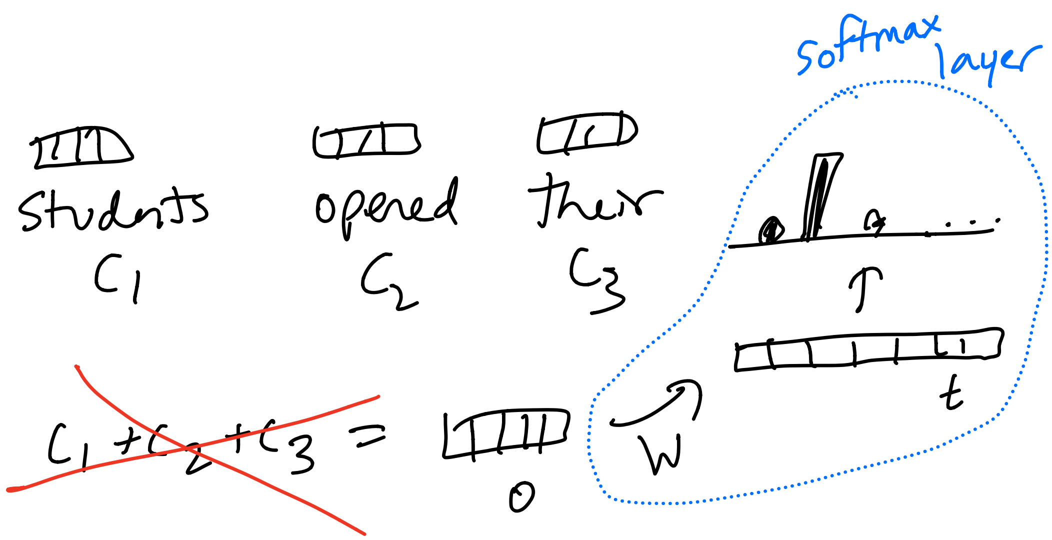
Logistics

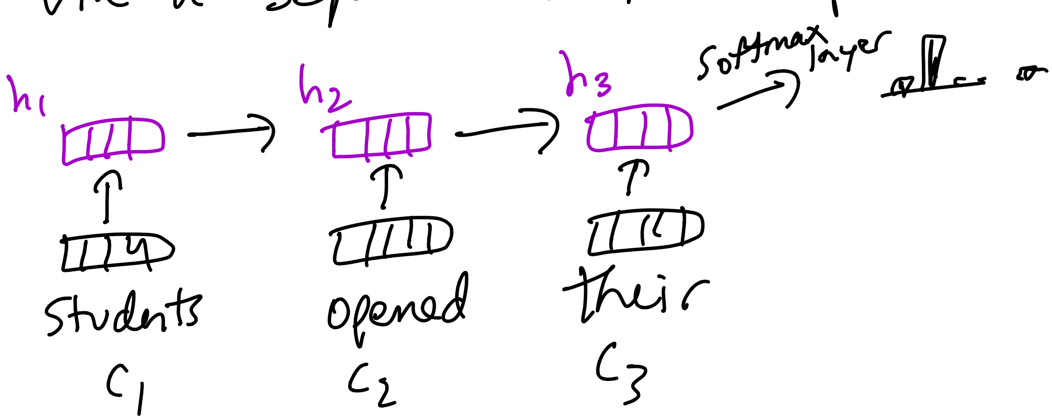
- ↳ group assignments due Friday
- ↳ first student presentation <u>2/20</u>
- ↳ HW 1 released Friday



---

recurrent neural networks

↳ explicitly model word order via a sequential composition process



→ h_t is called the hidden state at position (or time step) t of the seq

→ h_t is a function of c_t and h_{t-1}

→ h and c don't need same dim
d_h d_c

RNN composition fn:

$$h_t = f(W_h h_{t-1} + W_c c_t)$$

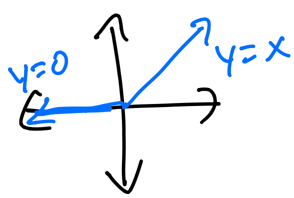
f: nonlinearity
"element-wise nonlinear fn"

