Attention

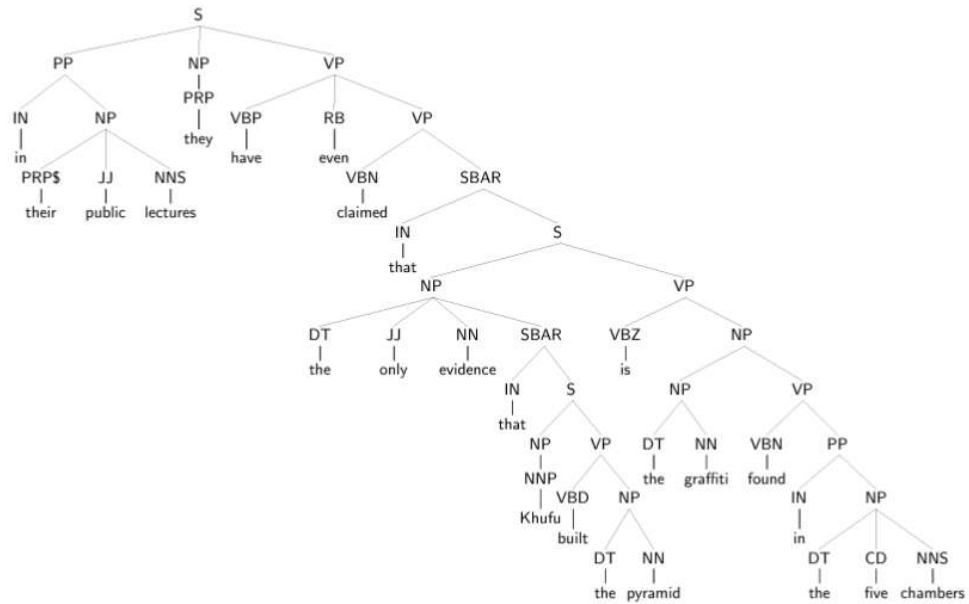
Recap: Language Modeling

Language is messy

- Syntax tree is *not* universal

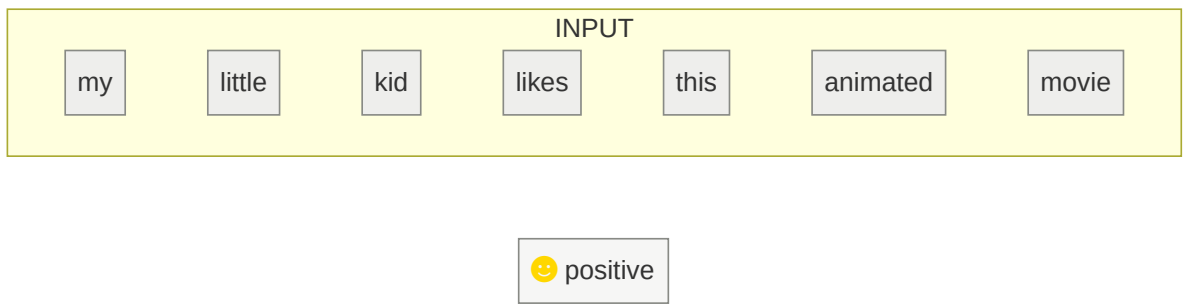
Solution:

- Treat the sentence as a flat sequence
- Use deep networks to parse language



Attention

A **set operator** that learns to reason about the structure of a set of elements



