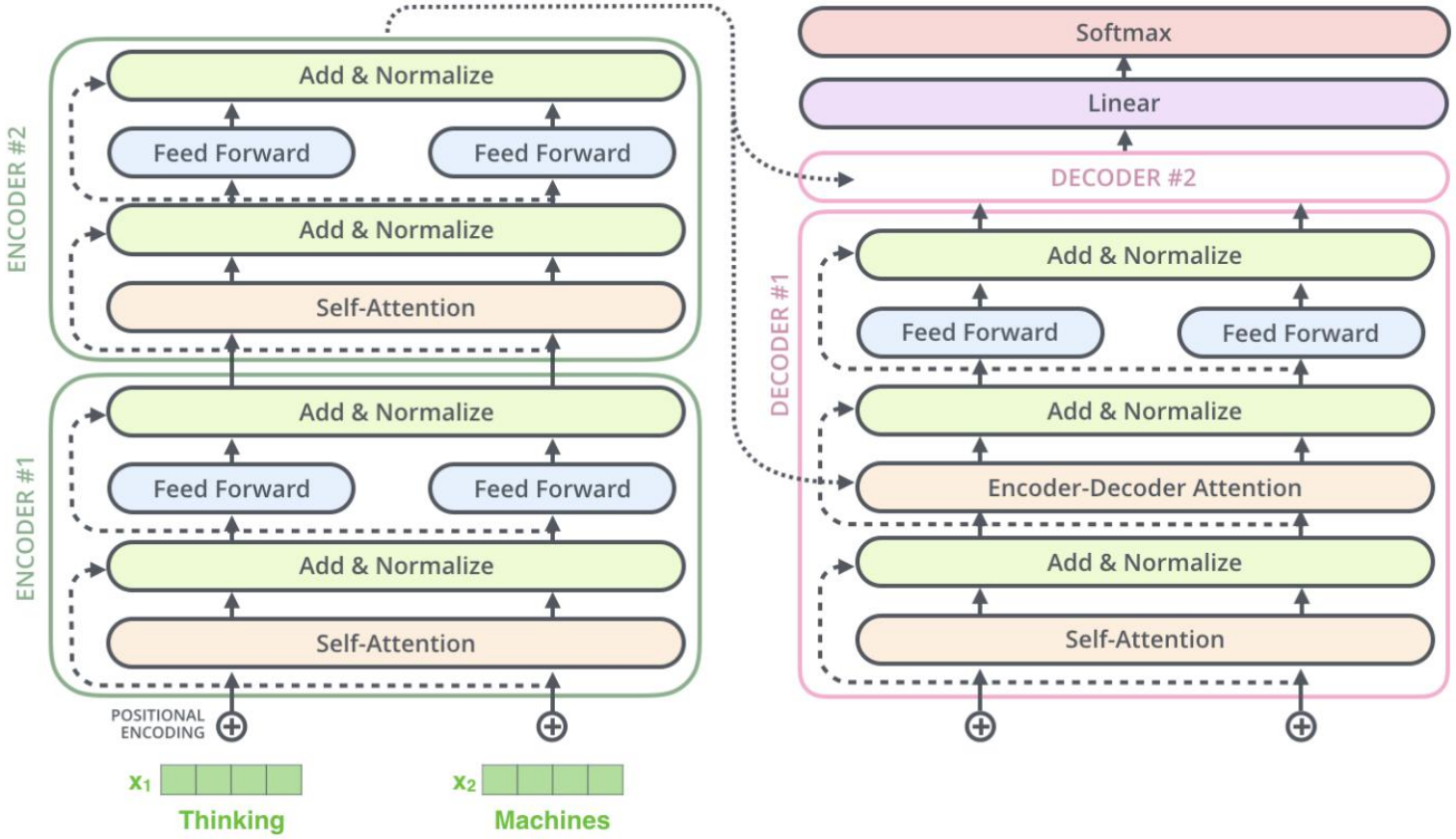
Transformer

- 传统的word2vec
  - 表示向量时有什么问题?
  - 如果‘干哈那’是一个词
  - 不同语境中相同的词如何表达
  - 预训练好的向量就永久不变了



Transformer

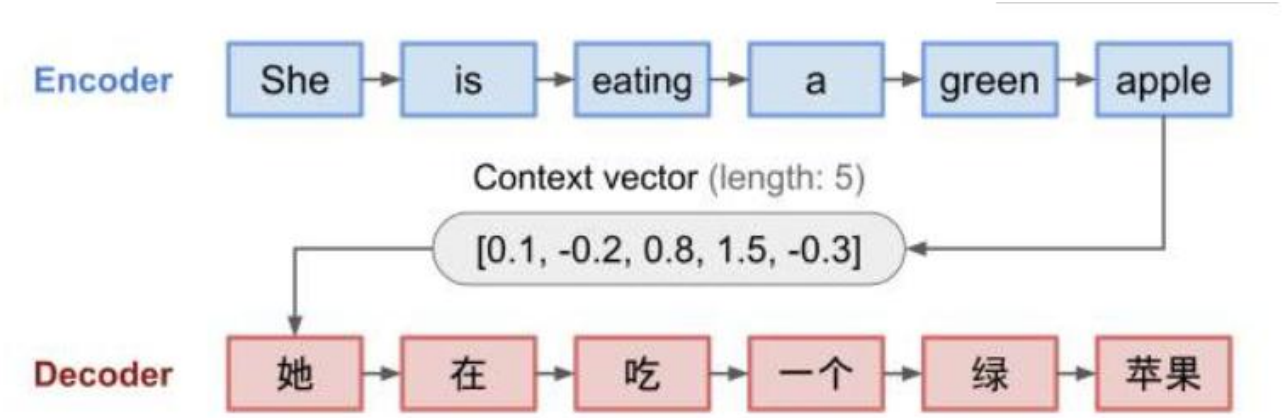
- 整体架构
  - 输入如何编码?
  - 输出结果是什么?
  - Attention的目的?
  - 怎样组合在一起?

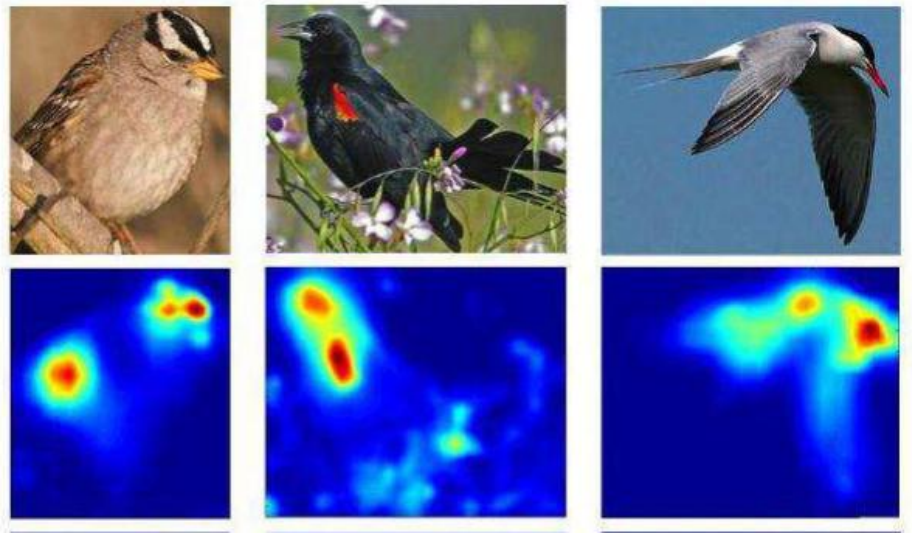


Transformer

- Attention是啥意思呢?

- 对于输入的数据，你的关注点是什么?
- 如何才能让计算机关注到这些有价值的信息?



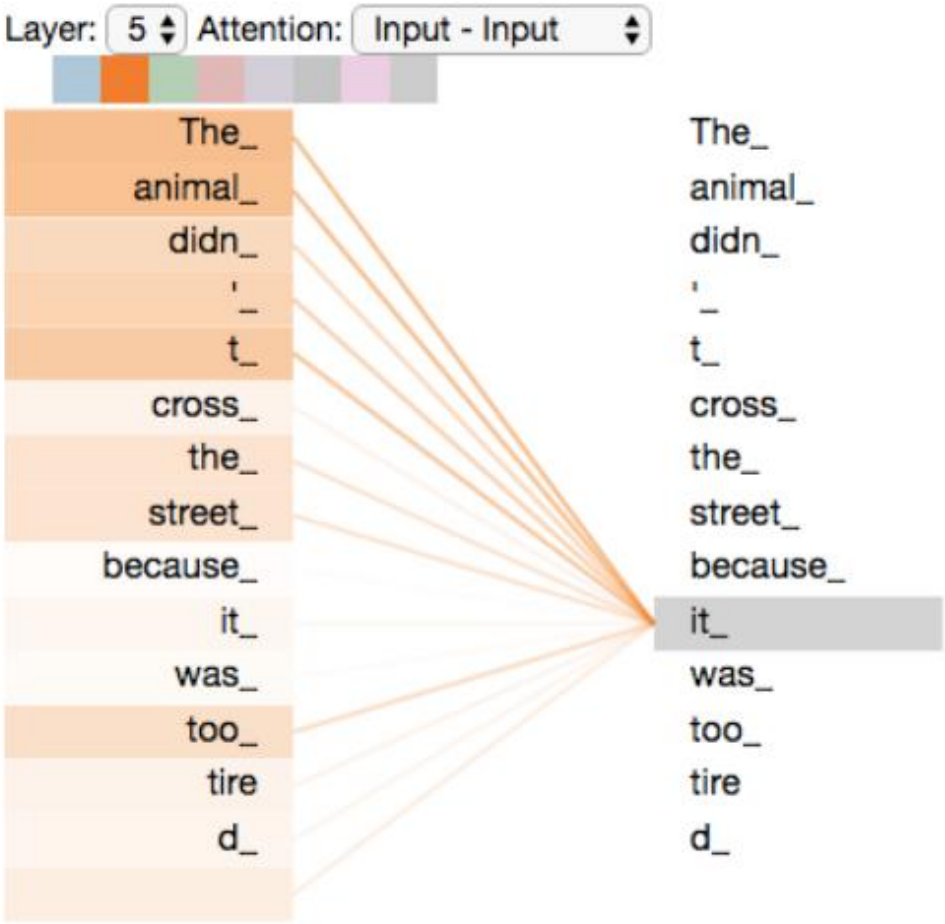


Transformer

- self-attention是什么?

The animal didn't cross the street because it was too tired.

