A Note on BPTT for LSTM LM

Tomonari MASADA @ Nagasaki University

January 15, 2015

1 Forward pass

K is the vocabulary size. N is the number of hidden layers. D_n is the number of hidden units at the nth layer. The input sequence is $\boldsymbol{X} \equiv (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$, where each $\boldsymbol{x}_t \equiv (x_{t1}, \ldots, x_{tK})^\top$ is a one-hot vector.

- ι : input gates
- ϕ : forget gates
- η : cells
- ω : output gates
- h : cell outputs

f and g are sigmoid functions.

$$\begin{aligned}
\boldsymbol{\iota}_t^n &= \boldsymbol{f}(\boldsymbol{W}_{x\iota}^n \boldsymbol{x}_t + \boldsymbol{W}_{h^-\iota}^n \boldsymbol{h}_t^{n-1} + \boldsymbol{W}_{h\iota}^n \boldsymbol{h}_{t-1}^n + \boldsymbol{w}_{\eta\iota}^n \odot \boldsymbol{\eta}_{t-1}^n + \boldsymbol{b}_\iota^n) \equiv \boldsymbol{f}(\boldsymbol{a}_{\iota,t}^n) && (1)\\
\boldsymbol{\phi}_t^n &= \boldsymbol{f}(\boldsymbol{W}_{x\phi}^n \boldsymbol{x}_t + \boldsymbol{W}_{h^-\phi}^n \boldsymbol{h}_t^{n-1} + \boldsymbol{W}_{h\phi}^n \boldsymbol{h}_{t-1}^n + \boldsymbol{w}_{\eta\phi}^n \odot \boldsymbol{\eta}_{t-1}^n + \boldsymbol{b}_\phi^n) \equiv \boldsymbol{f}(\boldsymbol{a}_{\phi,t}^n) && (2)\\
\boldsymbol{\eta}_t^n &= \boldsymbol{\phi}_t^n \odot \boldsymbol{\eta}_{t-1}^n + \boldsymbol{\iota}_t^n \odot \boldsymbol{g}(\boldsymbol{W}_{x\eta}^n \boldsymbol{x}_t + \boldsymbol{W}_{h^-\eta}^n \boldsymbol{h}_t^{n-1} + \boldsymbol{W}_{h\eta}^n \boldsymbol{h}_{t-1}^n + \boldsymbol{b}_\eta^n) \equiv \boldsymbol{\phi}_t^n \odot \boldsymbol{\eta}_{t-1}^n + \boldsymbol{\iota}_t^n \odot \boldsymbol{g}(\boldsymbol{a}_{\eta,t}^n) && (3)\\
\boldsymbol{\omega}_t^n &= \boldsymbol{f}(\boldsymbol{W}_{x\omega}^n \boldsymbol{x}_t + \boldsymbol{W}_{h^-\omega}^n \boldsymbol{h}_t^{n-1} + \boldsymbol{W}_{h\omega}^n \boldsymbol{h}_{t-1}^n + \boldsymbol{w}_{\eta\omega}^n \odot \boldsymbol{\eta}_t^n + \boldsymbol{b}_\omega^n) \equiv \boldsymbol{f}(\boldsymbol{a}_{\omega,t}^n) && (4)\\
\boldsymbol{h}_t^n &= \boldsymbol{\omega}_t^n \odot \boldsymbol{g}(\boldsymbol{\eta}_t^n) \;, && (5)
\end{aligned}$$

where $\odot$ is the element-wise product. The superscript n means the n-th layer.

These can be rewritten for $d = 1, \ldots, D_n$ separately:

$$\begin{aligned}
\iota_{t,d}^n &= f(\boldsymbol{w}_{x\iota,d}^{n\top} \boldsymbol{x}_t + \boldsymbol{w}_{h^-\iota,d}^{n\top} \boldsymbol{h}_t^{n-1} + \boldsymbol{w}_{h\iota,d}^{n\top} \boldsymbol{h}_{t-1}^n + w_{\eta\iota,d}^n \eta_{t-1,d}^n + b_{\iota,d}^n) \equiv f(a_{\iota,t,d}^n) && (6)\\
\phi_{t,d}^n &= f(\boldsymbol{w}_{x\phi,d}^{n\top} \boldsymbol{x}_t + \boldsymbol{w}_{h^-\phi,d}^{n\top} \boldsymbol{h}_t^{n-1} + \boldsymbol{w}_{h\phi,d}^{n\top} \boldsymbol{h}_{t-1}^n + w_{\eta\phi,d}^n \eta_{t-1,d}^n + b_{\phi,d}^n) \equiv f(a_{\phi,t,d}^n) && (7)\\
\eta_{t,d}^n &= \phi_{t,d}^n \eta_{t-1,d}^n + \iota_{t,d}^n g(\boldsymbol{w}_{x\eta,d}^{n\top} \boldsymbol{x}_t + \boldsymbol{w}_{h^-\eta,d}^{n\top} \boldsymbol{h}_t^{n-1} + \boldsymbol{w}_{h\eta,d}^{n\top} \boldsymbol{h}_{t-1}^n + b_{\eta,d}^n) \equiv \phi_{t,d}^n \eta_{t-1,d}^n + \iota_{t,d}^n g(a_{\eta,t,d}^n) && (8)\\
\omega_{t,d}^n &= f(\boldsymbol{w}_{x\omega,d}^{n\top} \boldsymbol{x}_t + \boldsymbol{w}_{h^-\omega,d}^{n\top} \boldsymbol{h}_t^{n-1} + \boldsymbol{w}_{h\omega,d}^{n\top} \boldsymbol{h}_{t-1}^n + w_{\eta\omega,d}^n \eta_{t,d}^n + b_{\omega,d}^n) \equiv f(a_{\omega,t,d}^n) && (9)\\
h_{t,d}^n &= \omega_{t,d}^n g(\eta_{t,d}^n) \;. && (10)
\end{aligned}$$

Word probabilities for each t are obtained as follows:

$$\hat{\boldsymbol{y}}_t = \boldsymbol{b}_y + \sum_{n=1}^N \boldsymbol{W}_{hy}^n \boldsymbol{h}_t^n \;,\;\; \boldsymbol{y}_t \equiv \frac{\exp(\hat{\boldsymbol{y}}_{t,k})}{\sum_{k'=1}^K \exp(\hat{\boldsymbol{y}}_{t,k'})} \;. \tag{11}$$

These can be rewritten for $k = 1, \ldots, K$ separately:

$$\hat{y}_{t,k} = b_{y,k} + \sum_{n=1}^N \sum_{d=1}^{D_n} w_{hy,kd}^n h_{t,d}^n \;,\;\; y_{t,k} \equiv \frac{\exp(\hat{y}_{t,k})}{\sum_{k'=1}^K \exp(\hat{y}_{t,k'})} \;. \tag{12}$$

2 Negative log likelihood

$$\mathcal{L}(\boldsymbol{X}) = -\sum_{t=1}^{T} \log y_{t,x_{t+1}} = -\sum_{t=1}^{T} \log \frac{\exp(\hat{y}_{t,x_{t+1}})}{\sum_{k=1}^{K} \exp(\hat{y}_{t,k})} = -\sum_{t=1}^{T} \hat{y}_{t,x_{t+1}} + \sum_{t=1}^{T} \log \Big\{ \sum_{k=1}^{K} \exp(\hat{y}_{t,k}) \Big\} \tag{13}$$

3 Backward pass

When $f(a) = \frac{1}{1+\exp(-a)}$, $f'(a) = f(a)(1 - f(a))$. When $f(a) = \tanh(a)$, $f'(a) = 1 - f(a)^2$.

3.1

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \hat{y}_{t,k}} &= -\frac{\partial \hat{y}_{t,x_{t+1}}}{\partial \hat{y}_{t,k}} + \frac{\partial}{\partial \hat{y}_{t,k}} \log \Big\{ \sum_{\bar{k}=1}^{K} \exp(\hat{y}_{t,\bar{k}}) \Big\} = -\frac{\partial \hat{y}_{t,x_{t+1}}}{\partial \hat{y}_{t,k}} + \frac{\sum_{\bar{k}=1}^{K} \exp(\hat{y}_{t,\bar{k}}) \frac{\partial \hat{y}_{t,\bar{k}}}{\partial \hat{y}_{t,k}}}{\sum_{k=1}^{K} \exp(\hat{y}_{t,k})} \\ &= -\Big\{ \delta(x_{t+1} = k) - y_{t,k} \Big\}\end{aligned} \tag{14}$$

