Generalizations of Softmax Attention

- Let sim$(q, k) \geq 0$ be an arbitrary function that measures similarity between the query q and key k

- Attention mechanism w.r.t sim is

- Softmax: $\mathrm{sim}(q, k) \doteq \exp(\langle q, k \rangle)$

15

$$o_j = \sum_{i \leq j} \frac{\mathrm{sim}(q_j, k_i)}{\sum_{i' \leq j} \mathrm{sim}(q_j, k_{i'})} v_i$$

$$\sum_{i \leq 7} \frac{\exp(\langle q_7, k_i \rangle)}{\sum_{i' \leq i} \exp(\langle q_7, k_{i'} \rangle)} v_i$$

Generalizations of Softmax Attention

- Let sim$(q, k) \geq 0$ be an arbitrary function that measures similarity between the query q and key k

- Attention mechanism w.r.t sim is

$$o_j = \sum_{i \leq j} \frac{\text{sim}(q_j, k_i)}{\sum_{i' \leq j} \text{sim}(q_j, k_{i'})} v_i$$

- Softmax: sim$(q, k) \doteq \exp(\langle q, k \rangle)$

$$\sum_{i \leq 7} \frac{\exp(\langle q_7, k_i \rangle)}{\sum_{i' \leq i} \exp(\langle q_7, k_{i'} \rangle)} v_i$$

Kernel View of Attention