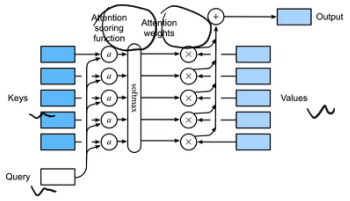
注意力分数



$f(x)=\sum_i \alpha(x,x_i)y_i=\sum_i softmax(-\frac{1}{2}(x-x_i)^2)y_i$

注意力权重　　注意力分数

- **拓展到高纬度**
  - 假设query　$q\in R^q$ m对key-value(ki, vi)　$k_i\in R^k,\ v_i\in R^v$
  - 注意力池化层:

$$f(q,(k_1,v_1),\dots(k_m,v_m))=\sum_{i=1}^{m}\alpha(q,k_i)v_i\ ,\quad \in R^v$$

$$\alpha(q,k_i)=softmax(a(q,k_i))=\frac{\exp(a(q,k_i))}{\sum_j \exp(a(q,k_j))}\ ,\quad \in R$$

注意力分数

a 函数要如何设计？

- **Additive Attention**
  - 可学参数:　$W_k\in R^{h\times k},\ W_q\in R^{h\times q},\ v\in R^h$

$$a(k,q)=v^T\cdot\tanh(W_k\cdot k+W_q\cdot q)\quad \in R^h$$

  - 等价于将key和query合并起来后放入一个隐藏大小为h，输出大小为1的单隐藏层MLP
- **Scaled Dot-Product Attention**
  - 若q和k都是同样长度d，那么可以　$a(q,k_i)=\langle q,k_i\rangle/\sqrt{d}$ → 对长度不敏感
  - 向量化版本:
    - $Q\in R^{n\times d},\ k\in R^{m\times d},\ V\in R^{m\times v}$
    - 注意力分数:　$a(Q,K)=QK^T/\sqrt{d},\ \in R^{n\times m}$
    - 注意力池化:　$f=Softmax(a(Q,K))V,\ \in R^{n\times v}$
- **总结**
  - 注意力分数是query和key的相似度，注意力权重是分数softmax的结果
  - 两种常见的分数计算：query和key合并放入单隐层MLP；query和key直接内积