13/10/23

Deep Learning

- Recap: Attention and Convolution
- Transformer

- Recap: $J(i,j) = \sum_c \sum_m \sum_n I(i+m, j-n, c) \cdot h(m,n,c)$ (Convolution)

$$J(i,j) = \sum_{a,b \in \mathcal{N}(i,j)} \text{softmax}_{ab}\left(\langle q(i,j), k(a,b)\rangle\right) v(a,b) \quad \text{(Attention)}$$

Self Attention:

$$q(i,j) = W_q\, x(i,j)$$
$$k(i,j) = W_k\, x(i,j)$$
$$v(i,j) = W_v\, x(i,j)$$

$x(i,j) \in \mathbb{R}^{1\times 1\times c}$

$W_q \in \mathbb{R}^{d\times c}$

$W_k \in \mathbb{R}^{d\times c}$

$W_v \in \mathbb{R}^{d\times c}$

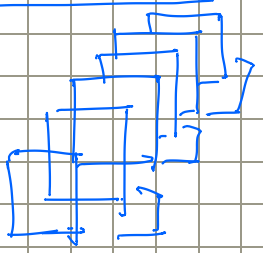
Note: If $W_q = W_k = W$,

- $\langle W x(i,j), W x(a,b)\rangle$
  $= (W x(i,j))^T \cdot W x(a,b)$,
- we would have a high value of the inner product at $(a,b) = (i,j)$
- This in turn would mean that $v(i,j)$ is always going to be assigned a high weight
- This is not ideal.

∴ We have W_q and W_k that are not the same.

- Multi Head Attention (MHA): In MHA, $c' = c/h$ where h is the number of heads

Ex:



$c = 6,\ h = 3 \Rightarrow c' = 6/3 = 2$

- We now work with c' channels per head.
- Suppose we have d' to be the output dimension, $\tilde{d} = h \cdot d'$

  i.e. the outputs of each head are concatenated. If $\tilde{d} \neq d$

  a projection matrix $W_0 \in \mathbb{R}^{d\times\tilde{d}}$ is applied

- The Transformer Architecture