Maths involved in Voice Recognition

Sanket Ranade
EE19BTECH11012

January 22,2020

Zero Padding

Zero padding consists of extending a signal with zeros. It maps a length N signal to a length M $>$ N signal, but N need not divide M.

Definition:

$$\textsc{ZeroPad}_{M,m}(x) \triangleq \begin{cases} x(m), & |m| < N/2 \\ 0, & \text{otherwise} \end{cases}$$

where $m = 0, \pm 1, \pm 2, \ldots, \pm M_h$, with $M_h \triangleq (M-1)/2$ for M odd, and $M/2-1$ for M even.

Using this, we converted originally recorded 80 files to 2000 in number.

MFCC

Mel Frequency Cepstral coefficients function gives us a matrix of dimension 49×39.
The file is broken down into 49 steps with each step having 39 features.
The formula for converting frequency to Mel scale is

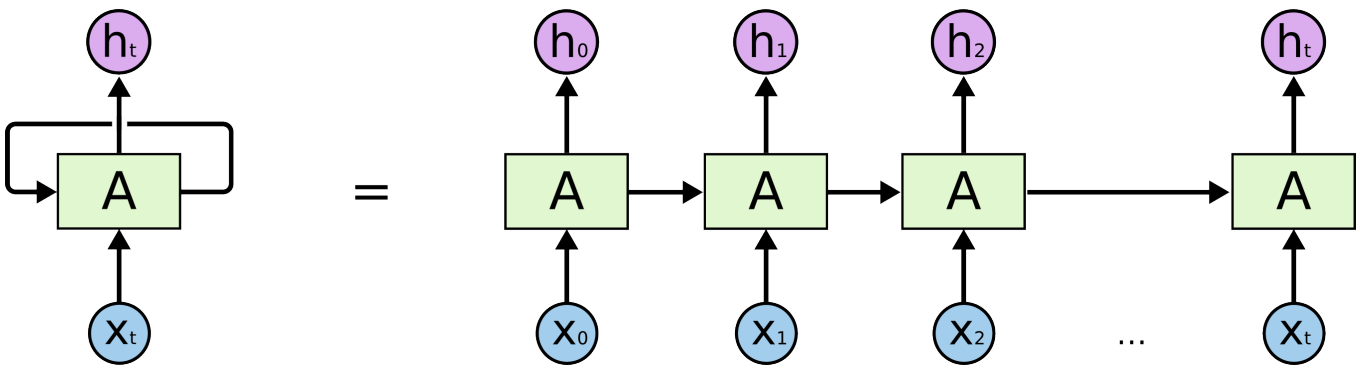
$$\mathbf{M}(\mathbf{f}) = 1125 * ln(1 + f/700)$$

Recurrent Neural Network

These are type of Neural Networks in which output of previous state is taken as input to current state.

.

The main and most important feature of RNN is Hidden state, which remembers some information about a sequence.

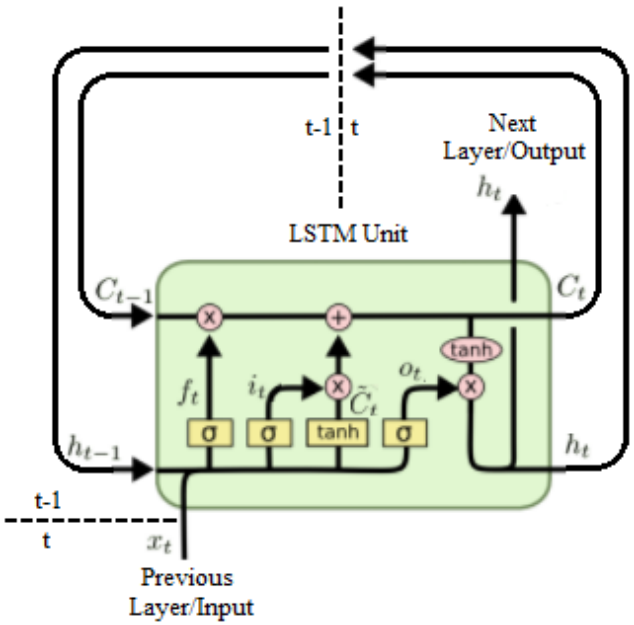


Long Short Term Memory

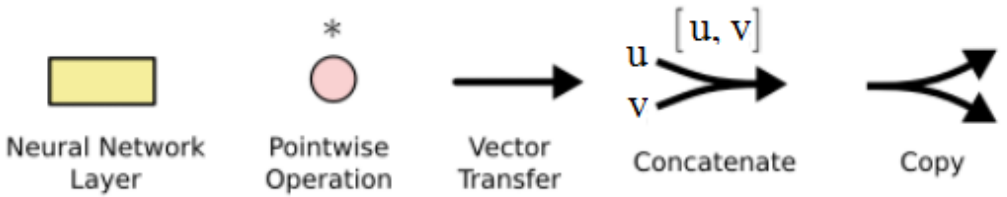
Standard RNNs suffer from vanishing and exploding gradient problems.

LSTMs deal with these problems by introducing new gates, such as input and forget gates which enable better preservation of long-range dependencies.

Understanding LSTM Networks



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$
$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$



Loss Function

The loss function in the LSTM is calculated as the categorical class entropy defined as follows

$$E = -\sum_{i}^{C} a_i * log(\hat{a}_i)$$

where C is the total number of classes and $\hat{a}_i$ is score of each sample of a_i in the softmax function.

$$f_i(\vec{a}) = \frac{e^{a_i}}{\sum_k e^{a_k}}$$