Recurrent Network

Anonymity
Department of Computer Science, Fudan University
May 12, 2019

Part I Differentiation of LSTM

1.1 Differentiation in one step of LSTM

Before our journey, we should have some preparations.

- First assume that $|h_i| = s$, $|x_{i-1}| = t$, we can easily get that

$$\frac{\partial \mathbf{z}}{\partial \mathbf{h}_{i-1}} = [I_{s\times s}, 0_{s\times t}] \tag{49}$$
$$\frac{\partial \mathbf{z}}{\partial \mathbf{x}_i} = [0_{t\times s}, I_{t\times t}] \tag{50}$$

- Second, since sigmoid and $\tanh$ functions occur frequently in our derivation, so we can get their derivatives.

$$\begin{aligned}\sigma'(x) &= (\frac{1}{1+e^{-x}})' \\ &= \frac{e^{-x}}{(1+e^{-x})^2} = \sigma(x)(1-\sigma(x))\end{aligned} \tag{1}$$
$$tanh'(x) = 1 - tanh^2(x) \tag{2}$$

Let's begin our derivation

- For $\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} &= \frac{\partial \mathbf{o}_t}{\partial \mathbf{h}_{t-1}} * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t))) \frac{\partial C_t}{\partial \mathbf{h}_{t-1}} \\ &= diag(\mathbf{o}_t(1-\mathbf{o}_t))W_o[I_{s\times s}, 0_{s\times t}]^T * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t))) \frac{\partial C_t}{\partial \mathbf{h}_{t-1}}\end{aligned} \tag{3}$$

$$\begin{aligned}\frac{\partial C_t}{\partial \mathbf{h}_{t-1}} &= \frac{\partial \mathbf{f}_t}{\partial \mathbf{h}_{t-1}} * C_{t-1} + \frac{\partial \mathbf{i}_t}{\partial \mathbf{h}_{t-1}} * \overline{C}_t + \mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial \mathbf{h}_{t-1}} \\ &= diag(\mathbf{f}_t(1-\mathbf{f}_t))W_f[I_{s\times s}, 0_{s\times t}]^T * C_{t-1} + diag(\mathbf{i}_t(1-\mathbf{i}_t))W_i[I_{s\times s}, 0_{s\times t}]^T * \overline{C}_t \\ &+ diag(\mathbf{i}_t * (1 - \overline{C}_t{}^2))W_C[I_{s\times s}, 0_{s\times t}]^T\end{aligned} \tag{4}$$

- For $\frac{\partial \mathbf{h}_t}{\partial \mathbf{f}_t}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial \mathbf{f}_t} &= diag(\mathbf{o}_t * (1 - tanh^2(C_t))) \frac{\partial C_t}{\partial \mathbf{f}_t} \\ &= diag(\mathbf{o}_t * (1 - tanh^2(C_t))) diag(C_{t-1})\end{aligned} \tag{5}$$

- For $\frac{\partial \mathbf{h}_t}{\partial \mathbf{i}_t}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial \mathbf{i}_t} &= diag(\mathbf{o}_t * (1 - tanh^2(C_t))) \frac{\partial C_t}{\partial \mathbf{i}_t} \\ &= diag(\mathbf{o}_t * (1 - tanh^2(C_t))) diag(\overline{C}_t)\end{aligned} \tag{6}$$

- For $\frac{\partial \mathbf{h}_t}{\partial \overline{C}_t}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial \overline{C}_t} &= diag(\mathbf{o}_t * (1 - tanh^2(C_t))) \frac{\partial C_t}{\partial \overline{C}_t} \\ &= diag(\mathbf{o}_t * (1 - tanh^2(C_t))) diag(\mathbf{i}_t)\end{aligned} \tag{7}$$

- For $\frac{\partial \mathbf{h}_t}{\partial C_t}$

$$\frac{\partial \mathbf{h}_t}{\partial C_t} = diag(\mathbf{o}_t * (1 - tanh^2(C_t))) \tag{8}$$

