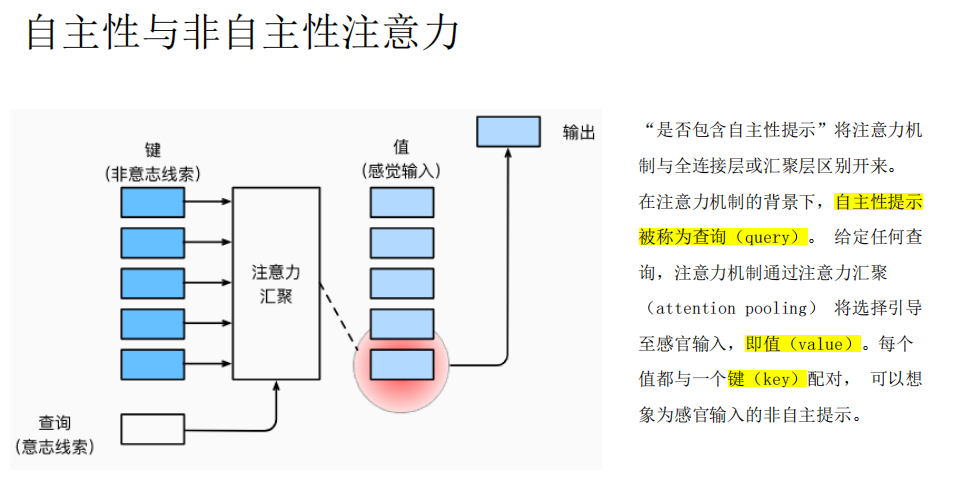
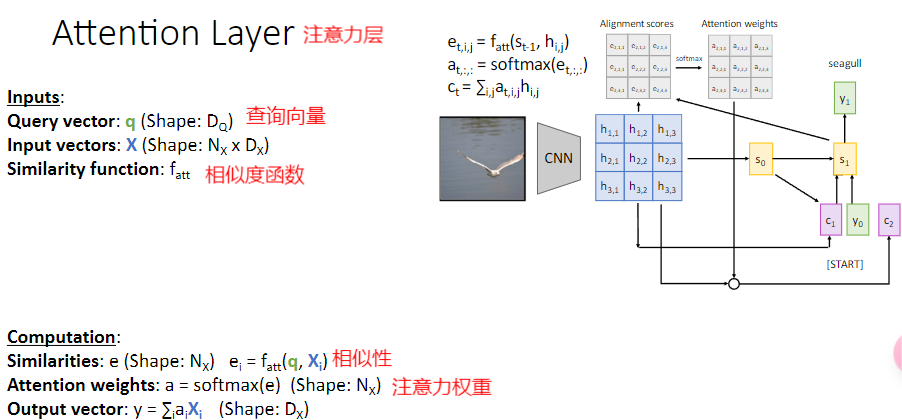
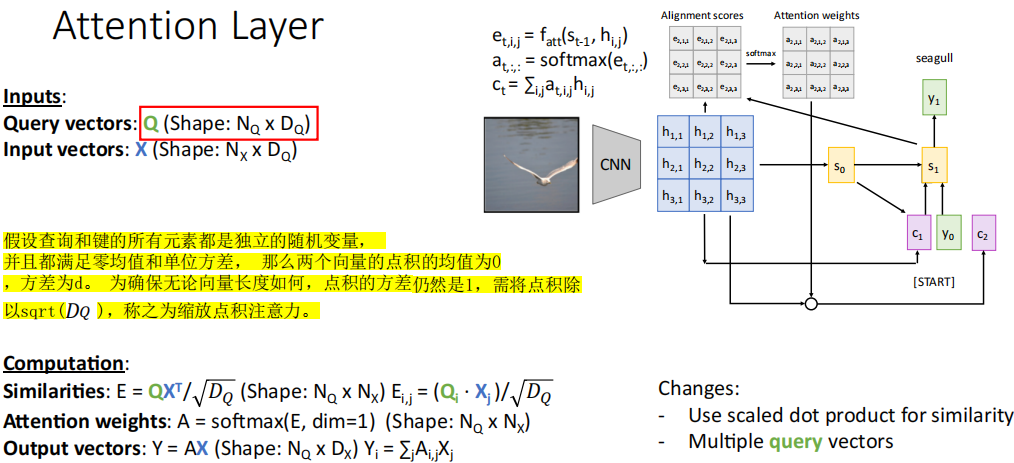
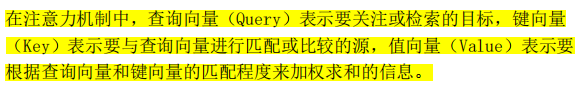
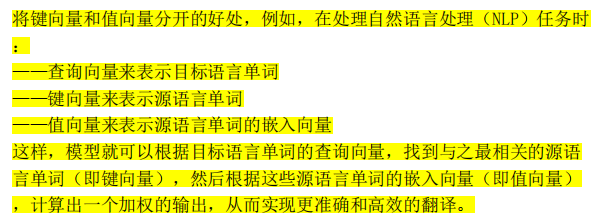
注意力机制

