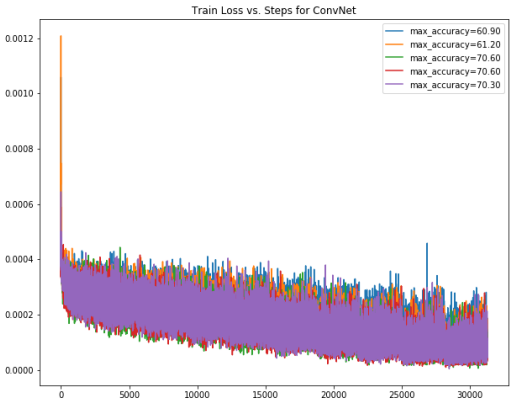
Homework 2
NLP and Representation Learning (DS-GA 1011)

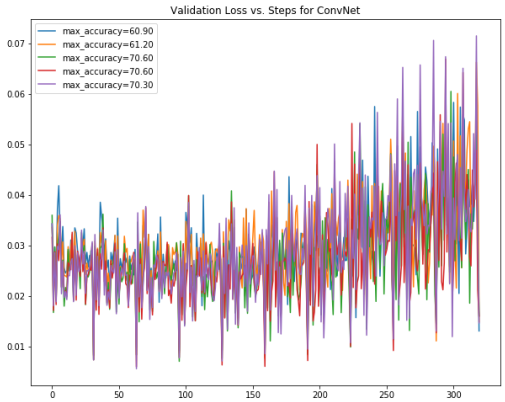
Name: Mihir Ujjwal Rana
NetID: mur214
MS in Data Science, New York University
Link to Notebook

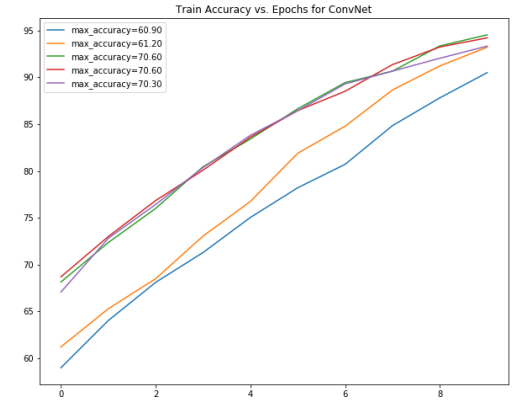
1 Training on SNLI

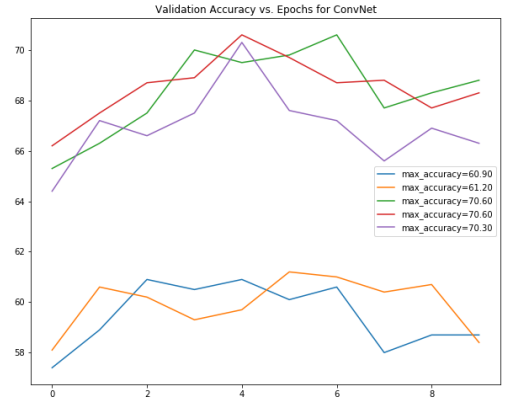
1.1 Results for Hyperparameter Tuning

The complete table of results for tuning with the training and validation losses and accuracies, as well as the number of parameters can be found in the notebook.
The plots for training and validation losses and accuracies for 5 example configurations (top 3 and bottom 2 based on validation accuracy) for the ConvNet are shown below:



