CS 6180 09/29

HW2 (I promise tomorrow)

Requested Google Cloud Credits (share instructions on how to retrieve them)

Review from last time

Long short-term Memory (LSTMs)

main motivation → vanishing gradients issues in RNNs.

$\vec{h}^{(t-1)}$ $\vec{c}^{(t-1)}$

$$\vec{f}^{(t)} = \sigma(W_f \vec{h}^{(t-1)} + U_f \vec{e}^{(t)} + \vec{b}_f)$$

$$\vec{i}^{(t)} = \sigma(W_i \vec{h}^{(t-1)} + U_i \vec{e}^{(t)} + \vec{b}_i)$$

$$\vec{o}^{(t)} = \sigma(W_o \vec{h}^{(t-1)} + U_o \vec{e}^{(t)} + \vec{b}_o)$$

$$\tilde{c}^{(t)} = \tanh(W_c \vec{h}^{(t-1)} + U_c \vec{e}^{(t)} + \vec{b}_c)$$

$$\vec{c}^{(t)} = \vec{f}^{(t)} * \vec{c}^{(t-1)} + \vec{i}^{(t)} * \tilde{c}^{(t)}$$

$$\vec{h}^{(t)} = \vec{o}^{(t)} * \tanh(\vec{c}^{(t)})$$

We will fix the issues of vanishing gradients

Let's focus on a LSTM with two time steps.

W_f	U_f	$\vec{b}_f$
W_i	U_i	$\vec{b}_i$
W_o	U_o	$\vec{b}_o$
W_c	U_c	$\vec{b}_c$

Let's figure out $\frac{\partial L}{\partial W_f}$.

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial \vec{f}^{(1)}} \frac{\partial \vec{f}^{(1)}}{\partial W_f} + \frac{\partial L}{\partial \vec{f}^{(2)}} \cdot \frac{\partial \vec{f}^{(2)}}{\partial W_f}$$

$\frac{\partial \vec{f}^{(1)}}{\partial W_f}$: we can compute $\vec{f}^{(1)} \cdot (1-\vec{f}^{(1)}) \cdot (\vec{h}^{(0)})^T$

$\frac{\partial \vec{f}^{(2)}}{\partial W_f}$: similar

$$\frac{\partial L}{\partial \vec{f}^{(1)}} = \frac{\partial L}{\partial \vec{c}^{(1)}} \frac{\partial \vec{c}^{(1)}}{\partial \vec{f}^{(1)}}$$

$$\frac{\partial L}{\partial \vec{f}^{(2)}} = \frac{\partial L}{\partial \vec{c}^{(2)}} \cdot \frac{\partial \vec{c}^{(2)}}{\partial \vec{f}^{(2)}}$$

$$\frac{\partial L}{\partial w_f} = \frac{\partial L}{\partial \vec{c}^{(1)}} \frac{\partial \vec{c}^{(1)}}{\partial \vec{f}^{(1)}} \frac{\partial \vec{f}^{(1)}}{\partial w_f} + \frac{\partial L}{\partial \vec{c}^{(2)}} \frac{\partial \vec{c}^{(2)}}{\partial \vec{f}^{(2)}} \frac{\partial \vec{f}^{(2)}}{\partial w_f}$$

$\vec{c}^{(0)}$ $\vec{c}^{(1)}$

We now need to figure out $\frac{\partial L}{\partial \vec{c}^{(1)}}$ and

$$\frac{\partial L}{\partial \vec{c}^{(2)}}$$

$$\frac{\partial L}{\partial \vec{C}^{(1)}} = \frac{\partial L}{\partial \vec{h}^{(1)}} \frac{\partial \vec{h}^{(1)}}{\partial \vec{C}^{(1)}} + \frac{\partial L}{\partial \vec{C}^{(2)}} \frac{\partial \vec{C}^{(2)}}{\partial \vec{C}^{(1)}}$$

$\frac{\partial \vec{h}^{(1)}}{\partial \vec{C}^{(1)}}$: $O^{(1)} * (1 - \tanh^2(C^{(1)}))$

