Attn

\[ \text{out} = \sum_i \text{softmax} \left( \frac{q_i^T k_i}{\sqrt{d}} \right) v_i \]

\[ \text{out} = \frac{\text{softmax} \left( \frac{q^T k}{\sqrt{d}} \right)}{L} v \]

Self-Attn:

- \( k_i = W_k s_i \)
- \( q_i = W_q s_i \)
- \( v_i = W_v s_i \)

X-Attn: q come from a different seq.

Transformers

Encoder

Decoder

\[ \text{FCN} \]

\[ \text{Self-Attn} \]

Input seq

Output seq

BERT

GPT
Not the best perf

New vocab items

Embedding

Are birds mammals?

Transformers

Fail