Attention (self/cross, hard/soft)

Dealing with sets

Self-attention (I)

$$\{oldsymbol{x}_i\}_{i=1}^t = \{oldsymbol{x}_1, oldsymbol{x}_2, \cdots oldsymbol{x}_t\} \sim oldsymbol{X} \in \mathbb{R}^{n imes t}, \quad oldsymbol{x}_i \in \mathbb{R}^n$$
 $egin{align*} oldsymbol{h} & oldsymbol{x}_1 & oldsymbol{x}_2 & oldsymbol{x}_1 & oldsymbol{x}_2 & oldsymbol{x}_1 & oldsymbol{x}_2 & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_2 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_t \\ oldsymbol{x} & oldsymbol{x}_2 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_2 \\ oldsymbol{x} & oldsymbol{x}_2 & \cdots & oldsymbol{x}_2 & \cdots & oldsymbol{x}_2 \\ & oldsymbol{x}_2 & \cdots & oldsymbol{x}_2 & \cdots & oldsymbol{x}_2 \\ & oldsymbol{x} & oldsymbol{x}_2 & \cdots & oldsymbol{x}_2 \\ & oldsymbol{x} & oldsymbol{x}_2 & \cdots & oldsymbol{x}_2 & \cdots & oldsymbol{x}_2 \\ & oldsymbol{x} & oldsymbol{x}_2 & \cdots & oldsymbol{x}_2$

Self-attention (II)

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

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

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

$$oldsymbol{H} = oldsymbol{X}oldsymbol{A} \in \mathbb{R}^{n imes t}$$

Key-value store

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

an associative array (dictionary / hash table)

Queries, keys, and values

$$\{oldsymbol{q}_i\}_{i=1}^t \leadsto oldsymbol{Q} \in \mathbb{R}^{d' imes t}$$

$$oldsymbol{q} = oldsymbol{W_q} oldsymbol{x}, \quad oldsymbol{k} = oldsymbol{W_k} oldsymbol{x}, \quad oldsymbol{v} = oldsymbol{W_v} oldsymbol{x}$$

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

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

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

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

$$\{\boldsymbol{q}_i\}_{i=1}^t \leadsto \{\boldsymbol{a}_i\}_{i=1}^t \leadsto \boldsymbol{A} \in \mathbb{R}^{t \times t}$$
 $\boldsymbol{H} = \boldsymbol{V}\boldsymbol{A} \in \mathbb{R}^{d'' \times t}$