As shown in figure the net contains 5 layers; Pre-LSTM, Dropout, Dense(1), Post-LSTM, Dense(6). The output of last dense layer is fed to 6-way softmax which produces a distribution over 6 class labels. Our minimizes the categorical cross-entropy loss or maximizes the average across all the training samples of the log probability of the correct label from the classification distribution.

The output of Pre-LSTM layer if fed to the attention module which calculates the attention weights or context and then it is fed to Post-LSTM followed by the dense layer through the 6 way softmax.

Attention Mechanism:

* There are two sperate layers of LSTM, one that comes before (Pre-LSTM 16 cell) Attention mechanism and one that comes after it (Post-LSTM 96 cell). The Pre-LSTM goes through Tx time steps and the Post-LSTM goes through Ty time step. (Tx = 500, Ty = 1)
* Here post-LSTM passes S (None,96) and C (None,96) i.e. output and hidden state at time step t to next time step t+1. The context that we calculate using DOT is passed on along with S and C.
* The RepeatVector repeats the output of Post-LSTM - (None,96) cell Tx times - (None,500,96) and the Concatenate joins it with the output of Pre-LSTM – (None,500,13) to output the joint vector (None,500,112) which is then fed to Dense(1).
* The Dense(1) is a one layer neural network which is trained to output the attention weights (None,500,1).
* The Dot then does the dot product between the attention weights (None,500,1) and output of Pre-LSTM (None,500,16) to give the context vector (None, 1,16).

(Note: The Pre-LSTM and Post-LSTM layers are CuDNNLSTM layers with tanh activation function.

S0 and C0 are the initial inputs or the -1 time step outputs of the Post-LSTM)

The architecture without attention is shown below which has the same architecture as above but without attention mechanism.

**Attention Mechanism**

The attention block shown in figure 1 shows the architecture of attention mechanism in the dotted line. Models without attention passes the output from dropout layer directly to post-LSTM layer. The similar attention mechanism is used in [a,b] for generating image captions while focusing on only parts of the image.

There are two types of LSTM layers used for pre-LSTM, one is uni-directional and other is bi-directional. The post-LSTM passes outputs o<t> and hidden cell state h<t> from one time step to next. The inputs to post-LSTM are s<t> , context and h<t>. The outputs of pre-LSTM are represented as p<t> for unidirectional and for bidirectional LSTM, the forward and backward direction, the outputs are concatenated p<t> = [p<t>(forward),p<t>(backward)]. The repeatVector copies o<t-1> for x times where x is the number of time frames used to extract the data and then Concatenate layer concatenates it with p<t> to compute e<t,t’> which is then passed to dense layer and then softmax layer to output a<t,t’> where the t represents the post-LSTM’s time step and t’ represents the pre-LSTM’s time step.

At any time step t, given the outputs of pre-LSTM ([p<1>,p<2>,…,p<x>] and the previous output of post-LSTM o<t-1>, the attention mechanism will compute attention vector or attention weights [a<t,1>,a<t,2>…a<t,x>] and output the context vector context<t> =t’Ʃx a<t,t’> p<t’>.

A] Xu, Kelvin & Ba, Jimmy & Kiros, Ryan & Cho, Kyunghyun & Courville, Aaron & Salakhutdinov, Ruslan & Zemel, Richard & Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.

B] Bahdanau, Dzmitry & Cho, Kyunghyun & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv. 1409.