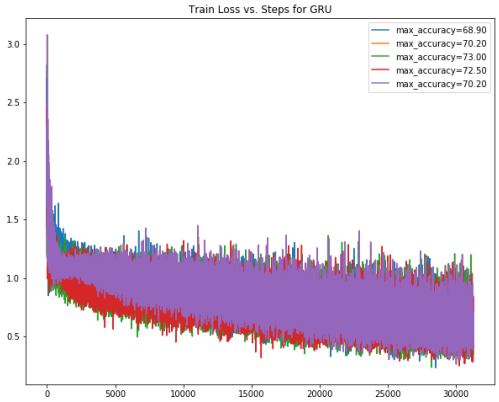


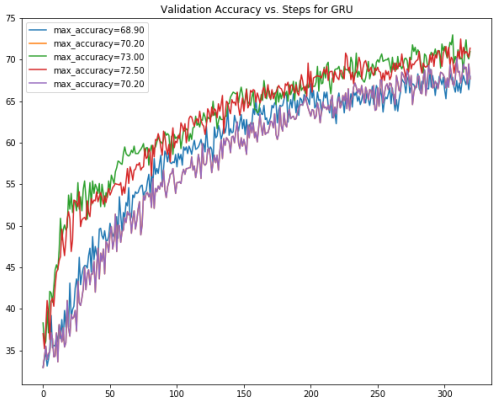




From the above graphs, we can see that the validation loss (and accuracy) starts worsening after the 4^{th} epoch (as is confirmed in the best ConvNet model below).

Similarly, the plots for training losses and validation accuracies for 5 example configurations (top 3 and bottom 2 based on validation accuracy) for the GRU are shown below:



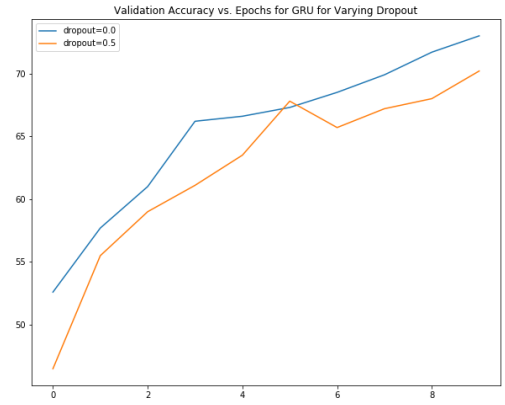


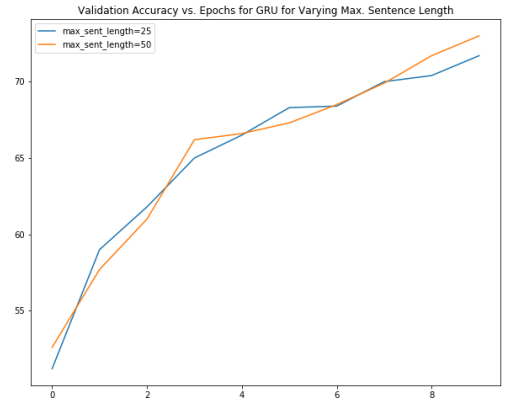
From above, we can see that the validation accuracy keeps improving till the 10^{th} epoch (as is confirmed in the best GRU model below), suggesting that training for a longer period could've improved the performance even further.

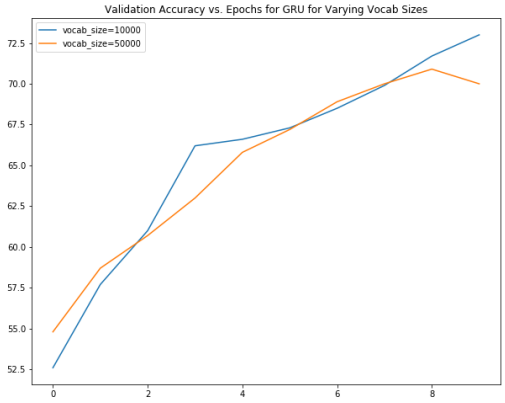
For hyperparameters, I tuned `dropout`, `vocab_size`, `max_sentence_length`, `aggregation_mode`, and `embedding_hidden_size` (for GRU/CNN). The intuition for some of them is written below:

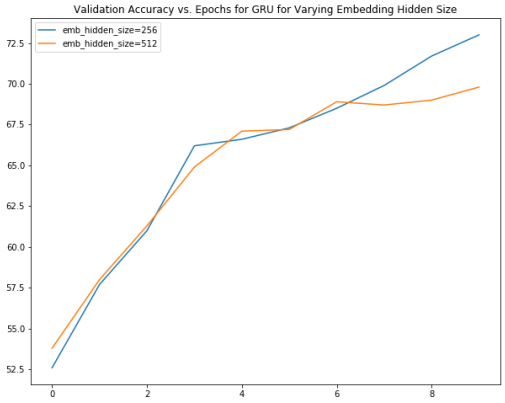
- Aggregation mode: There was clear trend in accuracy here, with concatenation > elementwise-multiplication > sum in most cases (*all* in case of ConvNet), presumably since the first takes into account all information, while a product usually captures better interactions compared to a sum.
- Dropout: The top spots were observed to be with no dropout, possibly because the embedding/classification hidden sizes were not too big and didn't need regularization.
- Max sentence length: On inspecting the lengths closely, I saw that the variance for the premise was just 6 (and 3 for hypothesis), while the 95^{th} percentile of the premise was 25, as against a maximum of 82 – which means there's not much variation in most of the lengths – a possible reason why no strong relationship with accuracy was observed.

