

编码和解码的問題

编码和解码的结构.

-token

sequence to sequence (seq2seq) 的基本問題

挑战 = seq2seq 生成问题

基于 RNN 的 seq2seq 例子



编码器解码器结构

注意力机制

在大框架下作出的特定位嵌入

什么是编码和解码

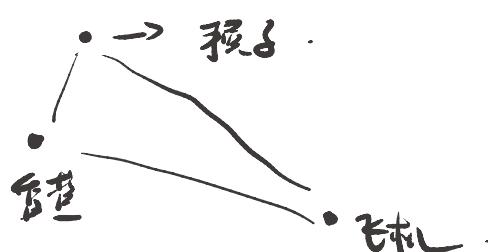


数值

语义关系值 (数字) · 单体识别语义关系 (相似)

↓ eg: 地面上坐标之间的距离
单体识别语义关系取值的大小

eg



两种“放端”编码例子

下图展示了对 tokenizer (分词器)

{ one-hot



每个 token -> 一个



->

信息量低且稀疏

token.

→ 对基础的语义单元进行表征

g: 词、单词、词根、子词

空间上相距：语义无关，非邻居特征

中文



潜语义

每个 tokenizer

共同的潜语义

每个 token -> one-hot

矩阵运算

- W

神经网络 sigmoid ($x^T M + b$)

矩阵相乘 矩阵计算
特征的提取

隐藏层加深：准确程度越高

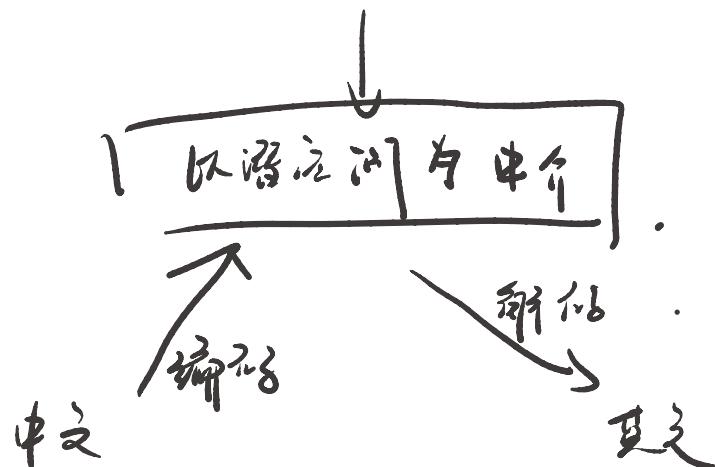
编码：token \rightarrow one-hot \rightarrow 由语义决定的词向量 embedding
 然后 \rightarrow

↓

嵌入矩阵

先嵌入再统一或先统一再嵌入

eg. 中、英之名句得到
 直接统一中英文可得到一个语义向量
 和之似语空间再联系起来



eg token $\xrightarrow{\text{embedding}}$ 向量 $\boxed{1}$, 2 , ..., $\boxed{10}$ 17
 苹果

每个维度都是该之某一个特征

基础语义

→ 国语中 每个词有一个向量 -一个词向量

词向量是矩阵相乘升维 or 降维。

→ 国语中 $I \times I$ 的卷积核，升维 or 降维

滑窗之间和字串之间和飞检

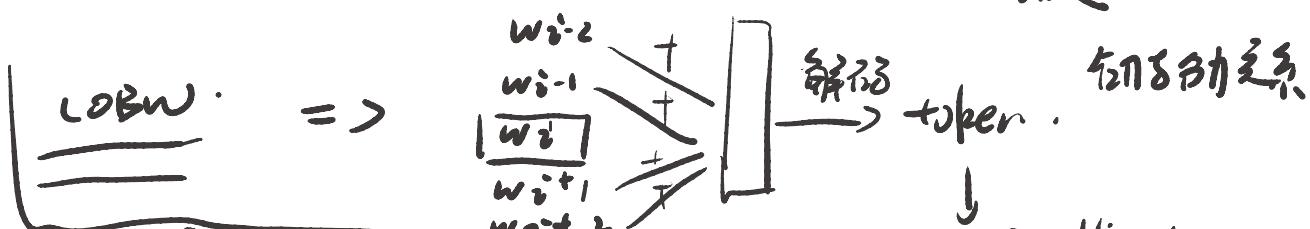
↓
逐维化
—— 对比

因此滑窗之间可以对训练中没有遇到的特征也可规划
规范化参数

是否真正理解了语言

word2vec 得到嵌入矩阵

通过上下文推断 token
语义



token 之间的语义关系
单词

e.g.: 训练数据 这是一个 苹果

一
绿
细
.

