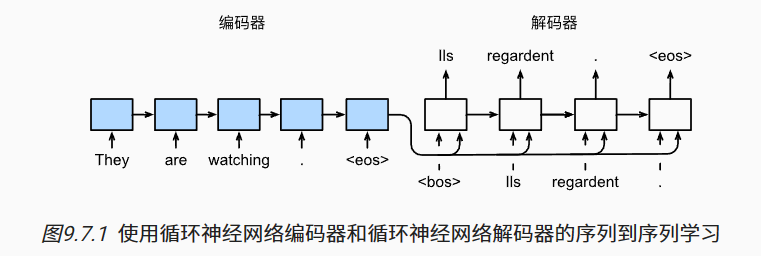
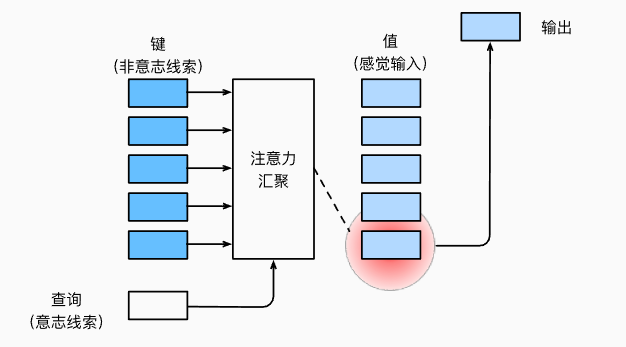
**Seq2Seq:**

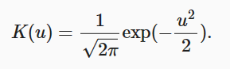
遵循编码器－解码器架构的设计原则，循环神经网络编码器使用长度可变的序列作为输入，将其转换为固定形状的隐状态。 换言之，输入序列的信息被编码到循环神经网络编码器的隐状态中。为了连续生成输出序列的词元，独立的循环神经网络解码器是基于输入序列的编码信息 和输出序列已经看见的或者生成的词元来预测下一个词元。

从源序列词元“<bos>”“Ils”“regardent”“.” 到新序列词元 “Ils”“regardent”“.”“<eos>”来移动预测的位置。

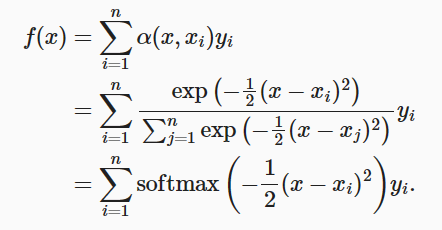
注意力机制（Attention Mechanism）：

“是否包含自主性提示”将注意力机制与全连接层或汇聚层区别开来。 在注意力机制的背景下，自主性提示被称为查询（query）。 给定任何查询，注意力机制通过注意力汇聚（attention pooling） 将选择引导至感官输入（sensory inputs，例如中间特征表示）。 在注意力机制中，这些感官输入被称为值（value）。 更通俗的解释，每个值都与一个键（key）配对， 这可以想象为感官输入的非自主提示。

根据输入位置对输出进行加权，下面考虑一个高斯核：



Nadaraya-Watson核回归：



注意力各种分类和原理，见

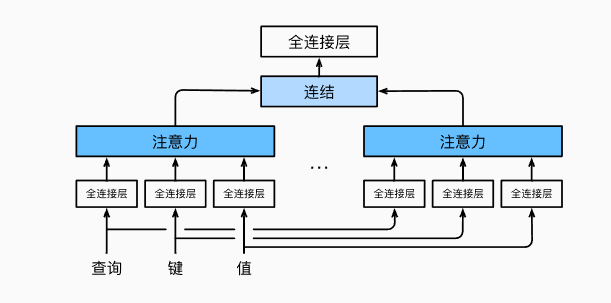
Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai

Zhang, Ralph R. Martin, Ming-Ming Cheng, Senior Member, IEEE, Shi-Min Hu, Senior Member, IEEE,