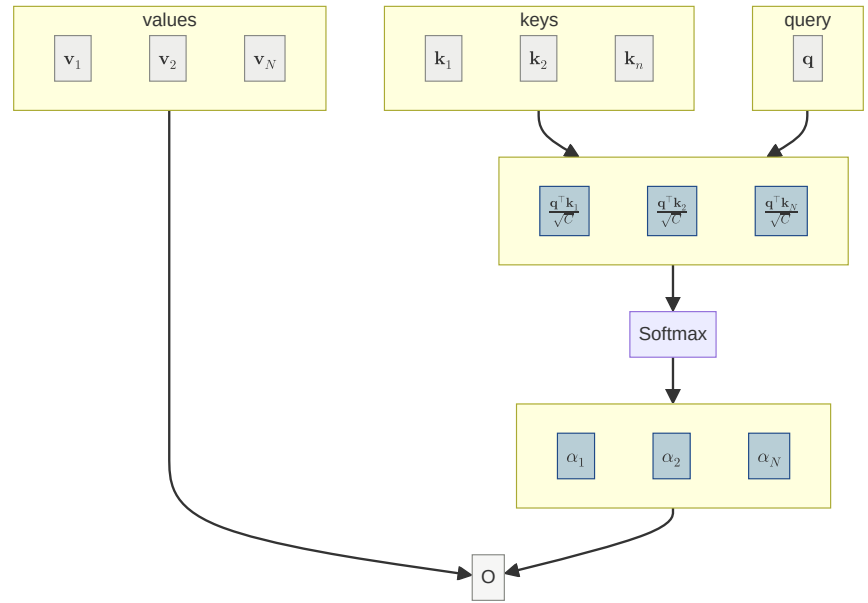
Attention

Inputs:

- query $\mathbf{q} \in \mathbb{R}^C$
- a set of keys $\mathbf{K} = \left[\mathbf{k}_1, \cdots, \mathbf{k}_N\right], \mathbf{k}_i \in \mathbb{R}^C$
- a set of values $\mathbf{V} = \left[\mathbf{v}_1, \cdots, \mathbf{v}_N\right], \mathbf{v}_i \in \mathbb{R}^C$

Output: $\mathbf{o} \in \mathbb{R}^C$

$$\mathbf{o} = \sum_i \alpha_i \mathbf{v}_i, \quad \text{where } \alpha_i = \frac{e^{\frac{\mathbf{q}^\top \mathbf{k}_i}{\sqrt{C}}}}{\sum_j e^{\frac{\mathbf{q}^\top \mathbf{k}_j}{\sqrt{C}}}}$$



Attention: Matrix Form

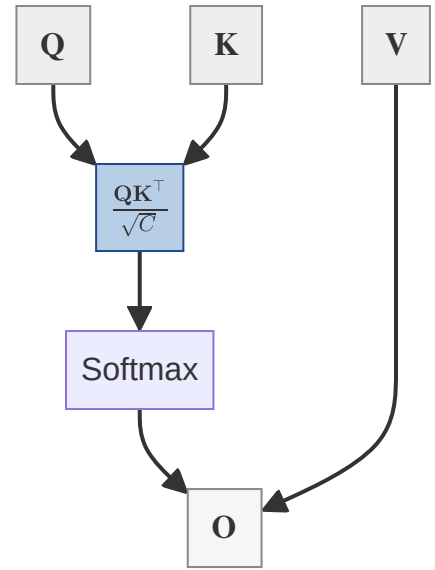
Inputs:

- queries $\mathbf{Q} \in \mathbb{R}^{M \times C}$
- keys $\mathbf{K} \in \mathbb{R}^{N \times C}$
- values $\mathbf{V} \in \mathbb{R}^{N \times C}$

Output: $\mathbf{O} \in \mathbb{R}^{M \times C}$

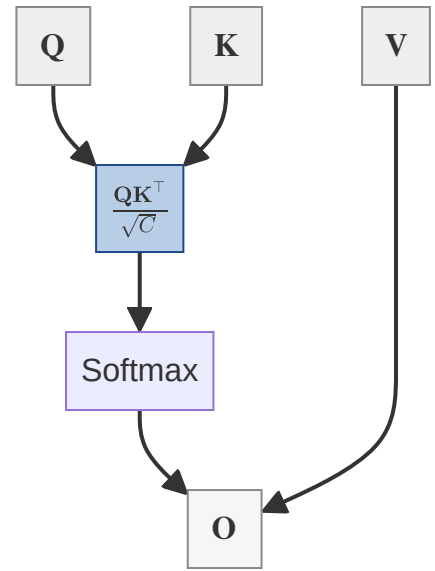
$$\mathbf{O} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{C}}\right)\mathbf{V}$$

$\text{softmax}(\cdot)$ is row-wise (each row sums to 1)



Benefits of Attention

Reasons about the interaction of variable sequence



How Does Attention Solve Language Tasks?

