12/10/23

Deep learning

- Introduction to Attention

Note: References posted on classroom

  - Recall Convolution
  - Attention mechanism

} Ramachandran et al NeurIPS 2019

- Transformer model - Vaswani et al. NeurIPS 2017

- Recall Convolution / Correlation

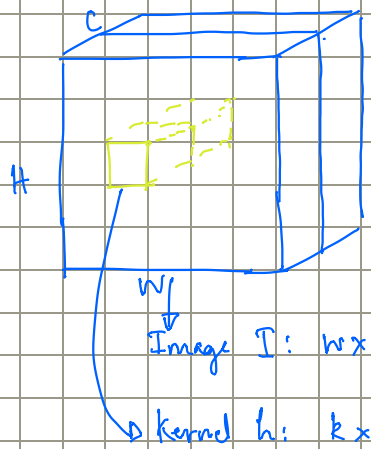
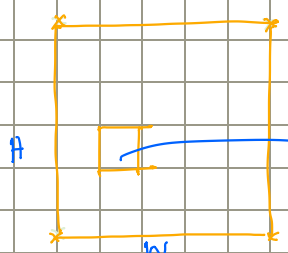


Image I: $W \times H \times C$

Kernel h: $k \times k \times C$



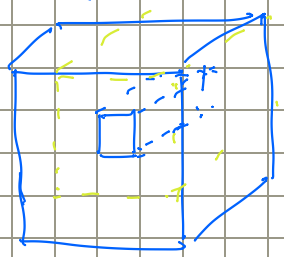
J: Activation map of size $W \times H$

Convolution Sum:

$$J(i,j) = \sum_{c} \sum_{m} \sum_{n} I(i-m, j-n, c) \cdot h(m,n,c)$$

W: width
H: height
C: channels
k: kernel width/height

- Attention (Self Attention):



$I: W \times H \times C$

$I(i,j) \in 1 \times 1 \times C$

$x(i,j) = I(i,j)$ (Input to the attention mechanism)

Note: we are dealing with vectors in the definitions

In the Self-attention mechanism, we derive three elements - query, key and value from the same input

query: $q(i,j) = W_Q \, x(i,j)$ $\quad W_Q \in \mathbb{R}^{d \times C}$

key: $k(i,j) = W_K \, x(i,j)$ $\quad W_K \in \mathbb{R}^{d \times C}$

value: $v(i,j) = W_V \, x(i,j)$ $\quad W_V \in \mathbb{R}^{d \times C}$

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

Convolution: $J(i,j) = \sum_{C} \sum_{m} \sum_{n} I(i-m, j-n, C) \cdot h(m,n,C)$