苹果

空间转换和
token 比较
修正参数

在训练后这些 token 和词语
会比较相似

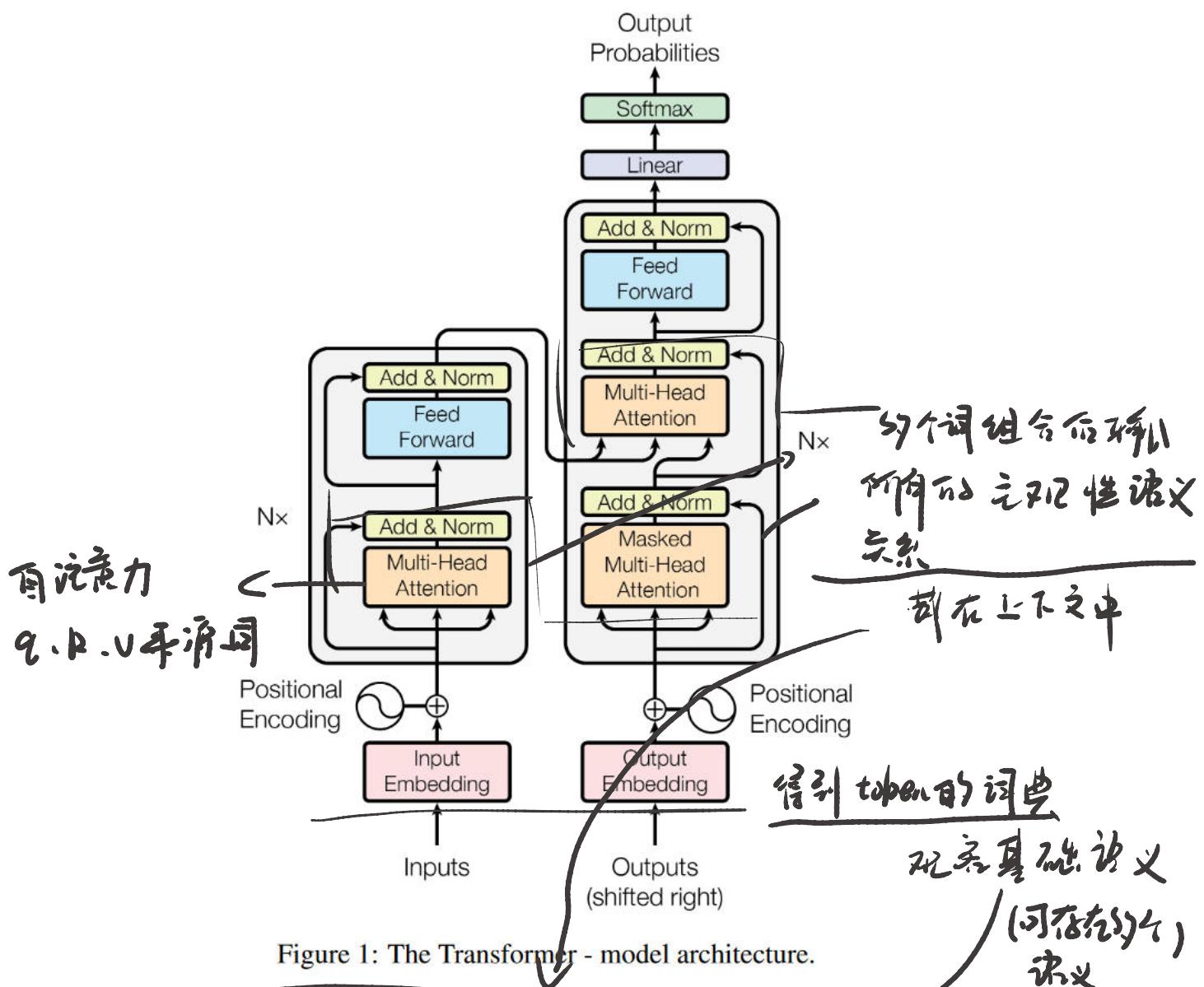
而用不同词典得不同映射和组合

体现的是泛观性

及语义关系

→ 该类方法集语义和事

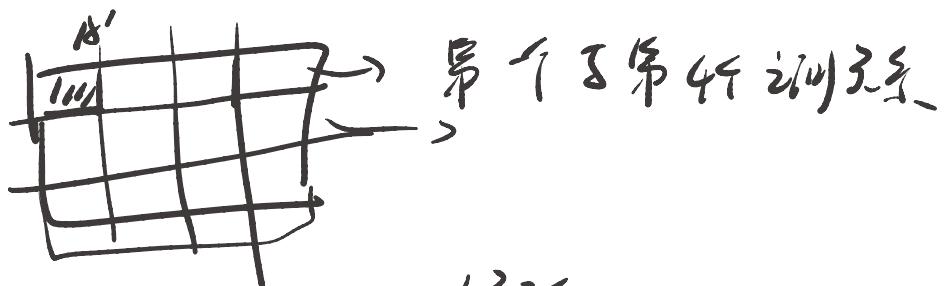
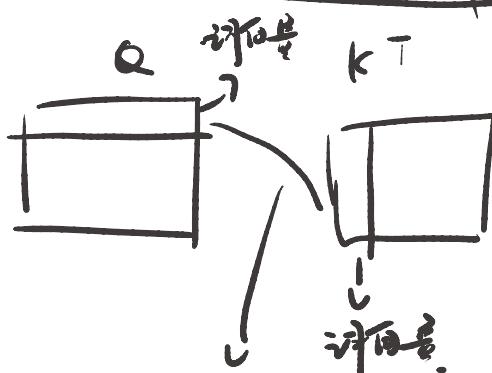
注意力机制[及原理]



