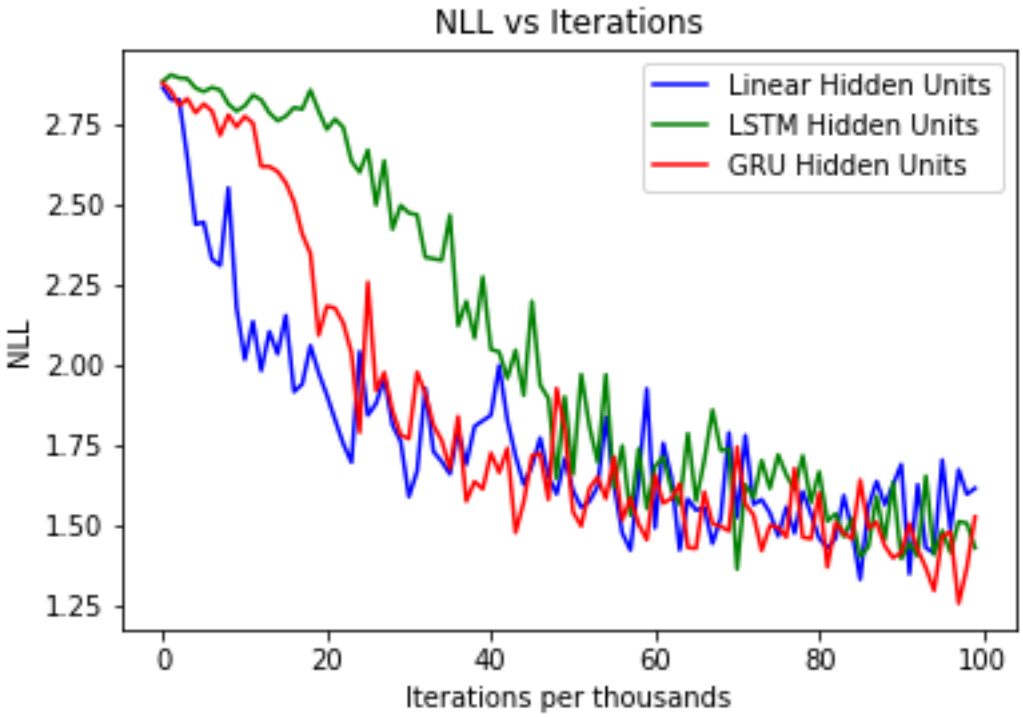
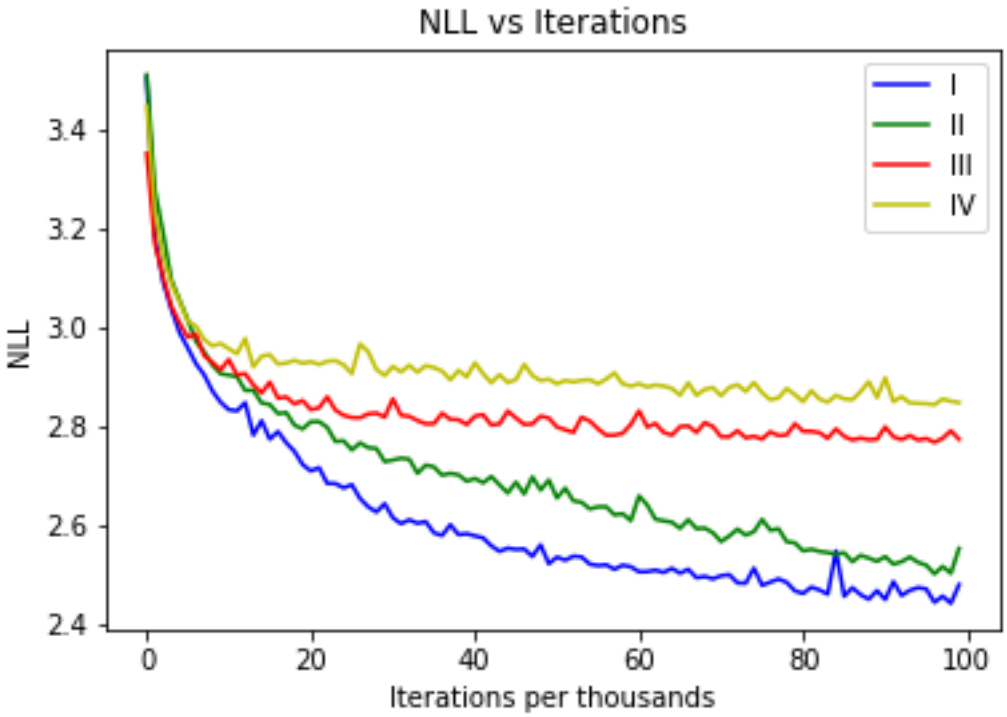
Part 1 (A):



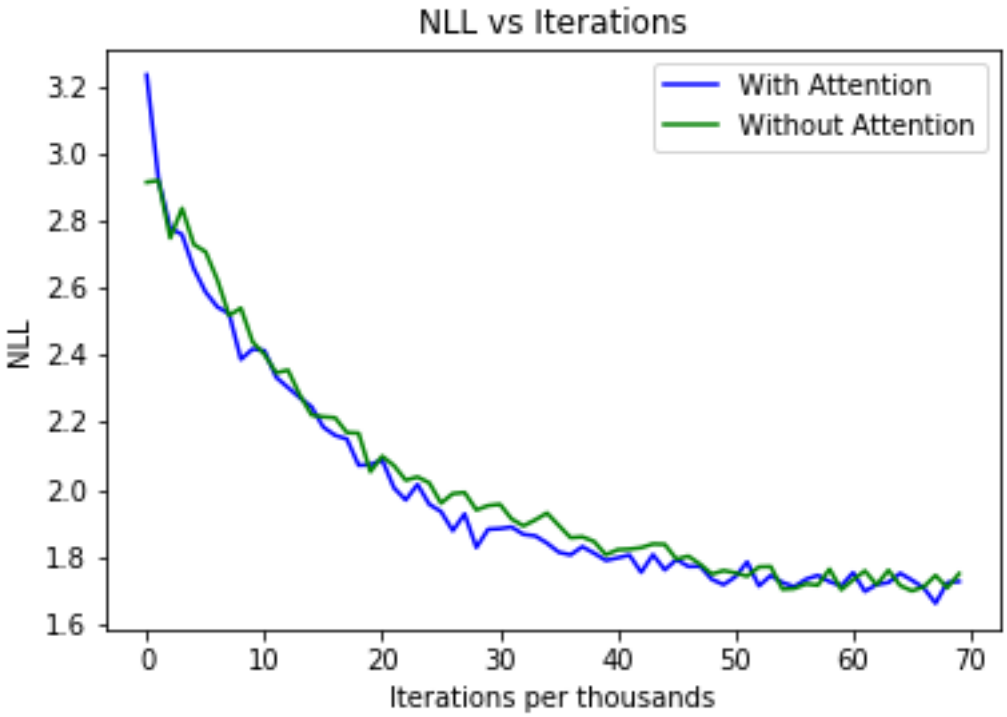
As We can see from the graph, in earlier iterations linear hidden units have lesser NLL than other two but as we proceed to further iterations GRU units gets better than the other ones. Even LSTM are performing better than linear hidden units in later iterations. This is because LSTM and GRU units solves the vanishing gradient problem present in linear hidden units. Also LSTM have two input and forget gate while GRU has only reset gate for both. In that case GRU is computationally cheaper than LSTM units.

Part 1 (B):



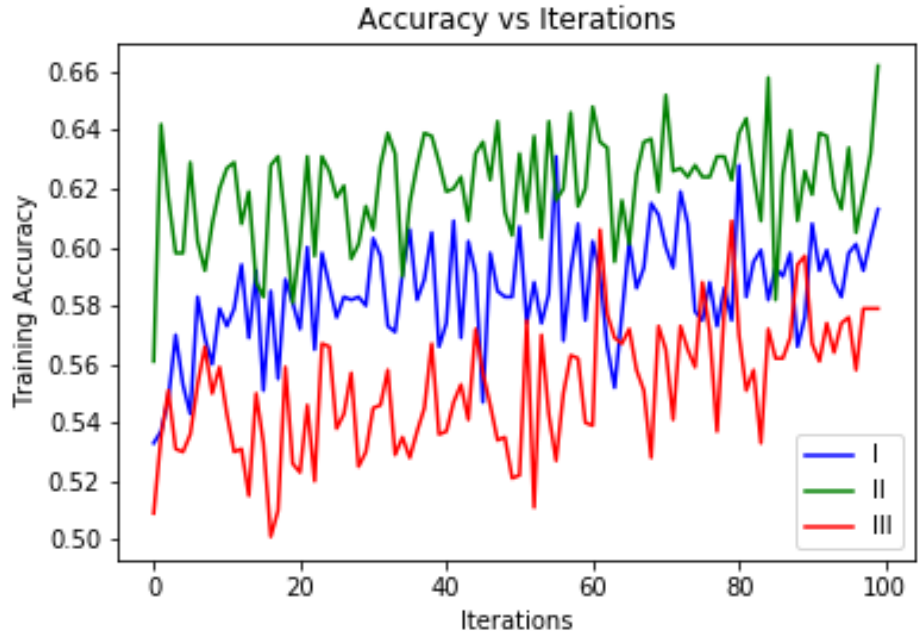
In the code we are using linear hidden units and then varying the type of input to the recurrent neural network. As we can see from the graph, as we are using more input information to the RNN we are getting less NLL and better model. Also if we are using previous character rather than category at each time step we can get lesser NLL loss. If we are using all information like previous character, previous hidden unit and category at each time step we get least NLL loss.

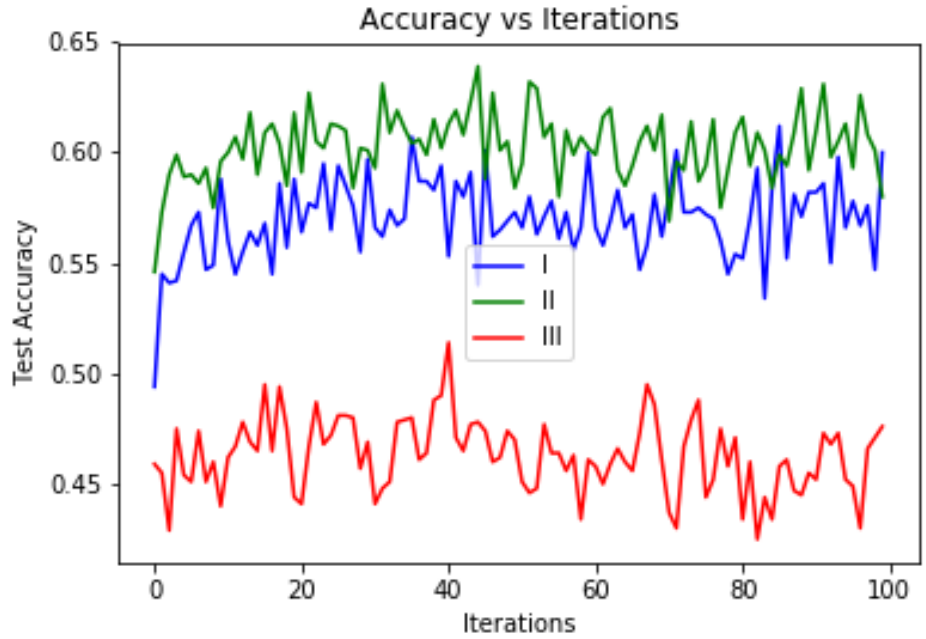
Part 1 (C):



As we can see from the graph model with attention mechanism is always performing better than model without iteration throughout the all iteration. But since number of training examples are less we cannot get the large difference between these two.