- For $\frac{\partial \mathbf{h}_t}{\partial C_{t-1}}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial C_{t-1}} &= diag(\mathbf{o}_t * (1 - tanh^2(C_t))) \frac{\partial C_t}{\partial C_{t-1}} \\ &= diag(\mathbf{o}_t * (1 - tanh^2(C_t))) diag(\mathbf{f}_t)\end{aligned} \tag{9}$$

- For $\frac{\partial \mathbf{h}_t}{\partial \mathbf{o}_t}$

$$\frac{\partial \mathbf{h}_t}{\partial \mathbf{o}_t} = diag(tanh(C_t)) \tag{10}$$

- For $\frac{\partial \mathbf{h}_t}{\partial \mathbf{x}_t}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial \mathbf{x}_t} &= \frac{\partial \mathbf{o}_t}{\partial \mathbf{x}_t} * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t))) \frac{\partial C_t}{\partial \mathbf{x}_t} \\ &= diag(\mathbf{o}_t(1-\mathbf{o}_t))W_o[0_{t\times s}, I_{t\times t}]^T * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t))) \frac{\partial C_t}{\partial \mathbf{x}_t}\end{aligned} \tag{11}$$

$$\begin{aligned}\frac{\partial C_t}{\partial \mathbf{x}_t} &= \frac{\partial \mathbf{f}_t}{\partial \mathbf{x}_t} * C_{t-1} + \frac{\partial \mathbf{i}_t}{\partial \mathbf{x}_t} * \overline{C}_t + \mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial \mathbf{x}_t} \\ &= diag(\mathbf{f}_t(1-\mathbf{f}_t))W_f[0_{t\times s}, I_{t\times t}]^T * C_{t-1} + diag(\mathbf{i}_t(1-\mathbf{i}_t))W_i[0_{t\times s}, I_{t\times t}]^T * \overline{C}_t \\ &+ diag(\mathbf{i}_t) * (1 - \overline{C}_t{}^2)W_C[0_{t\times s}, I_{t\times t}]^T\end{aligned} \tag{12}$$

- For $\frac{\partial \mathbf{h}_t}{\partial W_f}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial W_f} &= \frac{\partial \mathbf{o}_t}{\partial W_f} * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t)))\frac{\partial C_t}{\partial W_f} \\ &= diag(\mathbf{o}_t(1-\mathbf{o}_t))W_o\frac{\partial \mathbf{z}}{\partial W_f} * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t)))(\mathbf{f}_t * \frac{\partial C_{t-1}}{\partial W_f} + \frac{\partial \mathbf{f}_t}{\partial W_f} * C_{t-1} + \frac{\partial \mathbf{i}_t}{\partial W_f} * \overline{C}_t + \mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_f}) \\ &= diag(\mathbf{o}_t(1-\mathbf{o}_t))W_o[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_f} * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t)))(\mathbf{f}_t * \frac{\partial C_{t-1}}{\partial W_f} + \frac{\partial \mathbf{f}_t}{\partial W_f} * C_{t-1} + \frac{\partial \mathbf{i}_t}{\partial W_f} * \overline{C}_t + \mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_f})\end{aligned} \tag{13}$$

Here, we have a closer look at $\frac{\partial \mathbf{f}_t}{\partial W_f} * C_{t-1}$, $\frac{\partial \mathbf{i}_t}{\partial W_f} * \overline{C}_t$, $\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_f}$,

$$\frac{\partial \mathbf{f}_t}{\partial W_f} * C_{t-1} = diag(\mathbf{f}_t(1-\mathbf{f}_t))(diag(\mathbf{z}) + W_f[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_f}) * C_{t-1} \tag{14}$$

$$\frac{\partial \mathbf{i}_t}{\partial W_f} * \overline{C}_t = diag(\mathbf{i}_t(1-\mathbf{i}_t))W_i[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_f} * \overline{C}_{t-1} \tag{15}$$

$$\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_f} = \mathbf{i}_t * diag(1-\overline{C}_t^2)W_C[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_f} \tag{16}$$