For $w^n_{hy,kd}$,

$$\frac{\partial \hat{y}_{t,k}}{\partial w^n_{hy,kd}} = \frac{\partial b_{y,k}}{\partial w^n_{hy,kd}} + \sum_{n=1}^{N} \sum_{\bar{d}=1}^{D_n} \left(h^n_{t,\bar{d}} \frac{\partial w^n_{hy,k\bar{d}}}{\partial w^n_{hy,kd}} + w^n_{hy,k\bar{d}} \frac{\partial h^n_{t,\bar{d}}}{\partial w^n_{hy,kd}} \right) = h^n_{t,d} \tag{15}$$

$$\therefore \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial w^n_{hy,kd}} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \hat{y}_{t,k}} \frac{\partial \hat{y}_{t,k}}{\partial w^n_{hy,kd}} = -\sum_{t=1}^{T} h^n_{t,d} \Big\{ \delta(x_{t+1} = k) - y_{t,k} \Big\} \tag{16}$$

3.2 Output errors

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial h^n_{t,d}} &= \sum_{k=1}^{K} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \hat{y}_{t,k}} \frac{\partial \hat{y}_{t,k}}{\partial h^n_{t,d}} + \sum_{\bar{d}=1}^{D_n} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\omega,t+1,\bar{d}}} \frac{\partial a^n_{\omega,t+1,\bar{d}}}{\partial h^n_{t,d}} + \sum_{\bar{d}=1}^{D_n} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\eta,t+1,\bar{d}}} \frac{\partial a^n_{\eta,t+1,\bar{d}}}{\partial h^n_{t,d}} \\ &\quad + \sum_{\bar{d}=1}^{D_n} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\phi,t+1,\bar{d}}} \frac{\partial a^n_{\phi,t+1,\bar{d}}}{\partial h^n_{t,d}} + \sum_{\bar{d}=1}^{D_n} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\iota,t+1,\bar{d}}} \frac{\partial a^n_{\iota,t+1,\bar{d}}}{\partial h^n_{t,d}} \\ &\quad + \sum_{\bar{d}=1}^{D_{n+1}} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^{n+1}_{\omega,t,\bar{d}}} \frac{\partial a^{n+1}_{\omega,t,\bar{d}}}{\partial h^n_{t,d}} + \sum_{\bar{d}=1}^{D_{n+1}} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^{n+1}_{\eta,t,\bar{d}}} \frac{\partial a^{n+1}_{\eta,t,\bar{d}}}{\partial h^n_{t,d}} \\ &\quad + \sum_{\bar{d}=1}^{D_{n+1}} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^{n+1}_{\phi,t,\bar{d}}} \frac{\partial a^{n+1}_{\phi,t,\bar{d}}}{\partial h^n_{t,d}} + \sum_{\bar{d}=1}^{D_{n+1}} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^{n+1}_{\iota,t,\bar{d}}} \frac{\partial a^{n+1}_{\iota,t,\bar{d}}}{\partial h^n_{t,d}} \\ &= \sum_{k=1}^{K} w^n_{hy,kd} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \hat{y}_{t,k}} + \sum_{\bar{d}=1}^{D_n} w^n_{h\omega,\bar{d}d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\omega,t+1,\bar{d}}} + \sum_{\bar{d}=1}^{D_n} w^n_{h\eta,\bar{d}d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\eta,t+1,\bar{d}}} \\ &\quad + \sum_{\bar{d}=1}^{D_n} w^n_{h\phi,\bar{d}d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\phi,t+1,\bar{d}}} + \sum_{\bar{d}=1}^{D_n} w^n_{h\iota,\bar{d}d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\iota,t+1,\bar{d}}} \\ &\quad + \sum_{\bar{d}=1}^{D_{n+1}} w^{n+1}_{h^-\omega,\bar{d}d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^{n+1}_{\omega,t,\bar{d}}} + \sum_{\bar{d}=1}^{D_{n+1}} w^{n+1}_{h^-\eta,\bar{d}d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^{n+1}_{\eta,t,\bar{d}}} \\ &\quad + \sum_{\bar{d}=1}^{D_{n+1}} w^{n+1}_{h^-\phi,\bar{d}d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^{n+1}_{\phi,t,\bar{d}}} + \sum_{\bar{d}=1}^{D_{n+1}} w^{n+1}_{h^-\iota,\bar{d}d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^{n+1}_{\iota,t,\bar{d}}}\end{aligned} \tag{17}$$

