Speech Recognition

Tadipatri Uday Kiran Reddy
EE19BTECH11038
Dept. of Electrical Engg.,
IIT Hyderabad.

January 22, 2020

1. Data Augmentation
   - Zero Padding
   - Feature Extraction

2. Which Nerural Network??
   - Feedforward Neural Network
   - Recurrent Neural Network
   - Gated Recurrent Neural Networks
     - LSTM
     - GRU

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Zero Padding

This is a technique to increase no of samples of a given audio sample

- Declare an empty array of fixed length which is same length for training the data
- Add the data of the audio sample in the array
- Do this in a circular shift way until desired no of samples are required

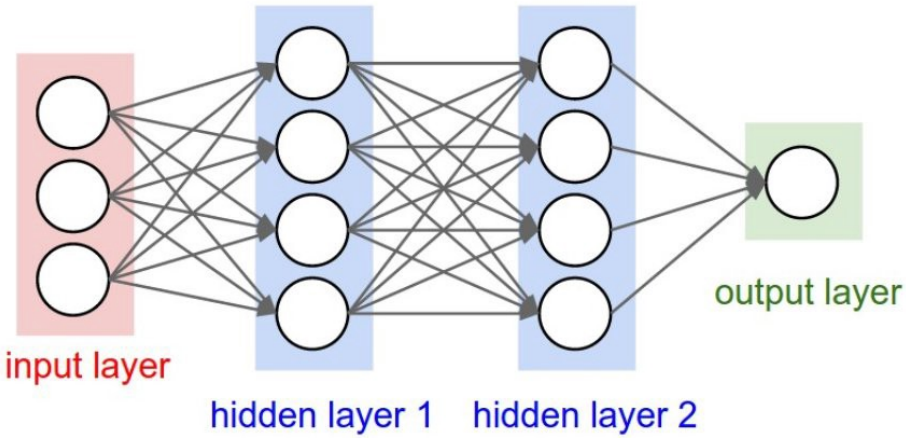
MFCC - Mel Frequency Cepstral Coefficient

MFCC takes into account human perception for sensitivity at appropriate frequencies by converting the conventional frequency to Mel Scale.

$$Mel(f) = 1125 \ln(1 + \frac{f}{700})$$

In our case, it returns a matrix of (49X39) i.e,49 time steps and each with 39 features.

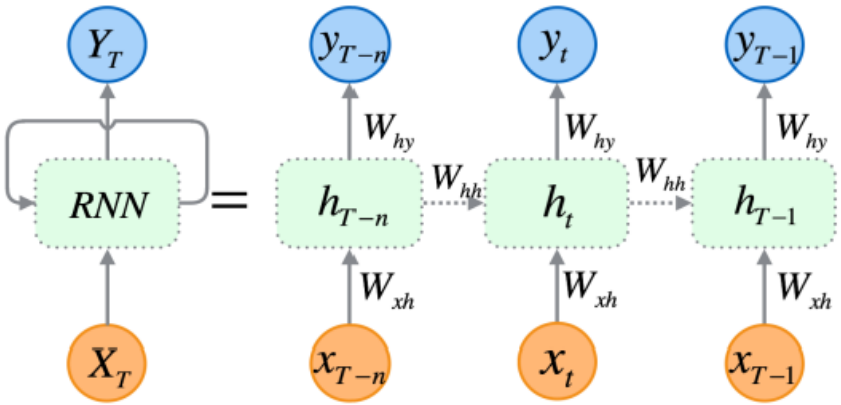
Feedforward Neural Network



Here in Feed forward Neural Networks such as MLPN's at time t for a input $x(f)_{t\geq 0}$
output would be $f(w,x(t))_{t\geq 0}$, **without regard to the previous history**.
i.e ,Ouput at t is independent of input at t-1 it only depends on the input at t
So this kind of neural network is not suitable for Speech recognition
The main advantage of the neural neural network over others is its memory for processing Temporal signals

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Recurrent Neural Networks



1 Standard RNN architecture and an unfolded structure with T time

RNN has an edge over FFNN because of its memory for processing Temporal Signals

Here h_t is the hidden state at time step t

$$h_t = f(x_t W_{xh} + W_{hh} h_{t-1})$$

where f is a non-linear function

For initial hidden state is initialised with all zeros,i.e,(for s_{-1})

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Cost Function

$$o_t = softmax(V_{st})$$

$$softmax(\mathbf{z})_i = \sigma(\mathbf{z}) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}};$$

