RNN/CNN Based Natural Language Inference

Manwen Li

https://github.com/manwenli/DS-GA-1011

1 Experiment Results on SNLI

1.1 RNN

For RNN, I changed the hidden sizes (100,200,300,400) as well as the way to combine the vectors for premise and hypothesis (concatenating the two vectors or performing element-wise multiplication). The constant learning rate is 5e-4. The first linear layer maps dimension of 2*hidden size to hidden size; after a relu, the second linear layer maps dimension of hidden layer to 3 (number of classes). The default setting is ID 0 in the table below.

ID	Learning Rate	Hidden Size	Kernel Size	Combine S1,S2	Max val acc.epoch	Max Val Acc.
0	5e-4	200	3	concat	9th	72.4
1	5e-4	300	3	concat	7th	72.5
2	5e-4	400	3	concat	7th	73.0
3	5e-4	100	3	concat	9th	72.5
4	5e-4	400	3	element-wise multiplication	5th	71.2

Figure 1: The table summarizes RNN's performance under different parameters

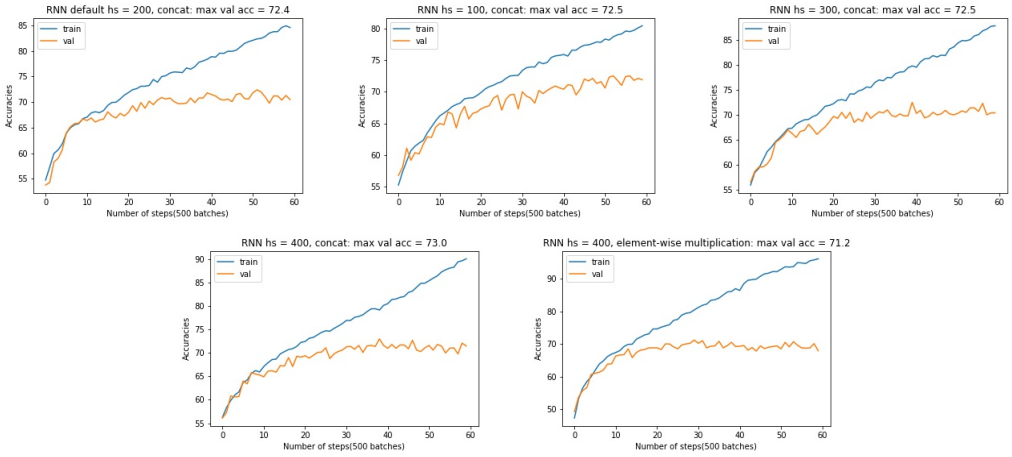

Figure 2: Tranining and Validation Accuracy for RNN

The best configuration is: hidden size = 400, kernel size = 3, which gives validation accuracy of 73 percent. The reason is that higher hidden sizes can extract more information or intrinsic features in the text. Thus, it would give a relative more accurate sentence representation. However, when the hidden size becomes too high (such as 600), the model will suffer from high variance and thus overfits.

In addition, I experimented two ways to determine the maximum sentence lengths. I first set the maximum sentence lengths to be the 99 percentile of the sentence lengths, and clip the sentences longer than that. Then I dynamically find the maximum sentence length for each batch. The results under those two preprocessing methods do not show huge differences. The experiment results below are based on the second method.

When hidden size is 400, the validation accuracy given by element-wise multiplying two sentences' representations is lower than concatenating the vector representations. The reason might be that after element-wise multiplication, the resulting dimension is only 1/2 of that given by concatenation and thus produce less accurate sentence representation according to the reasoning from the above paragraph. From the training and validation accuracy graph, the element wise multiplication will overfit more, since the training accuracy is very high and validation accuracy is lower than other configurations.

1.2 CNN

For CNN, I changed the kernel size (2,3,4) and hidden size (100,200,300,400). The default configuration is the same as the RNN default configuration, but the validation accuracy is lower than that of CNN's default configuration. The linear layers are the same as those in the RNN.