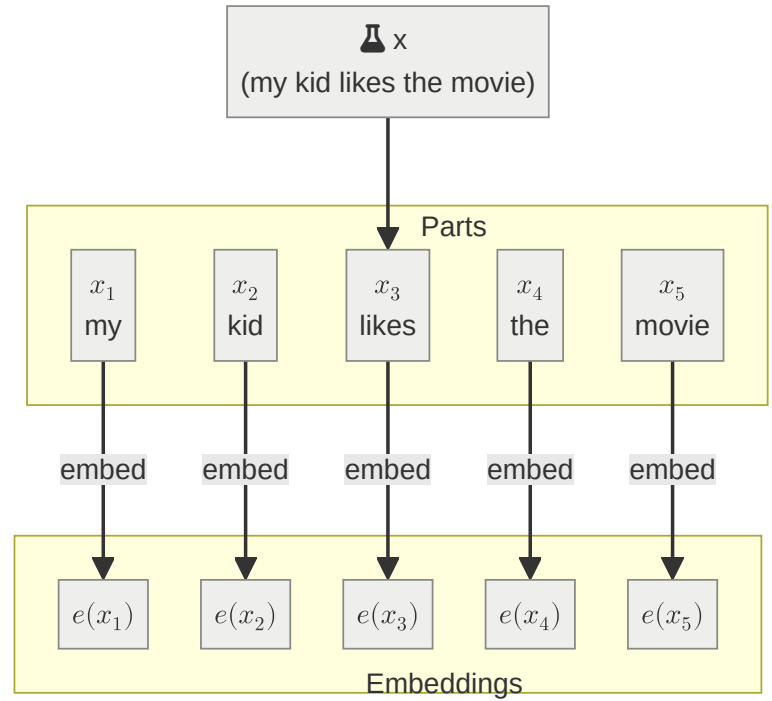
1. Split sentences into parts
   - characters
   - words
   - tokens

2. Embed each part x_i

$$x_i \to e(x_i) \in \mathbb{R}^C$$

3. Feed the embeddings to attention

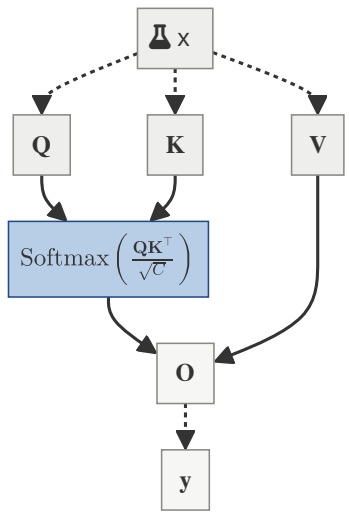
$$\mathbf{Q} = \mathbf{K} = \mathbf{V} = \{e(x_1), \cdots, e(x_N)\}$$



Self-Attention and Cross-Attention

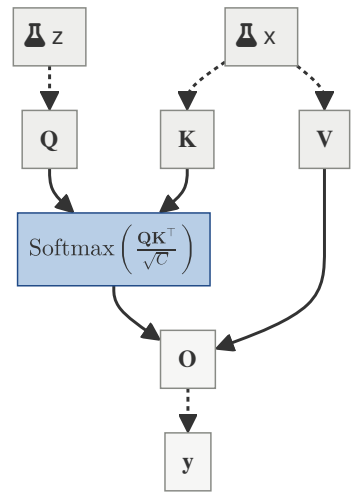
Self-Attention

- queries, keys & values from the same inputs



Cross-Attention

- keys and values come from the same inputs
- queries come from another set of inputs



Attention - TL;DR

Attention is a **set operation** which reasons about set elements

Attention takes three inputs - **queries**, **keys** and **values**