DS-GA 1011 NLP HW2 - Sylvie Shao (MS6771)

Github: https://github.com/sylvieshaonyu/DSGA-1011-NLP

1 Hyperparameter Tuning on SNLI Dataset

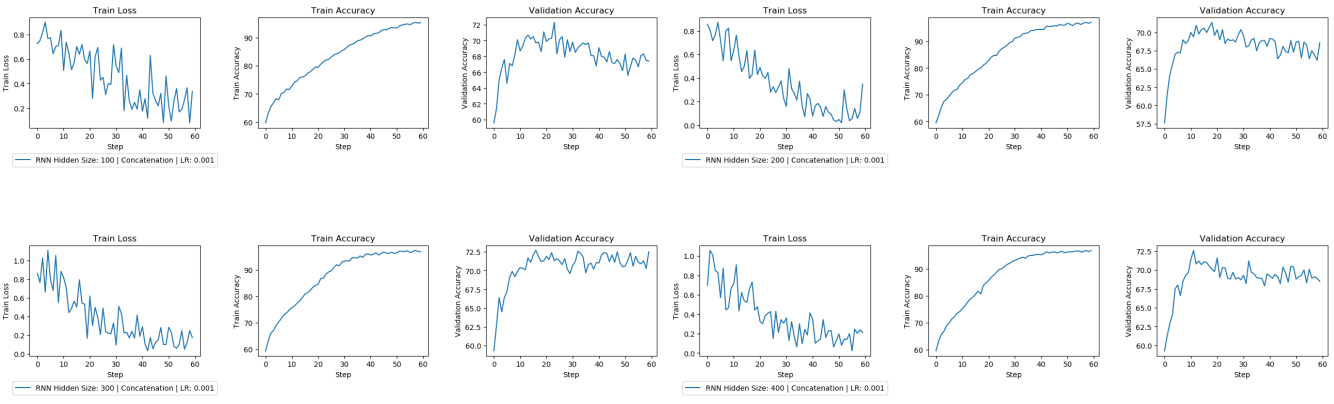
For the RNN and CNN models I trained on SNLI dataset, top popular 50,000 words were loaded from the Fasttext pretrained embedding matrix. A padding token with all zeros and an unknown token normally generated were manually added to the embedding matrix. In addition, the premise sentences tend to be longer than the hypothesis sentences. In order to capture this pattern in the training, two max sentence lengths of 99% length in the training set was defined, one for each sentence.

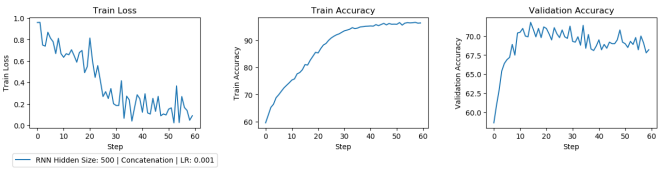
1. RNN

ID	Model	Hidden Dimension	Encoded Sentence Interaction	Learning Rate	Optimizer	Training Loss	Training Accuracy	Validation Accuracy
1	RNN	100	Concatenation	0.001	Adam	0.6946	82.10%	72.30%
2	RNN	100	Element-wise Multiplication	0.001	Adam	0.6313	78.34%	70.23%
3	RNN	200	Concatenation	0.001	Adam	0.6376	82.54%	71.00%
4	RNN	200	Element-wise Multiplication	0.001	Adam	0.6548	77.30%	72.56%
5	RNN	300	Concatenation	0.001	Adam	0.792	81.70%	**72.60%**
6	RNN	300	Element-wise Multiplication	0.001	Adam	0.2177	96.87%	71.20%
7	RNN	400	Concatenation	0.001	Adam	0.9103	75.67%	72.50%
8	RNN	400	Element-wise Multiplication	0.001	Adam	0.3948	88.57%	72.00%
9	RNN	500	Concatenation	0.001	Adam	0.6496	79.26%	71.00%
10	RNN	500	Element-wise Multiplication	0.001	Adam	0.8444	79.75%	72.58%

(a) Varying Hidden Dimension

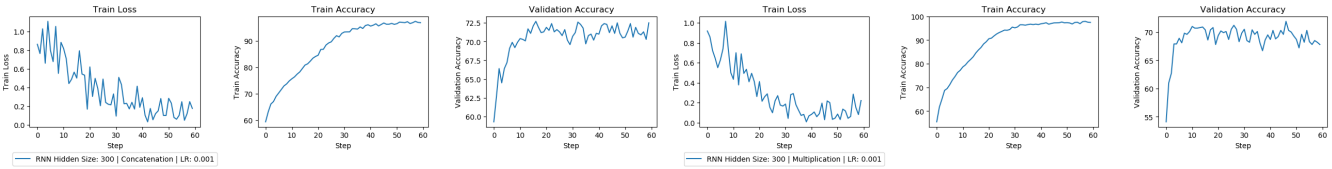
In order to fully understand how different hidden dimensions would impact RNN model performance, I experimented hidden dimensions of 100, 200, 300, 400 and 500. Hidden dimension of 300 achieved the best validation accuracy. With larger hidden dimension, the model is able to capture more information and achieve good results. However, with a hidden dimension being too high, the model could overfit the data and thus achieve suboptimal validation accuracy. In this case, hidden dimension of 300 seems to be the best since it allows the model to learn more features without overfitting the data.





