Self-attention:

- training time: parallelizable
- test time: sequentially

Training:

\[
\begin{array}{cccc}
q_1 & q_2 & q_3 \\
1 & 1 & 1 \\
\end{array}
\times
\begin{array}{cccc}
k_1 & k_2 & k_3 \\
1 & 1 & 1 \\
\end{array}
= \begin{array}{cccc}
q_1 & q_2 & q_3 \\
1 & 1 & 1 \\
\end{array}
\]

- we can only parallelize this computation because we know the full sequence ahead of time.

Test time: \[\frac{f_1}{f_1} f_2, k_2, q_2, q_2\]

My dog ?

\[
\text{Softmax}(\langle f_2, k_1, q_2, k_2 \rangle)
\]

Weighted sum of value vectors

\[
\text{dim}_d, \text{hyperparam}
\]
Decoding:

- dim of value vector \( a \)  
- \( \hat{o} = \text{softmax}(W \hat{x}_2) \)

Select a word from the pred. dist
- greedy
- Sampling

- my dog walked

\( \text{Softmax} \left( \langle q_3, k_1, k_2, 1, k_3 \rangle \right) \)

KV cache:
- Storing previously-computed key/value vectors so you don’t have to recompute them at every timestep

\( \text{test time} \)
- always start w/ a \( \langle \text{SOS} \rangle \)
  "Start of seq"

Transformer:
- neural LM built around
  multi-headed self-attention

\( \text{stacked layers of} \ a \) "deep"

multi-headed self-attention:
- Vaswani et al. 2017
- intuition: let’s have multiple sets of \( f(k), v \) vectors for each token
  then maybe each “head” can focus on a specific linguistic property
5. Syntactic structures (all verbs in prefix)
5. activate on certain words/phrases
5. entities/dates

\[
O = \text{softmax}(W_1z_1) \quad \text{predict opened}
\]

\[
\hat{z}_1 = f(W_2(h_1^1; h_2^1)) \quad \text{predict their}
\]

\[
\hat{z}_2 = f(W_2(h_2^1; h_3^1)) \quad \text{predict books}
\]

\[
h_1 \quad h_2^1 \quad h_2 \quad h_3^1 \quad h_3 \quad h_3^2
\]

\[
\text{Self-attn} \quad \text{masked}
\]

\[
\text{head 1} \quad \text{head 2}
\]

\[
P_1 \quad \text{Students} \quad P_2 \quad \text{opened} \quad P_3 \quad \text{their}
\]

\[
q_1 \quad k_1 \quad v_1 \quad \text{attention}
\]

\[
q_2 \quad k_2 \quad v_2
\]

\[
q_3 \quad k_3 \quad v_3
\]
\[ q_1' = f(W_q' [p_1+c_1]) \]
\[ q_2' = f(W_q' [p_1+c_1]) \]
\[ f = \text{relu}, \max(0,x) \]

Adding depth:

Transforming "block"/layer

Masked multi-head self-attention
Decoder language model:

Sequence to sequence models:

- Encoder \( \rightarrow \) source lang
- Decoder \( \rightarrow \) target lang