Attention Mechanisms in Computer Vision:A Survey

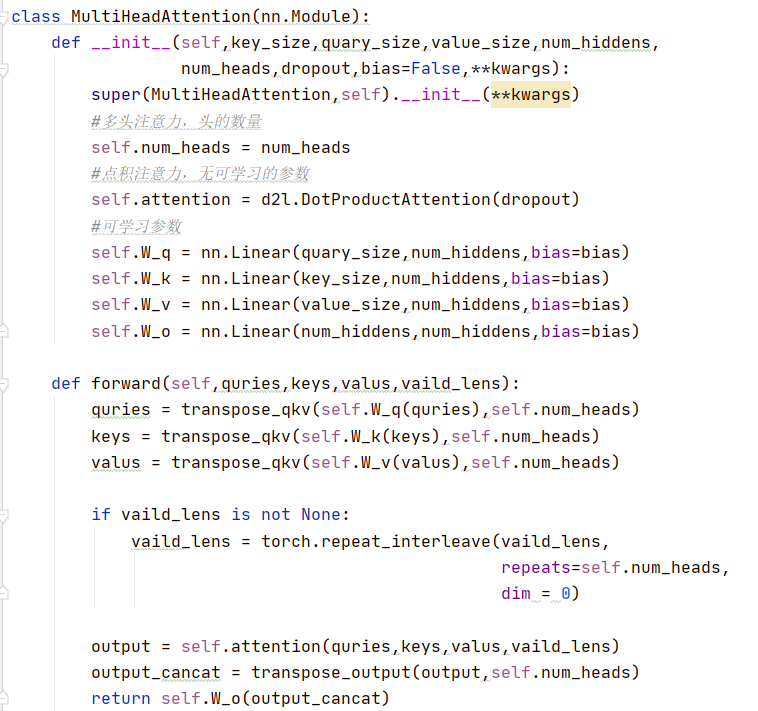
已做文献笔记

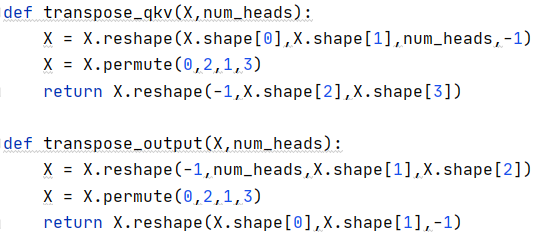
多头注意力：

在实践中，当给定相同的查询、键和值的集合时， 我们希望模型可以基于相同的注意力机制学习到不同的行为， 然后将不同的行为作为知识组合起来， 捕获序列内各种范围的依赖关系 （例如，短距离依赖和长距离依赖关系）。 因此，允许注意力机制组合使用查询、键和值的不同 子空间表示（representation subspaces）可能是有益的。

为此，与其只使用单独一个注意力汇聚， 我们可以用独立学习得到的h组不同的 线性投影（linear projections）来变换查询、键和值。 然后，这h组变换后的查询、键和值将并行地送到注意力汇聚中。 最后，将这h个注意力汇聚的输出拼接在一起， 并且通过另一个可以学习的线性投影进行变换， 以产生最终输出。 这种设计被称为多头注意力（multihead attention） (Vaswani et al., 2017)。 对于h个注意力汇聚输出，每一个注意力汇聚都被称作一个头（head）。

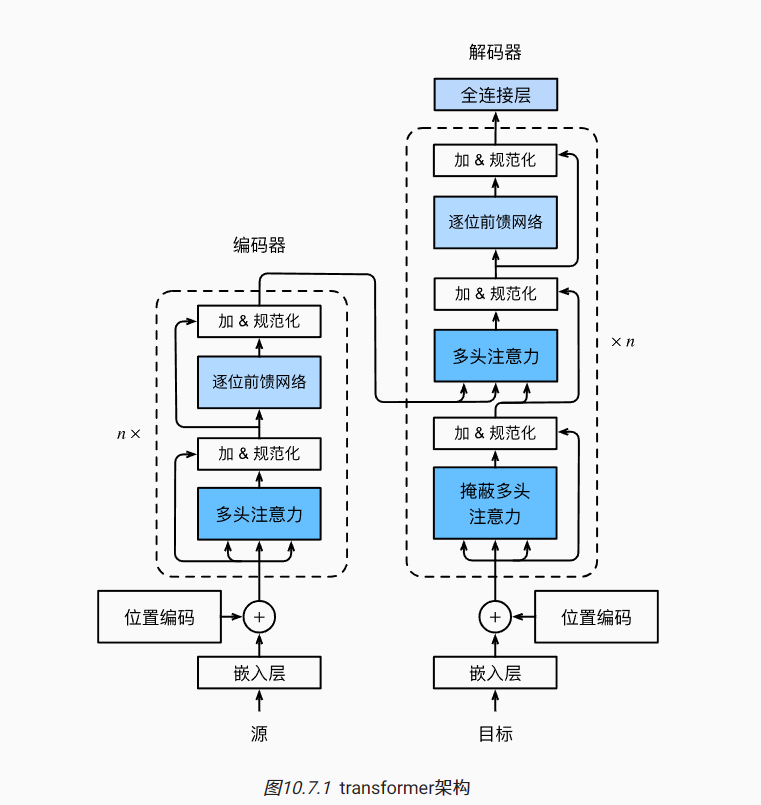
代码：



这个变换的意义在于，提高并行度，不要loop做h个头

此输入X和输出是同一个形状

Transformer:

从宏观角度来看，Transformer的编码器是由多个相同的层叠加而成的，每个层都有两个子层（子层表示为

）。第一个子层是多头自注意力（multi-head self-attention）汇聚；第二个子层是基于位置的前馈网络（positionwise feed-forward network）。具体来说，在计算编码器的自注意力时，查询、键和值都来自前一个编码器层的输出。受残差网络的启发，每个子层都采用了残差连接（residual connection）。