3.3 Output gates

$$\frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\omega,t,d}} = \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial h^n_{t,d}} \frac{\partial h^n_{t,d}}{\partial a^n_{\omega,t,d}} \tag{18}$$

$$\frac{\partial h^n_{t,d}}{\partial a^n_{\omega,t,d}} = g(\eta^n_{t,d}) f'(a^n_{\omega,t,d}) \tag{19}$$

3.4 States

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \eta^n_{t,d}} &= \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial h^n_{t,d}} \frac{\partial h^n_{t,d}}{\partial \eta^n_{t,d}} + \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \eta^n_{t+1,d}} \frac{\partial \eta^n_{t+1,d}}{\partial \eta^n_{t,d}} + \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\phi,t+1,d}} \frac{\partial a^n_{\phi,t+1,d}}{\partial \eta^n_{t,d}} + \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\iota,t+1,d}} \frac{\partial a^n_{\iota,t+1,d}}{\partial \eta^n_{t,d}} \\
&= \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial h^n_{t,d}} \frac{\partial h^n_{t,d}}{\partial \eta^n_{t,d}} + \phi^n_{t+1,d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \eta^n_{t+1,d}} + w_{\eta\phi,d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\phi,t+1,d}} + w_{\eta\iota,d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\iota,t+1,d}}
\end{aligned} \tag{20}$$

$$\begin{aligned}
\frac{\partial h^n_{t,d}}{\partial \eta^n_{t,d}} &= g(\eta^n_{t,d}) \frac{\partial \omega^n_{t,d}}{\partial \eta^n_{t,d}} + \omega^n_{t,d} g'(\eta^n_{t,d}) \\
&= w^n_{\eta\omega,d} f'(a^n_{\omega,t,d}) g(\eta^n_{t,d}) + \omega^n_{t,d} g'(\eta^n_{t,d}) \\
&= w^n_{\eta\omega,d} \frac{\partial h^n_{t,d}}{\partial a^n_{\omega,t,d}} + \omega^n_{t,d} g'(\eta^n_{t,d})
\end{aligned} \tag{21}$$

$$\begin{aligned}
\therefore \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \eta^n_{t,d}} &= \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial h^n_{t,d}} \left\{ w^n_{\eta\omega,d} \frac{\partial h^n_{t,d}}{\partial a^n_{\omega,t,d}} + \omega^n_{t,d} g'(\eta^n_{t,d}) \right\} + \phi^n_{t+1,d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \eta^n_{t+1,d}} + w_{\eta\phi,d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\phi,t+1,d}} + w_{\eta\iota,d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\iota,t+1,d}} \\
&= w^n_{\eta\omega,d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\omega,t,d}} + \omega^n_{t,d} g'(\eta^n_{t,d}) \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial h^n_{t,d}} + \phi^n_{t+1,d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \eta^n_{t+1,d}} + w_{\eta\phi,d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\phi,t+1,d}} + w_{\eta\iota,d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\iota,t+1,d}}
\end{aligned} \tag{22}$$

3.5 Cells, forget gates, and input gates

$$\frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\eta,t,d}} = \frac{\partial \eta^n_{t,d}}{\partial a^n_{\eta,t,d}} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \eta^n_{t,d}} = \iota^n_{t,d} g'(a^n_{\eta,t,d}) \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \eta^n_{t,d}} \tag{23}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\phi,t,d}} = \frac{\partial \eta^n_{t,d}}{\partial a^n_{\phi,t,d}} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \eta^n_{t,d}} = f'(a^n_{\phi,t,d}) \eta^n_{t-1,d} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \eta^n_{t,d}} \tag{24}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial a^n_{\iota,t,d}} = \frac{\partial \eta^n_{t,d}}{\partial a^n_{\iota,t,d}} \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \eta^n_{t,d}} = f'(a^n_{\iota,t,d}) g(\eta^n_{t,d}) \frac{\partial \mathcal{L}(\boldsymbol{X})}{\partial \eta^n_{t,d}} \tag{25}$$

References

[1] Thomas Breuel, Volkmar Frinken, and Marcus Liwicki. Long Short-Term Memory Managing Long-Term Dependencies within Sequences. `nntutorial-lstm.pdf` at `http://lstm.iupr.com/files`

[2] Alex Graves and J urgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18 (2005) 602–610.

[3] Alex Graves. Generating Sequences With Recurrent Neural Networks. *arXiv preprint arXiv:*/1308.0850, 2013.