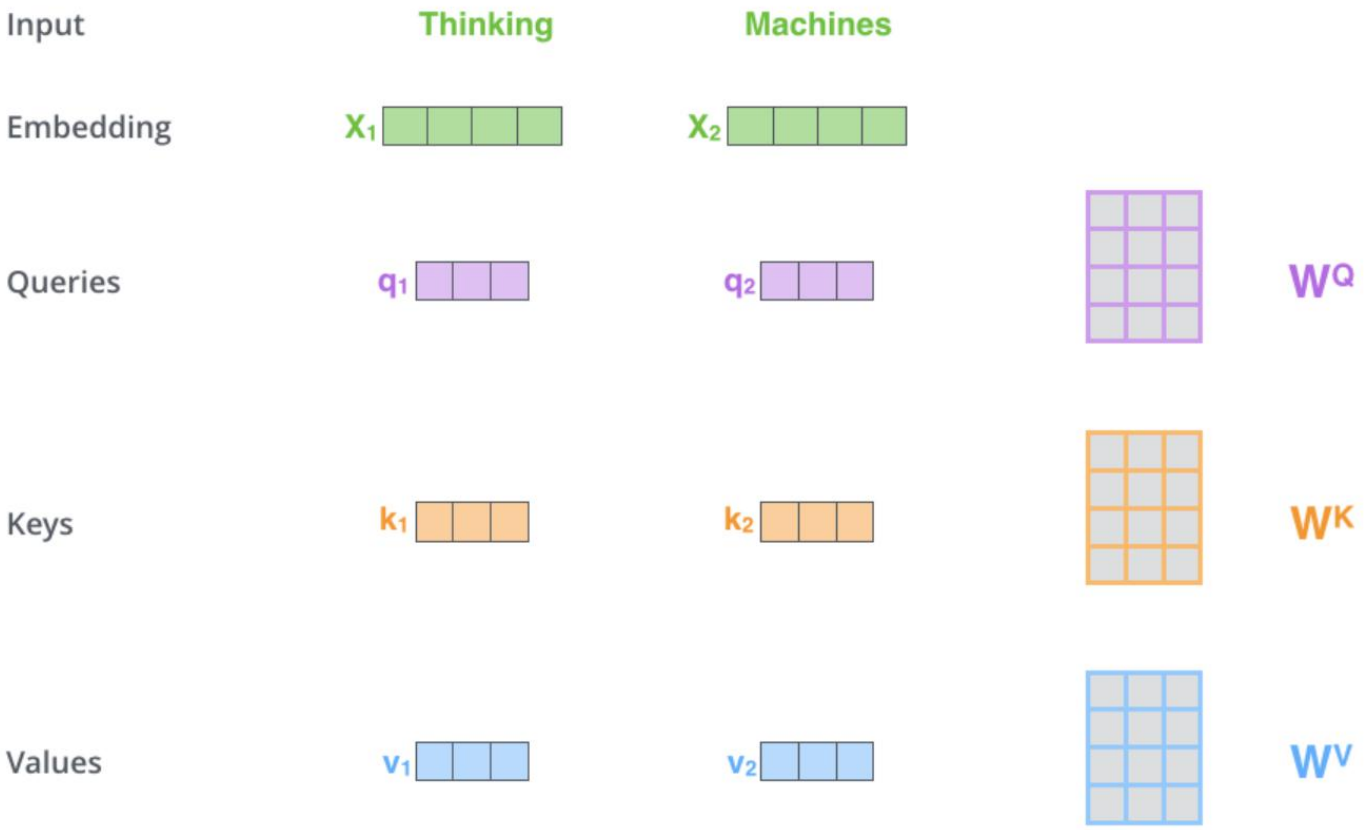
The animal didn't cross the street because it was too narrow.



Transformer

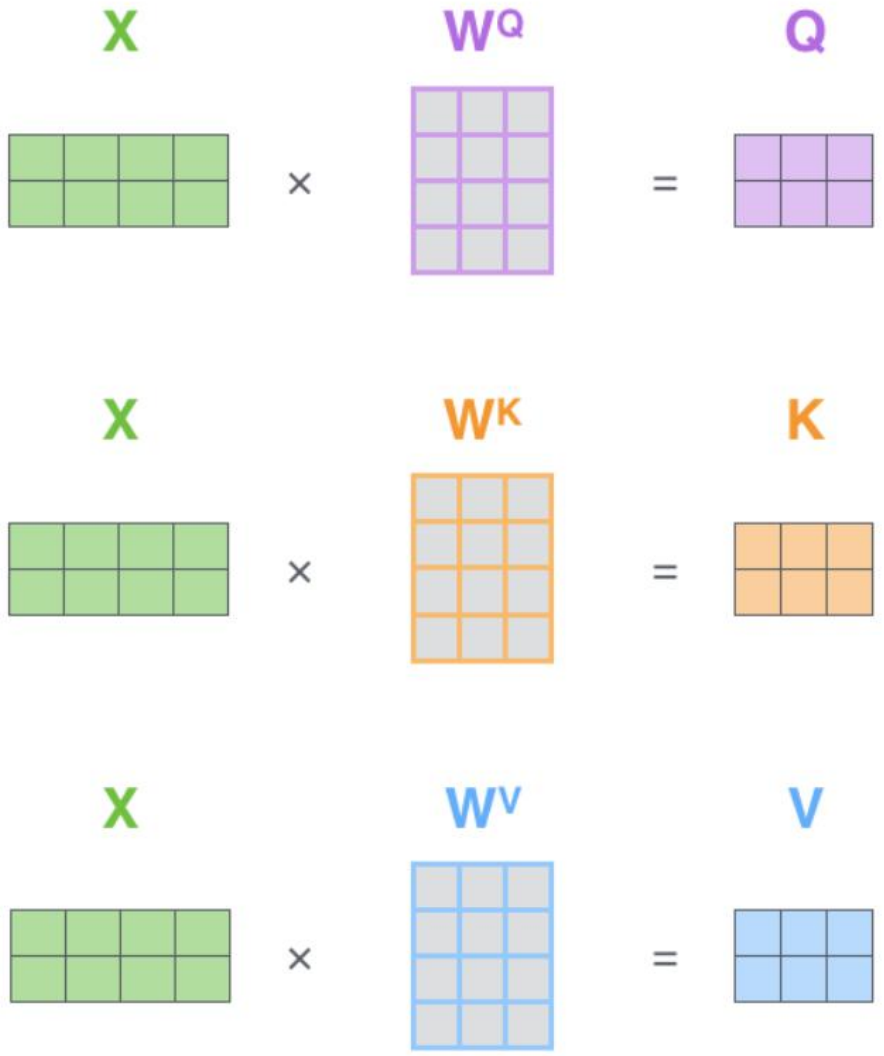
- self-attention如何计算?

  - 输入经过编码后得到向量
  - 想得到当前词语上下文的关系，可以当作是是加权
  - 构建三个矩阵分别来查询当前词跟其他词的关系，以及特征向量的表达。



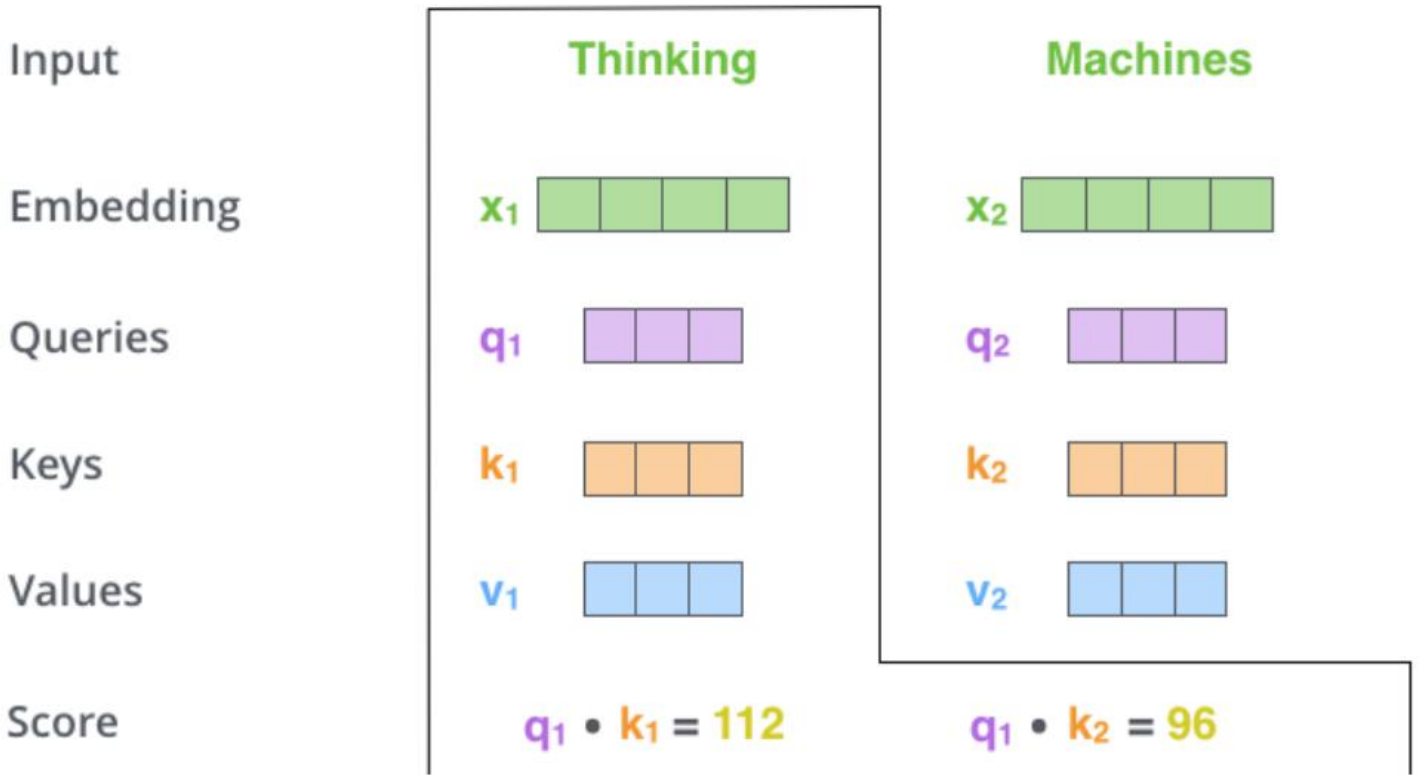
Transformer

- self-attention如何计算?
  - 三个需要训练的矩阵
  - Q: query，要去查询的
  - K: key，等着被查的
  - V: value，实际的特征信息



Transformer

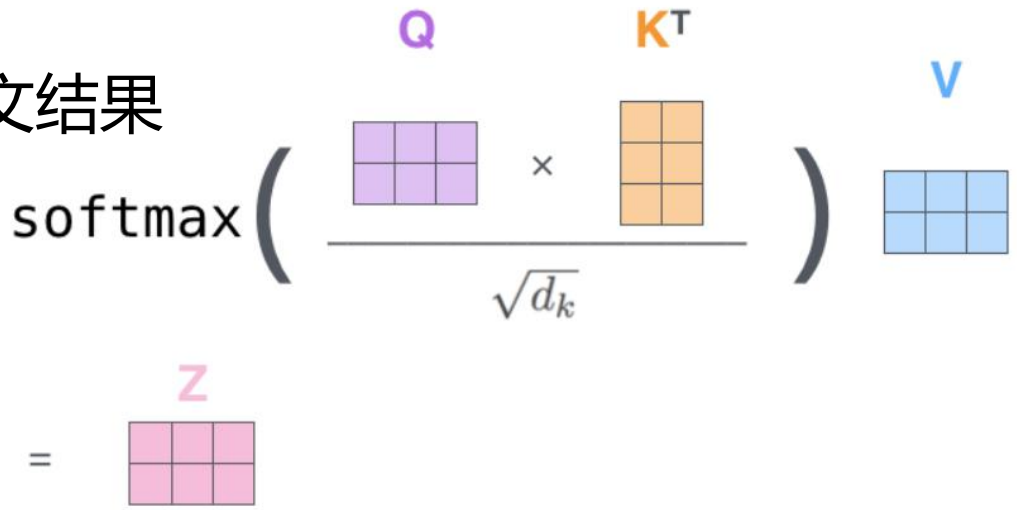
- self-attention如何计算?
  - q与k的内积表示有多匹配
  - 输入两个向量得到一个分值
  - K: key，等着被查的
  - V: value，实际的特征信息