ID	Learning Rate	Hidden Size	Kernel Size	Max val acc.epoch	Max Val Acc.
0	5e-4	200	3	8th	68.7
1	5e-4	300	3	7th	69.3
2	5e-4	400	3	10th	69.7
3	5e-4	100	3	9th	67.7
4	5e-4	400	2	9th	73.0
5	5e-4	400	4	7th	68.6

Figure 3: The table summarizes CNN's performance under different parameters

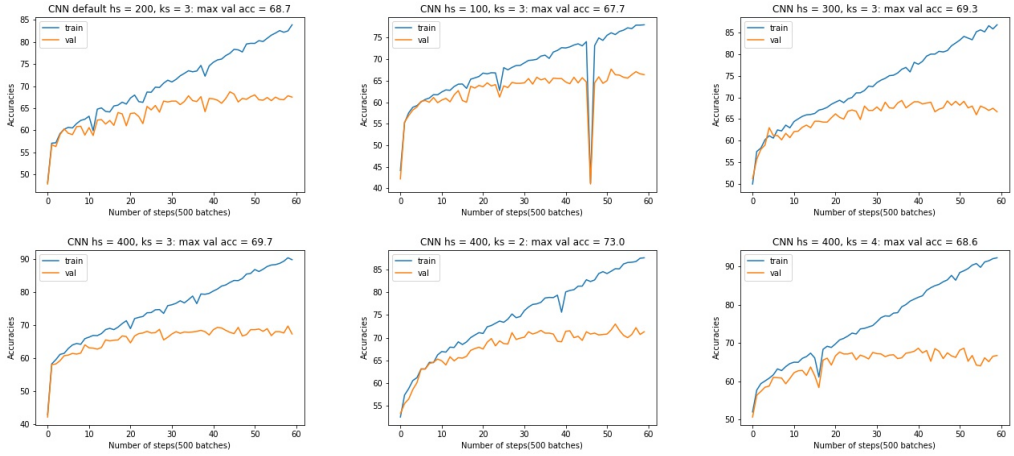

Figure 4: Tranining and Validation Accuracies for CNN

At the constant learning rate of 5e-4, the best configuration is: hidden size = 400, kernel size = 2, which gives validation accuracy of 73 percent. Here I notice a rising validation accuracy with respect to larger hidden size, for the same reasons described in the RNN section.

After I fix the best hidden size at 400, I find that the smaller the kernel size, the higher validation accuracy the model gives. Kernel size of 2,3,4 denotes the sliding windows over 2,3,4 words respectively, which is similar to the idea of n-gram. Here, when more words (e.g. 4) are taken into account together, the model is more precisely tailored to fit the training set; the train accuracy becomes larger, but at the same time the model variance becomes very large. As the result of overfitting, the validation accuracy decreases as kernel sizes increase from 2 to 4.

2 MNLI

The best RNN and CNN model configuration is summed up in the table below:

	Learning Rate	Hidden Size	Kernel Size	Combine S1, S2	Max Val Acc.
RNN	5e-4	400	-	Concat	73.0
CNN	5e-4	400	2	Concat	73.0

Figure 5: Best model configurations

	Fiction	Telephone	Slate	Government	Travel
RNN	50.55	49.15	46.21	48.33	48.17
CNN	46.73	48.45	43.81	49.51	45.82

Figure 6: Performance of RNN/CNN with the best config on mnli dataset across different genres

In general, RNN outperforms CNN in most of the genres (all genres except government). In addition, in the genre ”slate”, the accuracy for both RNN and CNN are significantly lower than those in other genres. RNN performs the best in fiction and telephone, while CNN performs the best in government and telephone. However, all the accuracy is much lower than the validation accuracy in the snli dataset, which indicates that the model lacks of generalization.

3 Correct/Incorrect Predictions

3.1 correct

S1: A man dressed all in white throws the first pitch at a baseball game .
S2: A man pitches at a baseball game .
Actual label: Entailment

S1: Two brown and white dogs , one jumping over a log while the one behind runs through the grass .
S2: The dogs are laying on a blanket .
Actual label: Contradiction

S1: A young girl sits on a blue bench drinking out of a straw .
S2: A girl is drinking through a straw on a humid day .
Actual label: Neutral

3.2 Incorrect

S1: A child in shorts throws a snowball at a mountain .
S2: A boy is playing in the snow without gloves .
Predicted label: Entailment
Actual label: Neutral

The model probably learns that people tend to wear shorts in summer, and that people only wear gloves in cold weather. Thus, the model captured the ”snow” and judges that if the boy is wearing shorts, he must be ”without gloves”; but it is confused by the fact the the child is wearing shorts in winter.

S1: A town worker working on electrical equipment .
S2: The worker is off work for the day .
Predicted label: Neutral
Actual label: Contradiction

The model probably fails to learn the word ”off”. In addition, it recognized the part that the worker is working on something, but it doesn't know if the work will last for a day, so it classifies this case as neutral. The reason why the word ”off” is ignored might be that ”off work” or ”off” is not common in the training set.

S1: A woman in a blue shirt and black workout pants practicing martial arts in front of a house .
S2: A woman has a white shirt .
Predicted label: Entailment
Actual label: Contradiction

The model has a poor judgment regarding colors - it might be due to lack of color related texts in the training samples.