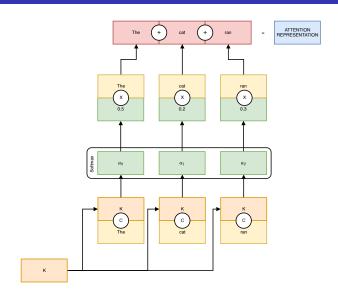
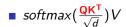
Lab class Computational Semantics 2021

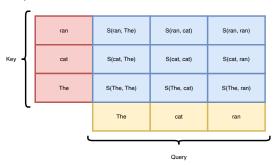
Adam Ek

Attention



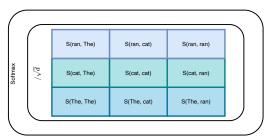
Self-Attention: A visual explanation





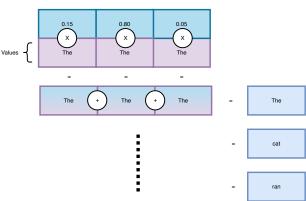
Normalization

• softmax $(\frac{QK^T}{\sqrt{\mathbf{d}}})V$

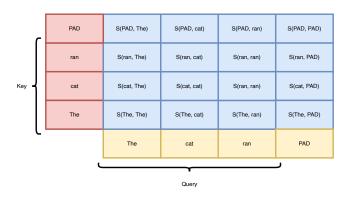


Getting word representations

• $softmax(\frac{QK^T}{\sqrt{d}})V$



Padding



Attention mask

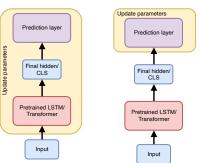
S(P)	AD, The)	S(PAD, cat)	S(PAD, ran)	S(PAD, PAD)	
S(ra	an, The)	S(ran, cat)	S(ran, ran)	S(ran, PAD)	
S(c	at, The)	S(cat, cat)	S(ran, ran)	S(cat, PAD)	
S(T	he, The)	S(The, cat)	S(The, ran)	S(The, PAD)	



0	0	0	0
1	1	1	0
1	1	1	0
1	1	1	0

About fine-tuning

- Fine-tuning: We update the parameters of a pretrained model to adapt it to new data (a new task, a new domain, ...)
- We do not finetune if we add a prediction layer on top of a model and only update its parameters.



Building neural nets: some mental tips

- Always keep in mind:
 - What's the goal?
 - What's the expected input?
 - 3 What's the expected output?
- How we go from input to output is in 93.333...% of the cases conventions or what people previously have found to work.
 - As long as you can motivate the transition from input to output, you're good!
- When debugging: focus on the cause not the symptoms

Questions???