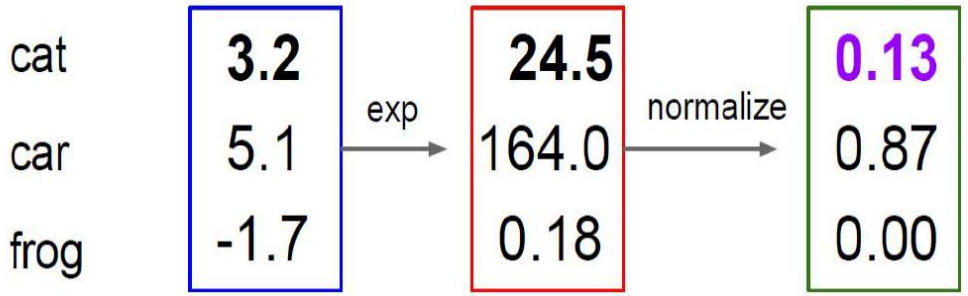
Transformer

- self-attention如何计算?
  - 最终的得分值经过softmax就是最终上下文结果
  - Scaled Dot-Product Attention

    不能让分值随着向量维度的增大而增加



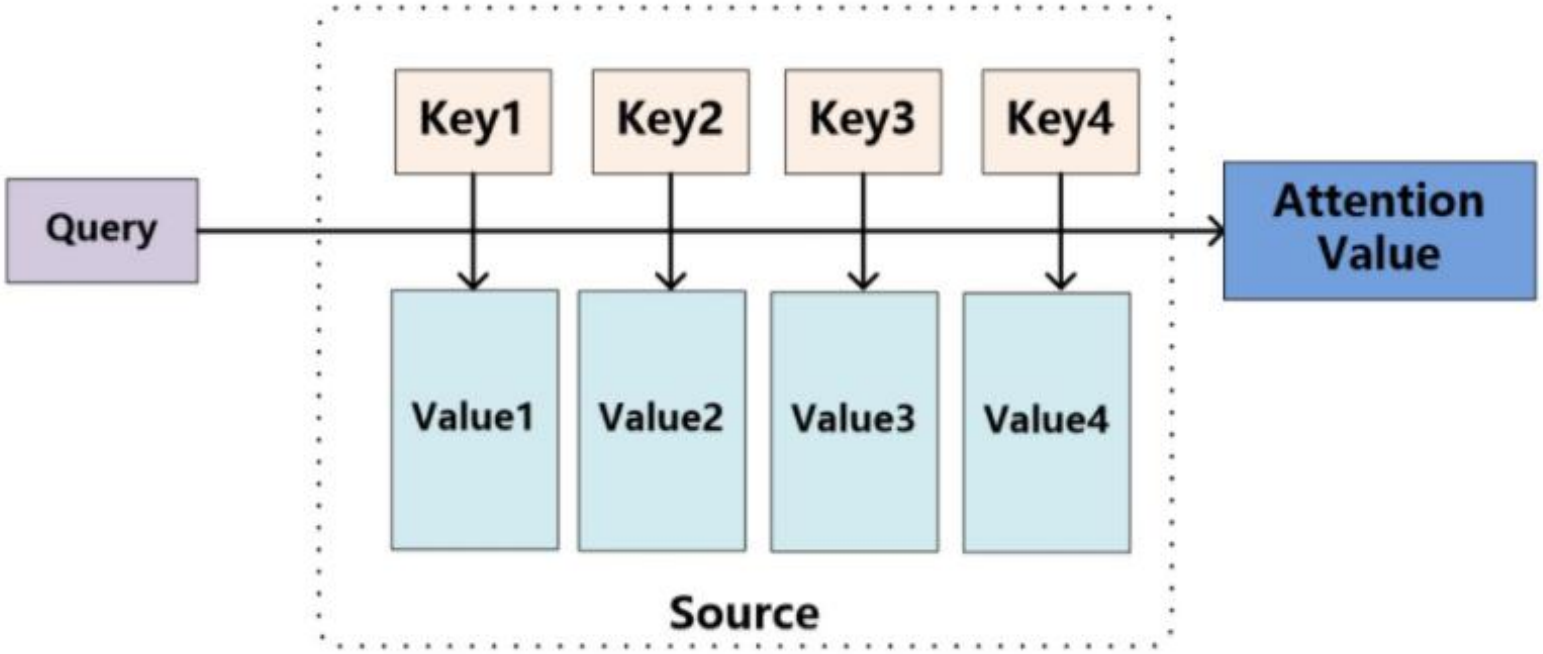
- softmax回忆:



Transformer

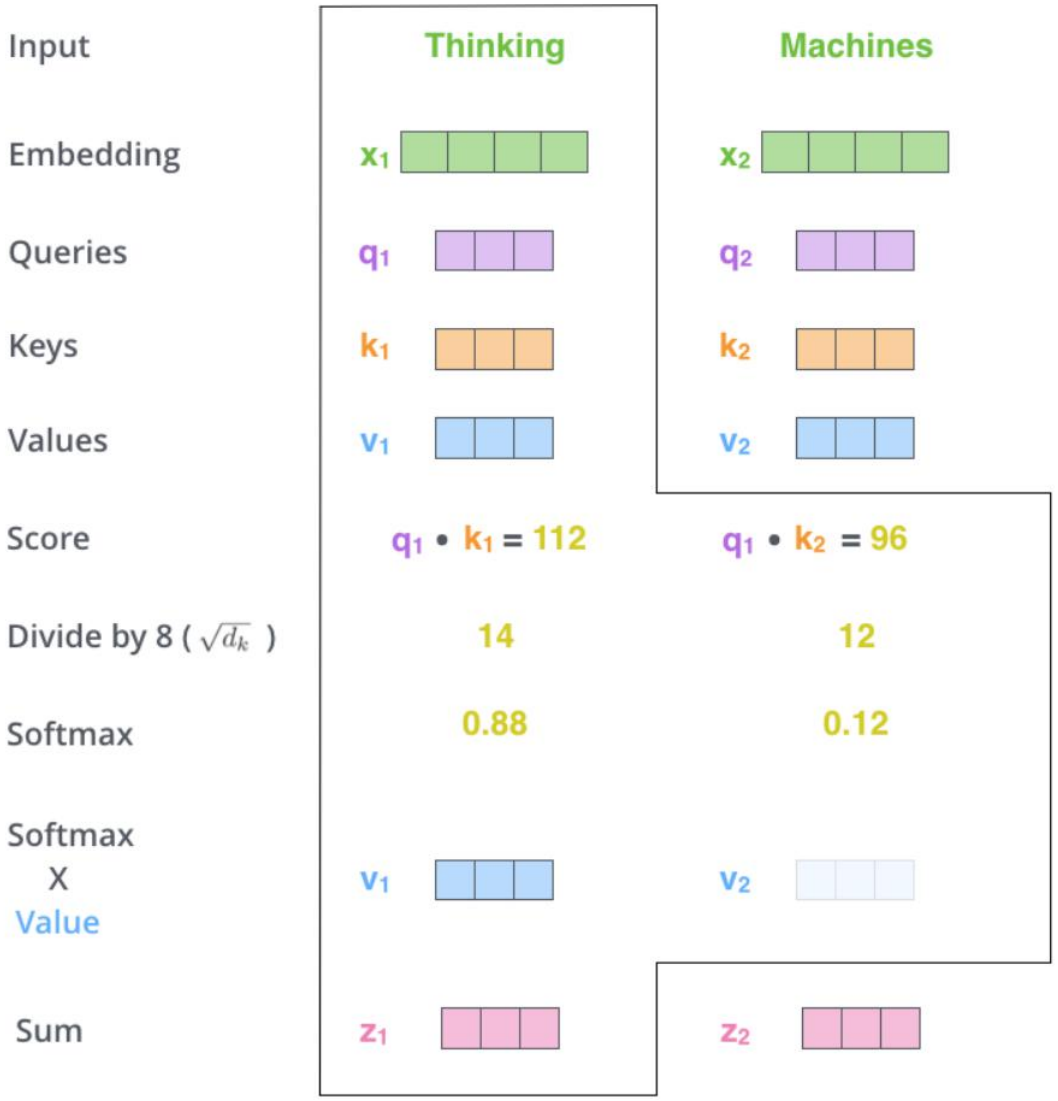
- 每个词的Attention计算

  - 每个词的Q会跟整个序列中每一个K计算得分，然后基于得分再分配特征



Transformer

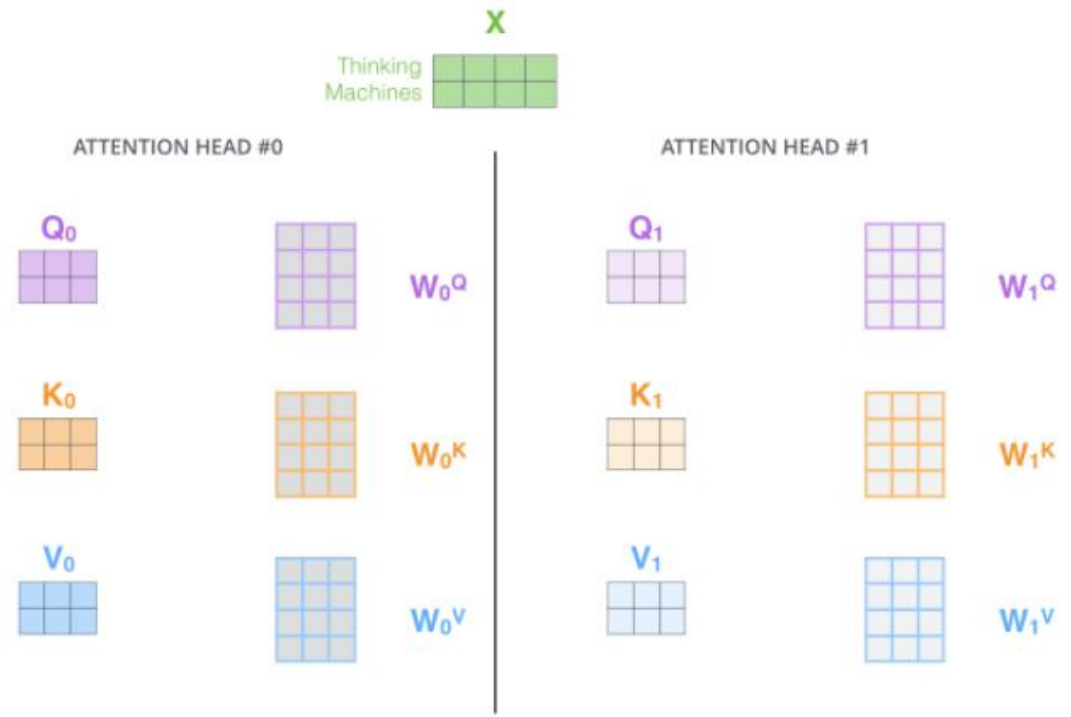
- Attention整体计算流程
  - 每个词的Q会跟每一个K计算得分
  - Softmax后就得到整个加权结果
  - 此时每个词看的不只是它前面的序列而是整个输入序列
  - 同一时间计算出所有词的表示结果



