Problem 4

1. Derive LSTM step backward.

First we prepare the derivative of sigmoid (σ) and tanh.

$$\frac{\partial \sigma(\theta)}{\partial \theta} = \sigma(\theta) \cdot [1-\sigma(\theta)] \qquad \frac{\partial \tanh \theta}{\partial \theta} = 1 - \tanh^2 \theta$$

Then we prepare the derivatives w.r.t f_t and $\tilde{C}_t$.

$$\frac{\partial f_t}{\partial x_t} = (W_x^f)^T [f_t * (1-f_t)] \qquad \frac{\partial f_t}{\partial W_x^f} = [f_t * (1-f_t)] x_t^T$$

$$\frac{\partial f_t}{\partial h_{t-1}} = (W_h^f)^T [f_t * (1-f_t)] \qquad \frac{\partial f_t}{\partial W_h^f} = [f_t * (1-f_t)] h_{t-1}^T$$

$$\frac{\partial f_t}{\partial b^f} = f_t * (1-f_t)$$

$$\frac{\partial \tilde{C}_t}{\partial x_t} = (W_x^c)^T (1-\tilde{C}_t^2) \qquad \frac{\partial \tilde{C}_t}{\partial W_x^c} = (1-\tilde{C}_t^2) x_t^T$$

$$\frac{\partial \tilde{C}_t}{\partial h_{t-1}} = (W_h^c)^T (1-\tilde{C}_t^2) \qquad \frac{\partial \tilde{C}_t}{\partial W_h^c} = (1-\tilde{C}_t^2) h_{t-1}^T$$

$$\frac{\partial \tilde{C}_t}{\partial b^c} = 1-\tilde{C}_t^2$$

We first obtain the easy one:

$$\frac{\partial L}{\partial C_{t-1}} = \frac{\partial L}{\partial C_t} \cdot \frac{\partial C_t}{\partial C_{t-1}} = \frac{\partial L}{\partial C_t} * f_t .$$

f_t family:

$$\frac{\partial L}{\partial W_x^f} = \frac{\partial L}{\partial c_t} \cdot \frac{\partial c_t}{\partial f_t} \cdot \frac{\partial f_t}{\partial W_x^f} = \left\{ \frac{\partial L}{\partial c_t} * c_{t-1} * [f_t * (1-f_t)] \right\} x_t^T$$

$$\frac{\partial L}{\partial W_h^f} = \frac{\partial L}{\partial c_t} \cdot \frac{\partial c_t}{\partial f_t} \cdot \frac{\partial f_t}{\partial W_h^f} = \left\{ \frac{\partial L}{\partial c_t} * c_{t-1} * [f_t * (1-f_t)] \right\} h_{t-1}^T$$

$$\frac{\partial L}{\partial b^f} = \frac{\partial L}{\partial c_t} \cdot \frac{\partial c_t}{\partial f_t} \cdot \frac{\partial f_t}{\partial b^f} = \frac{\partial L}{\partial c_t} * c_{t-1} * [f_t * (1-f_t)]$$

i_t family

$$\frac{\partial L}{\partial W_x^i} = \frac{\partial L}{\partial c_t} \cdot \frac{\partial c_t}{\partial i_t} \cdot \frac{\partial i_t}{\partial W_x^i} = \left\{ \frac{\partial L}{\partial c_t} * \tilde{c}_t * [i_t * (1-i_t)] \right\} x_t^T$$

$$\frac{\partial L}{\partial W_h^i} = \frac{\partial L}{\partial c_t} \cdot \frac{\partial c_t}{\partial i_t} \cdot \frac{\partial i_t}{\partial W_h^i} = \left\{ \frac{\partial L}{\partial c_t} * \tilde{c}_t * [i_t * (1-i_t)] \right\} h_{t-1}^T$$

$$\frac{\partial L}{\partial b^i} = \frac{\partial L}{\partial c_t} \cdot \frac{\partial c_t}{\partial f_t} \cdot \frac{\partial f_t}{\partial b^i} = \frac{\partial L}{\partial c_t} * \tilde{c}_t * [i_t * (1-i_t)]$$

o_t family

$$\frac{\partial L}{\partial W_x^o} = \frac{\partial L}{\partial h_t} \cdot \frac{\partial h_t}{\partial o_t} \cdot \frac{\partial o_t}{\partial W_x^o} = \left\{ \frac{\partial L}{\partial h_t} * \tanh(c_t) * [o_t * (1-o_t)] \right\} x_t^T$$

$$\frac{\partial L}{\partial W_h^o} = \frac{\partial L}{\partial h_t} \cdot \frac{\partial h_t}{\partial o_t} \cdot \frac{\partial o_t}{\partial W_h^o} = \left\{ \frac{\partial L}{\partial h_t} * \tanh(c_t) * [o_t * (1-o_t)] \right\} h_{t-1}^T$$

$$\frac{\partial L}{\partial b^o} = \frac{\partial L}{\partial h_t} \cdot \frac{\partial h_t}{\partial o_t} \cdot \frac{\partial o_t}{\partial b^o} = \frac{\partial L}{\partial h_t} * \tanh(c_t) * [o_t * (1-o_t)]$$