- For $\frac{\partial \mathbf{h}_t}{\partial W_i}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial W_i} &= \frac{\partial \mathbf{o}_t}{\partial W_i} * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t)))\frac{\partial C_t}{\partial W_i} \\ &= diag(\mathbf{o}_t(1-\mathbf{o}_t))W_o\frac{\partial \mathbf{z}}{\partial W_i} * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t)))(\mathbf{f}_t * \frac{\partial C_{t-1}}{\partial W_i} + \frac{\partial \mathbf{f}_t}{\partial W_i} * C_{t-1} + \frac{\partial \mathbf{i}_t}{\partial W_i} * \overline{C}_t + \mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_i}) \\ &= diag(\mathbf{o}_t(1-\mathbf{o}_t))W_o[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_i} * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t)))(\mathbf{f}_t * \frac{\partial C_{t-1}}{\partial W_i} + \frac{\partial \mathbf{f}_t}{\partial W_i} * C_{t-1} + \frac{\partial \mathbf{i}_t}{\partial W_i} * \overline{C}_t + \mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_i})\end{aligned} \tag{17}$$

Similarly, for $\frac{\partial \mathbf{f}_t}{\partial W_i} * C_{t-1}$, $\frac{\partial \mathbf{i}_t}{\partial W_i} * \overline{C}_t$, $\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_i}$

$$\frac{\partial \mathbf{f}_t}{\partial W_i} * C_{t-1} = diag(\mathbf{f}_t(1-\mathbf{f}_t))W_f[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_i} * C_{t-1} \tag{18}$$

$$\frac{\partial \mathbf{i}_t}{\partial W_i} * \overline{C}_t = diag(\mathbf{i}_t(1-\mathbf{i}_t))(diag(\mathbf{z}) + W_i[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_i}) * \overline{C}_{t-1} \tag{19}$$

$$\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_i} = \mathbf{i}_t * diag(1-\overline{C}_t^2)W_C[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_i} \tag{20}$$

- For $\frac{\partial \mathbf{h}_t}{\partial W_C}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial W_C} &= \frac{\partial \mathbf{o}_t}{\partial W_C} * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t)))\frac{\partial C_t}{\partial W_C} \\ &= diag(\mathbf{o}_t(1-\mathbf{o}_t))W_o\frac{\partial \mathbf{z}}{\partial W_i} * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t)))(\mathbf{f}_t * \frac{\partial C_{t-1}}{\partial W_C} + \frac{\partial \mathbf{f}_t}{\partial W_C} * C_{t-1} + \frac{\partial \mathbf{i}_t}{\partial W_C} * \overline{C}_t + \mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_C}) \\ &= diag(\mathbf{o}_t(1-\mathbf{o}_t))W_o[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_i} * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t)))(\mathbf{f}_t * \frac{\partial C_{t-1}}{\partial W_C} + \frac{\partial \mathbf{f}_t}{\partial W_C} * C_{t-1} + \frac{\partial \mathbf{i}_t}{\partial W_C} * \overline{C}_t + \mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_C})\end{aligned} \tag{21}$$

Similarly, for $\frac{\partial \mathbf{f}_t}{\partial W_C} * C_{t-1}$, $\frac{\partial \mathbf{i}_t}{\partial W_C} * \overline{C}_t$, $\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_C}$

$$\frac{\partial \mathbf{f}_t}{\partial W_C} * C_{t-1} = diag(\mathbf{f}_t(1-\mathbf{f}_t))W_f[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_C} * C_{t-1} \tag{22}$$

$$\frac{\partial \mathbf{i}_t}{\partial W_C} * \overline{C}_t = diag(\mathbf{i}_t(1-\mathbf{i}_t))W_i[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_C} * \overline{C}_{t-1} \tag{23}$$

$$\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_C} = \mathbf{i}_t * diag(1-\overline{C}_t^2)(diag(\mathbf{z}) + W_C[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_C}) \tag{24}$$