(b) Varying Encoded Sentences Interaction

In order to fully understand how different encoded sentence interaction methods would impact RNN model performance, I experimented concatenation and element-wise multiplication of the two encoded sentences. Concatenation seemed to perform better. Concatenation allows for higher dimensions for the upcoming two fully connected layers, which enables the model to learn more features. In comparison, element-wise multiplication gives a lower dimension output for the two fully connected layers, which leads to the model achieving far higher training accuracy much earlier than concatenation. The model learns too quickly in the training set and overfitted the data.

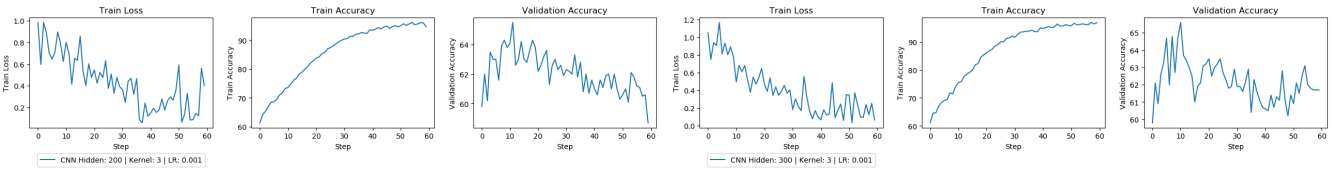


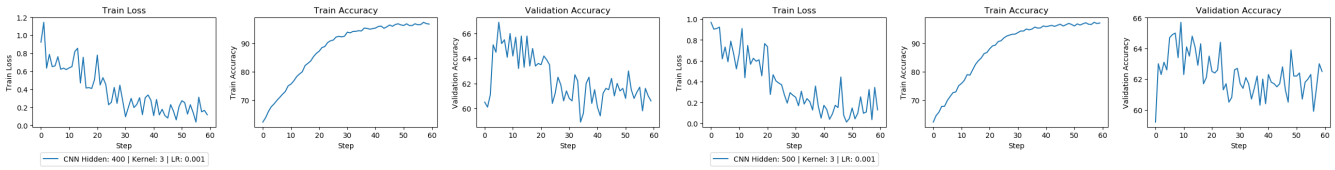
2. CNN

ID	Model	Hidden Dimension	Kernel Size	Learning Rate	Optimizer	Training Loss	Training Accuracy	Validation Accuracy
1	CNN	200	3	0.001	Adam	0.5985	76.80%	65.50%
2	CNN	200	4	0.001	Adam	0.3711	73.67%	64.80%
3	CNN	200	5	0.001	Adam	0.7229	78.86%	64.90%
4	CNN	300	3	0.001	Adam	0.5791	76.47%	65.50%
5	CNN	300	4	0.001	Adam	0.7206	71.35%	65.80%
6	CNN	300	5	0.001	Adam	0.6356	72.36%	65.80%
7	CNN	400	3	0.001	Adam	0.5995	71.40%	**66.90%**
8	CNN	400	4	0.001	Adam	0.5539	81.58%	65.80%
9	CNN	400	5	0.001	Adam	0.8698	71.42%	66.00%
10	CNN	500	3	0.001	Adam	0.6766	74.84%	65.70%
11	CNN	500	4	0.001	Adam	0.7956	72.89%	66.20%
12	CNN	500	5	0.001	Adam	0.9476	69.43%	64.30%

(a) Varying Hidden Dimension

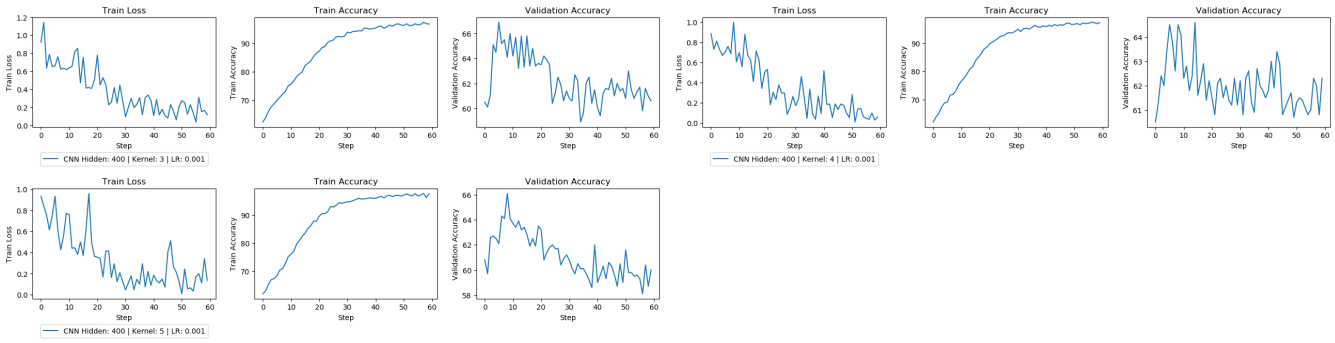
In order to fully understand how different hidden dimensions would impact CNN model performance, I experimented hidden dimensions of 200, 300, 400 and 500. Hidden dimension of 400 achieved the best validation accuracy. Same explanation applies to the CNN model. With larger hidden dimension, the model is able to capture more information and achieve good results. However, with a hidden dimension being too high, the model could overfit the data and thus achieve suboptimal validation accuracy. In this case, hidden dimension of 400 seems to be the best since it allows the model to learn more features without overfitting the data.





