Assignment 3

170050024,170070007,170070046

1 RNN Transducer Model

The model consists of two different parts the prediction network and the transcription network. The prediction networks takes as input a $U+1$ length vector sequence with elements $\mathbf{y}^* = (\phi, \mathbf{y_1}, ..., \mathbf{y_U})$ where $\mathbf{y} = (\mathbf{y_1}, ...\mathbf{y_U})$ is the output sequence and each y_u of them is size K(number of different outputs and they are one hot encoded) and outputs prediction vectors $\mathbf{g} = (\mathbf{g_0}, \mathbf{g_1}, ..., \mathbf{g_U})$ where each vector g_u is of size $K+1$ as now ϕ is also a part of the output space. It is a LSTM network with the standard input,forget,memory and output gates(In the paper 128 LSTM were used). This network tries to guess the next char given the previous ones but now it can also be ϕ along with the other elements of Y.
The transcription network takes as input the original input in the question that is $(x_1, x_2, ..., x_T)$ and returns as output the transcription vectors $(f_1, f_2, ..., f_T)$ where each vector (f_t) is of size of $K+1$ where K is as above. It is a bidirectional LSTM network with 128 LSTM's.
There is also an output probability lattice of size $(U+1)*T$ where the horizontal and vertical transition probabilities are given by $\phi(t,u)$ and $y(t,u)$ respectively.The total probability of reaching the final nodes is given by product of all the transmission probabilities which is also the product of forward and backward probabilities.
We train this model by maximizing the probability $Pr(y|x)$ which is given by the sum of probabilities encountered while traversing all the paths in the grid which can also be said as sum over all the possible alignments a such that $B(a) = y_*$. If we go via the first definition

$$Pr(y|x) = \sum_{(t,u):t+u=n} \alpha(t,u)\beta(t,u) \tag{1}$$

Here $\alpha(t,u)$ and $\beta(t,u)$ are dependent on the $P(k|t,u)$ and hence we can now calculate the gradient of the Loss function $L = -log(Pr(y|x))$ with respect to $Pr(k|t,u)$ for use in the gradient descent algorithm.

2 Relaxation of frame independence

The prediction network removes the independence between the output frames as we output g_u by having considered all the previous $(\phi, y_0, ..., y_{u-1}, y_u)$ frames and use these $g_u's$ for further work.

3 Extension to Online streaming

In the transcription network we had used bidirectional LSTM's, but for using it in online streaming we will have to convert it into a single direction LSTM because when a new input vector x_{T+1} vector comes we don't have to recalculate all the f_t's and the new f_{T+1} can be calculated by using the hidden output and cell state from the time step T as there will be no change in them.

4 Limitation

When the output space size (K) is very large the beam search will take a very long time. As during the testing time, to find the next y_u it will have to iterate through all the K different states which will be a large number.