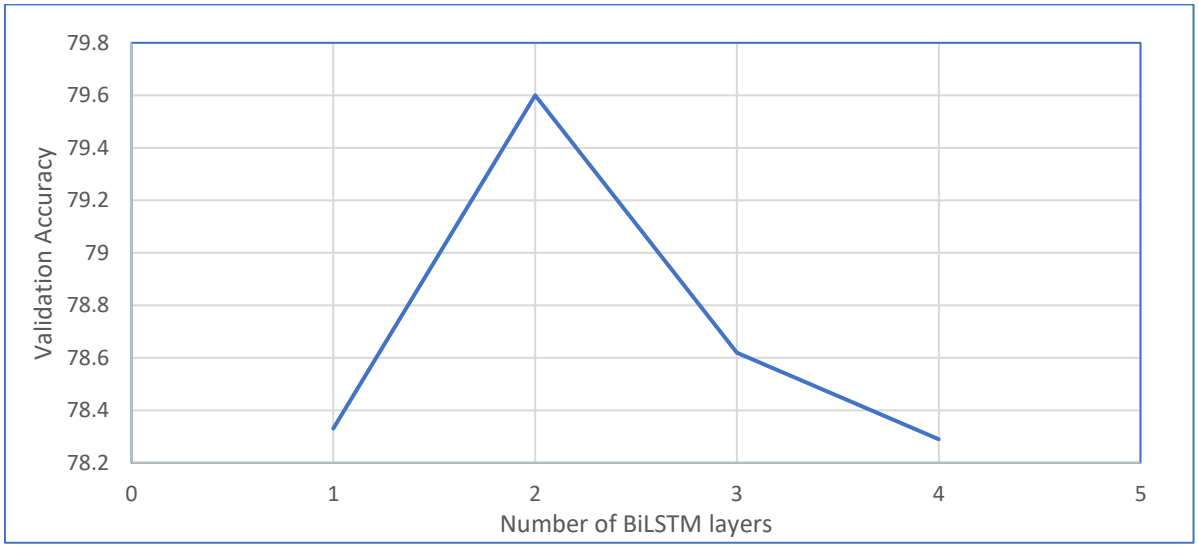
Natural Language Inference on SNLI Dataset

I decided to go with a model with an embedding layer followed by a linear translation layer and then two layers of bi-directional LSTM. After passing the premise and hypothesis through these layers, they were concatenated and then passed through 5 dense layers with dropout.

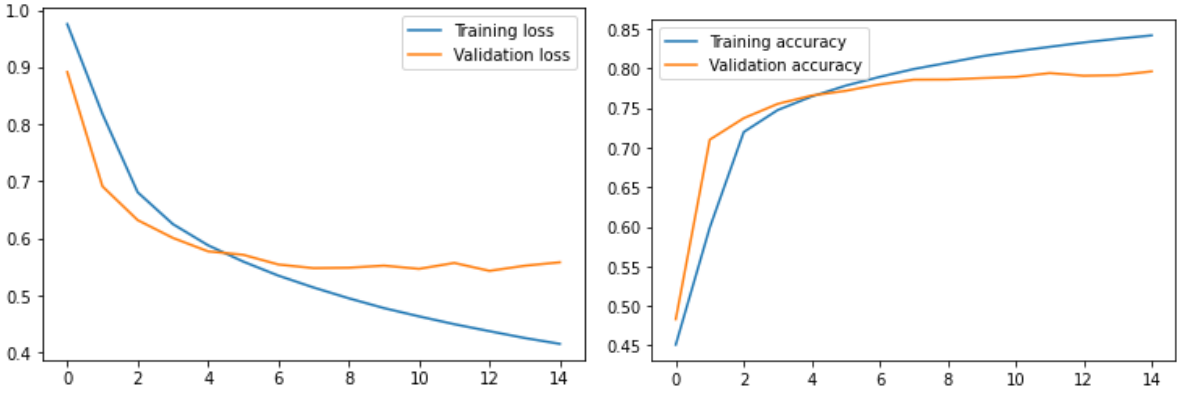
In the model I chose, using Bidirectional LSTMs are very important. In the chosen number of epochs (15) a model with Non- Bidirectional LSTMs showed no learning with accuracy stagnating in and around 33%

	Validation accuracy
BiLSTM	79.60
LSTM	~33%

I tried using various number of layers for BiLSTM, and found that accuracy peaked at a 2 layered model.



The model only had to be trained for 15 epochs after which it started to overfit/plateau.



Probe

I constructed a POS probe to analyse the model. It took the output from the different internal layers of the inference model, passed it through a dense neural net to predict the POS tag of the words given as input.

To extract the POS tags from the dataset I iterated over the sentences given in the parsed tree format. And tokenized both the words and the tags before passing it through the aforementioned model.

I carried out probing on 3 different models I had tried while training for inference so as to analyse the difference and understand how the model calculates and interprets English language

	Validation Accuracy	Validation Loss
Embedding Layer	68.13	1.977
Translation Layer	64.85	2.216
2 layered BiLSTM Layer	86.23	1.428

The above tabulated result is an expected outcome given that sequential information is processed in the LSTM layers and not before thus showing a major rise in POS tagging accuracy. However, the embeddings itself gives us a pretty high accuracy. That can be explained given the fact that a lot of prepositions, articles and nouns can be interpreted as such without requiring context. However, adverbs, adjectives and such will need context and memory-based calculation which can only be done by the LSTM layers.

	Validation Accuracy	Validation Loss
2 layered BiLSTM	86.23	1.428
1 layered BiLSTM	83.76	1.533
4 layered BiLSTM	87.89	1.363
2 layered LSTM	70.78	3.296

From the given observations it is clear that a Mono Directional LSTM does not help the model to interpret English Language since it has a similar accuracy and in fact worse loss than a POS tagger implemented after the embedding layer. This is very obvious since words in the latter part of a sentence very often changed the meaning of the initial words to a great extent thus changing which POS tag they belong to.

On the other side of the spectrum, we can see that the 4 layered BiLSTM actually has better results in identifying POS tags but as we have observed before, gives us a worse result on our inference model. From this we can hypothesise that POS tags are not the only measure of understanding language. There are other metrics which help the model understand language and consequently interpret the inference between two sentences and if a model can interpret POS tags better doesn't necessarily mean it is going to be perform better on the actual task.