RNN/CNN-based Natural Language Inference

Alexandra Simonoff

2018/10/31

1 Introduction

For assignment 2 we were tasked with Stanford Natural Language Inference task. Namely, we were to classify whether two sentences were neutral, entailed or contradictory to each other. We did this using a bi-directional GRU RNN and a CNN each with two fully connected layers and using pre-trained embeddings.

Full analysis and code can be found at `https://github.com/AMS889/DS-GA1011_Homework2`

2 Analysis of Results

For hyperparameter tuning we chose to downsample to $\frac{1}{4}$ of the SNLI dataset for time's sake (using a local machine) and then training the best model on the full dataset. Each model uses an Adam optimizer, a learning rate of 0.0003 (as recommended by lab 4), 15 epochs and 3 classes.

The first hyperparameter we tuned was the hidden size for both the RNN and CNN models. We tested 50, 100 and 200 on both models and found that the highest validation accuracy was achieved with a hidden size of 100. It appears 50 underfits sightly and 200 might be beginning to overfit the data as the training accuracy was highest (on both CNN and RNN) with hidden size 200 but validation was marginally better with 100. This makes sense as growing our hidden size means the model will become more complex and with more complexity often comes less generalization. When we increase complexity we fit our training data better at the expense of generalization hence our validation set might not perform very well. In addition it appears the training accuracy is much higher for our CNN than our RNN but the validation accuracies are comparable.

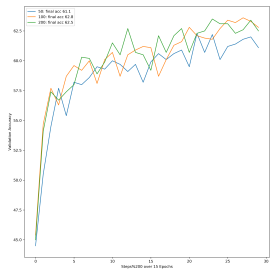

Figure 1: RNN Hidden Size Validation Accuracy Training Curves

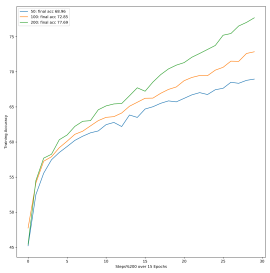

Figure 2: RNN Hidden Size Training Accuracy Training Curves

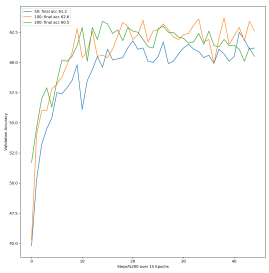

Figure 3: CNN Hidden Size Training Accuracy Training Curves

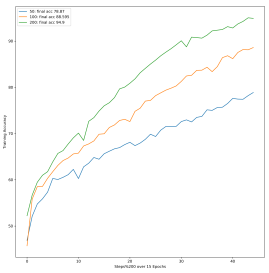

Figure 4: CNN Hidden Size Training Accuracy Training Curves

Next we tested varying the dropout rate on our CNN model using 0, 0.2, 0.5 and 0.8. Similarly to the incredibly high training accuracies of 100 and 200 hidden sizes we see a dropout rate of 0 has an incredibly high training accuracy and increasing the dropout rate decreases the training accuracy. With that said, the validation accuracy tells a different story. A dropout rate of 0 very clearly begins to overfit

our data as the validation is actually decreasing after epoch 2. Dropout rates of 0.2 and 0.5 show the highest validation accuracies with 0.2 just edging out 0.5. At these rates the model appears to fit the data well without overfitting. This is similar to our previous hyperparameter in that we will likely overfit the data if our dropout rate is low and likely underfit if our dropout rate is high. When we adjust our dropout, we force our model to ignore a random set of the neurons it's training on. Higher drop out rates mean more neurons will be dropped which in turn lowers complexity and makes our model more generalizable. With that said we see a decrease in validation accuracy with a dropout rate of 0.8 which suggests this value is past the point where (given our number of epochs, data size and hidden size) we begin to underfit the data.

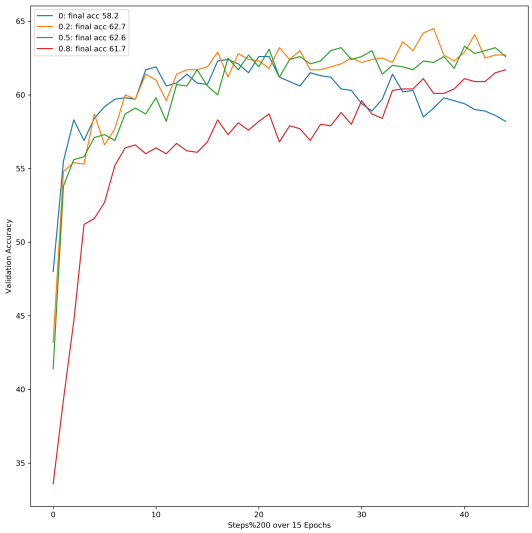

Figure 5: CNN Dropout Rate Validation Accuracy Training Curves

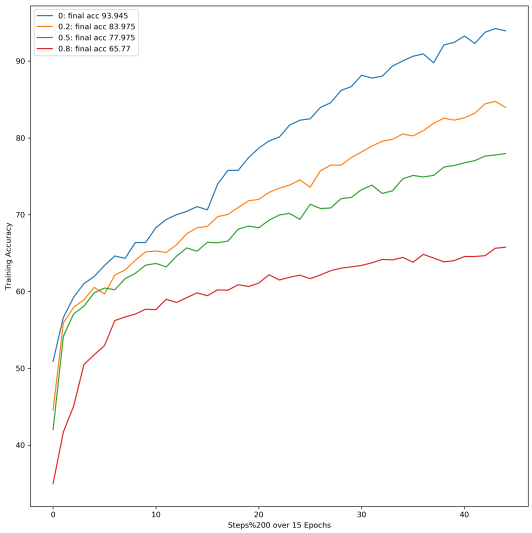

Figure 6: CNN Dropout Rate Training Accuracy Training Curves

3 Final Model

As our best CNN performs at the same level as our best RNN but appears to have more stability across steps (lower variance), we have decided to select a CNN with a 0.2 dropout rate and a hidden size of 100 as our best model. This model is then trained for 15 epochs over our full training set (less a 20% hold out test set) and achieves a validation accuracy of 68.0% and a test accuracy of 68.7% on the 20% hold out sample.

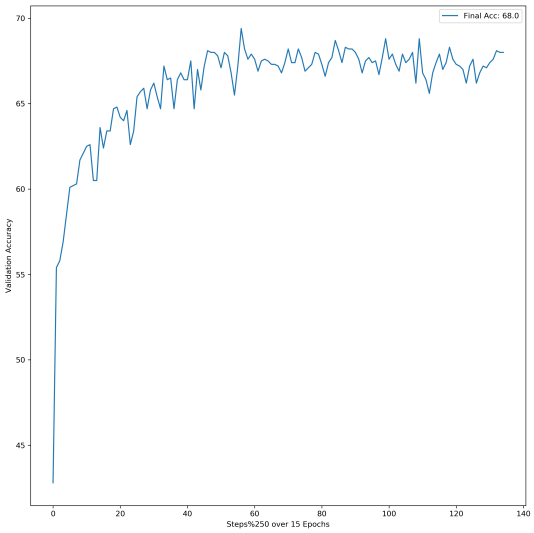

Figure 7: Best CNN Model Validation Accuracy Training Curve

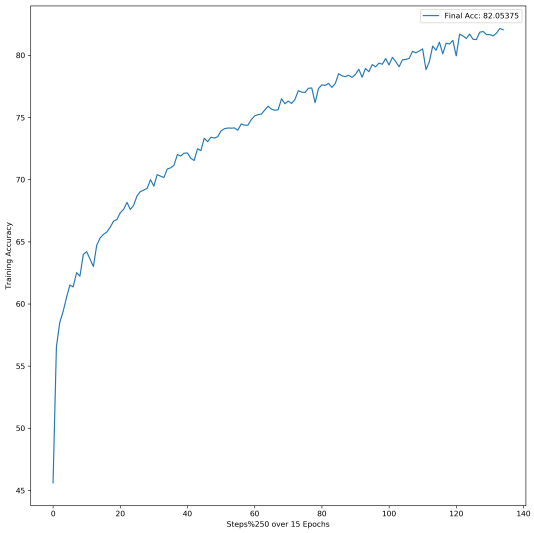

Figure 8: Best CNN Model Training Accuracy Training Curves

Three examples that were correctly classified in our validation set are:
Sentence 1: A woman holding a red cup with a straw in it sits in front of a man and a woman looking at a book.
Sentence 2: A woman sits.
Label: Entailment