$$z_1 = 0.88v_1 + 0.12v_2$$

Transformer

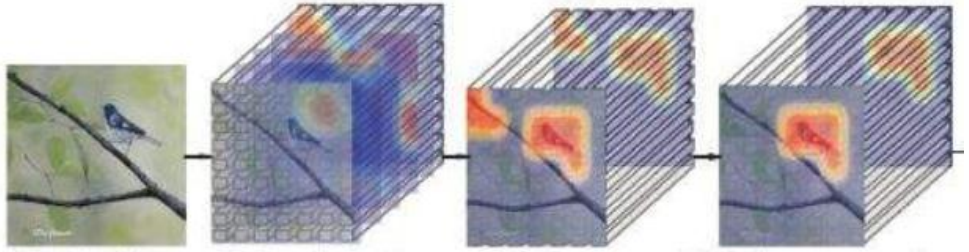
- multi-headed机制



- 一组q,k,v得到了一组当前词的特征表达
- 类似卷积神经网络中的filter

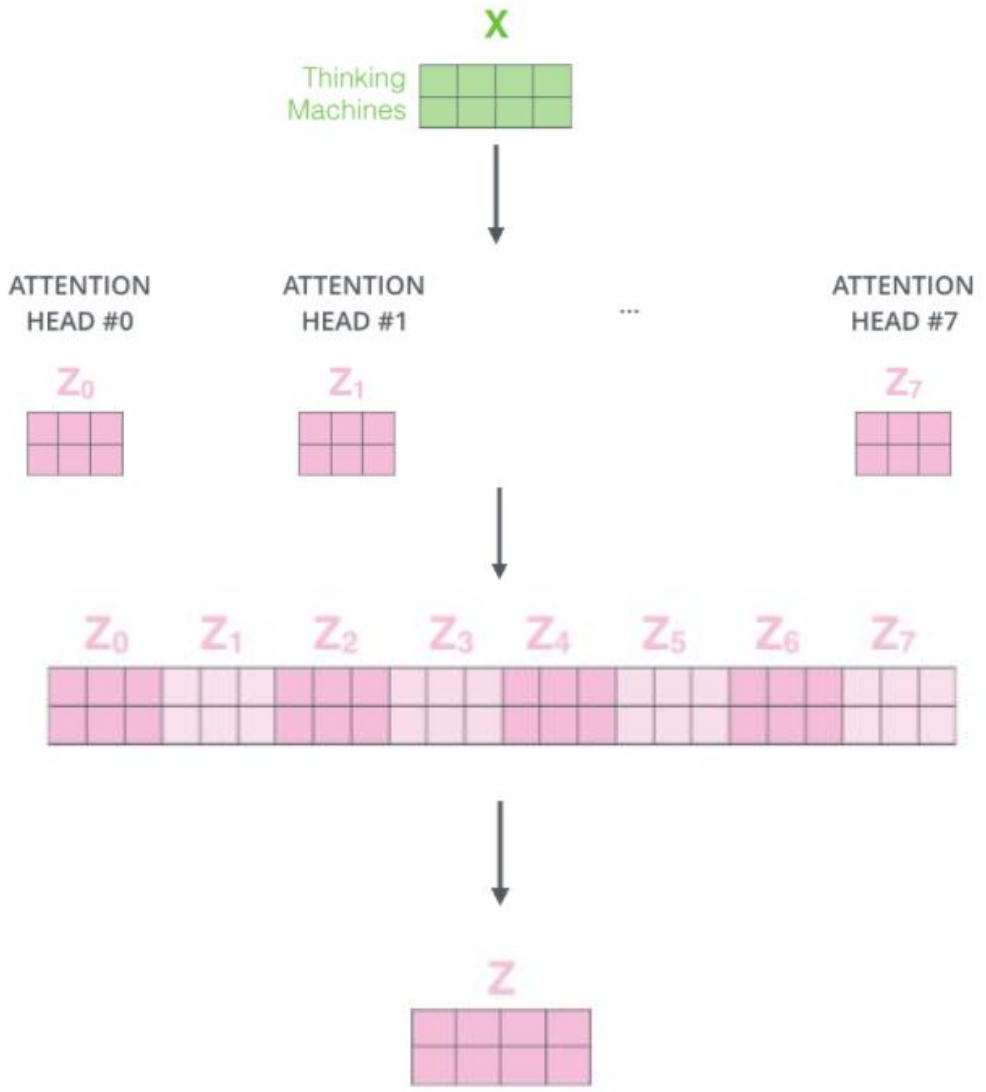
  能不能提取多种特征呢?

- 卷积中的特征图:



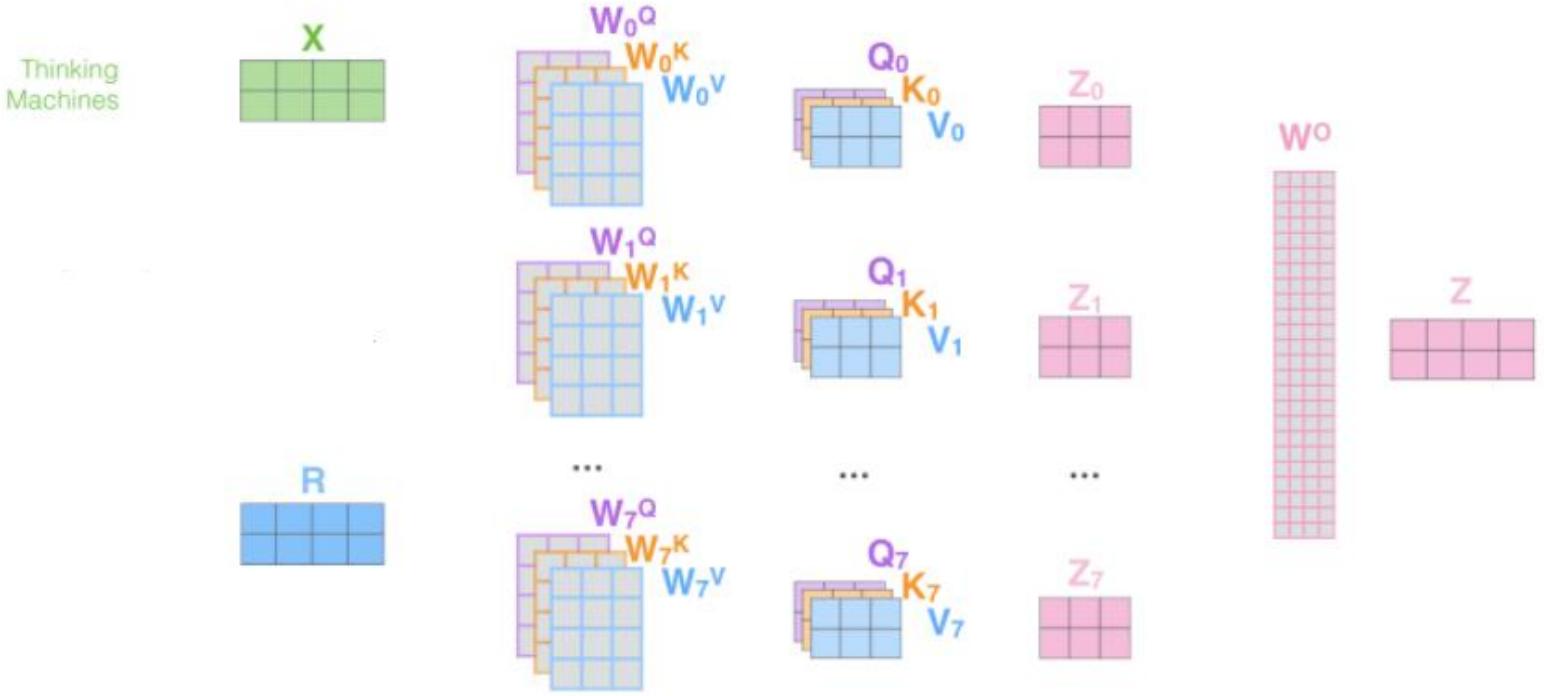
Transformer

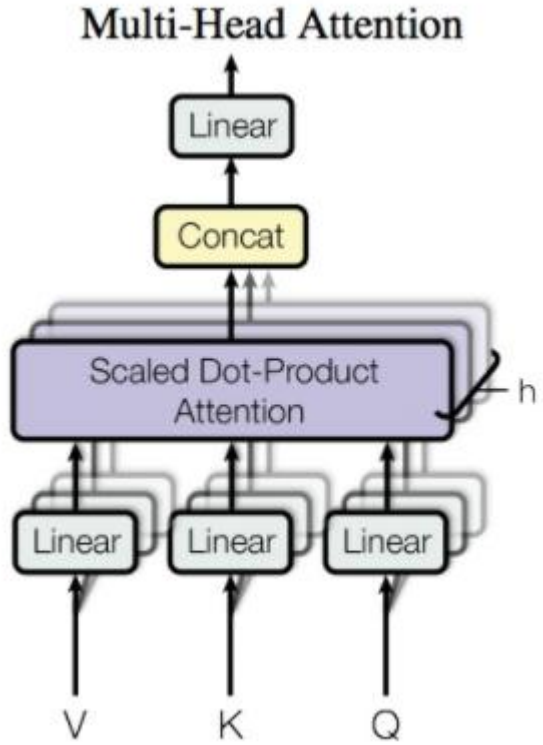
- multi-headed机制
  - 通过不同的head得到多个特征表达
  - 将所有特征拼接在一起
  - 可以通过再一层全连接来降维



