DS-GA 1011 NLP Assignment 2
Zihao Zhao (zz913)

RNN/CNN-based Natural Language Inference

### Hyperparameter Tuning

model 1 is the CNN model with hidden size = 200, interacting the two encoded sentences by concatenation.

model 2 is the CNN model with hidden size = 100, interacting the two encoded sentences by concatenation.

model 3 is the CNN model with hidden size = 200, interacting the two encoded sentences by element-wise multiplication.

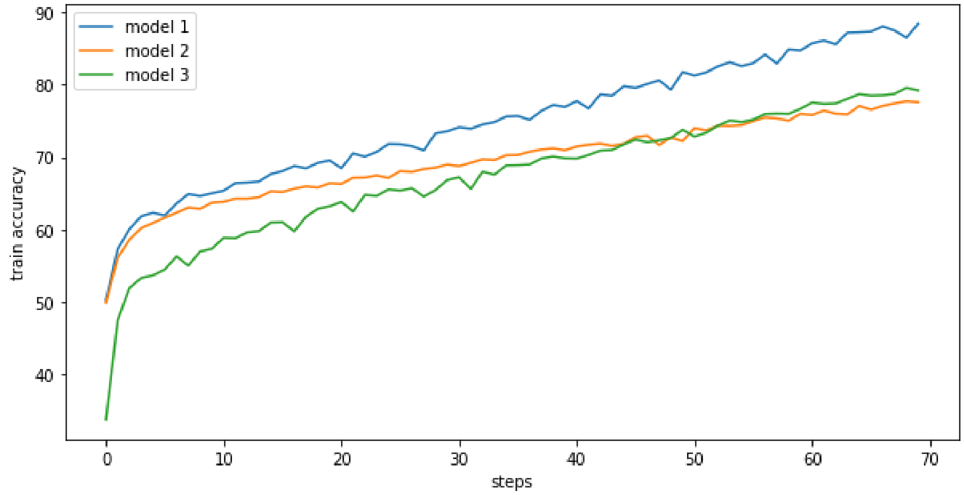

Figure 1: Train curve for different CNN models

From the training curve we see that model 1 has the highest training accuracy over model 1 and 2. This indicates that concatenation may be better in fitting the training data, and higher hidden size helps describe the training set much better.

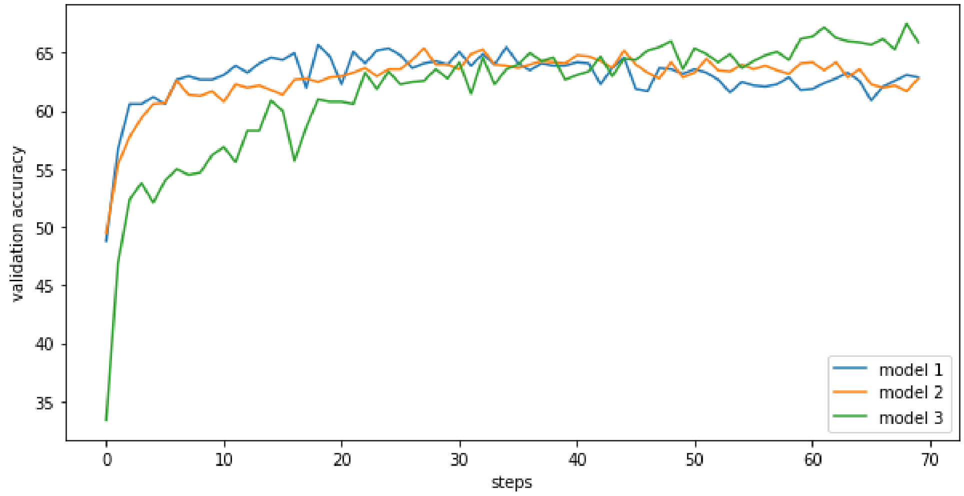

Figure 2: Validation curve for different learning rates

From the validation curve we see that although the validation accuracy increase slowly for model 3, it finally turns out to have the highest validation accuracy among the three models. This shows that interacting the two encoded sentences by element-wise multiplication may help the model to generalize, and thus fit better into validation data.

model 4 is the RNN model with hidden size = 200, interacting the two encoded sentences by concatenation.
model 5 is the RNN model with hidden size = 100, interacting the two encoded sentences by concatenation.
model 6 is the RNN model with hidden size = 200, interacting the two encoded sentences by element-wise multiplication.
model 7 is the RNN model with hidden size = 200, interacting the two encoded sentences by element-wise multiplication, with dropout regularization.

