Transformer: self attention

\[ A_{m \rightarrow t} = \frac{\exp(\text{affinity}(h_m, h_t))}{\sum_{t' = 1}^{M} \exp(\text{affinity}(h_m, h_{t'}))} \]

\[ \sum_{t = 1}^{M} A_{m \rightarrow t} = 1 \]

Wencode

one hot

encoder

Wdecode

<start>

decoder

Problem with LSTM: need to do in a sufficient way, slow

Embedding

\[ I \rightarrow \begin{array}{c}
10 \\
-5 \\
\end{array} \hspace{1cm}
\begin{array}{c}
10 \\
-5 \\
\end{array} \rightarrow
\begin{array}{c}
\text{love} \\
\end{array} \]

continuous vector

operation

\[ W \]

graph

continuous vector

entity/agent
Affinity \((h_m, h_t) = \langle q_m, k_t \rangle / \sqrt{\text{dim}}\)

### Step 1

\[
\Delta_m = \sum_{t=1}^{M} A_{m,t} V_t
\]

- higher weight given to value vector \(V_t\) with more attention
- yields weighted average

\[
h_m \leftarrow h_m + W \Delta_m
\]

contextualize → each vector has info from all other vectors

### Step 2

Each \(h_m\) alone

\[
h_m \leftarrow h_m + \Delta_m
\]

layer normalization after each residual (not batch normalization) (so, after step 1 & step 2 residuals)

\[
h_1, h_2, \ldots, h_m, \ldots, h_M
\]

normalization to mean 0 variance 1
Multi-head

\[
\begin{align*}
  h_i & \rightarrow \text{Wkey} \\
  h_m & \rightarrow \text{Wquery} \\
  v_m & \rightarrow \text{Wvalue} \\
\end{align*}
\]

Initial embedding

\[
\begin{align*}
  h_m & + P_m \\
\end{align*}
\]

Fourier basis

\[
\begin{align*}
  \sin w; m & \\
  \cos w; m & \\
\end{align*}
\]

Position embedding can also be learned

Decoding

\[
\begin{align*}
  \text{Masked} \\
  \text{start} \rightarrow \text{Amo} \rightarrow \text{machingas} \rightarrow ? \rightarrow \text{aprendido}
\end{align*}
\]