Transformer

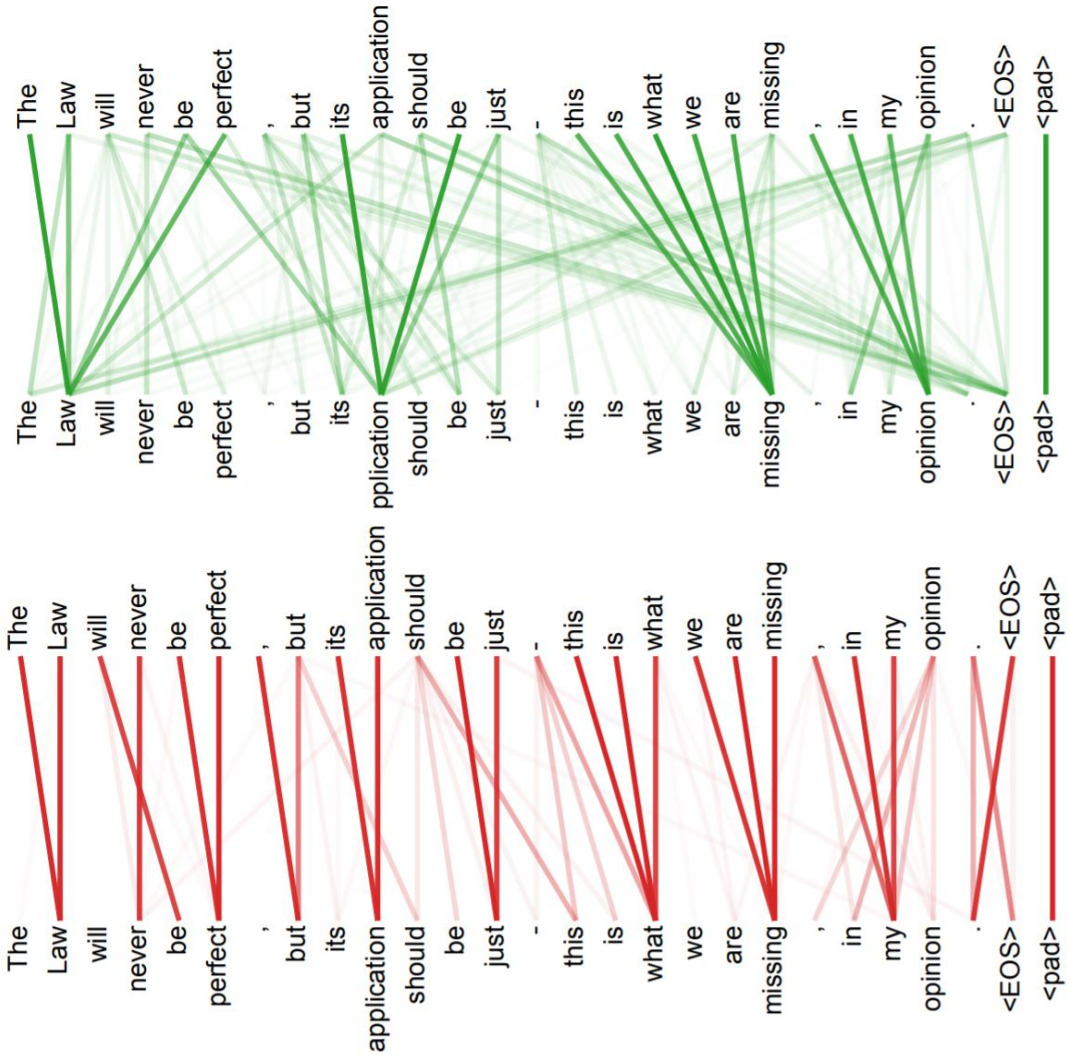
- multi-headed机制





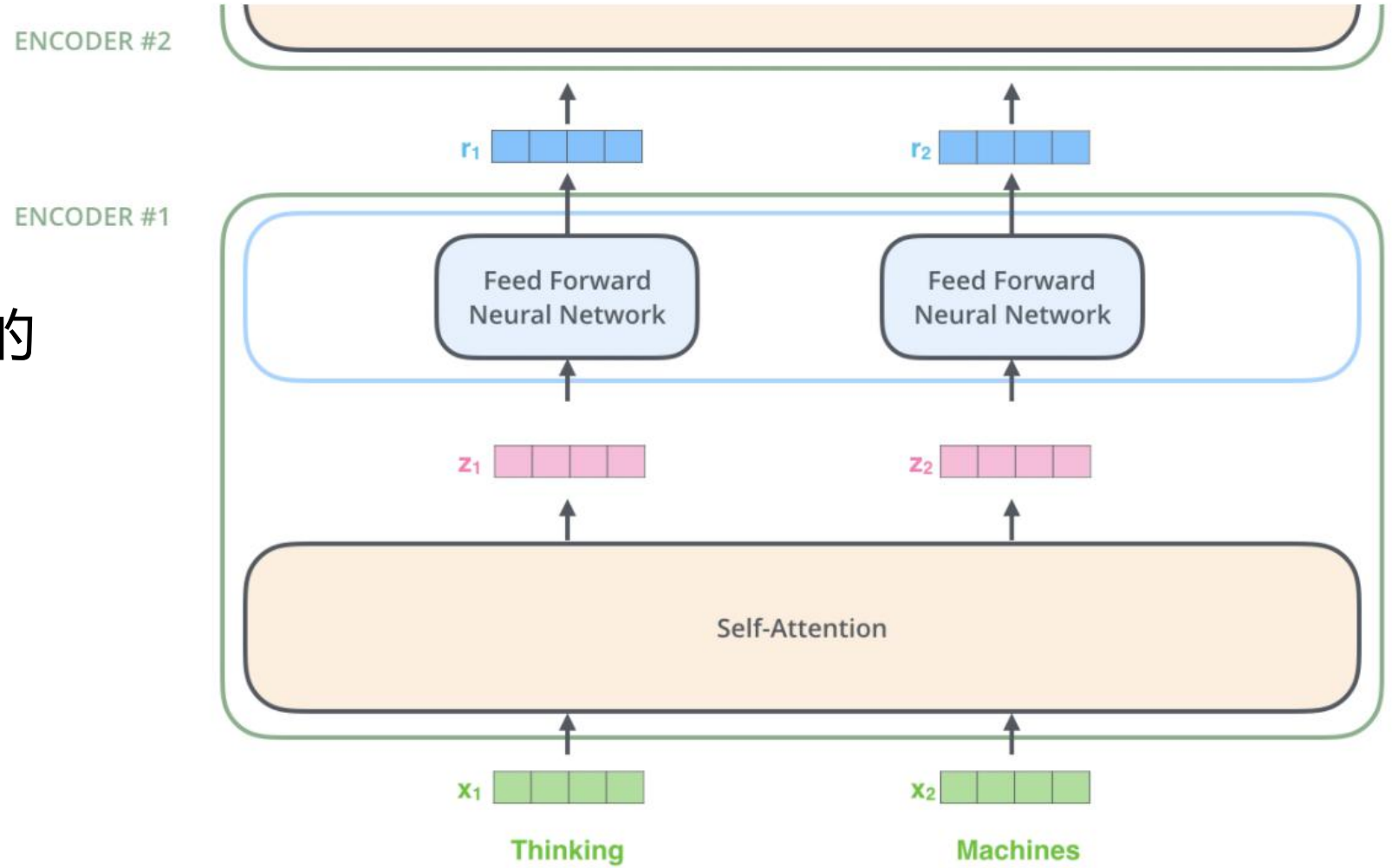
Transformer

- multi-headed结果
  - 不同的注意力结果
  - 得到的特征向量表达也不相同



Transformer

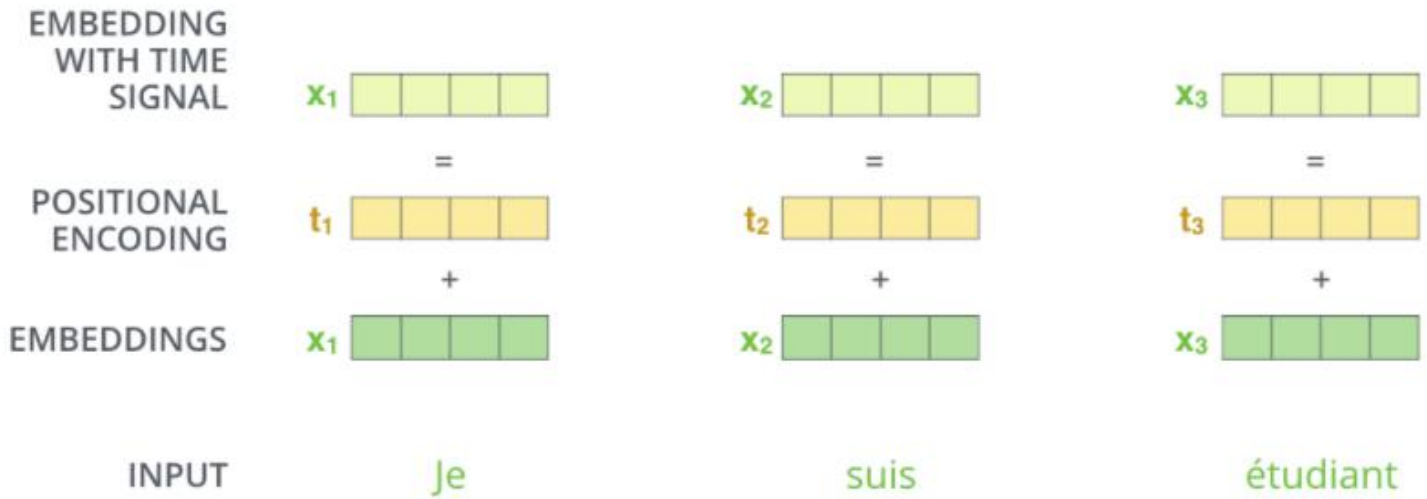
- 堆叠多层
- 一层怎能够
- 计算方法都是相同的



Transformer

- 位置信息表达

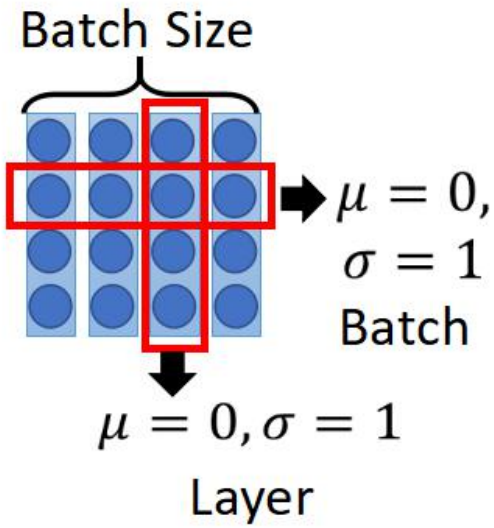
在self-attention中每个词都会考虑整个序列的加权，所以其出现位置并不会对结果产生什么影响，相当于放哪都无所谓，但是这跟实际就有些不符合了，我们希望模型能对位置有额外的认识。



