注意力机制

- 神经网络
  - 全连接神经网络FNN：每一层的神经元都和上下两层的每一个神经元完全连接
  - 卷积神经网络CNN：即训练参数量远小于全连接神经网络的卷积层来进行特征提取和学习
  - 循环神经网络RNN：能够使用历史信息作为输入、包含环和自重复的网络
- 仅将注意力集中在重点部分
  - Query（查询值）
  - Key（键值）
  - Value（真值）

  计算 Q与K的相关性为V加权求和，从而拟合序列中每个词同其他词的相关关系
- 自注意力：计算本身序列中每个元素对其他元素的注意力分布
- 多头注意力：同时对一个语料进行多次注意力计算
  - 将原始的输入序列进行多组的自注意力处理；然后再将每一组得到的自注意力结果拼接起来，再通过一个线性层进行处理，得到最终的输出
  - $\mathrm{MultiHead}(Q,K,V)=\mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_\mathrm{h})W^O$
    $\mathrm{where\ head}_\mathrm{i}=\mathrm{Attention}(QW_i^Q,KW_i^K,VW_i^V)$
- Key 与 Query 相关性越高，则其所应该赋予的注意力权重就越大
  - $attention(Q,K,V)=softmax(\frac{QK^T}{\sqrt{d_k}})V$
  - 4 再次放缩保证稳定性
  - 3 注意力分数*值向量得到标量
  - 2 放缩权重得到注意力分数
  - 1 点积度量词向量相似性
- 掩码自注意力：让模型只能使用历史信息进行预测而不能看到未来信息
  - 和文本序列等长的上三角矩阵