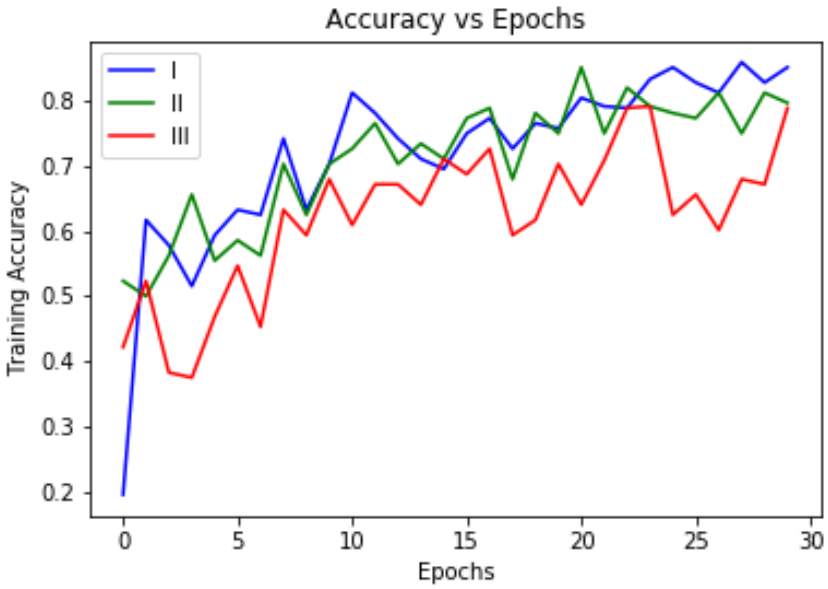
Part 2 (A):

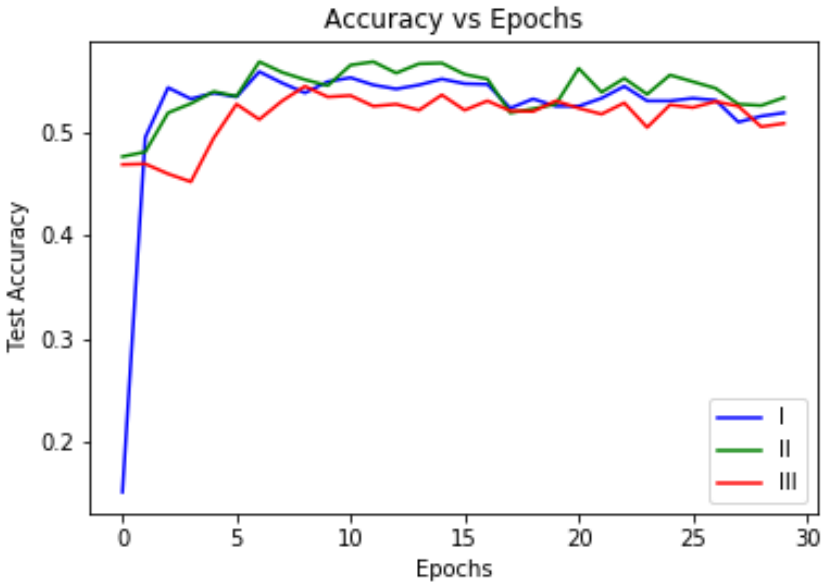




Comparing accuracy for both training and test dataset, it can easily be seen that when we are using claim and claimant both we are getting better results. The problem with type III when we are using claim, claimant and relevant sentences, the input size gets quite large. Since we are using recurrent neural network with linear hidden units it is difficult for the model to remember such large sequences. That's why it is performing worse.

Part 2 (B):





As the dataset contains 15000 examples, I have used only one layer, 256 cells, and 8 heads. Here I have tried different number of layers but as the dataset is small there was not much difference in using more layers.

Also for the different cases for inputs, if we are using claim, claimant and relevant sentences, it has been seen that accuracy does not gets up to that level because input size gets too high which causes vanishing gradient problem. There can be different number of hyper parameter combination to be tried.