$\frac{\partial \vec{C}^{(2)}}{\partial \vec{C}^{(1)}}$: $f^{(2)}$

$$\frac{\partial L}{\partial \vec{C}^{(2)}} = \frac{\partial L}{\partial \vec{h}^{(2)}} \frac{\partial \vec{h}^{(2)}}{\partial \vec{C}^{(2)}}$$

$$= \frac{\partial L}{\partial \vec{y}^{(2)}} \cdot \frac{\partial \vec{y}^{(2)}}{\partial \vec{h}^{(2)}} \cdot \frac{\partial \vec{h}^{(2)}}{\partial \vec{C}^{(2)}}$$

we're getting a recurrence relationship linking

$$\frac{\partial L}{\partial \vec{c}^{(t)}} \quad \text{to} \quad \frac{\partial L}{\partial \vec{c}^{(t+1)}}$$

$$\rightarrow \frac{\partial L}{\partial \vec{c}^{(\text{last } t)}}$$

→ will be "easy" to compute

$$\frac{\partial L}{\partial \vec{h}^{(1)}} = \frac{\partial L}{\partial L^{(1)}} \frac{\partial L^{(1)}}{\partial \hat{y}^{(1)}} \frac{\partial \hat{y}^{(1)}}{\partial \vec{h}^{(1)}}$$

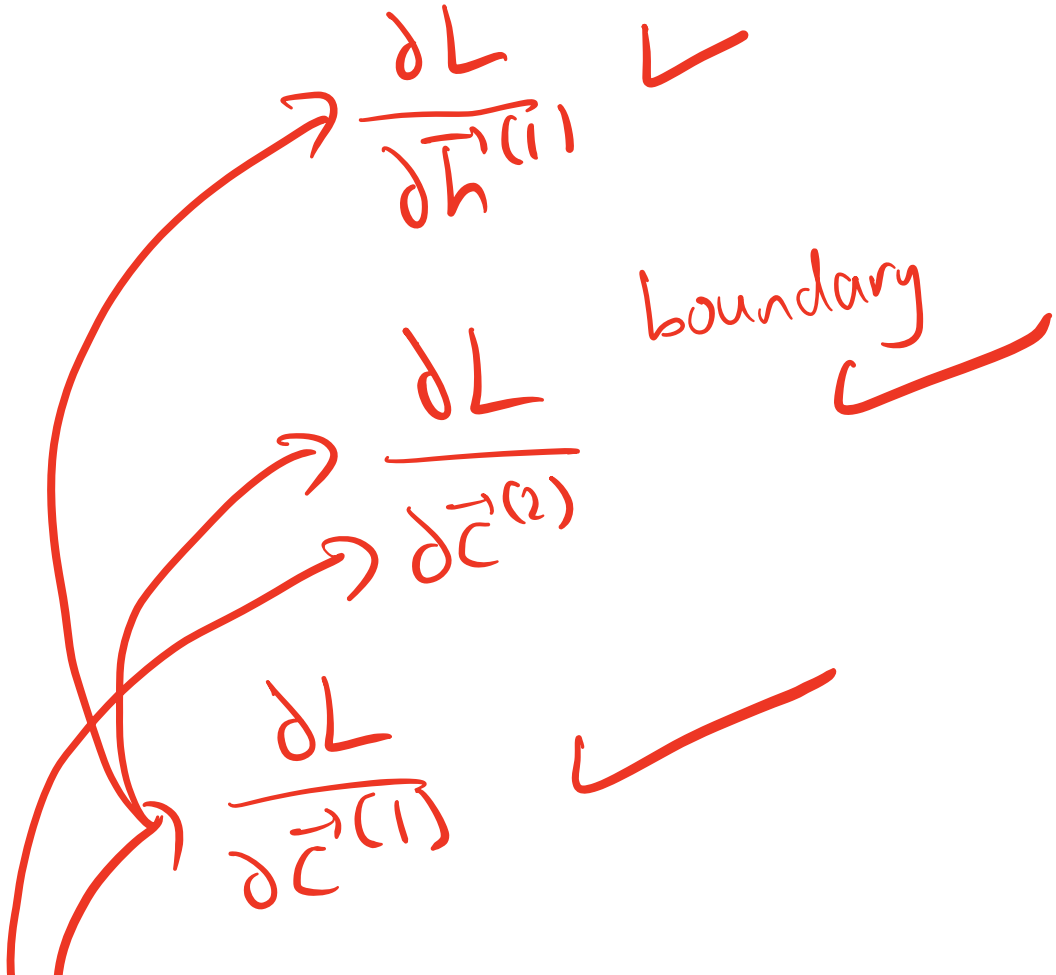
$$+ \frac{\partial L}{\partial \vec{h}^{(2)}} \left(\frac{\partial \vec{h}^{(2)}}{\partial \vec{o}^{(2)}} \frac{\partial \vec{o}^{(2)}}{\partial \vec{h}^{(1)}} + \frac{\partial \vec{h}^{(2)}}{\partial \vec{c}^{(2)}} \cdot \frac{\partial \vec{c}^{(2)}}{\partial \vec{f}^{(2)}} \frac{\partial \vec{f}^{(2)}}{\partial \vec{h}^{(1)}} \right.$$