Sentence 1: A group of dancers with green shirts are all holding hands in a circle, one lady has a white shirt.
Sentence 2: A group of dancers is about to dance for a school competition
Label: Neutral

Sentence 1: A man is sitting outside wearing blue pants.
Sentence 2: The person is inside.
Label: Contradiction

Three examples that were incorrectly classified in our validation set are:
Sentence 1: Two women on the dance floor with other people dancing.
Sentence 2: Two women are dancing on the dance floor.
Correct Label: Entailment
Predicted Label: Contradiction
Hypothesis: Our model likely interpreted sentence 1 to mean that other people were dancing, not the two women.

Sentence 1: A young boy is on a patio crowded with metal tables and chairs and 2 yellow umbrellas.
Sentence 2: A young boy is on a patio with lots of UNK.
Correct Label: Entailment
Predicted Label: Neutral
Hypothesis: Since the last word in sentence 2 was not in our vocabulary we do not know its relationship with sentence 1 so we can't determine the implication of sentence one and whether sentence 2 fulfills or refutes it.

Sentence 1: A man at the beach, building a sand castle.
Sentence 2: The waves have torn down a sand castle that the man is standing near.

Label: Contradiction
Predicted Label: Neutral
Hypothesis: As sentence 2 does not identify the sand castle torn down to be the mans so the model might view the two castles separately and see no connection between the events of sentences 1 and 2.

4 Applying to MNLI Genres

Now we took the CNN model we tested above as well as our best RNN model with hidden size 100 (our best from RNN training) and tested the models on the validation sets of the MNLI data by genre. The RNN model achieves a 67.9% validation accuracy and 70.1% test accuracy after 15 epochs training on the full SNLI dataset (less the hold out). In order to better understand the MNLI dataset we referenced *A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference* by Bowman, Nangia and Williams.

We find the the RNN actually outperforms the CNN in generalizing to the MNLI data consistently having a higher validation accuracy in all genres. Perhaps our RNN model does a better job learning general sentence structure that can generalize across many types of writing.

The worst performance on both models was seen in the slate genre; this genre contains pop-culture articles from Slate magazine from 1996-2000. As pop culture covers many different different topics and the colloquial/slang used in pop culture shifts often it's not too surprising our model underfits this data.

We see the best performance on the fiction genre; this genre contains sentences from many fiction works in the 20th century. It's likely our model performs well on these as fiction works tend to simulate conversations and natural language to tell a story and we trained our model on human written sentence pairs. It's not too hard a stretch to think the fiction genre contains sentences somewhat similar to the data we trained on. Similarly if trained on classical literature it's not too hard to believe a lot of the language has aged well and can be interpreted by our model in modern day. Compared to travel, telephone and government, fiction is much more likely to be telling a story through natural language.

The model performs moderately well on government, telephone and travel genres. The government genre contains reports and press releases from public government websites, telephone contains two-way telephone conversations from the early 90's and travel contains travel guides from the early 2000's. Government reports tend to be very formal works with many words that are unknown to our vocabulary so it's not too surprising the performance is not high. Travel works as well tend to refer to specific monuments that might use language that is understandable in a specific context of a travel guide for a location but not generalizing well to works not focused on exploring a city. Lastly, telephone calls are often very colloquial and unplanned so the sentences from these calls could be very complicated and confusing by nature so a model would struggle to relate two oddly structured and colloquial sentences.

	Genre				
Model	Government	Telephone	Slate	Fiction	Travel
RNN	45.37%	45.57%	44.61%	47.14%	45.62%
CNN	43.41%	42.89%	41.01%	43.22%	43.58%
Average	44.39%	44.23%	42.81%	45.18%	44.60%