- For $\frac{\partial \mathbf{h}_t}{\partial W_o}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial W_o} &= \frac{\partial \mathbf{o}_t}{\partial W_o} * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t)))\frac{\partial C_t}{\partial W_o} \\ &= diag(\mathbf{o}_t(1-\mathbf{o}_t))(diag(\mathbf{z}) + W_o[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_o}) * tanh(C_t) + diag(\mathbf{o}_t * (1 - tanh^2(C_t)))(\mathbf{f}_t * \frac{\partial C_{t-1}}{\partial W_o} + \frac{\partial \mathbf{f}_t}{\partial W_o} * C_{t-1} + \frac{\partial \mathbf{i}_t}{\partial W_o} * \overline{C}_t + \mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_o})\end{aligned} \tag{25}$$

Similarly, for $\frac{\partial \mathbf{f}_t}{\partial W_o} * C_{t-1}$, $\frac{\partial \mathbf{i}_t}{\partial W_o} * \overline{C}_t$, $\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_o}$

$$\frac{\partial \mathbf{f}_t}{\partial W_o} * C_{t-1} = diag(\mathbf{f}_t(1-\mathbf{f}_t))W_f[I_{s\times s}, 0_{s\times t}]^T\frac{\partial \mathbf{h}_{t-1}}{\partial W_o} * C_{t-1} \tag{26}$$

$$\frac{\partial \mathbf{i}_t}{\partial W_o} * \overline{C}_t = diag(\mathbf{i}_t(1-\mathbf{i}_t))W_i[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial W_o} * \overline{C}_{t-1} \quad (27)$$

$$\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial W_o} = \mathbf{i}_t * diag(1-\overline{C}_t^2)W_C[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial W_o} \quad (28)$$

- For $\frac{\partial \mathbf{h}_t}{\partial b_f}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial b_f} &= \frac{\partial \mathbf{o}_t}{\partial b_f} * tanh(C_t) + diag(\mathbf{o}_t * (1-tanh^2(C_t)))\frac{\partial C_t}{\partial b_f} \\ &= diag(\mathbf{o}_t(1-\mathbf{o}_t))W_o[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_f} * tanh(C_t) + diag(\mathbf{o}_t * (1-tanh^2(C_t)))(\mathbf{f}_t * \frac{\partial C_{t-1}}{\partial b_f} + \frac{\partial \mathbf{f}_t}{\partial b_f} * C_{t-1} + \frac{\partial \mathbf{i}_t}{\partial b_f} * \overline{C}_t + \mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial b_f})\end{aligned} \quad (29)$$

Similarly, for $\frac{\partial \mathbf{f}_t}{\partial b_f} * C_{t-1}$, $\frac{\partial \mathbf{i}_t}{\partial b_f} * \overline{C}_t$, $\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial b_f}$

$$\frac{\partial \mathbf{f}_t}{\partial b_f} * C_{t-1} = diag(\mathbf{f}_t(1-\mathbf{f}_t))(W_f[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_f} + I_{s\times s}) * C_{t-1} \quad (30)$$

$$\frac{\partial \mathbf{i}_t}{\partial b_f} * \overline{C}_t = diag(\mathbf{i}_t(1-\mathbf{i}_t))W_i[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_f} * \overline{C}_{t-1} \quad (31)$$

$$\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial b_f} = \mathbf{i}_t * diag(1-\overline{C}_t^2)W_C[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_f} \quad (32)$$

