图 3-16 和式（3-64）展示了自注意力的具体计算方法，其中Q、K和V表示同一个输入向量。

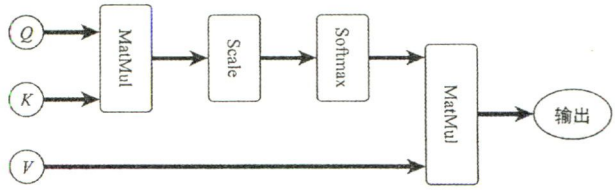

图 3-16　多头注意力机制中的注意力计算方法

$$f(Q,K,V)=\text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{3-64}$$

直观来说，Q、K和V是同一个句子，通过向量乘积获得词元之间的关系矩阵并归一化后，作用于输入句子自身。多头注意力机制将“多头”和“自注意力”两种方法组合，其计算过程如图 3-17 所示。

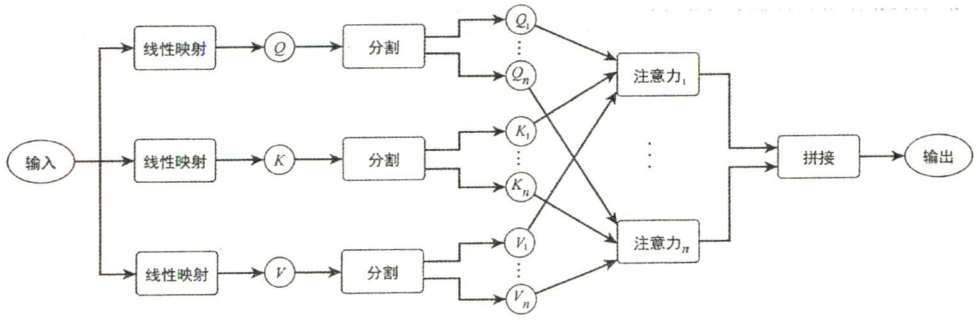

图 3-17　多头自注意力机制的计算过程

在多头注意力机制之外，为了实现语言序列中不同词元位置具备不同语义，BERT 引入了位置嵌入（Position Embedding）。事实上，位置嵌入实现了类似 RNN 或 LSTM 等网络结构的序列建模功能。相比于 LSTM，使用位置嵌入能够更好地实现 GPU 的并行计算，为大规模预训练模型实现高效训练提供必不可少的基础支撑。同时，BERT 模型精巧地设计了“上下句预测”任务，比如判断“春色满园关不住”和“一枝红杏出墙来”是合理的上下句，而“回首向来萧瑟处”和“一枝红杏出墙来”则不是合理的上下句。为了实现上下句表达，BERT 模型引入了区分上下句的片段嵌入（Segmentation Embedding）。片段嵌入除了用于上下句预测，在具体任务中还可以用于阅读理解（Reading Comprehension）、问答（Question Answering）等任务。比如在问答任务中，上下句分别表示问题和答案。如图 3-18 所示，BERT 在嵌入层对位置嵌入、