$$\text{softmax} \left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_{\text{out}}}} \right) \cdot \mathbf{V}$$

(Q · P) 分解为两个向量内积的计算 \rightarrow 技术 \hookrightarrow 特殊性

2. 可能能 A' 中为
正负的量相交生
大得分离



$$\begin{matrix} t_1 & \begin{matrix} \text{---} \\ \text{---} \end{matrix} \\ t_2 & \begin{matrix} \text{---} \\ \text{---} \end{matrix} \end{matrix} \cdot \begin{matrix} v \\ \begin{matrix} \text{---} \\ \text{---} \end{matrix} \end{matrix} = \begin{matrix} \text{结果} \\ \begin{matrix} \text{---} \\ \text{---} \end{matrix} \end{matrix}$$

\vec{v}_1, \vec{v}_2 很大
 得到很大
 (很复杂)

Q · k

为什么训练这个是好的而 A 为什么直接训练一个 A

$$A = X \cdot WA$$

$$\delta = XWq [x \cdot w_k]^T$$

$$= XWq Wk^T x^T$$

$$= XWx x^T$$

更强大能
加非线性

为什么不用 $w_q \cdot w_q^T B P_6$?

$$A = X \cdot [w_q \cdot w_q^T] \cdot X^T$$

$$B' = X \cdot [w_q \cdot w_b^T] \cdot X^T = X \cdot B \cdot X^T$$

$$B = \frac{1}{2} (w_q \cdot w_b^T + w_b \cdot w_q^T)$$

对称矩阵

在数学上省 . 一样

钱 .

将主观语义与客观语义

前面表达语义的语境设置的语义
设置语义

在设置语境下表达语义
表达语义

| 2| 当两者进行相乘时 .