for i = 1,2,3,,,,K and z = $(z_1, z_2, z_3, , , , z_K)$

- Categorical Cross Entropy

$$E(y_t, \hat{y}_t) = -y_t \log \hat{y}_t$$

$$E(y, \hat{y}) = \sum_{n=1}^{T} E(y_t, \hat{y}_t)$$

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

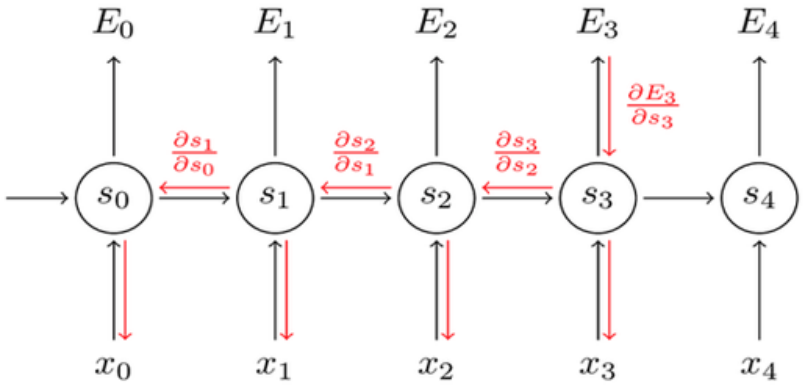
Optimiser

We use Root Mean Square Error

$$J(\mathbf{w}) = \frac{1}{m}\frac{1}{T}\sum_{i=1}^{m}\sum_{t=1}^{T}(y_i(t) - f(\mathbf{w}, \mathbf{x}_i(1:t)))^2$$

Here the error is calclulated based on previous history also

Back Propagation



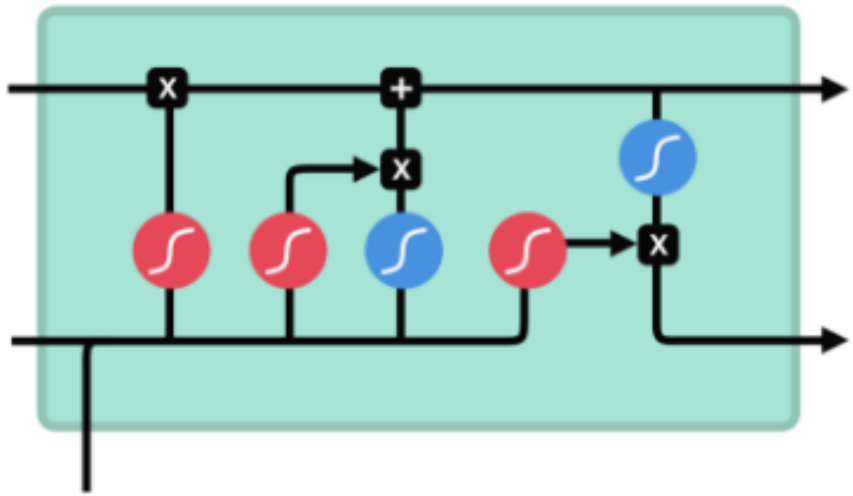
Backpropagation Through Time

$$\frac{\partial E}{\partial W} = \sum_{i=0}^{3} \frac{\partial E_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial s_3} \frac{\partial s_3}{\partial s_k} \frac{\partial s_k}{\partial W}$$

Now we have learn the data with the gradient $\frac{\partial E}{\partial W}$ back along the chain
But the gradient becomes negligible over the chain.
So RNN fails in storing the gradient.

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

LSTM - Long short Term Memory



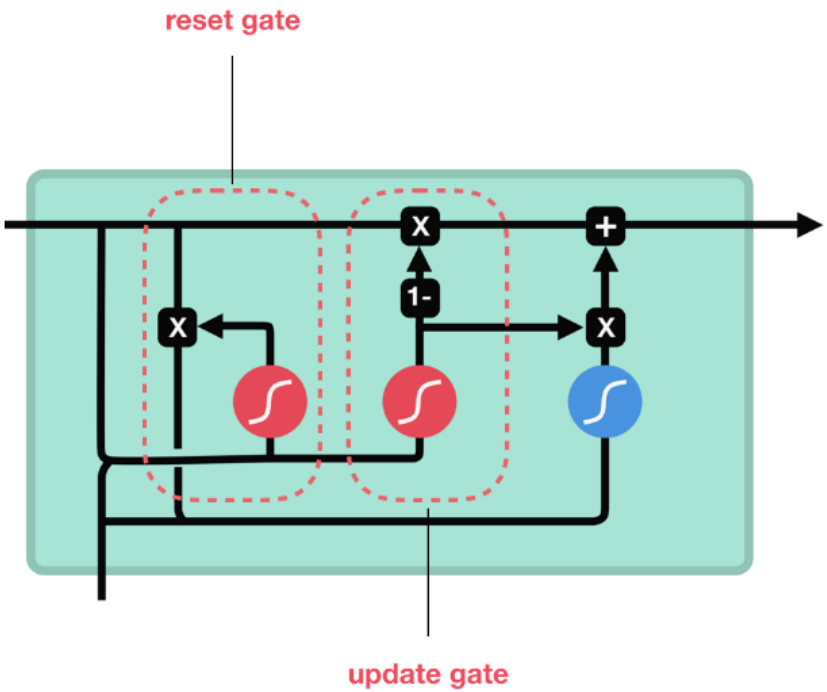
These operations are used to allow the LSTM to keep or forget information.

It has forget gate which decides whether to hold the information or not.

This ensures that learning rate does not decrease down the chain.

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

GRU - Gated Recurrent Units



The GRU controls the flow of information like the LSTM unit, but without having to use a memory unit. It just exposes the full hidden content without any control.

My assumption is that GRU has only two gates(i.e,**Reset gate** and **Update gate**) and also it has less fewer tensor operation,So this may be more efficient and faster compared to LSTM.

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad