50.039 Theory and Practice of Deep Learning
Theory Homework 6

Joel Huang 1002530

March 26, 2019

Cell state and hidden state

Is c_{t-1} a function of h_{t-1}? The previous cell state c_{t-1} is carried over from the previous LSTM cell. h_{t-1}, the hidden state carried over from the previous LSTM cell, is used to compute c_t, the cell state for the current cell; therefore only c_t is a function of h_{t-1}, and c_{t-1} never is.

Hidden state derivative

Taking the derivative of the output hidden state equation for time t,

$$h_t = o_t \circ \tanh(c_t)$$

$$\frac{\partial h_t}{\partial h_{t-1}} = \tanh(c_t) \circ \frac{\partial o_t}{\partial h_{t-1}} + o_t \circ \frac{\partial \tanh(c_t)}{\partial h_{t-1}}$$

Since

$$\tanh(c_t) = \tanh(f_t \circ c_{t-1} + i_t \circ u_t),$$

$$\frac{\partial \tanh(c_t)}{\partial h_{t-1}} = (1 - \tanh^2(c_t)) \left(f_t \circ \frac{\partial c_{t-1}}{\partial h_{t-1}} + c_{t-1} \circ \frac{\partial f_t}{\partial h_{t-1}} + u_t \circ \frac{\partial i_t}{\partial h_{t-1}} + i_t \circ \frac{\partial u_t}{\partial h_{t-1}} \right)$$

And since in the previous section we show that c_{t-1} is not a function of h_{t-1},

$$\frac{\partial \tanh(c_t)}{\partial h_{t-1}} = (1 - \tanh^2(c_t)) \left(c_{t-1} \circ \frac{\partial f_t}{\partial h_{t-1}} + u_t \circ \frac{\partial i_t}{\partial h_{t-1}} + i_t \circ \frac{\partial u_t}{\partial h_{t-1}} \right)$$

Substituting all values found, we have

$$\frac{\partial h_t}{\partial h_{t-1}} = \tanh(c_t) \circ \frac{\partial o_t}{\partial h_{t-1}} + o_t \circ \left[(1 - \tanh^2(c_t)) \left(c_{t-1} \circ \frac{\partial f_t}{\partial h_{t-1}} + u_t \circ \frac{\partial i_t}{\partial h_{t-1}} + i_t \circ \frac{\partial u_t}{\partial h_{t-1}} \right) \right]$$

Sigmoid derivative

$$\begin{aligned}
\sigma(z) &= \frac{1}{1+e^{-z}} = (1+e^{-z})^{-1} \\
\sigma'(z) &= e^{-z} \cdot (1+e^{-z})^{-2} \\
&= \frac{e^{-z}}{1+e^{-z}} \cdot \frac{1}{1+e^{-z}} \\
&= \left(\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}} \right) \cdot \sigma(z) \\
&= (1 - \sigma(z)) \cdot \sigma(z)
\end{aligned}$$

Forget gate derivative

$$
\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1}) \\
\frac{\partial f_t}{\partial h_{t-1}} &= U_f \cdot \sigma'(W_f x_t + U_f h_{t-1}) \\
&= U_f \cdot [\,(1 - \sigma(W_f x_t + U_f h_{t-1})) \cdot \\
&\quad \sigma(W_f x_t + U_f h_{t-1})\,]
\end{aligned}
$$

Forget gate activation

Which vector h_{t-1} among all the vectors of euclidean length 1 maximize the values of $f_t^{(d)}$?

$$\underset{h_{t-1}:\|h_{t-1}\|_2=1}{\operatorname{argmax}} f_t^{(d)}$$

$$\underset{h_{t-1}:\|h_{t-1}\|_2=1}{\operatorname{argmax}} \sigma(W_f^{(d)} \cdot x_t + U_f^{(d)} \cdot h_{t-1})$$

The logistic function is monotonic, so maximizing $f_t^{(d)}$ is equivalent to maximizing $U_f^{(d)} \cdot h_{t-1}$. Let θ be the angle between the two vectors $U_f^{(d)}, h_{t-1}$. Since $\|h_{t-1}\| = 1$ and $x_t = 0$:

$$
\begin{aligned}
\underset{h_{t-1}:\|h_{t-1}\|_2=1}{\operatorname{argmax}} U_f^{(d)} \cdot h_{t-1} &= \underset{h_{t-1}:\|h_{t-1}\|_2=1}{\operatorname{argmax}} \|U_f^{(d)}\|\|h_{t-1}\| \cos(\theta) \\
&= \underset{\theta}{\operatorname{argmax}} \|U_f^{(d)}\| \cos(\theta)
\end{aligned}
$$

The value of θ that maximizes $\|U_f^{(d)}\| \cos(\theta)$ is 0. Given some $U_f^{(d)} \neq 0$ and $x_t = 0$, $f_t^{(d)}$ attains its maximum at $\|U_f^{(d)}\|$ when $\theta = 0$, subject to $\|h_{t-1}\| = 1$. Therefore the greatest activation of the forget gate component $f_t^{(d)}$, when no bias is used, is acheieved when h_{t-1} is aligned with weight vector $U_f^{(d)}$. The direction where $\theta = 0$ is:

$$\hat{h}_{t-1} = \frac{h_{t-1}}{\|h_{t-1}\|} = \frac{U_f^{(d)}}{\|U_f^{(d)}\|}$$

Given $x_t = 0$, $W_f^{(d)} \cdot x_t$ does not affect the argmax nor the max, as the expression $\operatorname{argmax}_{h_{t-1}:\|h_{t-1}\|_2=1} \sigma(W_f^{(d)} \cdot x_t + U_f^{(d)} \cdot h_{t-1})$ reduces to $\operatorname{argmax}_{h_{t-1}:\|h_{t-1}\|_2=1} \sigma(U_f^{(d)} \cdot h_{t-1})$ when $W_f^{(d)} \cdot x_t = 0$.