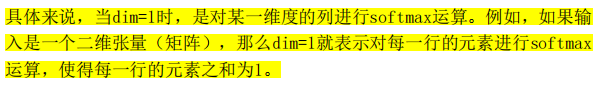






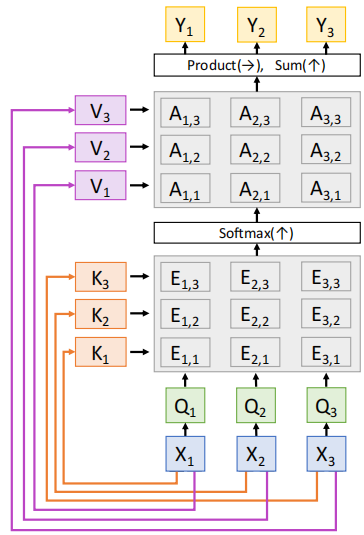
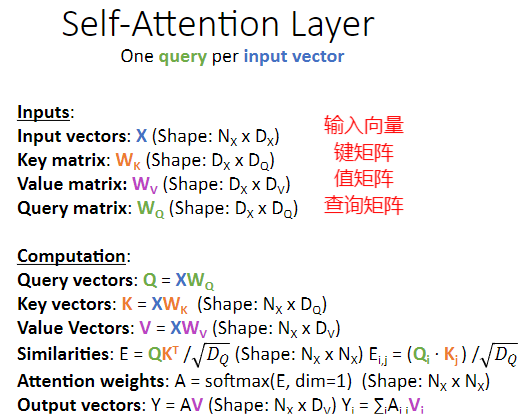












自注意力层具有置换同变性。置换同变性：改变输入数据顺序，输出顺序也会改变。

自注意层主要作用于向量集合。

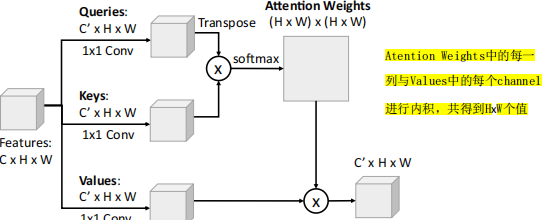
自注意力机制不“了解”它正在处理的向量的顺序。为了使处理位置感知，可以将位置编码与输入进行拼接或相加。

遮蔽自注意层（Masked Self-Attention Layer）：某些文本序列被填充了没有意义的特殊词元。为了仅将有意义的词元作为值来获取注意力汇聚，可以指定一个有效序列长度，以便在计算softmax时过滤掉超出指定范围的位置。

多头注意力机制（Multi-Head Attention）:是一种在Transformer模型中被广泛采用的注意力机制扩展形式。其基本思想是通过并行地运行多个独立的注意力机制来获取输入序列的不同子空间的注意力分布，从而更全面地捕获序列中潜在的多种语义关联。

将输入的向量映射到多个不同的子空间，多头注意力并行学习，发现不同语义对应的信息，从而提取更多的特征。

CNN with self-attention



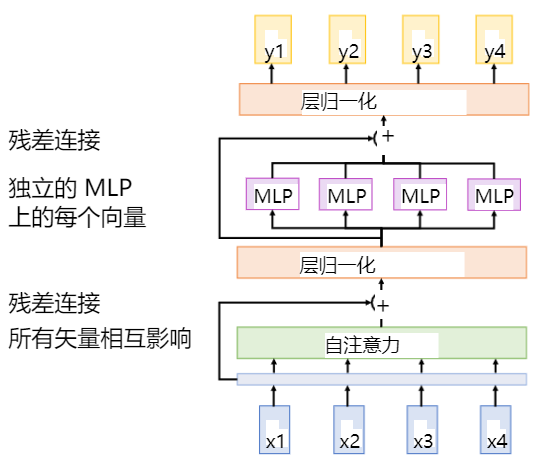
处理序列的三种方法：

循环神经网络（有序序列）、一维卷积（多维网格）、自注意力机制（向量集）

RNN：适用于有序序列 (+) 适用于长序列：在一个 RNN 层之后，就能 "看到 "整个序列 (-) 不可并行：需要按顺序计算隐藏状态

一维卷积：在多维网格上工作 (-) 不擅长长序列：高度并行：每个输出可并行计算

自注意力：适用于向量集 (-) 擅于处理长序列：在一个自我层之后，每个输出都能 "看到 "所有输入！(+) 高度并行：每个输出都可以并行计算 (-) 非常耗费内存



The Transformer

Transformer块：

输入：向量集合x

输出：向量集合y

自注意力是向量之间唯一的交互！

层归一化和MLP独立地作用于每个向量

高度可扩展，高度可并行化

Transformer：迁移学习

预训练：从互联网下载大量文本数据；使用巨型Transformer模型进行语言建模训练

微调：在自己的自然语言处理任务上对Transformer进行微调