- For $\frac{\partial \mathbf{h}_t}{\partial b_i}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial b_i} &= \frac{\partial \mathbf{o}_t}{\partial b_i} * tanh(C_t) + diag(\mathbf{o}_t * (1-tanh^2(C_t)))\frac{\partial C_t}{\partial b_i} \\ &= diag(\mathbf{o}_t(1-\mathbf{o}_t))W_o[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_i} * tanh(C_t) + diag(\mathbf{o}_t * (1-tanh^2(C_t)))(\mathbf{f}_t * \frac{\partial C_{t-1}}{\partial b_i} + \frac{\partial \mathbf{f}_t}{\partial b_i} * C_{t-1} + \frac{\partial \mathbf{i}_t}{\partial b_i} * \overline{C}_t + \mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial b_i})\end{aligned} \quad (33)$$

Similarly, for $\frac{\partial \mathbf{f}_t}{\partial b_i} * C_{t-1}$, $\frac{\partial \mathbf{i}_t}{\partial b_i} * \overline{C}_t$, $\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial b_i}$

$$\frac{\partial \mathbf{f}_t}{\partial b_i} * C_{t-1} = diag(\mathbf{f}_t(1-\mathbf{f}_t))W_f[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_i} * C_{t-1} \quad (34)$$

$$\frac{\partial \mathbf{i}_t}{\partial b_i} * \overline{C}_t = diag(\mathbf{i}_t(1-\mathbf{i}_t))(W_i[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_i} + I_{s\times s}) * \overline{C}_{t-1} \quad (35)$$

$$\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial b_i} = \mathbf{i}_t * diag(1-\overline{C}_t^2)W_C[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_i} \quad (36)$$

- For $\frac{\partial \mathbf{h}_t}{\partial b_C}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial b_C} &= \frac{\partial \mathbf{o}_t}{\partial b_C} * tanh(C_t) + diag(\mathbf{o}_t * (1-tanh^2(C_t)))\mathbf{i}_t * \frac{\partial C_t}{\partial b_C} \\ &= diag(\mathbf{o}_t(1-\mathbf{o}_t))W_o[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_i} * tanh(C_t) + diag(\mathbf{o}_t * (1-tanh^2(C_t)))(\mathbf{f}_t * \frac{\partial C_{t-1}}{\partial b_i} + \frac{\partial \mathbf{f}_t}{\partial b_i} * C_{t-1} + \frac{\partial \mathbf{i}_t}{\partial b_i} * \overline{C}_t + \mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial b_i})\end{aligned} \quad (37)$$

Similarly, for $\frac{\partial \mathbf{f}_t}{\partial b_C} * C_{t-1}$, $\frac{\partial \mathbf{i}_t}{\partial b_C} * \overline{C}_t$, $\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial b_C}$

$$\frac{\partial \mathbf{f}_t}{\partial b_C} * C_{t-1} = diag(\mathbf{f}_t(1-\mathbf{f}_t))W_f[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_C} * C_{t-1} \quad (38)$$

$$\frac{\partial \mathbf{i}_t}{\partial b_C} * \overline{C}_t = diag(\mathbf{i}_t(1-\mathbf{i}_t))W_i[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_C} * \overline{C}_{t-1} \quad (39)$$

$$\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial b_C} = \mathbf{i}_t * diag(1-\overline{C}_t^2)(W_C[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_C} + I_{s\times s}) \quad (40)$$

- For $\frac{\partial \mathbf{h}_t}{\partial b_o}$

$$\begin{aligned}\frac{\partial \mathbf{h}_t}{\partial b_o} &= \frac{\partial diag(\mathbf{o}_t}{\partial b_o} * tanh(C_t) + \mathbf{o}_t * (1-tanh^2(C_t)))\frac{\partial C_t}{\partial b_o} \\ &= diag(\mathbf{o}_t(1-\mathbf{o}_t))(W_o[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_o} + I_{s\times s}) * tanh(C_t) + diag(\mathbf{o}_t * (1-tanh^2(C_t)))(\mathbf{f}_t * \frac{\partial C_{t-1}}{\partial b_o} + \frac{\partial \mathbf{f}_t}{\partial b_o} * C_{t-1} + \frac{\partial \mathbf{i}_t}{\partial b_o} * \overline{C}_t + \mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial b_o})\end{aligned} \quad (41)$$

