

# Week9习题课讲义

Topic:线性代数在语言模型中的应用（由于4.1的内容已经在week7习题课讲完，本周讲一些扩展性的知识，供大家开阔视野）

## 从向量的本质到词汇编码

Week1习题课：“什么是向量？” $\Rightarrow$  任何事物都需要通过数值编码进行量化的处理，编码而成的一系列数码组合就是向量

- 欧式空间中的坐标即为“欧式空间中的一点”的编码
- $A\vec{x} = \vec{b}$ 中的 $\vec{x}$ 即为 $A$ 矩阵列向量线性组合方式的编码

自然语言文本由一个个词汇构成，这些不同的词汇如何进行编码？

### Idea1：词汇到单个数值的一一映射

但这无法很好地表征不同词汇之间的语义关系

### Idea2：将自然语言的语义视为一个空间

假设有一系列国家名称词汇：

China Russia Canada Liechtenstein

简化情况：一个二维空间，第一维度表示“领土面积”，第二维度表示“人口数量”

- 每个词汇（token）编码成二维向量，可以看出不同词汇之间存在的语义特点
- 向量之间也存在线性运算：两词向量的差往往代表“语义之间存在的差别” $\Rightarrow \vec{Italy} - \vec{German} = \vec{Mussolini} - ?$

GPT-3中，使用2048维语义空间，即2048维词向量进行token编码

## 如何进行词汇编码？

一个语言学规律：一个词本身的词义可以通过上下文的词汇，而非自身的特征定义。  
我们可以设计一种机制，使得原始的词向量通过综合上下文的信息以丰富词义。

原始词向量：一系列随机向量

词义填充的步骤：

- 文本中的每个词原始词向量为 $\vec{x}_i$
- 每个词向量：通过三个矩阵 $W^Q, W^K, W^V$ ，分别得到 $\vec{q}_i, \vec{k}_i, \vec{v}_i$ 
  - 这里 $W^Q, W^K$ 矩阵的每一列代表：对于第 $i$ 语义维度，单位语义强度所具备的“问题”编码、“特征”编码、“返回信息”编码
  - 例如：语义空间第 $i$ 维代表“词性是动词”。 $W_Q$ 该列向量代表：“这个词是不是动词”这一信息的编码， $W_K$ 该列向量代表：“这个词有动词词性特征”这一信息的编码， $W_V$ 该列向量代表：“这个词是动词时返回的信息”编码。如果语义空间其他维度代表“此行是名词”、“词语表示这是一个人名”、“词语表示某种颜色”……同理。
- 对于第 $i$ 位置的token，使用 $\vec{q}_i$ 与其他位置的 $\vec{k}_j$ 进行点积得到“吻合程度”，作为 $j$ 位置“返回信息”的权重 $\Rightarrow \vec{q}_i K^T \Rightarrow Normalization(\vec{q}_i K^T) \Rightarrow Normalization(\vec{q}_i K^T) V$
- 对于所有位置： $Normalization(QK^T)V$

Example: "The curious cat quietly climbed the tree to chase a butterfly."

## 注意力机制

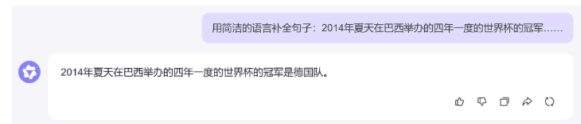
补全句子：2014 年夏天在巴西举办的四年一度的世界杯的冠军是( )



$\Rightarrow$ 2014 年夏天在巴西举办的四年一度的世界杯的冠军是( ) 我们该如何“注意”到上文的重要信息？

## 注意力机制是怎么实现的？

补全句子：2014 年夏天在巴西举办的四年一度的世界杯的冠军是( )



对于待填的词：

- 提出问题：这个词的词性如何？指代实际含义还是表示语气？词义由前文哪些线索推断？……
- 回看上文：上文每一个词都有什么样的特征（词性？词义？）？这个词的性质和我的问题吻合度多高？…… $\Rightarrow$  以此找到吻合度高的“关键词”
- 总结信息：这些高吻合度的词综合起来，为待填词提供了什么信息？

## 注意力机制是怎么实现的？

补全句子：2014 年夏天在巴西举办的四年一度的世界杯的冠军是( )

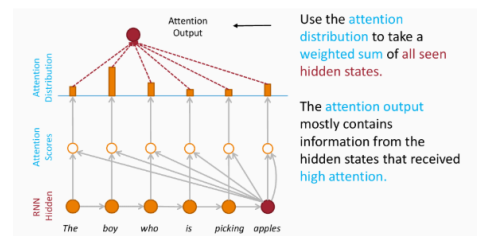
对于待填的词：

- 提出问题：这个词的词性如何——由每个词向量 $\vec{x}_i$ 推断出向量 $\vec{q}_i$ ，代表“问题”的编码
- 回看上文：上文每一个词都有什么样的特征——由每个词向量 $\vec{x}_i$ 推断出向量 $\vec{k}_i$ ，代表“特征”的编码
- 根据当前词的 $\vec{q}_i$ 和之前词的 $\vec{k}_j$ 找到吻合度高的“关键词”
- 总结信息：由每个词向量 $\vec{x}_i$ 推断出向量 $\vec{v}_i$ ，所有词的 $\vec{v}_i$ 根据“吻合度”大小加权累加即为“注意力的总结信息”，作为信息参考

Q：一个 $\vec{q}_i$ 够不够？

## 注意力机制

对于当前待生成的词：



通过注意力权重选择性、有差别地吸收上文的信息，为下一词生成提供参考