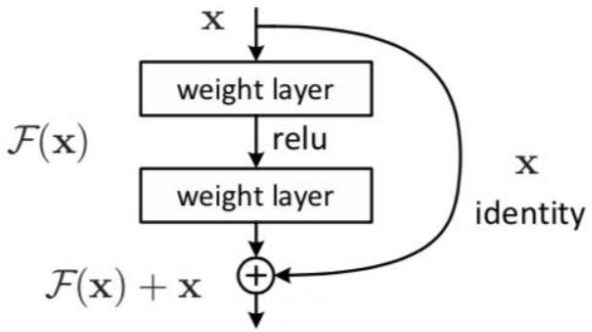
Transformer

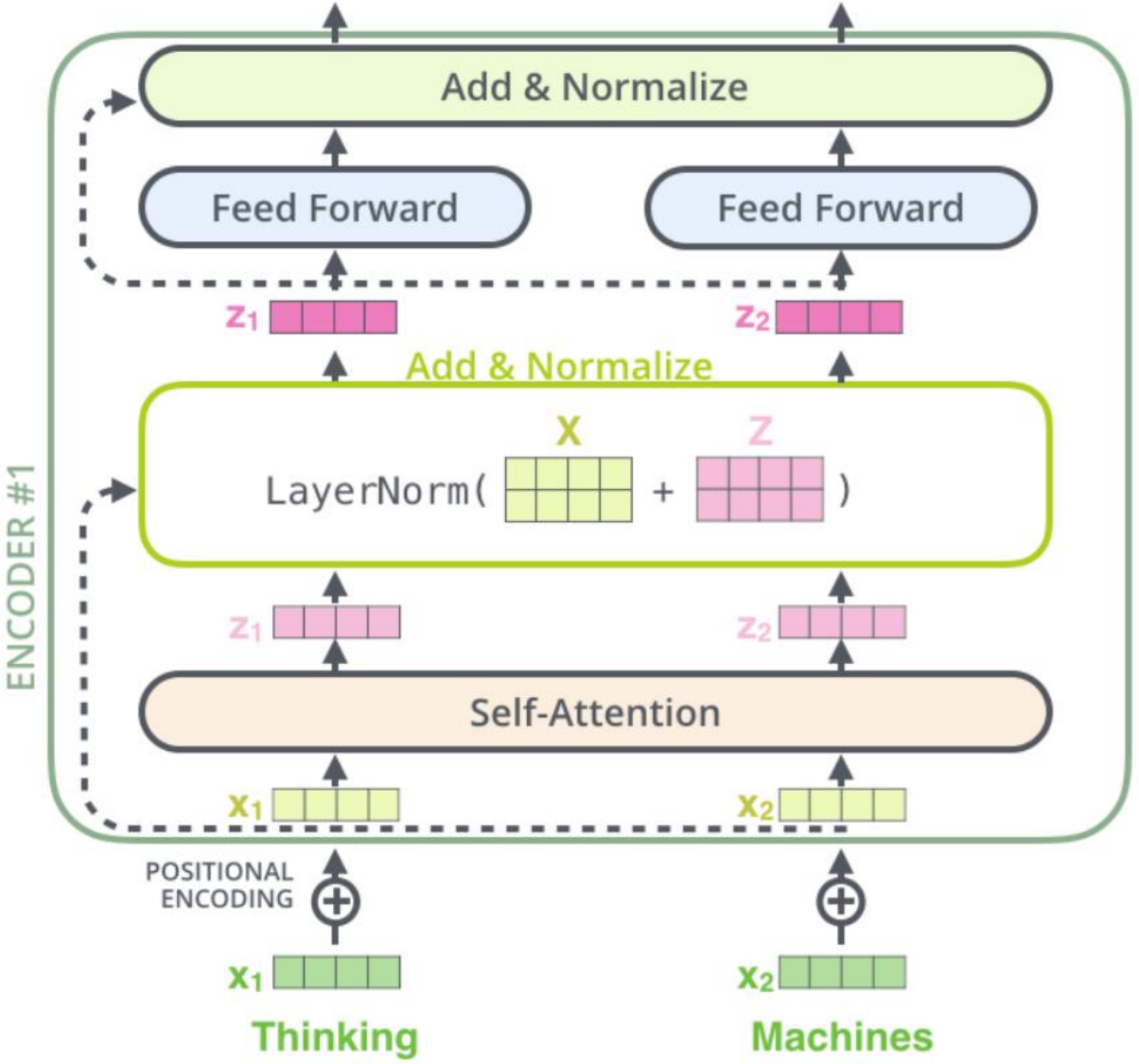
- Add与Normalize

归一化：



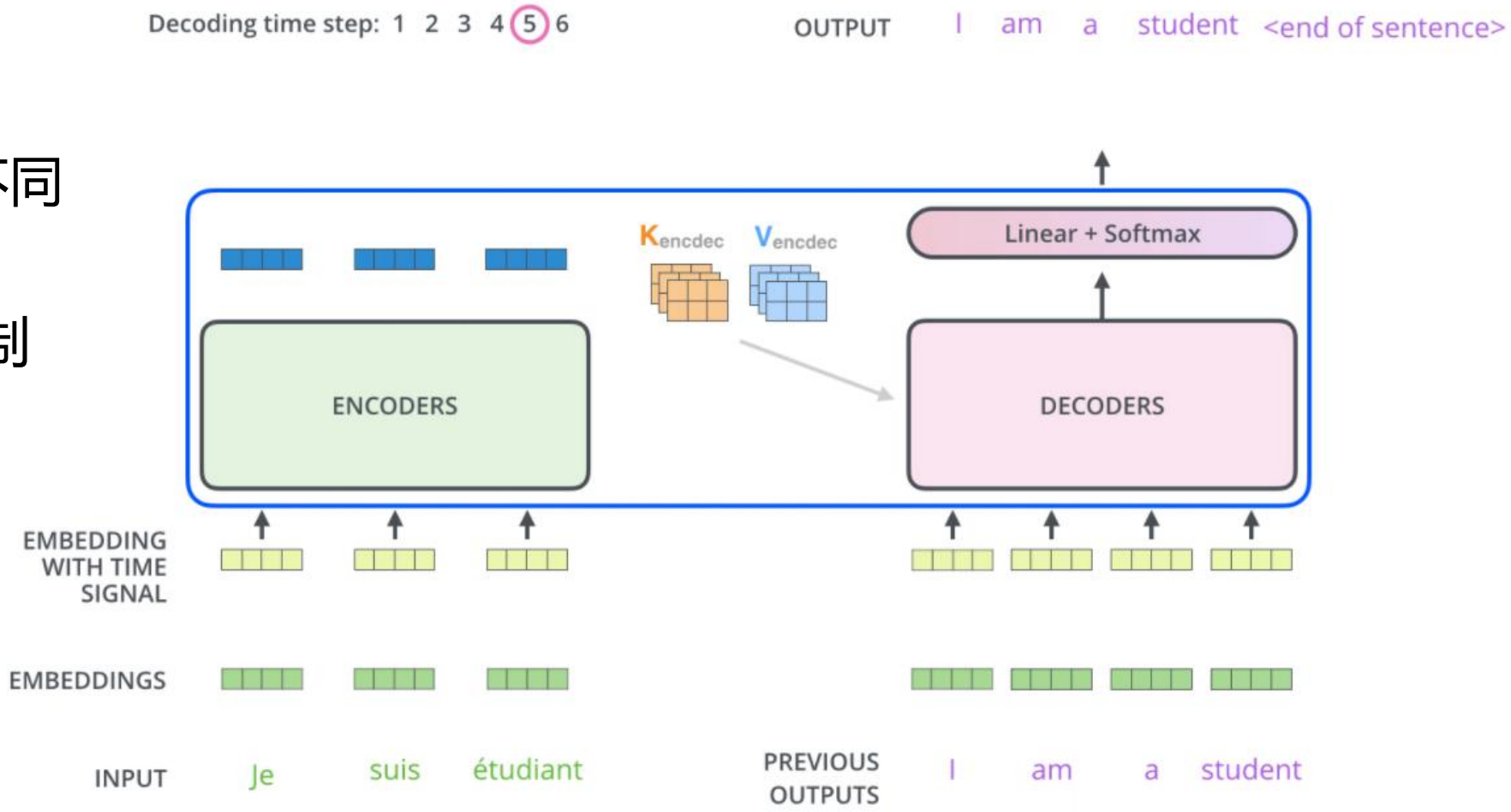
连接：基本的残差连接方式





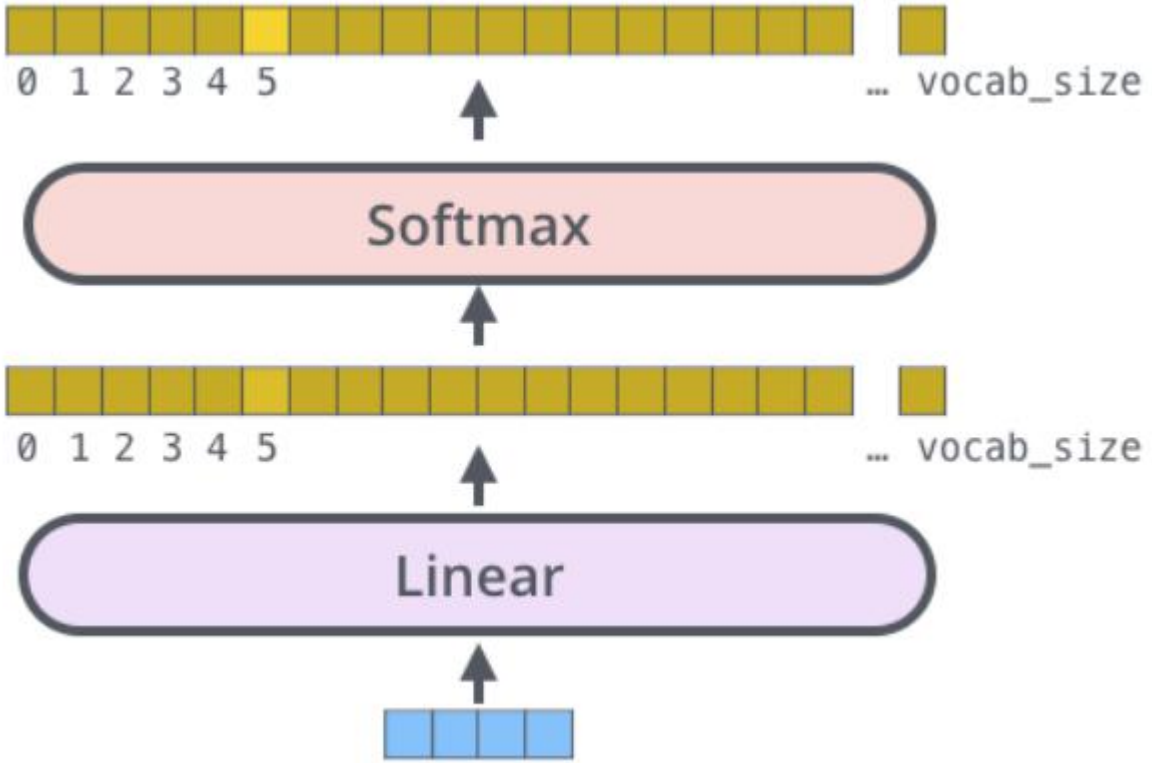
Transformer

- Decoder
  - Attention计算不同
  - 加入了MASK机制



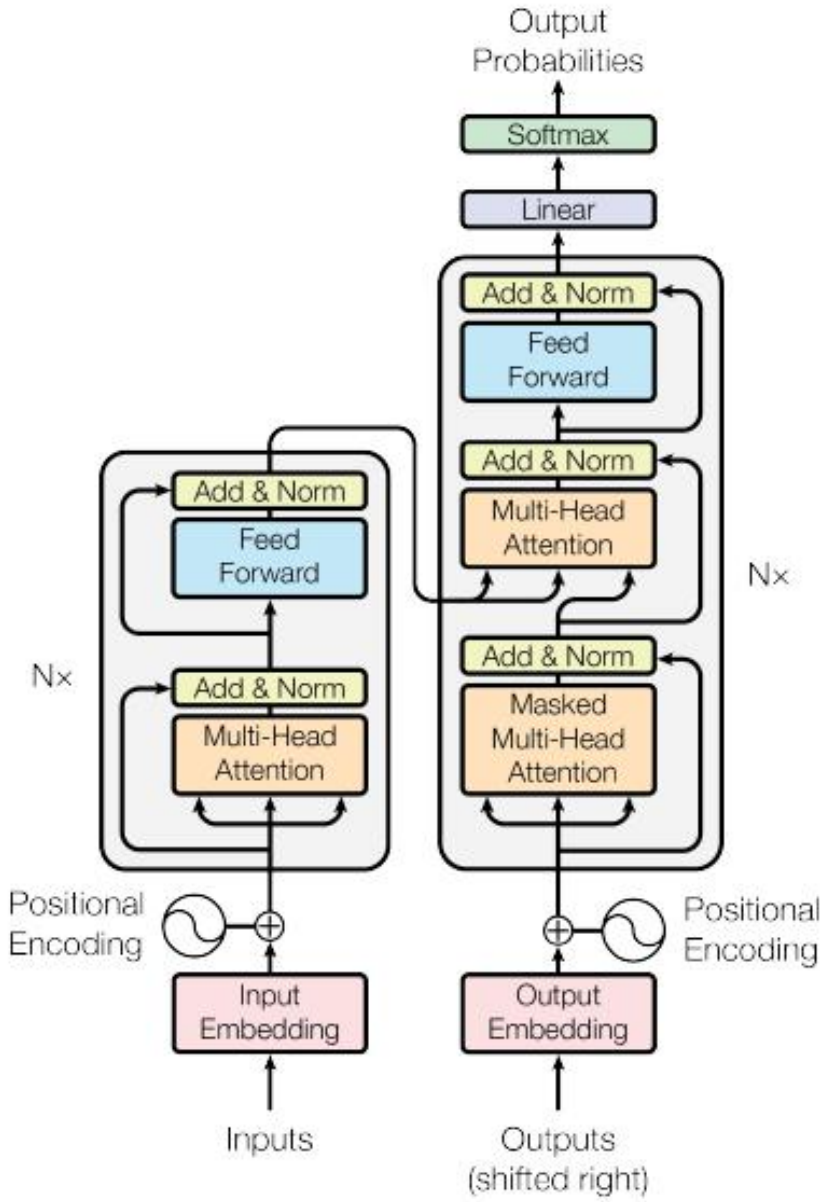
Transformer

- 最终输出结果
  - 得出最终预测结果
  - 损失函数 cross-entropy即可