Similarly, for $\frac{\partial \mathbf{f}_t}{\partial b_o} * C_{t-1}$, $\frac{\partial \mathbf{i}_t}{\partial b_o} * \overline{C}_t$, $\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial b_o}$

$$\frac{\partial \mathbf{f}_t}{\partial b_o} * C_{t-1} = diag(\mathbf{f}_t(1-\mathbf{f}_t))W_f[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_o} * C_{t-1} \tag{42}$$

$$\frac{\partial \mathbf{i}_t}{\partial b_o} * \overline{C}_t = diag(\mathbf{i}_t(1-\mathbf{i}_t))W_i[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_o} * \overline{C}_{t-1} \tag{43}$$

$$\mathbf{i}_t * \frac{\partial \overline{C}_t}{\partial b_o} = \mathbf{i}_t * diag(1-\overline{C}_t^2)W_C[I_{s\times s}, 0_{s\times t}]^T \frac{\partial \mathbf{h}_{t-1}}{\partial b_o} \tag{44}$$

1.2 Backpropagation Through Time

1.2.1 Motivation

The goal of the BPTT is to modify the weights of a neural network in order to minimize the error of the network outputs compared to some expected output in response to corresponding inputs.

1.2.2 Procedure

- Present a sequence of timesteps of input and output pairs to the network.
- Unroll the network then calculate and accumulate errors across each timestep.
- Roll-up the network and update weights.
- Repeat.

1.2.3 Specification in LSTM

To make things easier, we define our loss function to be the cross entropy, given by

$$E_t(\mathbf{h}_t, \widehat{\mathbf{h}_t}) = -\sum \widehat{\mathbf{h}_t} \log(\mathbf{h}_t) \tag{45}$$

$$\begin{aligned} E(\mathbf{h}, \widehat{\mathbf{h}}) &= \sum_t E_t(\mathbf{h}_t, \widehat{\mathbf{h}_t}) \\ &= -\sum_t \sum \widehat{\mathbf{h}_t} \log(\mathbf{h}_t) \end{aligned} \tag{46}$$

Here, $\widehat{\mathbf{y}_t}$ is the correct word at time step t, and $\mathbf{y}_t$ is our prediction. We typically treat the full sequence as one training example, so the total error is just the sum of the errors at each time step.

Remember our goal is to calculate the gradients of the error with respect of our parameters W and b, and then learn good parameters using the gradient. Just like we sum up the errors, we also sum up the gradients at each time step for one training example $\frac{\partial E}{\partial W} = \sum_t \frac{\partial E_t}{\partial W}$, we can use the chain rule to calculate these gradients and just.

To further simplify the derivation, we only take $\frac{\partial E_3}{\partial W_f}$ as an example. And before reaching to our result, we can clearly observe that

$$\begin{aligned} \mathbf{h}_t &= \mathbf{o}_t * tanh(C_t) \\ &= \sigma(W_o\mathbf{z} + b_t) * tanh(\mathbf{f}_t * C_{t-1} + \mathbf{i}_t * \overline{C}_t) \\ &= \sigma(W_o[\mathbf{h}_{t-1}, \mathbf{x}_t] + b_t) * tanh(\sigma(W_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + b_f) * C_{t-1} + \sigma(W_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + b_i) * tanh(W_C[\mathbf{h}_{t-1}, z] + b_C)) \end{aligned} \tag{47}$$

So we can conclude that $\mathbf{h}_t = \mathbf{h}_t(\mathbf{h}_{t-1}, W, b)$. In this case, we can reach to the derivation below

$$\begin{aligned}\frac{\partial E_3}{\partial W_f} &= \frac{\partial E_3}{\partial \mathbf{h}_3}\frac{\partial \mathbf{h}_3}{\partial W_f} \\ &= \sum_{k=0}^{3}\frac{\partial E_3}{\partial \mathbf{h}_3}\frac{\partial \mathbf{h}_3}{\partial \mathbf{h}_k}\frac{\partial \mathbf{h}_k}{\partial W_f}\end{aligned} \tag{48}$$