W_h: $d_h \times d_h$ matrix

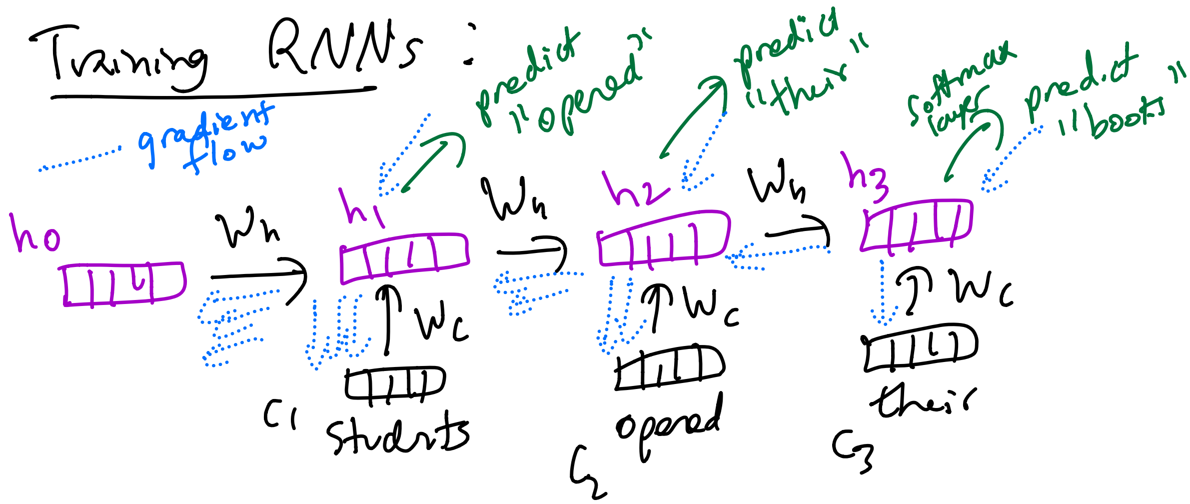
W_c: $d_h \times d_c$ matrix

$f(x) = \tanh(x)$

$f(x) = \mathrm{ReLU}(x) = \max(0, x)$



Training RNNs:



$$L_3 = -\log p(\text{books} \mid \text{"students opened their"})$$

$$L_2 = -\log p(\text{their} \mid \text{"students opened"})$$

$$L_1 = -\log p(\text{opened} \mid \text{students})$$

$$L = \left.\frac{L_1 + L_2 + L_3}{3}\right\} \text{ ave NLL of ground-truth next word}$$

example batch:

1. <bos> students opened their books <eos>
2. <bos> people walked their dogs <eos>
3. <bos> the classroom fell silent <eos>

issues w/ RNNs:

- slow, we have to compute h_{t-1} before computing h_t, no way to parallelize
  - you can parallelize a linear RNN
    - $f(x) = x$
    - key insight for state space models (e.g. Mamba)
- "bottleneck"
  - entire prefix represented by a single vector