$\tilde{c}_t$ family

$$\frac{\partial L}{\partial W_x^c} = \frac{\partial L}{\partial c_t} \cdot \frac{\partial c_t}{\partial \tilde{c}_t} \cdot \frac{\partial \tilde{c}_t}{\partial W_x^c} = \left[\frac{\partial L}{\partial c_t} * i_t * (1-\tilde{c}_t^2) \right] x_t^T$$

$$\frac{\partial L}{\partial W_h^c} = \frac{\partial L}{\partial c_t} \cdot \frac{\partial c_t}{\partial \tilde{c}_t} \cdot \frac{\partial \tilde{c}_t}{\partial W_h^c} = \left[\frac{\partial L}{\partial c_t} * i_t * (1-\tilde{c}_t^2) \right] h_{t-1}^T$$

$$\frac{\partial L}{\partial b^c} = \frac{\partial L}{\partial c_t} \cdot \frac{\partial c_t}{\partial \tilde{c}_t} \cdot \frac{\partial \tilde{c}_t}{\partial b^c} = \left[\frac{\partial L}{\partial c_t} * i_t * (1-\tilde{c}_t^2) \right]$$

Finally we get

$$\frac{\partial L}{\partial x_t} = \frac{\partial L}{\partial C_t}\cdot\frac{\partial C_t}{\partial f_t}\cdot\frac{\partial f_t}{\partial x_t} + \frac{\partial L}{\partial C_t}\cdot\frac{\partial C_t}{\partial i_t}\cdot\frac{\partial i_t}{\partial x_t}$$

$$+ \frac{\partial L}{\partial C_t}\cdot\frac{\partial C_t}{\partial \tilde{C}_t}\cdot\frac{\partial \tilde{C}_t}{\partial x_t} + \frac{\partial L}{\partial h_t}\cdot\frac{\partial h_t}{\partial o_t}\cdot\frac{\partial o_t}{\partial x_t}$$

$$= (W_x^f)^T\left\{\frac{\partial L}{\partial C_t} * C_{t-1} * [f_t * (1-f_t)]\right\}$$

$$+ (W_x^i)^T\left\{\frac{\partial L}{\partial C_t} * \tilde{C}_t * [i_t * (1-i_t)]\right\}$$

$$+ (W_x^c)^T\left\{\frac{\partial L}{\partial C_t} * i_t * (1-\tilde{C}_t^2)\right\}$$

$$+ (W_x^o)^T\left\{\frac{\partial L}{\partial h_t} * \tanh(C_t) * [o_t * (1-o_t)]\right\}$$

$$\frac{\partial L}{\partial h_{t-1}} = \frac{\partial L}{\partial C_t}\cdot\frac{\partial C_t}{\partial f_t}\cdot\frac{\partial f_t}{\partial h_{t-1}} + \frac{\partial L}{\partial C_t}\cdot\frac{\partial C_t}{\partial i_t}\cdot\frac{\partial i_t}{\partial h_{t-1}}$$

$$+ \frac{\partial L}{\partial C_t}\cdot\frac{\partial C_t}{\partial \tilde{C}_t}\cdot\frac{\partial \tilde{C}}{\partial h_{t-1}} + \frac{\partial L}{\partial h_t}\cdot\frac{\partial h_t}{\partial o_t}\cdot\frac{\partial o_t}{\partial h_{t-1}}$$

$$= (W_h^f)^T\left\{\frac{\partial L}{\partial C_t} * C_{t-1} * [f_t * (1-f_t)]\right\}$$

$$+ (W_h^i)^T\left\{\frac{\partial L}{\partial C_t} * \tilde{C}_t * [i_t * (1-i_t)]\right\}$$

$$+ (W_h^c)^T\left\{\frac{\partial L}{\partial C_t} * i_t * (1-\tilde{C}_t^2)\right\}$$

$$+ (W_h^o)^T\left\{\frac{\partial L}{\partial h_t} * \tanh(C_t) * [o_t * (1-o_t)]\right\}.$$