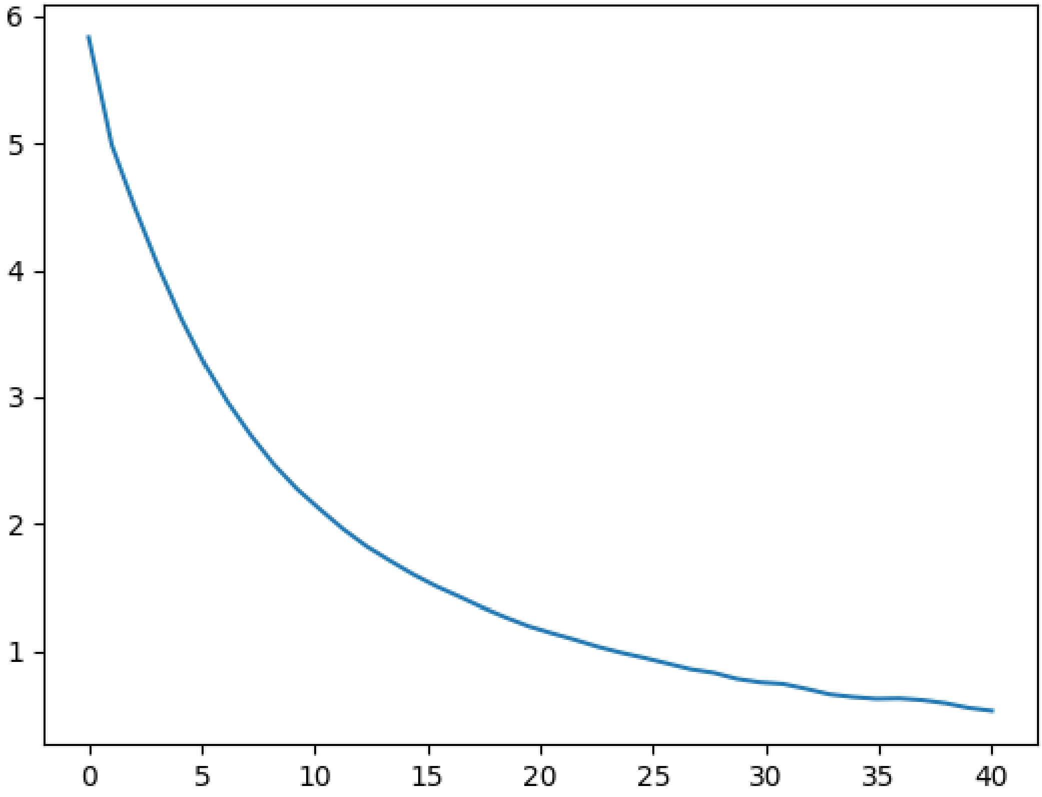
Part II Autograd Training of LSTM

2.1 Initialization

The embedding layer and the parameters for the model should not be initialized to zero. If not, the output of $\mathbf{f}_t$ and $\mathbf{i}_t$ is equal and their corresponding gradient is equal too, therefore, contributing to the same update, which will go on for all the n iterations we run. All the factors will lead to symmetric ways which will dramatically lower the ability of the model to capture the underlying pattern. If the parameters is initialized too large, it may cause another problems, for example, the vanishing gradient for Sigmoid function. A heuristics way is to scale our random normal weight initializations by $\frac{1}{\sqrt{n}}$ in order to have unit variance which will greatly help convergence during training.