---

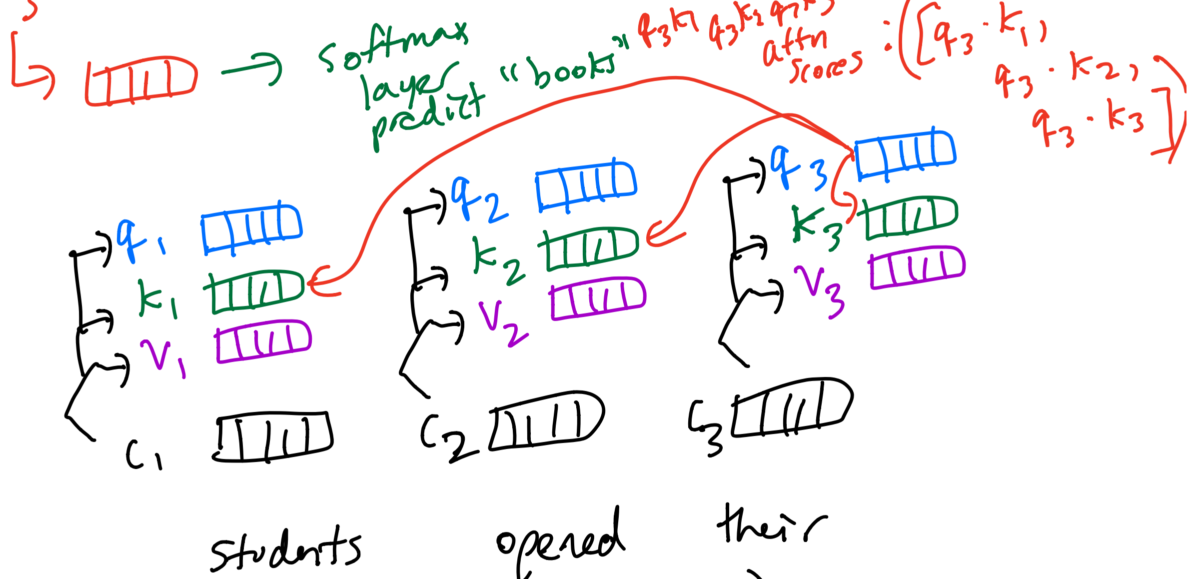
attention mechanism

- Bahdanau, Cho et al 2014
- Transformer, Vaswani et al, 2017
  - hidden state at each timestep is independent of prev. hidden states

Self-attention

Compute hidden state @ 3^{rd} timestep

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



query $q_1 = f(W_q c_1)$ $\quad$ $q_2 = f(W_q c_2)$

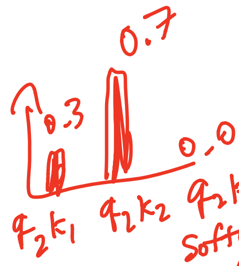
key $k_1 = f(W_k c_1)$

value $v_1 = f(W_v c_1)$

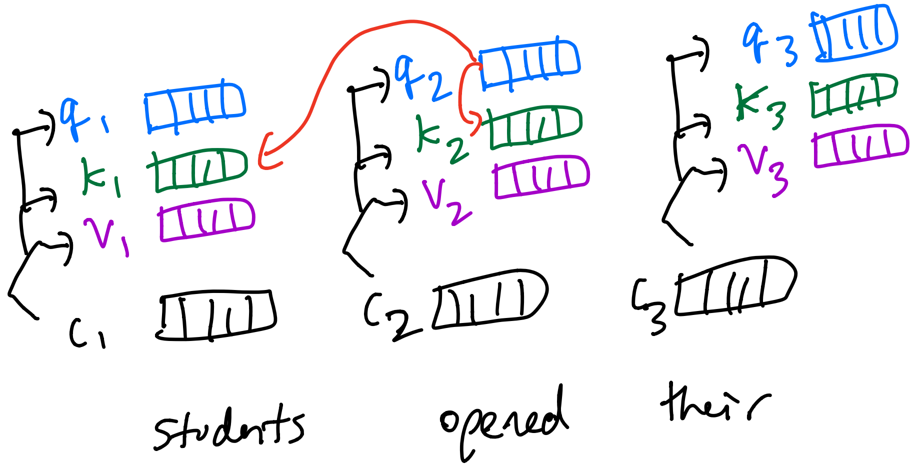
computation of h_2

$h_2 = 0.3v_1 + 0.7v_2$

↳ [ ] $\xrightarrow{\text{softmax}}$ predict "their"



attn scores: $\text{softmax}([q_2k_1, q_2k_2])$



next class

↳ how to parallelize attention computations

↳ multi-head attn

↳ position embeddings