$$\left. + \frac{\partial \vec{h}^{(2)}}{\partial \vec{c}^{(2)}} \frac{\partial \vec{c}^{(2)}}{\partial \vec{i}^{(2)}} \frac{\partial \vec{i}^{(2)}}{\partial \vec{h}^{(1)}} \right.$$

$$+ \frac{\partial \vec{h}^{(2)}}{\partial \vec{c}^{(2)}} \frac{\partial \vec{c}^{(2)}}{\partial \tilde{\vec{c}}^{(2)}} \frac{\partial \tilde{\vec{c}}^{(2)}}{\partial \vec{h}^{(1)}} \Bigg)$$

$$\frac{\partial L}{\partial \vec{h}^{(2)}} = \frac{\partial L}{\partial L^{(2)}} \frac{\partial L^{(2)}}{\partial \hat{y}^{(2)}} \frac{\partial \hat{y}^{(2)}}{\partial \vec{h}^{(2)}}$$

First compute $\frac{\partial L}{\partial \vec{h}^{(2)}}$ ← boundary



$\frac{\partial L}{\partial W_h}$ we can finally compute

pytorch does all of that for you

we love pytorch

Suppose we have 10 words and word 2 is relevant for word 10

with vanilla RNNs, we would have gotten vanishing gradients

$$\vec{C}^{(0)} = 0$$

$$\vec{C}^{(1)} = \beta^{(1)} * C^{(0)} + \lambda^{(1)} * \tilde{C}^{(1)}$$

$$= 1 * 0 + 0 * \tilde{C}^{(1)}$$

$$= 0$$

$$\lambda^{(1)} = 0$$

$$\beta^{(1)} = 1$$

$$\vec{C}^{(2)} = \beta^{(2)} * C^{(1)} + \lambda^{(2)} * \tilde{C}^{(2)}$$

$$= 1 * 0 + 1 * \tilde{C}^{(2)}$$

$$= \tilde{C}^{(2)}$$

$$\lambda^{(2)} = 1$$

$$\beta^{(2)} = 1$$

$$\vec{C}^{(3)} = \beta^{(3)} * C^{(2)} + \lambda^{(3)} * \tilde{C}^{(3)}$$

$$\approx \tilde{C}^{(2)}$$

$$\lambda^{(3)} = 0$$

$f^{(3)} = 1$

$\forall t > 3$ $\quad i^{(t)} = 0$

$f^{(t)} = 1$

$c^{(10)} = \tilde{c}^{(2)}$

$\frac{\partial L^{(10)}}{\partial c^{(2)}}$

$c^{(2)} \to c^{(3)} \to c^{(4)}$

$\cdots \to c^{(10)} \to h^{(10)}$

$\to L^{(10)}$

path: $c^{(2)} \to c^{(3)} \to \cdots \to c^{(10)} \to h^{(10)} \to L^{(10)}$

$$\frac{\partial L^{(10)}}{\partial c^{(2)}} = \frac{\partial L^{(10)}}{\partial h^{(10)}} \frac{\partial h^{(10)}}{\partial c^{(10)}} \frac{\partial c^{(10)}}{\partial c^{(9)}} \cdots \frac{\partial c^{(3)}}{\partial c^{(2)}}$$

$$= \frac{\partial L^{(10)}}{\partial h^{(10)}} * O^{(10)} * \left(1 - \tanh^2\left(C^{(10)}\right)\right) * f^{(10)} * f^{(9)} * f^{(8)} * \cdots * f^{(3)}$$

$$= \frac{\partial L^{(10)}}{\partial h^{(10)}} * O^{(10)} * \left(1 - \tanh^2\left(C^{(10)}\right)\right) * \underbrace{\prod_{i=3}^{10} f^{(i)}}$$

$$(0.95)^8$$

$$\simeq 0.663$$

Transformers

→ attention

"Attention is all you need"

Paper that introduced transformers

(by Google Folks)