Attention (self/cross, hard/soft)

Dealing with sets

$$\boldsymbol{h} = \boldsymbol{X}\boldsymbol{a}$$

Self-attention (I)

$$\{\boldsymbol{x}_i\}_{i=1}^{t} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_t\} \rightsquigarrow \boldsymbol{X} \in \mathbb{R}^{n \times t}, \quad \boldsymbol{x}_i \in \mathbb{R}^n$$

$$\boldsymbol{h} = \alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2 + \cdots + \alpha_t \boldsymbol{x}_t = \boldsymbol{X}\boldsymbol{a} \in \mathbb{R}^n$$

$$\alpha_i > 0$$

$$\boldsymbol{X} \doteq \begin{bmatrix} | & | & & | \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_t \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times t}$$

soft attention: $\boldsymbol{a}^\top \mathbf{1} = 1$

hard attention: $\boldsymbol{a} \in \mathbb{I}_t$

$$\boldsymbol{h} = \boldsymbol{X}\boldsymbol{a}$$

Self-attention (II)

$$\boldsymbol{a} = \mathrm{softargmax}_{\beta}(\boldsymbol{X}^{\top}\boldsymbol{x}) \in \mathbb{R}^{t}$$

$$\{\boldsymbol{x}_i\}_{i=1}^{t} \rightsquigarrow \{\boldsymbol{a}_i\}_{i=1}^{t} \rightsquigarrow \boldsymbol{A} \in \mathbb{R}^{t \times t}$$

$$\{\boldsymbol{a}_i\}_{i=1}^{t} \rightsquigarrow \{\boldsymbol{h}_i\}_{i=1}^{t} \rightsquigarrow \boldsymbol{H} \in \mathbb{R}^{n \times t}$$

$$\boldsymbol{H} = \boldsymbol{X}\boldsymbol{A} \in \mathbb{R}^{n \times t}$$

Key-value store

- Paradigm for
  - storing (saving)
  - retrieving (querying)
  - managing

  an associative array (dictionary / hash table)

Queries, keys, and values

$$\{\boldsymbol{q}_i\}_{i=1}^t \rightsquigarrow \boldsymbol{Q} \in \mathbb{R}^{d' \times t}$$

$$\boldsymbol{q} = \boldsymbol{W}_{\boldsymbol{q}}\boldsymbol{x}, \quad \boldsymbol{k} = \boldsymbol{W}_{\boldsymbol{k}}\boldsymbol{x}, \quad \boldsymbol{v} = \boldsymbol{W}_{\boldsymbol{v}}\boldsymbol{x}$$

$$\beta = \frac{1}{\sqrt{d'}}$$

$$\boldsymbol{q}, \boldsymbol{k} \in \mathbb{R}^{d'}, \quad \boldsymbol{v} \in \mathbb{R}^{d''}$$

$$\{\boldsymbol{x}_i\}_{i=1}^t \rightsquigarrow \{\boldsymbol{q}_i\}_{i=1}^t, \{\boldsymbol{k}_i\}_{i=1}^t, \{\boldsymbol{v}_i\}_{i=1}^t \rightsquigarrow \boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$$

$$\boldsymbol{a} = \mathrm{softargmax}_{\beta}(\boldsymbol{K}^\top \boldsymbol{q}) \in \mathbb{R}^t$$

$$\boldsymbol{h} = \boldsymbol{V}\boldsymbol{a} \in \mathbb{R}^{d''}$$

$$\{\boldsymbol{q}_i\}_{i=1}^t \rightsquigarrow \{\boldsymbol{a}_i\}_{i=1}^t \rightsquigarrow \boldsymbol{A} \in \mathbb{R}^{t \times t}$$

$$\boldsymbol{H} = \boldsymbol{V}\boldsymbol{A} \in \mathbb{R}^{d'' \times t}$$