(b) Varying Kernel Size

In order to fully understand how different kernel sizes would impact CNN model performance, I experimented kernel size of 3, 4 and 5. Kernel size of 3 achieved the best validation accuracy. Kernel size can be thought of as n-grams so that the width of the filter corresponds to bigrams, trigrams etc. Having the network learn larger n-grams early exposes it to fewer examples, which leads to lower dimension in the following layers, and thus could cause overfitting. However, having the kernel size too small, one might lose the context information as parts of phrases could be separated by several other words.



3. Three Correct and Three Incorrect Predictions on the Validation Set Using Best Model

Overall, RNN achieved better results than CNN. RNN make more intuitive sense. They resemble how we process language - reading sequentially from left to right. With a bidirectional GRU unit in our RNN model, it enables the model to look from future time steps to better understand the context and eliminate ambiguity.

```
Three Correct Predictions:
Sentence 1: Two girls laying in the grass smile as their picture is taken . <pad> <pad> <pad> <pad> <pad> <pad> <pad>
<pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad>
Sentence 2: The girls are sisters . <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pa
d>
Label: neutral
Predicted: neutral

Three Correct Predictions:
Sentence 1: A man with a mustache who is wearing white and beige carefully <unk> a sand castle . <pad> <pad> <pad> <p
ad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad>
Sentence 2: A man building a sand castle <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad
>
Label: entailment
Predicted: entailment

Three Correct Predictions:
Sentence 1: The view of a man with a shaved head and blue shirt through a <unk> fence . <pad> <pad> <pad> <pad> <pad>
<pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad>
Sentence 2: A man with a shaved head behind a <unk> fence . <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad>
Label: entailment
Predicted: entailment
```

```
Three Incorrect Predictions:
Sentence 1: A little girl offers a ball to an upset toddler on a grassy field while a man in <unk> shorts stands behi
nd them . <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad>
Sentence 2: A man watches two children on a grassy field . <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad>
Label: entailment
Predicted: contradiction

Three Incorrect Predictions:
Sentence 1: A man with a beard , curly hair and a beard wearing a green shirt and navy blue jacket standing still loo
king through his red sunglasses . <pad> <pad> <pad> <pad> <pad> <pad> <pad>
Sentence 2: A man is <unk> down . <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad>
Label: contradiction
Predicted: entailment

Three Incorrect Predictions:
Sentence 1: <unk> flower cart vendor . <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad>
<pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad>
Sentence 2: <unk> vendor by the curb . <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad> <pad>
Label: neutral
Predicted: entailment
```

Reasons for Wrong Predictions:

1. The premise sentence separately described the actions of the two children and a man. The hypothesis sentence summarized the premise sentence by only mentioning the man and omitting the separate action of the two children. This may cause confusion to the model, and thus predicted contradiction.
2. The premise sentence is too short compared to the hypothesis sentence. Also, in the awfully short hypothesis sentence, there's also an unknown token which should be an adjective containing key information.
3. Both premise and hypothesis sentences are too short. In addition, they both contain unknown tokens which might confuse the model.

2 Evaluating on MultiNLI Dataset

Using the best model - RNN - trained on SNLI dataset, the validation accuracies on MNLI dataset by genre differs significantly from the one on SNLI dataset. The validation accuracies all decreased by a lot, regardless of genre. This means that RNN model is highly dependent on the dataset and hard to generalize to a different dataset.

In addition, comparing the results within MNLI across genres, both RNN and CNN achieved lowest validation accuracy in Slate genre. There could be some sarcasm in the sentences of Slate genre, which might not be easily trained with only SNLI dataset. Both RNN and CNN achieved among the highest validation accuracy in Government genre. Most government documents always provides context and plenty of explanations, which enables both models to learn well with only SNLI datas

Genre	RNN Validation Accuracy	CNN Validation Accuracy
Telephone	49.75%	45.97%
Fiction	48.34%	43.62%
Slate	46.71%	41.72%
Government	50.39%	44.39%
Travel	47.45%	45.01%