2. Derive LSTM backward.

We first derive the easy part. (simple sum of each step)

f_t family

$$\frac{\partial L}{\partial W_x^f} = \sum_{t=1}^{T} \left\{ \frac{\partial L}{\partial c_t} * c_{t-1} * [f_t * (1-f_t)] \right\} x_t^T$$

$$\frac{\partial L}{\partial W_h^f} = \sum_{t=1}^{T} \left\{ \frac{\partial L}{\partial c_t} * c_{t-1} * [f_t * (1-f_t)] \right\} h_{t-1}^T$$

$$\frac{\partial L}{\partial b^f} = \sum_{t=1}^{T} \left\{ \frac{\partial L}{\partial c_t} * c_{t-1} * [f_t * (1-f_t)] \right\}$$

i_t family

$$\frac{\partial L}{\partial W_x^i} = \sum_{t=1}^{T} \left\{ \frac{\partial L}{\partial c_t} * \tilde{c}_t * [i_t * (1-i_t)] \right\} x_t^T$$

$$\frac{\partial L}{\partial W_h^i} = \sum_{t=1}^{T} \left\{ \frac{\partial L}{\partial c_t} * \tilde{c}_t * [i_t * (1-i_t)] \right\} h_{t-1}^T$$

$$\frac{\partial L}{\partial b^i} = \sum_{t=1}^{T} \left\{ \frac{\partial L}{\partial c_t} * \tilde{c}_t * [i_t * (1-i_t)] \right\}$$

$\tilde{c}_t$ family

$$\frac{\partial L}{\partial W_x^c} = \sum_{t=1}^{T} \left[\frac{\partial L}{\partial c_t} * i_t * (1-\tilde{c}_t^2) \right] x_t^T$$

$$\frac{\partial L}{\partial W_h^c} = \sum_{t=1}^{T} \left[\frac{\partial L}{\partial c_t} * i_t * (1-\tilde{c}_t^2) \right] h_{t-1}^T$$

$$\frac{\partial L}{\partial b^c} = \sum_{t=1}^{T} \left[\frac{\partial L}{\partial c_t} * i_t * (1-\tilde{c}_t^2) \right]$$

O_t family

$$\frac{\partial L}{\partial W_x^o} = \sum_{t=1}^{T} \left\{ \frac{\partial L}{\partial h_t} * \tanh(c_t) * [o_t * (1-o_t)] \right\} x_t^T$$

$$\frac{\partial L}{\partial W_h^o} = \sum_{t=1}^{T} \left\{ \frac{\partial L}{\partial h_t} * \tanh(c_t) * [o_t * (1-o_t)] \right\} h_{t-1}^T$$

$$\frac{\partial L}{\partial b^o} = \sum_{t=1}^{T} \left\{ \frac{\partial L}{\partial h_t} * \tanh(c_t) * [o_t * (1-o_t)] \right\}$$

Finally, $\frac{\partial L}{\partial x_t}$ is the same as previous step.

$$\frac{\partial L}{\partial x_t} = (W_x^f)^T \left\{ \frac{\partial L}{\partial c_t} * c_{t-1} * [f_t * (1-f_t)] \right\}$$
$$+ (W_x^i)^T \left\{ \frac{\partial L}{\partial c_t} * c_{t-1} * [f_t * (1-f_t)] \right\}$$
$$+ (W_x^c)^T \left\{ \frac{\partial L}{\partial c_t} * i_t * (1-\tilde{c}_t^2) \right\}$$
$$+ (W_x^o)^T \left\{ \frac{\partial L}{\partial h_t} * \tanh(c_t) * [o_t * (1-o_t)] \right\}$$

$\frac{\partial L}{\partial h_0}$ is the case $t=1$ for $\frac{\partial L}{\partial h_{t-1}}$ in the previous seep.

$$\frac{\partial L}{\partial h_0} = (W_h^f)^T \left\{ \frac{\partial L}{\partial c_1} * c_0 * [f_1 * (1-f_1)] \right\}$$
$$+ (W_h^i)^T \left\{ \frac{\partial L}{\partial c_1} * \tilde{c}_1 * [i_1 * (1-i_1)] \right\}$$
$$+ (W_h^c)^T \left\{ \frac{\partial L}{\partial c_1} * i_1 * (1-\tilde{c}_1^2) \right\}$$
$$+ (W_h^o)^T \left\{ \frac{\partial L}{\partial h_1} * \tanh(c_1) * [o_1 * (1-o_1)] \right\}$$