Transformers:

- A neural LM whose component is **multi-head self-attention**

---

**multi-head self-atttn**:

- Instead of just one set of query/key/value vectors, let's have many sets (heads).

- Intuition: having multiple sets of Q, K, V can allow each head to attend to different linguistic properties of the prefix.

- n-gram windows, subject/object of sentence, discourse (global) context, entities, verbs...

---

**Single-head self-atttn**: predict opened

\[ h_1 = \text{f}(W_q c_1) \]

\[ k_1 = \text{f}(W_k c_1) \]

\[ v_1 = \text{f}(W_v c_1) \]

**Multi-head self-attention**: predict this

\[ h_2 = \text{f}(W_q c_2) \]

\[ k_2 = \text{f}(W_k c_2) \]

\[ v_2 = \text{f}(W_v c_2) \]

**Multi-head self-attention**: predict book

\[ h_3 = \text{f}(W_q c_3) \]

\[ k_3 = \text{f}(W_k c_3) \]

\[ v_3 = \text{f}(W_v c_3) \]
multi-head Self-attn

Predict Opened

$z_1 \rightarrow h'_1 \rightarrow h'_2 \rightarrow h_1 \rightarrow h_2$

Predict their

$z_2 \rightarrow h'_2 \rightarrow h_2 \rightarrow h'_3 \rightarrow h_3$

Predict book

$z_3 \rightarrow h'_3 \rightarrow (\mathbf{w} \mathbf{h_1}^1, h_1^2)$

Masked Self-attn

$\{\text{head 1}\}$

$\{\text{head 2}\}$

$c_1$: Students

$q'_1 = f(W^1 q_1 c_1)$

$q_2' = f(W^2 q_1 c_2)$

$c_2$: Opened

$c_3$: their
Adding depth

$z^L$ predict opened

$z^2 \rightarrow z^3 \rightarrow \text{PReLU} \rightarrow z^2 \rightarrow \text{predict then}$

$z^L \rightarrow z^3 \rightarrow \text{Final layer token-level representation}$

$z^2 \rightarrow z^3 \rightarrow \text{Second layer}$

$z^1 \rightarrow z^2 \rightarrow \text{First layer}$

Residual connection: input to next layer is $z^1 + z^2$

Residual connection: input to next layer is $z^1 + z^2 + z_3$

masked multi-head self-attention

embedding layer

Student

$P_1 \rightarrow P_2 \rightarrow P_3$
What if we want to give the model some input and have it generate a completion? Let's say we're translating from French to English.

Encoder:
(no need to predict the next word)

Decoder:
(responsible for generating text)

\[ p(e_n | e_1 \ldots e_{n-1}, f) \]

(conditional LM)