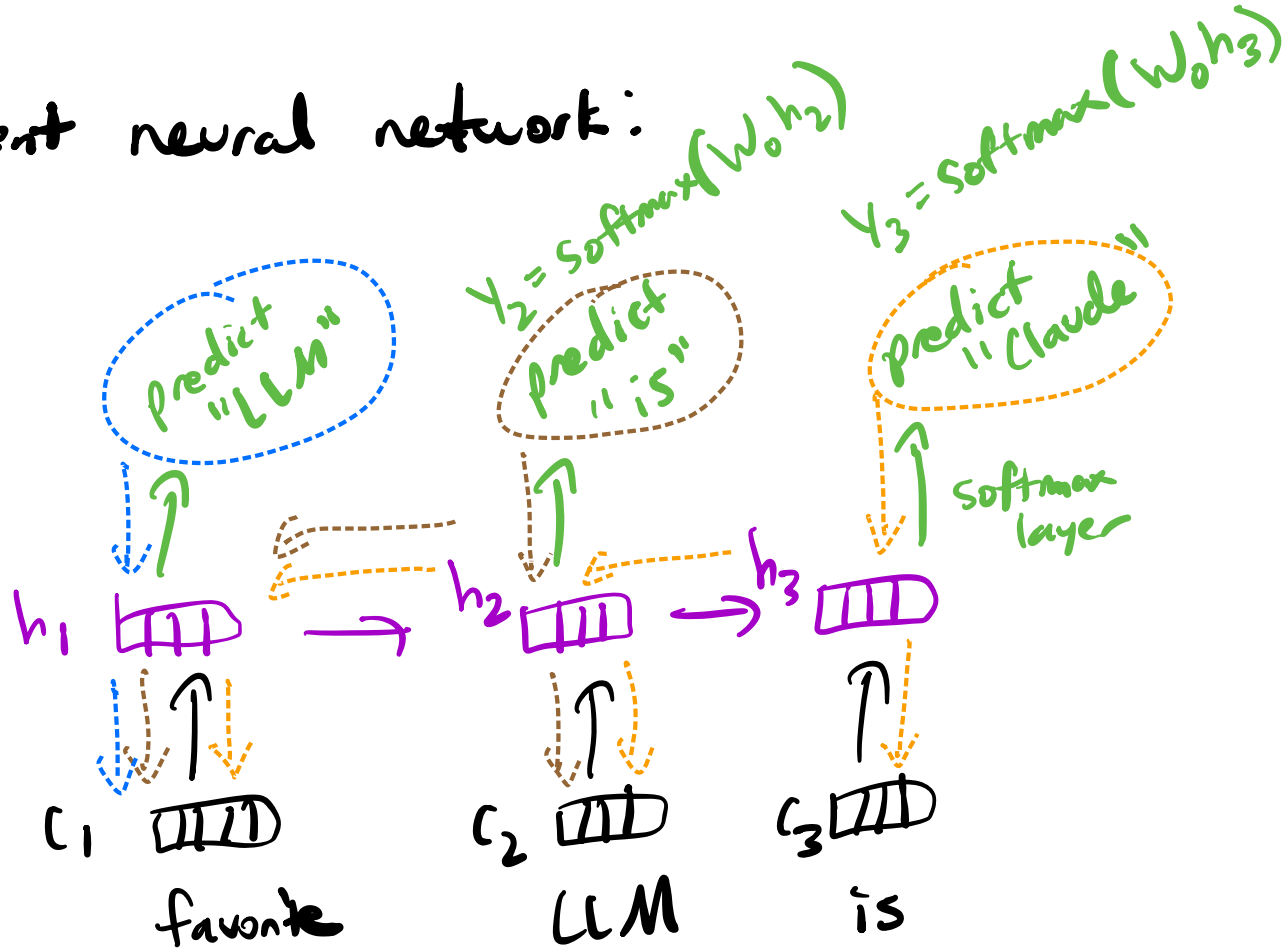
Recurrent neural network:



$$L_3 = -\log p(\text{claude} \mid \text{"favorite LLM is"})$$

$$L_2 = -\log p(\text{is} \mid \text{favorite LLM})$$

$$L_1 = -\log p(\text{LLM} \mid \text{favorite})$$

$$L = \frac{L_3 + L_2 + L_1}{3} \Bigg\} \text{ ave. token-level neg. log likelihood}$$

batch:

<sos> my favorite LLM is claude <eos>

<sos> i am a zebra <eos>

⋮

issues w/ RNNs:

↳ vanishing / exploding gradient

$$\left.\begin{aligned} h_2 &= W_h h_1 + W_c c_2 \\ h_3 &= W_h h_2 + W_c c_3 \\ &\vdots \end{aligned}\right\} \text{linear RNN}$$

$$h_3 = W_h W_h h_1$$

$$h_n = W_h^{n-1} h_1$$

↳ bottleneck

↳ $h_t \Rightarrow$ current hidden state

↳ entire prefix is represented with a single vector

↳ cannot parallelize computation of h_t

↳ h_t depends on h_{t-1}

attention mechanism:

- ↳ originated as a way to alleviate bottleneck in RNNs
  - ↳ Bahdanav, Cho et al 2014
- ↳ Transformer, Vaswani et al 2017
  - ↳ drop recurrence
  - ↳ "self-attention"

self-attention

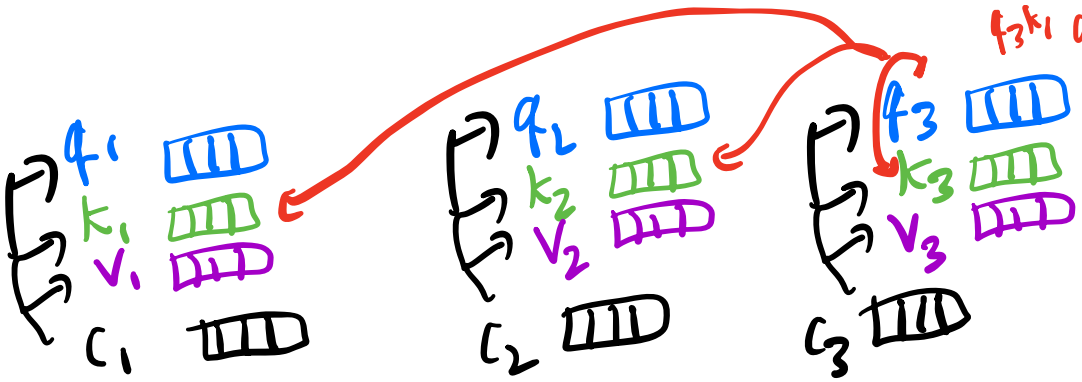
computation of h_3

$h_3 =$ [III] → softmax layer → predict claude

$h_3 = 0.3v_1 + 0.5v_2 + 0.2v_3$

$\text{softmax}(\langle q_3 \cdot k_1, q_3 \cdot k_2, q_3 \cdot k_3 \rangle)$

attention scores



favorite LLM is

query $q_1 = f(W_q c_1)$
key $k_1 = f(W_k c_1)$
value $v_1 = f(W_v c_1)$

} linear/ff layers, $f = \text{ReLU}$

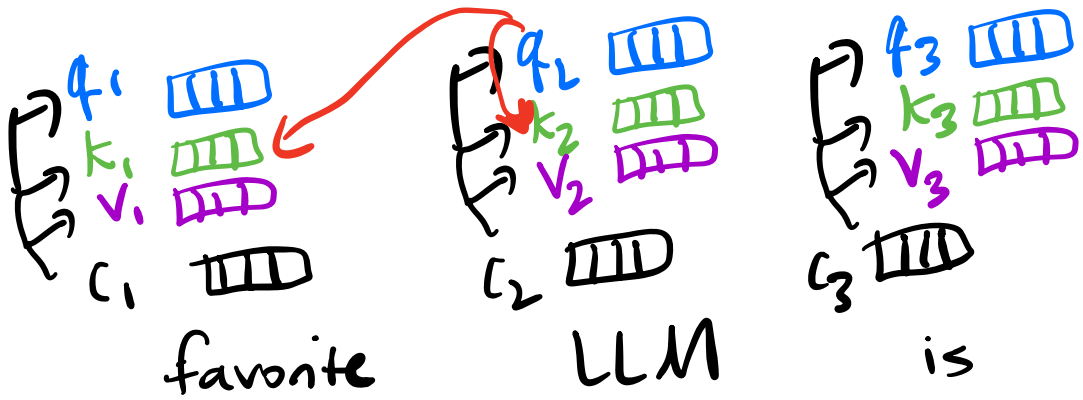
$q_2 = f(W_q c_2)$

Computation of h_2

$$h_2 = 0.3 \cdot v_1 + 0.7 v_2$$

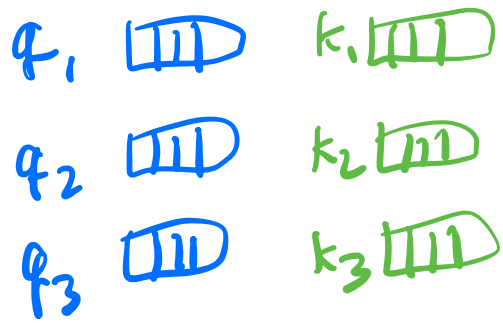
