

Thesis intermediate Presentation

Moritz Wolter

August 7, 2016

$$B(f) = 1125 \ln(1 + f/700) \quad (1)$$

$$H_m = 0 \quad \text{if } k < f[m-1] \quad (2)$$

$$H_m = \frac{k - f[m-1]}{f[m] - f[m-1]} \quad \text{if } f[m-1] \leq k \leq f[m] \quad (3)$$

$$H_m = \frac{f[m+1] - k}{f[m+1] - f[m]} \quad \text{if } f[m] \leq k \leq f[m+1] \quad (4)$$

$$H_m = 0 \quad \text{if } k > f[m+1] \quad (5)$$

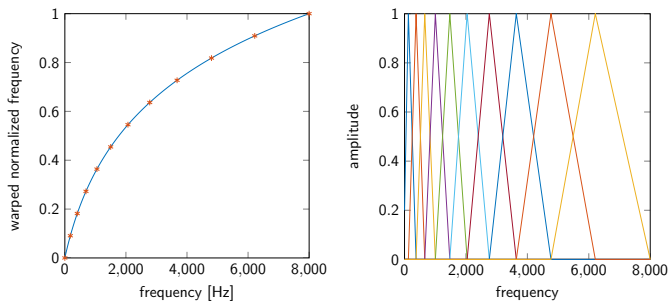


Figure 1 : The Mel scale and mel banks.

Long Short Term Memory (LSTM)

$$\mathbf{i}_t = \sigma(\mathbf{W}_i[\mathbf{x}_t \mathbf{h}_{t-1} \mathbf{c}_{t-1}]^T + \mathbf{b}_i) \quad (6)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f[\mathbf{x}_t \mathbf{h}_{t-1} \mathbf{c}_{t-1}]^T + \mathbf{b}_f) \quad (7)$$

$$\mathbf{c}_t = \mathbf{f}_t \mathbf{c}_{t-1} + \mathbf{i}_t \tanh(\mathbf{W}_c[\mathbf{x}_t \mathbf{h}_{t-1}]^T + \mathbf{b}_c) \quad (8)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o[\mathbf{x}_t \mathbf{h}_{t-1} \mathbf{c}_t]^T + \mathbf{b}_o) \quad (9)$$

$$\mathbf{h}_t = \mathbf{o}_t \tanh(\mathbf{c}_t) \quad (10)$$

Long Short Term Memory (LSTM)