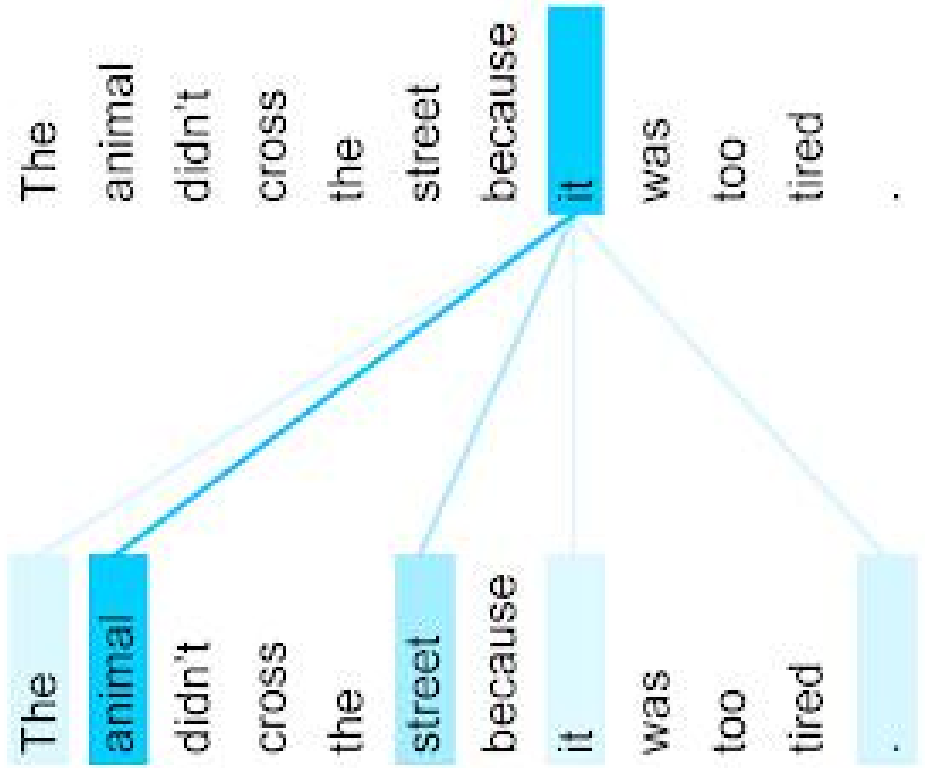
Transformer

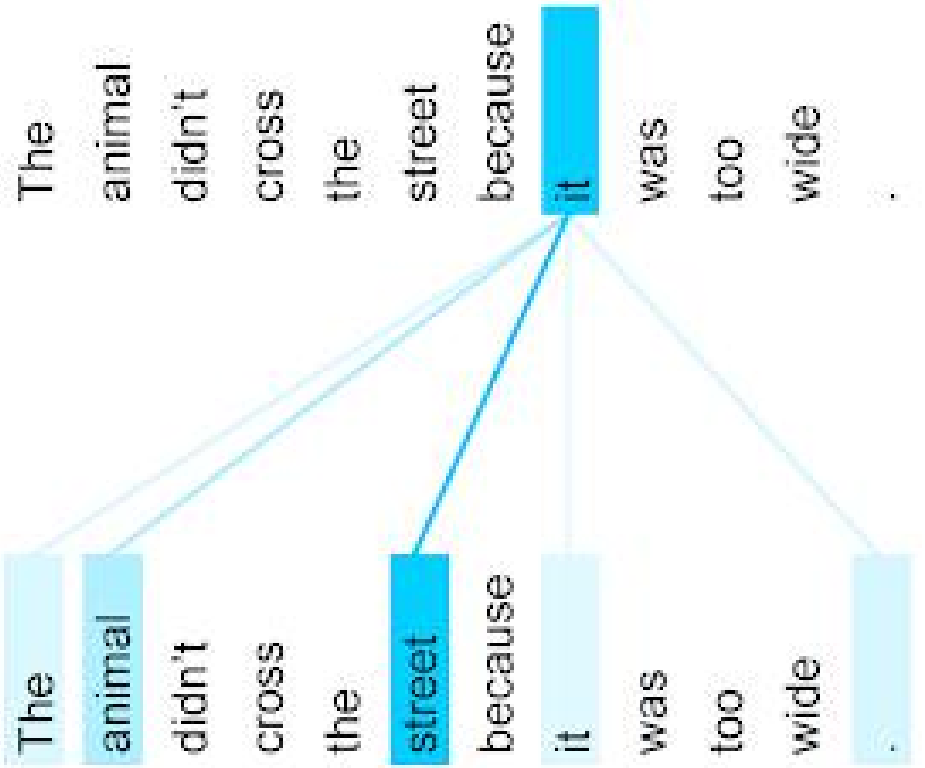
- 整体梳理
  - Self-Attention
  - Multi-Head
  - 多层堆叠，位置编码
  - 并行加速训练



Transformer

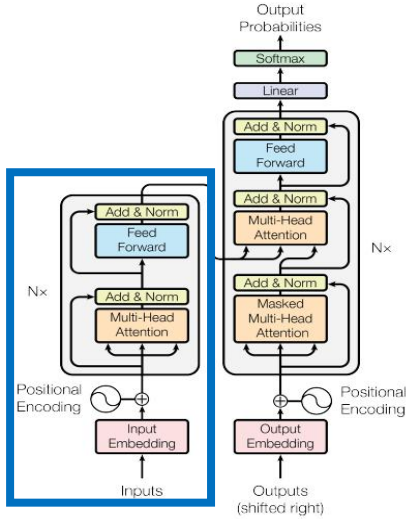
- 效果展示





BERT

- BERT训练的词向量有什么不同？
  - 在word2vec中，相同词对应的向量训练好后就固定了
  - 但在不同的场景中，‘干哈呢’的意思会相同吗？
  - 这俩兄弟都叫transformer:



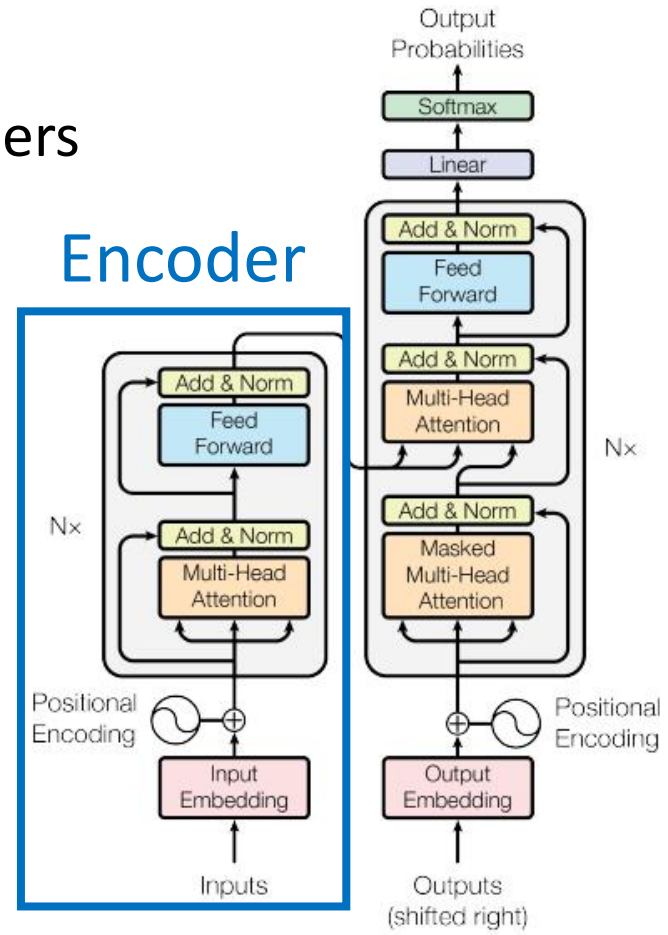
BERT

- 这名字该怎么解释?

