Today
- RNNs
- LSTMs (the type of RNN you will be using)
- Implementation

Announcements
- All out later today
- Midterm back Monday

Recap: RNNs + Language modeling

\[
p(w | I \text{ saw the dog}) = \text{softmax} \left( Z \mathbf{h} \right)
\]
Training: "Backpropagation through time" = backprop

\[ \text{loss} = \log p(w^*) \]

Backprop: loss \rightarrow gradient for each param

Multiple updates for \( W, V \), sum them together

This still works!
LSTMs

What we've seen so far is a recipe for RNNs.

Many types of RNNs

Long short-term memory (LSTM) 1998

RNN state is a "short-term memory".
LSTMs are better at remembering into far more timesteps

Problem with Elman networks:
vanishing gradient problem (exploding)

Elman: \( h_3 = \tanh(Wx_3 + V - \tanh(Wx_2 + V - \tanh(Wx_1))) \)
**LSTM definition**

Elman: \( \overline{h}_i = \tanh (W \overline{x}_i + V \overline{h}_{i-1}) \)

Gated: \( \overline{h}_i = \overline{h}_{i-1} \odot f + \text{function}(\overline{x}_i, \overline{h}_{i-1}) \odot i \)

\( f \): forget gate  \( i \): input gate

vector of values in \([0, 1]\)

\[
\begin{bmatrix}
\overline{h}_{i-1} \\
0 \\
1 \\
0 \\
\end{bmatrix}

\begin{bmatrix}
f \\
0.5 \\
0 \\
0 \\
\end{bmatrix}

= 

\begin{bmatrix}
\overline{h}_i \\
0 \\
0 \\
0 \\
\end{bmatrix}

\]

\[ f = \text{sigmoid} \left( W^{(1)} \overline{x}_i + W^{(2)} \overline{h}_{i-1} \right) \]

\[ i = \text{sigmoid} \left( W^{(3)} \overline{x}_i + W^{(a)} \overline{h}_{i-1} \right) \]

\[ o : \frac{e^x}{1 + e^x} \]
$h$ is "memory"

\[
\overline{h}_3 \rightarrow \text{[diagram]} + \frac{Xy}{\text{stuff}} \Rightarrow \overline{h}_y
\]

Real LSTM: next page
Chris Olah blog

LSTM: 8 weight matrices

$h, c$

hidden cell

state state

in Pytorch, these are dealt with as a tuple

$h_i, c_i$
Poll:

\[ f = \text{sigmoid} \left( W^{(1)} x_i + W^{(2)} h_{i-1} + b_{\text{forget}} \right) \]
\[ i = \text{sigmoid} \left( W^{(3)} x_i + W^{(4)} h_{i-1} + b_{\text{input}} \right) \]

\([0,1,2]\) closer to \([1,1,2]\) than \([0,1,3]\)

Most recent stuff is more relevant

Forget gate high: charge further back matter more