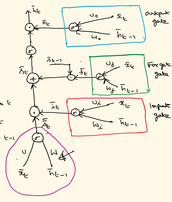
LSTMs

Wednesday, 15 April 2020 09:24

Long Short-Term Memory (LSTM) Networks

LSTM Layer:



σ : Sigmoid

$\bar{c}_t$: cell state at t

$\bar{c}_{t-1}$: cell state at t-1

$\bar{h}_t$: Temporary state at t

$\bar{h}_t$: Layer state at t

$\bar{h}_{t-1}$: Layer state at t-1

4J recurrent

4U weights

4 W matrix

$U_{d \times J}$

$W_{J \times J}$

Selective Read:

Temporary state $\quad \tilde{h}_t = \sigma(W_h \bar{h}_{t-1} + U_h \bar{x}_t + \bar{b}_h)$ tanh as in RNN

Input gate $\quad \bar{i}_t = \sigma(W_i \bar{h}_{t-1} + U_i \bar{x}_t + \bar{b}_i)$

Selective Forget:

Forget gate $\quad \bar{f}_t = \sigma(W_f \bar{h}_{t-1} + U_f \bar{x}_t + \bar{b}_f)$

cell state $\quad \bar{c}_t = \underbrace{\bar{c}_{t-1} \odot \bar{f}_t}_{\text{selective forget}} + \underbrace{\bar{i}_t \odot \tilde{h}_t}_{\text{selective read}}$

$\odot$ Element-wise multiplication

Selective Write:

Output gate $\quad \bar{o}_t = \sigma(W_o \bar{h}_{t-1} + U_o \bar{x}_t + \bar{b}_o)$

Layer state $\quad \bar{h}_t = \underbrace{\sigma(\bar{c}_t)}_{\text{selective write}} \odot \bar{o}_t$

Parameter updates in LSTM:

$$\Delta w_{jk}^{L} = -\eta \frac{\partial \bar{E}_t}{\partial w_{jk}}$$

$$\Delta w_{jk}^{L} = \eta \left(\sum_{t=1}^{N} \gamma_{kt} \bar{h}_t^{L} \right) \cdot \frac{\partial h_{kt}}{\partial w_{jk}}$$

$$\Delta w_{jk}^{L} = -\eta \frac{\partial \bar{E}_t}{\partial w_{jk}}$$

$$= \frac{\partial}{\partial w_{jk}} \left[\frac{1}{2} \sum_{k=1}^{N} \left(\bar{h}_{tk} - y_{tk} \right)^2 \right]$$

$$= \eta \sum_{t=1}^{N} \left(\bar{h}_{tk} - y_{tk} \right) \cdot \frac{\partial \sigma(a_{tk})}{\partial a_{tk}} \cdot v_{jk} \cdot \frac{\partial h_{tk}}{\partial w_{jk}}$$

$$= \eta \left(\sum_{t=1}^{N} \gamma_{kt} \bar{h}_t^{L} \right) \cdot \frac{\partial h_{tk}}{\partial w_{jk}}$$

$$a_{tk} = \sum_{j=1}^{M} v_{kj} \cdot h_{tj} + b_{tk}$$

Dependency graph:

$$\frac{\partial h_{tk}}{\partial w_{jk}}$$

$$h_{tk} = o_{tk} \cdot \sigma(c_{tk})$$

$$\frac{\partial h_{tk}}{\partial w_{jk}} = o_{tk} \cdot \frac{\partial \sigma(c_{tk})}{\partial w_{jk}} + \sigma(c_{tk}) \cdot \frac{\partial o_{tk}}{\partial w_{jk}}$$

$$\frac{\partial \sigma(c_{tk})}{\partial w_{jk}} = \frac{\partial \sigma(c_{tk})}{\partial c_{tk}} \cdot \frac{\partial c_{tk}}{\partial w_{jk}} \quad \text{— (1)}$$

$$\frac{\partial c_{tk}}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}} \left[f_{tk} \cdot c_{t-1,k} + i_{tk} \cdot \tilde{h}_{tk} \right]$$

$$= f_{tk} \cdot \frac{\partial c_{t-1,k}}{\partial w_{jk}} + c_{t-1,k} \cdot \frac{\partial f_{tk}}{\partial w_{jk}}$$

