

Generalization After Attention

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

• **Attention mechanism** **works in**

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



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

Generalizations of Softmax Attention

- Let $\text{sim}(q, k) \geq 0$ 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: $\text{sim}(q, k) \doteq \exp(\langle q, k \rangle)$

Kernel View of Attention