- Bidirectional Encoder Representations from Transformers

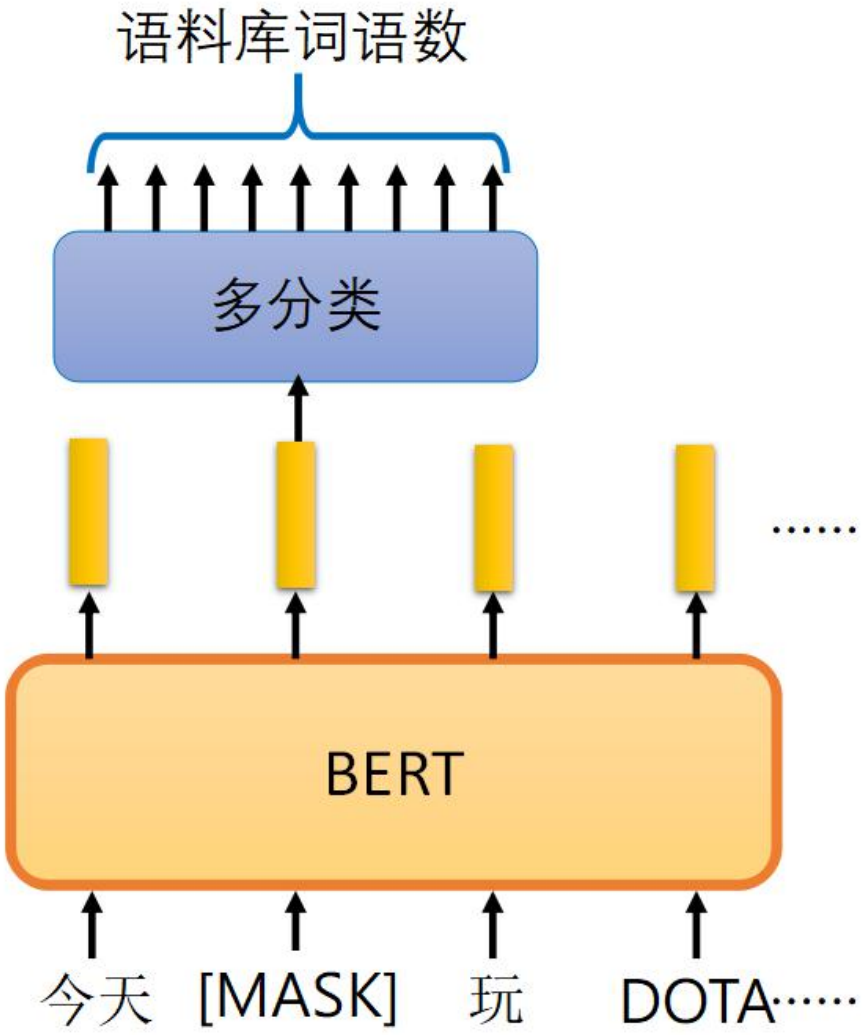
- 说白了就是transformer的encoder部分

- 并不需要标签，有预料就能训练了



BERT

- 如何训练BERT
  - 方法1：句子中有15%的词汇被随机mask掉
  - 交给模型去预测被mask的家伙到底是什么
  - 词语的可能性太多了，中文一般是字
  - 如果BERT训练的向量好，那分类自然OK

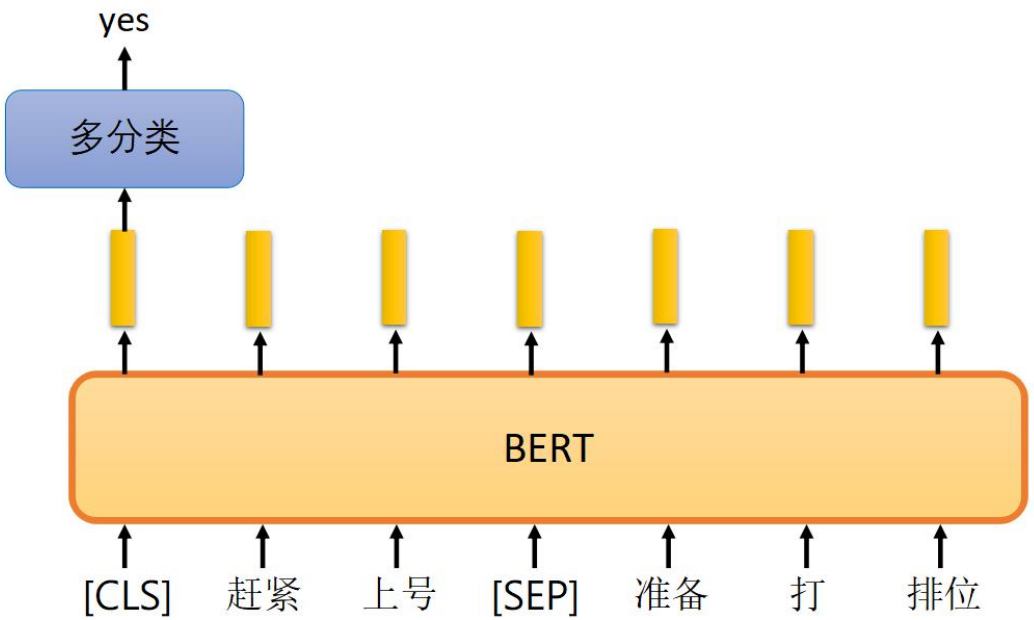


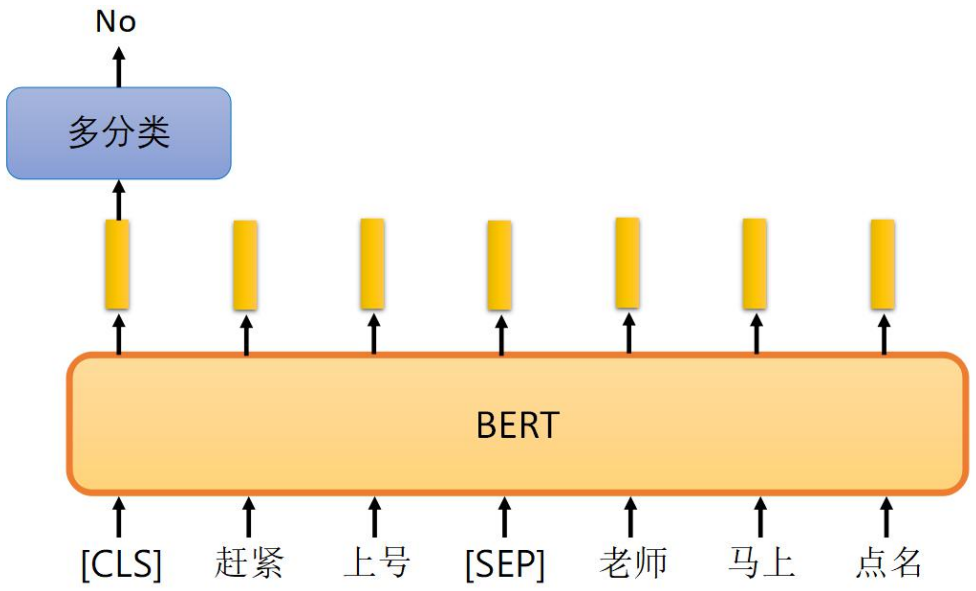
BERT

- 如何训练BERT

  - 方法2：预测两个句子是否应该连在一起

  - [seq]：两个句子之前的连接符，[cls]：表示要做分类的向量





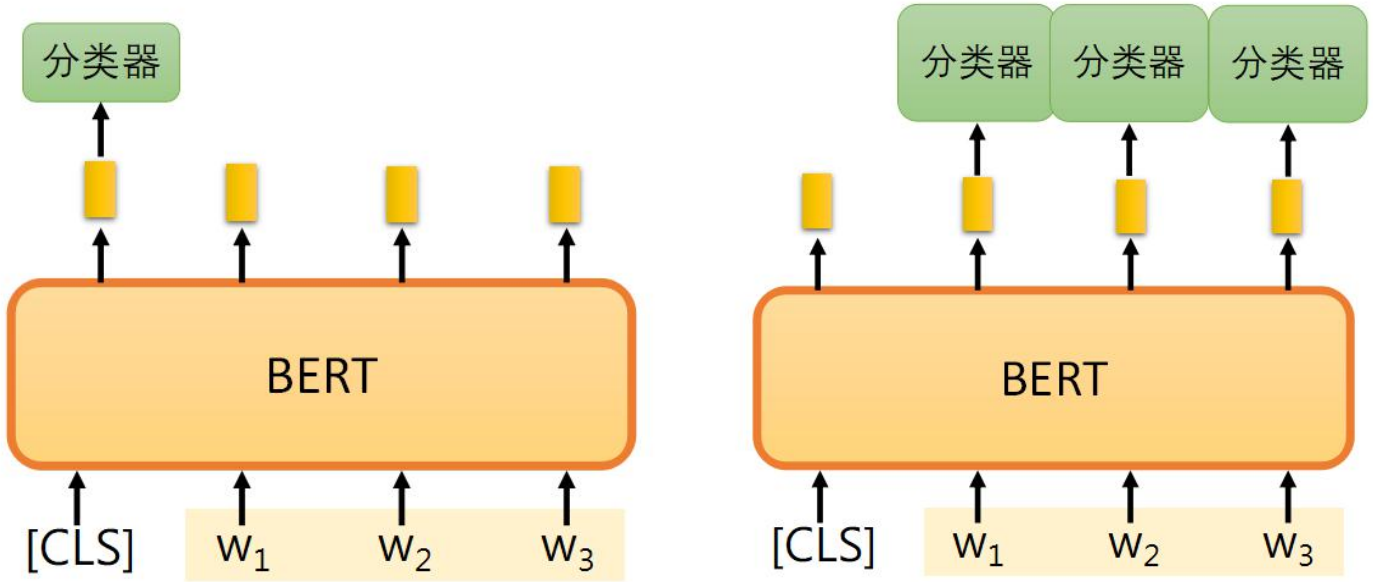
BERT

- 如何使用BERT

  - 是不是需要先训练好向量的表达，然后再训练需要的模型呢?

  - 所需的任务融入BERT中即可，它俩一起训练的!

  - 分类任务:



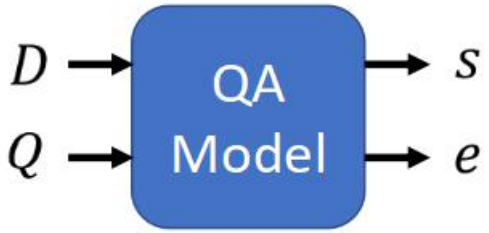
BERT

- 如何使用BERT

  - 阅读理解题，输入是文章和问题，输出是理解的答案位置。

$$\underline{\textit{文章}}: \quad D = \{d_1, d_2, \cdots, d_N\}$$

$$\underline{\textit{问题}}: \quad Q = \{q_1, q_2, \cdots, q_N\}$$



$$\underline{\textit{结果}}: \quad A = \{q_s, \cdots, q_e\}$$

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

BERT

- 如何使用BERT

- 如何设计网络呢？需要分别计算答案的起始和终止位置