Attention (p-5) (LLM notes)

p-57

- in the original paper: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

- $Q = \underset{\text{query}}{[q_1\ q_2 \cdots q_n]} \cdot K = \underset{\text{key}}{[k_1\ k_2 \cdots k_n]}$ $\cdot\ QK^T =$

	Q_1	Q_2 ... Q_n
K_1	$K_1 Q_1$	$K_1 Q_2$
K_2	$K_2 Q_1$	$K_2 Q_2$
$\vdots$		
K_n		

} Attention pattern

- a helpful thing is to devide all dot products in Attention patter by $\sqrt{d_k}$ the sqrt of dimention of key query space

	Q_1	Q_2 .
K_1	$\frac{K_1 Q_1}{\sqrt{d_k}}$	$\frac{K_2 Q_2}{\sqrt{d_k}}$ . . .
K_2	.	. . .

- Softmax (applied column by column) is to normalize values into prob dist.

$\begin{bmatrix} 30 \\ 20 \\ 10 \end{bmatrix} \rightarrow \text{Softmax} \rightarrow \begin{bmatrix} 0.9 \\ .2 \\ .1 \\ \vdots \\ . \end{bmatrix}$

- in the training process when the weights updated, to either reward or punish it based on how high a prob it assigns to true next word in the passage it turns out to make the training more efficient you can have it simultaniously predict every possible next token following each initial subsequence of tokens in the passage. This makes it so a single training Ex is Now a lot.

the → ??

the fluffy ?? → {dog 0.1, cat 0.1, ..., blue 0.5} → {creature 0.1, animal 0.5, ..., doll 0}

the fluffy blue ?? → ??

the fluffy blue creature.

- for our attention pattern this means never allow later words to influence earlier words since they can give away the ans for what comes next

- this means all the dot products in attention pattern with **red dot** • the ones representing later tokens influencing earlier ones, to somehow be 0 (ignored)

we cannot set them to 0 as the cols would not add to 1, so instead before softmax you set all of those entries to $-\infty$ now after softmax the cols add to 1 and range $[0, 1)$ this is called **Masking**

Unnormalized attention pattern

3.5	0.9	1.9	4.4
$-\infty$	-0.3	-0.2	0.8
$-\infty$	$-\infty$	0.89	0.67
$-\infty$	$-\infty$	$-\infty$	1.31

Softmax →

1	0.25	0.69	0.92
0	0.25	0.08	0.02
0	0	0.24	0.02
0	0	0	0.04