Recurrent neural network:

\[ h_n = f(W_h h_{n-1} + W_c c_n) \]

RNN gradient \( \Rightarrow \) "backpropagation thru time"

\[ L_3 = -\log P(\text{books} | \text{students opened their}) \]
\[ L_2 = -\log P(\text{their} | \text{students opened}) \]
\[ L_1 = -\log P(\text{opened} | \text{students}) \]

\[ L = \frac{L_1 + L_2 + L_3}{3} \]

average neg. log likelihood of the ground-truth next word over all tokens in the batch

batch:
1. students opened their books
2. people walked their dog
3. the classroom fell silent
Issues w/ RNNs:

1. "bottleneck"
   - Ray Mooney, 2014
   "you can’t represent the meaning of a sentence in a BLEEPING vector."

2. lack of parallelism across timesteps
   - I can only compute $h_n$ after computing $h_{n-1}$

Attention mechanism:

history: - developed initially for RNNs and for machine translation
   - Bahdanau, Cho et al, 2014
   - Vaswani et al. 2017 dropped the "recurrent" aspect and created a fully attention-based architecture
   - Transformer
   - for machine translation
   - introduced by Google
   - hidden state at each timestep is independent of $h_1...n-1$
Self-attention:

computation of hidden state at timestep 3:

\[ h_3 = 0.3v_1 + 0.5v_2 + 0.2v_3 \]

\[ \text{softmax} \]

\[ \text{predict "books"} \]

\[ \text{attn scores: } \langle q_3 \cdot k_1, q_3 \cdot k_2, q_3 \cdot k_3 \rangle \]

Query: \( q_1 = f(W_q c_1) \), \( q_2 = f(W_q c_2) \)

Key: \( k_1 = f(W_k c_1) \)

Value: \( v_1 = f(W_v c_1) \)

\( W_q, W_k, W_v \) are randomly initialized parameters learned during training!
computations at second time step:

\[ h_2 = 0.3 \cdot v_1 + 0.7 \cdot v_2 \]

\[ \text{Softmax} \rightarrow \text{predict layer} \]

\[ \text{attn: Softmax}(q_2 \cdot k_1, q_2 \cdot k_2) \]

no dependencies between \( h_1, h_2, h_3 \)

\[ \Rightarrow \text{parallelize} \]

\[ \Rightarrow \text{reduce bottleneck} \]
how to parallelize:

\[ f_1, f_2, f_3 \quad k_1, k_2, k_3 \]

**Attention vectors**

\[ a_1 = \langle q_1, k_1 \rangle \]
\[ a_2 = \langle q_2 k_1, q_2 k_2 \rangle \]
\[ a_3 = \langle q_3 k_1, q_3 k_2, q_3 k_3 \rangle \]

\[ f_1 \quad f_2 \quad f_3 \]
\[ k_1 \quad k_2 \quad k_3 \]

These cells have info about the future, we need to mask.

Softmax

\[ \begin{pmatrix}
  f_1 \\
  f_2 \\
  f_3 
\end{pmatrix}
\]

\[ \begin{pmatrix}
  1 & -\infty & -\infty \\
  1 & 1 & -\infty \\
  1 & 1 & 1 
\end{pmatrix}
\]

**Attention scores**

\[ \begin{pmatrix}
  k_1 \\
  k_2 \\
  k_3 
\end{pmatrix}
\]
\[ \begin{pmatrix}
  v_1 \\
  v_2 \\
  v_3 
\end{pmatrix}
\]

= 

\[ \begin{pmatrix}
  h_1 \\
  h_2 \\
  h_3 
\end{pmatrix} \]