$$+ i_{tk} \cdot \frac{\partial \tilde{h}_{tk}}{\partial w_{jk}} + \tilde{h}_{tk} \cdot \frac{\partial i_{tk}}{\partial w_{jk}} \quad \text{— (2)}$$

$$\frac{\partial o_{tk}}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}} \left[\sum_{m=1}^{J} w_{omk} \cdot h_{t-1,m} + \sum_{l=1}^{d} u_{olk} x_{tl} \right]$$

$$= \frac{\partial \sigma(a_{otk})}{\partial a_{otk}} \cdot \frac{\partial a_{otk}}{\partial w_{jk}}$$

$$\frac{\partial o_{tk}}{\partial w_{jk}} = \frac{\partial \sigma(a_{otk})}{\partial a_{otk}} \cdot w_{jkk} \cdot \frac{\partial h_{t-1,k}}{\partial w_{jk}} \quad \text{— (3)}$$

$$\frac{\partial f_{tk}}{\partial w_{jk}} = \frac{\partial \sigma(a_{ftk})}{\partial a_{ftk}} \cdot w_{jkk} \cdot \frac{\partial h_{t-1,k}}{\partial w_{jk}} \quad \text{— (4)}$$

$$\frac{\partial i_{tk}}{\partial w_{jk}} = \frac{\partial \sigma(a_{itk})}{\partial a_{itk}} \cdot w_{jkk} \cdot \frac{\partial h_{t-1,k}}{\partial w_{jk}} \quad \text{— (5)}$$

$$\frac{\partial \tilde{h}_{tk}}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}} \sigma(\tilde{a}_{tk})$$

$$= \frac{\partial \sigma(\tilde{a}_{tk})}{\partial \tilde{a}_{tk}} \cdot \frac{\partial \tilde{a}_{tk}}{\partial w_{jk}}$$

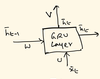
$$\tilde{a}_{tk} = \sum_{m=1}^{J} w_{mk} \cdot h_{t-1,m} + \sum_{n=1}^{d} u_{nk} x_{tn}$$

$$\frac{\partial \tilde{a}_{tk}}{\partial w_{jk}} = h_{t-1,j} + w_{jk} \cdot \frac{\partial h_{t-1,j}}{\partial w_{jk}}$$

$$\frac{\partial \tilde{h}_{tk}}{\partial w_{jk}} = \left[h_{t-1,j} + w_{jk} \cdot \frac{\partial h_{t-1,j}}{\partial w_{jk}} \right] \cdot \frac{\partial \sigma(\tilde{a}_{tk})}{\partial \tilde{a}_{tk}} \quad \text{— (6)}$$

Gated Recurrent Unit (GRU)

- No forget gate
- No cell state

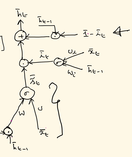


$$\bar{r}_t = \sigma(W_r \bar{h}_{t-1} + U_r \bar{x}_t + \bar{b}_r)$$

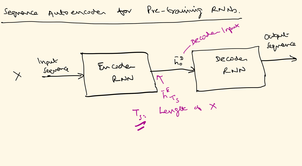
$$\bar{z}_t = \sigma(W_z \bar{h}_{t-1} + U_z \bar{x}_t + \bar{b}_z)$$

$$\tilde{h}_t = \sigma\left[W (\bar{r}_t \odot \bar{h}_{t-1}) + U \cdot \bar{x}_t + \bar{b}_h \right]$$

$$\bar{h}_t = \bar{z}_t \odot \tilde{h}_t + \underbrace{(\bar{1} - \bar{z}_t) \cdot \bar{h}_{t-1}}_{\bar{h}_t}$$



Sequence Autoencoder for Pre-training RNNs.



Metrics for evaluation of text generation system

- Captioning
- Machine translation

Measure similarity of a ground truth sentence and a sentence generated

1. Bilingual Evaluation Understudy (BLEU)
   - Presence of common n-grams in two sentences
2. Metric for Evaluation of Translation with Explicit Ordering (METEOR)
3. Recall Oriented Understudy of Gisting Evaluation (ROUGE)
4. Consensus-based Image Description Evaluation (CIDEr)

n-gram : Group of adjacent n words

n = 1

I am born Chennai

He is born Delhi

She is born Bangalore

n = 2

BLEU - 1

- 2
- 3
- 4