

Causal Language Modeling



I work in the city of New

1 2 3 4 5 6 7

Querye

q_1 q_2 q_3 q_4 q_5 q_6 q_7

Keys

k_1 k_2 k_3 k_4 k_5 k_6 k_7

Values

v_1 v_2 v_3 v_4 v_5 v_6 v_7

Task: Compute a representation of the context to be able to predict the next word





R *h*

Answer: York

Causal Language Modeling

I work in the city of New _____ Answer: York

1 2 3 4 5 6 7

Query	q_1	q_2	q_3	q_4	q_5	q_6	q_7
Keys	k_1	k_2	k_3	k_4	k_5	k_6	k_7
Values	v_1	v_2	v_3	v_4	v_5	v_6	v_7

} \mathbb{R}^h

Task: Compute a representation of the context to be able to predict the next word

Attention Mechanism

- Representation of the context:

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