如果乘积太小，证明 设定语义与表达语义有相对
背离的地方，则需进行参数更新

而在同一个语境下，输出表示的是在同一主观语义框架下、
的结果

物理模型

所以两个矩阵 \rightarrow 可以在不同的设置语义下进行乘积
可能得到不一样的结论。 \rightarrow 结果太增加了可解释性

“人-机-模型.” 可能不太适合

自注意力：Q, K, V 相同。

交叉注意力 Q 的来源不同：相当于引入了学习过去词频统计的知识。

例 在表达语义这个很浅的层面上学习新知识。

哪怕只替换一个词，就出现问题

但在翻译里面是一种枢纽。

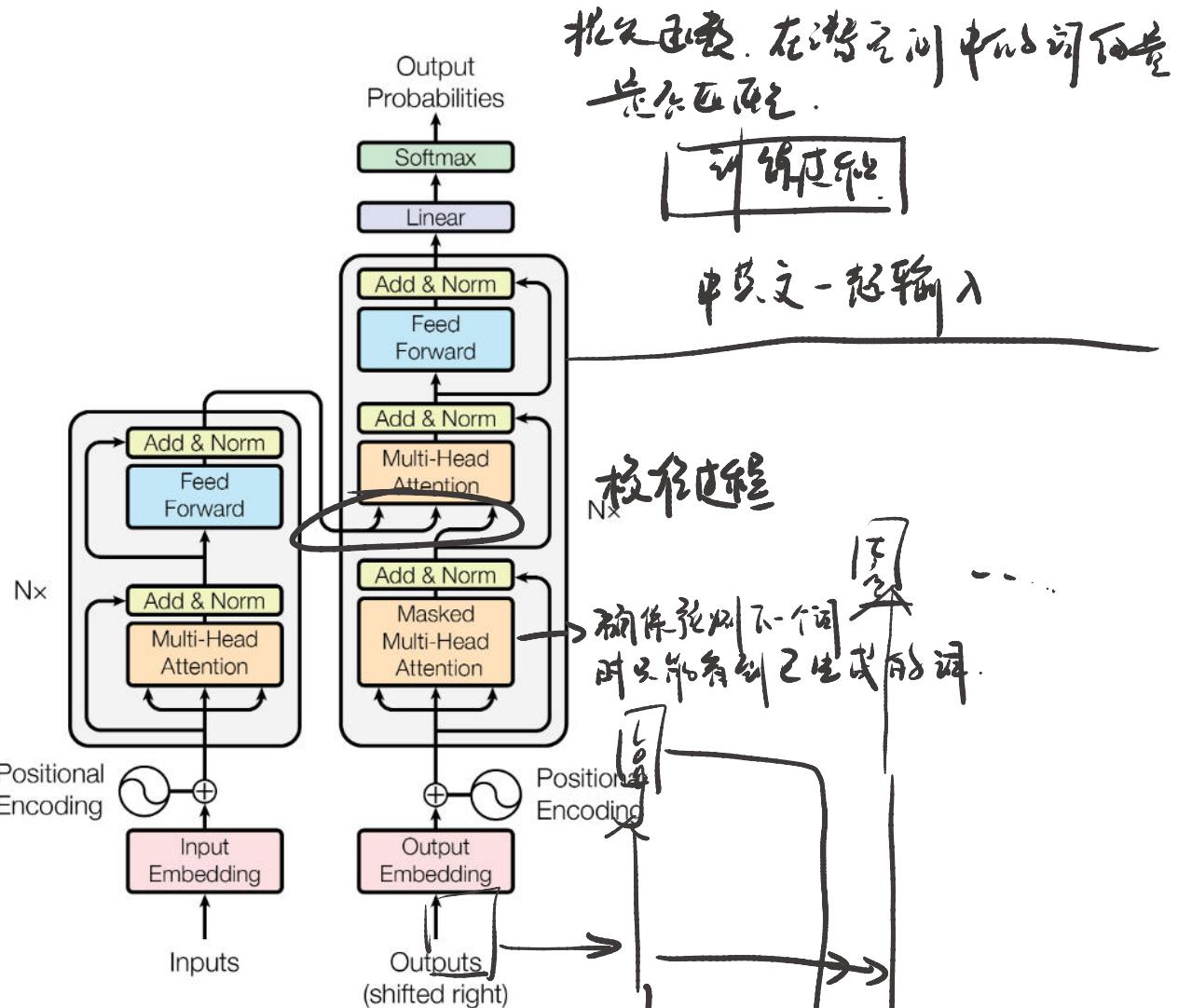


Figure 1: The Transformer - model architecture.

推理过程

从左到右

开始
long

位置角 α

航行计算同时和体积变化无关

加速度记时

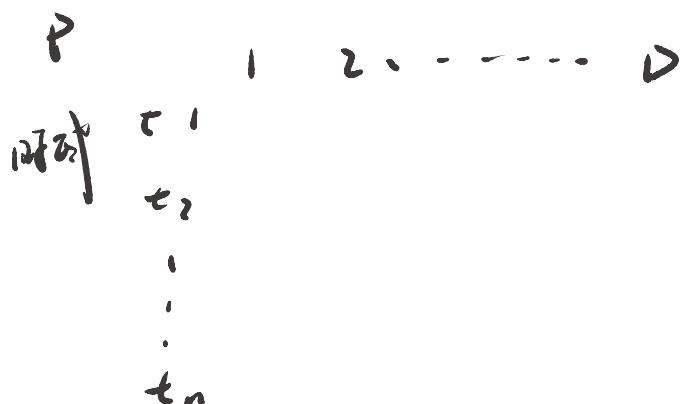
绝对位置角 α : 将

表算绝对到了时间量 t_1/t_2 从同向加速度区间中

位置下标二元坐标

从 FS 的航速程解。

时间 和 航线上 离差。



时间周期变大

时间

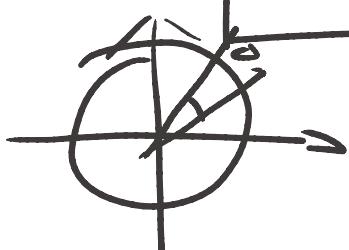
相对关系从固定变成了随轻快度

故此，阶段之间相互正交

{ 在更大范围内有更小的“基浪频率” }

旋转型阵

$$\begin{bmatrix} \cos\omega_i & \sin\omega_i \\ -\sin\omega_i & \cos\omega_i \end{bmatrix} \begin{bmatrix} \sin\omega_i x \\ \cos\omega_i x \end{bmatrix}$$



相对位置，对 A' (注意力得分) 进行操作

$f(a, k) \rightarrow g_1(q, q-k) \rightarrow g_1(q)g_2(q-k)$
构造出来的

$g_1(q), g_2(q-k)$

$\frac{V^T W^T R^T q - k}{\text{归一化}}$

$A = (x+p)W_q r_y W^T R^T (x+p)^T$



生成相对位置 $= R^T q - k$ 相关。

解决流能力

属性 1 CNN 对比 \rightarrow 捕捉依赖关系。

不同的头从不同的维度捕捉

顶部
侧面

本质上利用相对位置编码和源。