softmax layer predict "is"

$$\text{softmax}(\langle q_2 \cdot k_1, q_2 \cdot k_2 \rangle)$$



↳ in self-attn, no dependency between h_t and h_{t-1}

parallelizing self-attn:



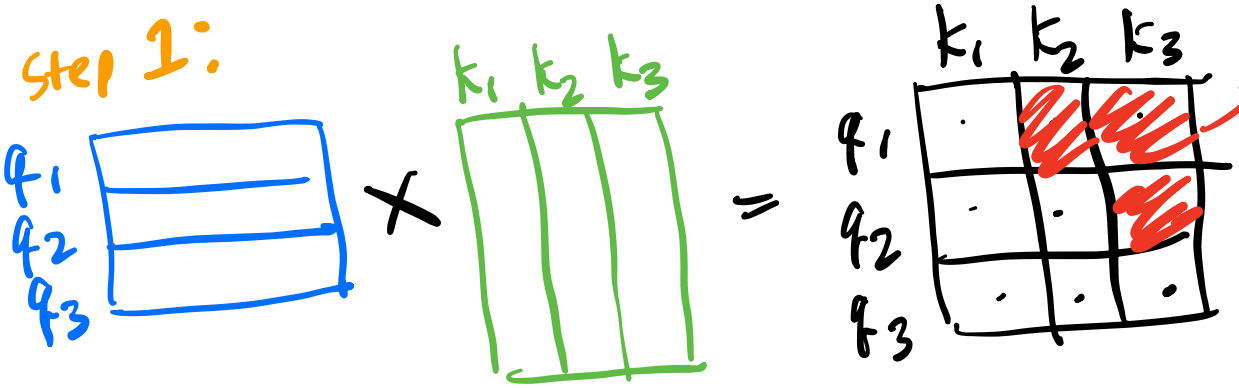
attn vectors:

$a_1 : \langle q_1 \cdot k_1 \rangle$

$a_2 : \langle q_2 \cdot k_1, q_2 \cdot k_2 \rangle$

$a_3 : \langle q_3 \cdot k_1, q_3 \cdot k_2, q_3 \cdot k_3 \rangle$

step 1:



step 2:

$$\text{Softmax}\left(\begin{array}{c|ccc} & k_1 & k_2 & k_3 \\ \hline q_1 & & & \\ q_2 & & & \\ q_3 & & & \end{array} \cdot \begin{array}{ccc} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{array} \right)$$

↳ PyTorch mask_fill (0, -∞)

↓

$$\text{Softmax}\left(\begin{array}{c|ccc} & k_1 & k_2 & k_3 \\ \hline q_1 & \cdot & -\infty & -\infty \\ q_2 & \cdot & \cdot & -\infty \\ q_3 & \cdot & \cdot & \cdot \end{array} \right)$$

step 3

$$\begin{array}{c|ccc} & k_1 & k_2 & k_3 \\ \hline q_1 & 1.0 & 0 & 0 \\ q_2 & 0.3 & 0.7 & 0 \\ q_3 & 0.3 & 0.5 & 0.2 \end{array} \times \begin{array}{c} v_1 \\ v_2 \\ v_3 \end{array}$$

$$= \begin{array}{c} h_1 \\ h_2 \\ h_3 \end{array} \begin{bmatrix} \quad & \quad & \quad \\ \quad & \quad & \quad \\ \quad & \quad & \quad \end{bmatrix}$$