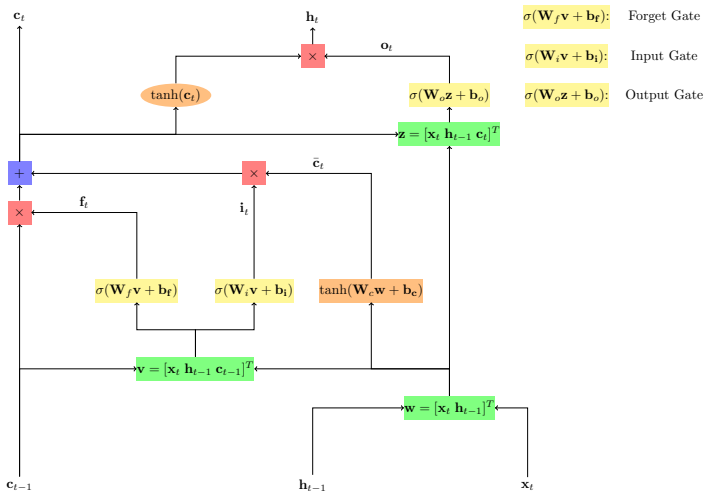


Figure 2 : Visualization of the LSTM architecture

Bidirectional BLSTM

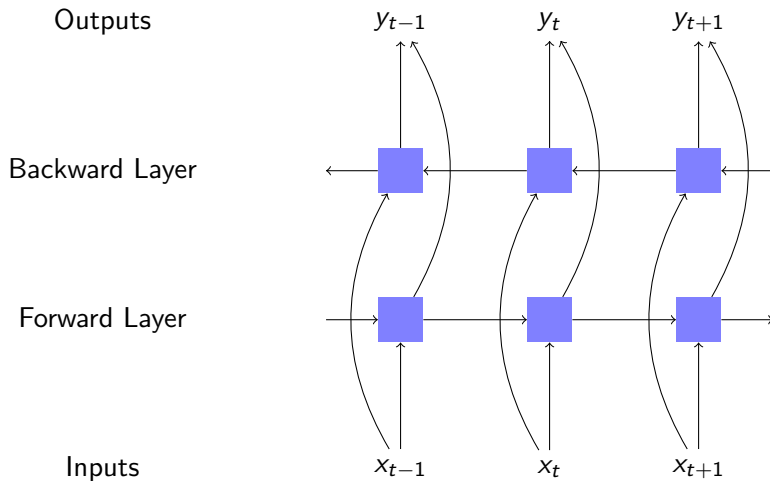


Figure 3 : Bidirectional LSTM architecture

Listener

- Pyramidal Bidirectional long short term memory (*pBLSTM*).
- Pyramid structure compresses the features.
- Three *pBLSTM*s on top of a *BLSTM* layer \Rightarrow compression factor $2^3 = 8$.
- Pyramidal inputs concatenate the out put from previous layers:

$$\mathbf{h}_i^j = \text{pBLSTM}(\mathbf{h}_{i-1}^j, [\mathbf{h}_{2i}^{j-1}, \mathbf{h}_{2i+1}^{j-1}]) \quad (11)$$

- i denotes the time step (from 0) and j the layer.

Attend and Spell

- attention based *LSTM* transducer.
- Find the most likely character given the features and previously found letters.

$$\mathbf{c}_i = \text{AttentionContext}(\mathbf{s}_i, \mathbf{H}) \quad (12)$$

$$\mathbf{s}_i = \text{RNN}(\mathbf{s}_i, \mathbf{H}) \quad (13)$$

$$P(\mathbf{y}_i | \mathbf{x}, y_{<i}) = \text{CharacterDistribution}(\mathbf{s}_i, \mathbf{c}_i) \quad (14)$$

Attention Context

- Produce a context vector \mathbf{c}_i , with alignment information.

$$e_{i,u} = \phi(\mathbf{s}_i)^T \psi(\mathbf{h}_u) \quad (15)$$

$$\alpha_{i,u} = \frac{\exp(e_{i,u})}{\sum_u \exp(e_{i,u})} \quad (16)$$

$$\mathbf{c}_i = \sum_u \alpha_{i,u} \mathbf{h}_u \quad (17)$$

- ϕ and ψ are feed-forward MLP networks.
- \mathbf{s}_i is the decoder state.
- The α s work like a sliding window.
- U denotes the total number of feature vectors.

The LAS-Architecture

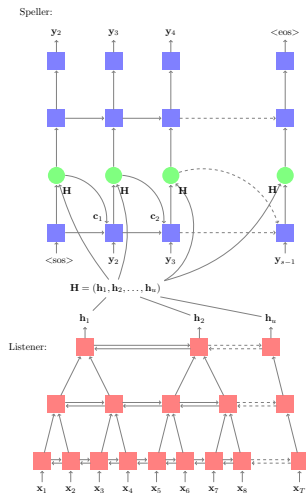


Figure 4 : The LAS architecture

Decoding and Rescoring

- Left to right beam search.

Tensorflow

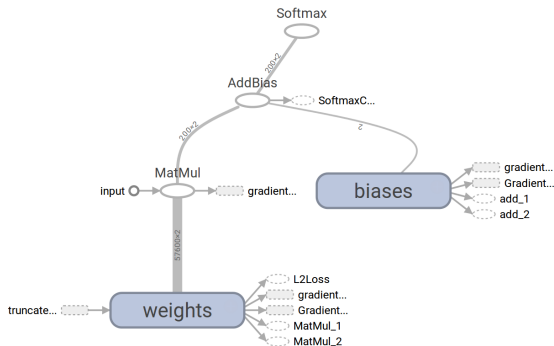


Figure 5 : A simple linear node in tensorboard