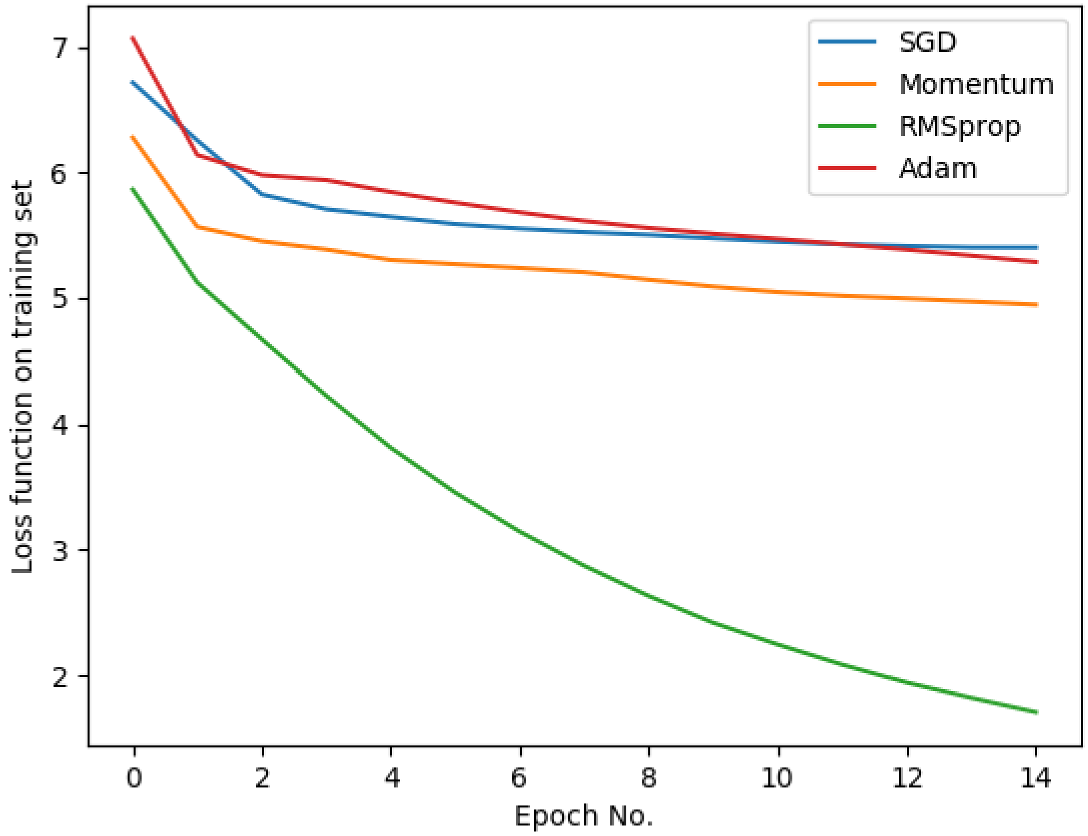
2.2 Training

To ease my computer, I only run 40 epochs on near 1000 poems using RMSprop optimization with 16 poems per batch and my perplexity on development dataset is 9.39.



2.3 Optimization

In this section, I only compare 4 different kinds of optimization : SGD, Momentum, RMSprop, Adam, and clearly we can see that RMSprop stand out from them in terms of rate of convergence and result.



Acknowledgement

During my journal, I've referred to many excellent blogs to finish my work, including but not limited within below.

- I began to understand LSTM through Colah's introduction
- To finish the derivation of BPTT, I referred WILDML's Blog
- And for initialization of parameters for the model, I've referred to Zuo bai Zhang's work, and details are also available in Andrew Ng's course on Coursera

3 Appendix

Poems generated from the training

- 日暮驚沙亂雪飛，傍垂相與銷殘枝。既修昭持，垂涕念生還。
- 山桃紅數莫折伊。看花子萬同在一行歸望，樓日月龍悲移人。
- 夜靜聽天生。今朝韓信是眼關月，黃金枝映洛陽橋。

- 湖天上皇休，年度寒浪不移。
- 海底飛塵終有日，永安宮外最年不歎覺心。
- 月落轅門鼓角鳴，玉枕終年對離別。陳觴仙衛邕，此昭眠時餘。