Attention

- “Attention is all you need”: title of the Transformer paper by Vaswani et al.
- Key mechanism for accessing relevant information in a context to make predictions
- This segment: example of how attention can impact language modeling

Attention: Running Example

- Fixed-length sequence of As and Bs

AAAAAAA

ABAAAAB

ABAABAB

AAAABAB

BAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAB

- All As = last letter is A; any B = last letter is B
- **Attention**: method to access arbitrarily* far back in context from this point
- RNNs generally struggle with this; remembering context for many positions is hard (though of course they can do this simplified example — you can even hand-write weights to do it!)

Keys and Query

Attention

Attention

keys k_i

[1, 0] [1, 0] [0, 1] [1, 0]

A A B A

query: q = [0, 1] (we want to find Bs)

We can make attention more peaked by amplifying the embeddings

$k_i = W^K e_i$ $W^K = \begin{matrix} 10 & 0 \\ 0 & 10 \end{matrix}$ [10, 0][10, 0][0, 10][10, 0]

0 0 1 0

What will new attention values be with these keys?

Attention, Formally

- Original “dot product” attention: $s_i = k_i^T q$
- Scaled dot product attention: $s_i = k_i^T W q$
- Equivalent to having two weight matrices: $s_i = (W^K k_i)^T (W^Q q)$

- Other forms exist: Luong et al. (2015), Bahdanau et al. (2014) present some variants (originally for machine translation)
- We will see that the real attention computation actually has three matrices (one for values as well); we’ll come back to this later