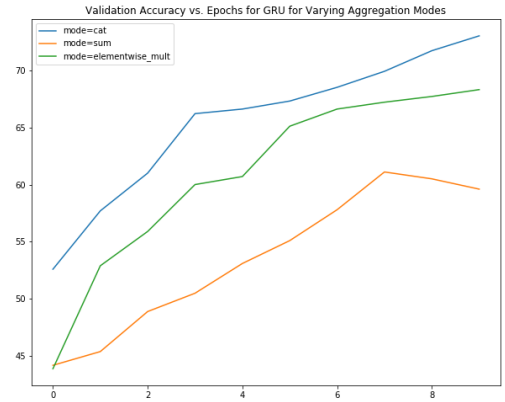
The effect of each hyperparameter is shown in the graphs below, where I've plotted the validation accuracy for the GRU model (with the non-tuned parameters set to those corresponding to the best GRU model in each case):











1.2 Best Model

The best model obtained for both the GRU and ConvNet had the following configurations:

	gru	cnn
lr	0.0003	0.0003
emb_hidden_size	256	256
class_hidden_size	256	256
vocab_size	10000	10000
max_sent_length	50	50
agg_mode	cat	cat
dropout	0	0
num_params	4098359	3537719
max_val_accuracy	73	70.6
n_epochs	10	4
kernel_size	-	3

The list of 3 correct and incorrect predictions on the validation set can be found here and here respectively. For the incorrect examples, possible reasons why the model mis-classifies each of them are:

1. In this case, the hypothesis and premise have a large set of common words, which is probably the model thinks they're in agreement.
2. The model possibly faces a problem in this example since the operating keyword here, which will likely determine the relationship, is an unknown token.
3. This is an confusing example even for a human, and may even be mis-labelled since it seems to be a contradiction.

2 Evaluating on MultiNLI

	accuracy_cnn	accuracy_gru
fiction	43.819094	44.522613
government	40.649605	41.633859
slate	39.920160	40.818363
telephone	43.283585	44.875625
travel	44.093689	43.177190

Clearly, the results for the MultiNLI data set are relatively much worse for every genre compared to the SNLI data set, understandably so, since the kind of language in the two corpora are very different. The networks learnt are not transferrable directly to other data sets, and performs marginally better than predicting at random.
Among the different genres, both GRU and CNN seem to perform best on `fiction` and `travel`, and worst on `slate`. On inspecting the counts for each genre, I found the following:

Genre	government	telephone	slate	fiction	travel
Count	1016	1005	1002	995	982

This means the model is performing best on the columns with the least amount of data (which is a little surprising) – although, the type of language in government documents, for example, is very different than the normal (casual) human language, which may be a reason for the worse performance on it.

3 Fine-tuning on MultiNLI

The results for each genre without any finetuning can be found in Section 2.
I trained my best GRU mentioned in 1.2 on the MultiNLI data set separately for each genre for 2 epochs, and obtained the following results on testing it for each genre :

	fiction_val	government_val	slate_val	telephone_val	travel_val
telephone_trained	48.341709	46.751967	43.213573	53.432840	46.334013
fiction_trained	50.653267	52.263778	43.712574	50.149256	51.527500
slate_trained	51.758796	51.476377	46.506986	52.537316	51.221997
government_trained	49.748743	53.346455	46.706587	51.343286	50.203669
travel_trained	48.542714	50.492126	46.606785	52.039802	48.472509

Not surprisingly, the results improve drastically even with very little training (just 2 epochs). Not only does the accuracy improve for the genre that I finetuned on, but it also generalizes pretty well to other genres, improving their accuracies by substantial amounts in all cases.