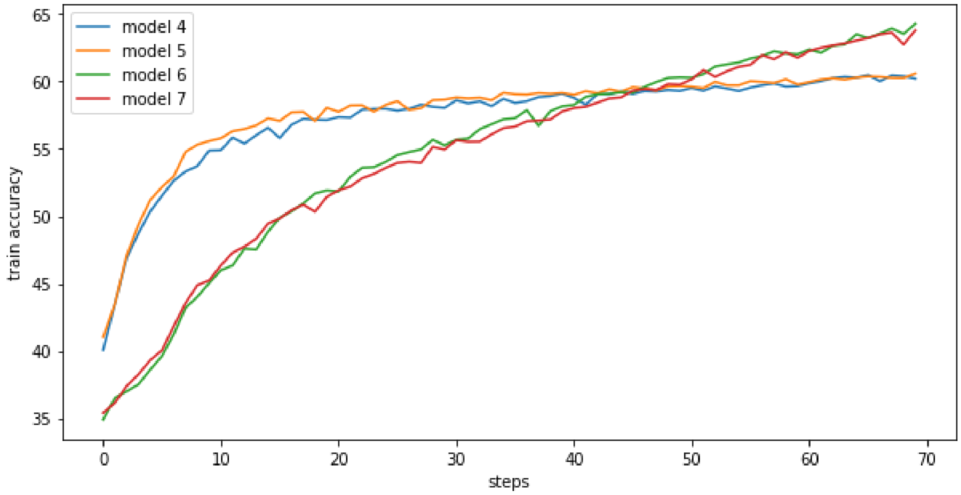

Figure 3: Validation curve for different vocabulary sizes

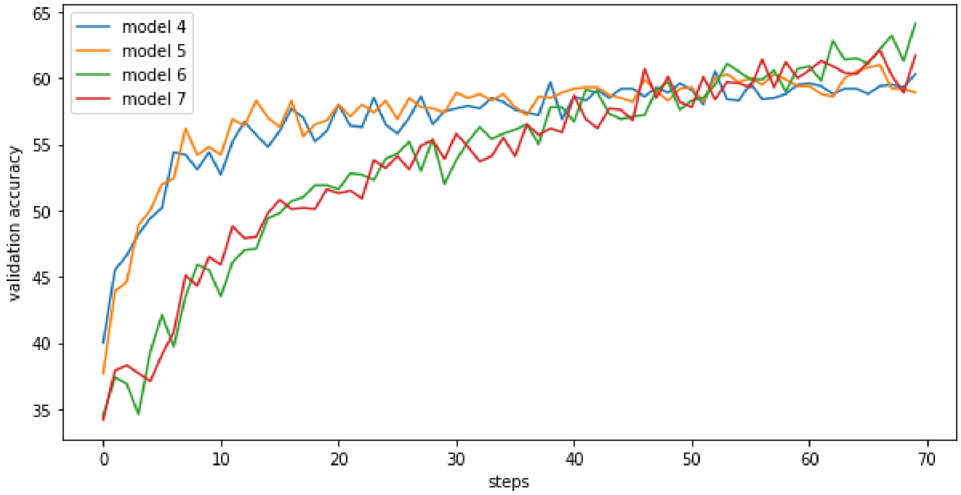

Figure 4: Validation curve for different vocabulary sizes

From both curves above we see that model 4 & 5 have similar performance and 6 & 7 have another similar performance. This shows that hidden size and regularization does not make significant difference on the model, while the interactive method of two encoded sentences has great effect on the accuracy of the model.

Although the accuracy of element-wise multiplication of two encoded sentences increases slowly at the beginning, it turns out to have much higher training and validation accuracy at the end, and the increasing rate remains for a relatively long time. In comparison, the increasing rate for model 4 & 5 becomes close to 0 at the early stage of training. We could deduce that model 6 & 7 is possible to have higher accuracy if we train for more epochs.

In conclusion, we find model 3 is the best for CNN and model 6 is the best for RNN.

**Number of trained parameters**

number of parameters in model 1 is 381203

number of parameters in model 2 is 140603

number of parameters in model 3 is 320803

number of parameters in model 4 is 924003

number of parameters in model 5 is 322003

number of parameters in model 6 is 683203

number of parameters in model 7 is 683203

**Correct and incorrect predictions**

Then we highlight 3 correct and 3 incorrect predictions in the validation set:

model 3 correct predictions:

1. The label is entailment

”Man and a woman walking on the street ”

”There are at least two people in the picture . ”

2. The label is neutral

”A young girl is swimming in a pool . ”

”The girl is practicing for a swim meet . ”

3. The label is contradiction

”A swimming dog with a small branch in its mouth . ”

”A dog is ice skating . ”

model 3 incorrect predictions:

1. The predicted label is contradiction, but the true label is neutral.

”The player from the black team is telling something to the player of the red team .”

”The two teams are relaxing before the match . ”

Analysis: Probably this is because ”telling something” and ”relaxing” are considered to be contradicted in the model.

2. The predicted label is entailment, but the true label is neutral.

”A man doing maintenance on the railroad tracks ”

”There is an older man doing work on the railroad tracks .”

Analysis: Probably this is because the training data contains many examples with maintenance and work at the same time and the labels are entailment.

3. The predicted label is entailment, but the true label is contradiction.

”A horse and rider on a ¡unk¿(steeplechase) course .”

”Three horses are playing in a field together . ”

Analysis: The vocabulary is not big enough to contain ”steeplechase”, and this may lead to misunderstand of the first sentence.

model 6 correct predictions:

1. The label is contradiction

”A man is photographing a small staked camel and a woman is trying to walk past it . ”

”A man and a woman are riding a camel . ”

2. The label is entailment

"A band performing on the corner of the street . "
"The band is performing outdoors . "
3. The label is entailment
"Two Asian women talking and having drinks at a small round table . "
"Two ladies are sitting together at the table . "

model 6 incorrect predictions:
1. The predicted label is neutral, but the true label is entailment
"A kid in a red and black coat is laying on his back in the snow with his arm in the air and a red sled is next to him . "
"It is a cold day . "
Analysis: Probably because the first sentence is much longer and the information like "snow" and "sled" cannot easily be referred to as "cold".
2. The predicted label is contradiction, but the true label is neutral
"A man who has a gray beard and gray hair laughs while wearing a purple shirt . "
"A man is laughing at a woman who has fallen over . "
Analysis: Since the part after "laugh" are totally different for these two sentences, so the model may takes these as contradiction.
3.The predicted label is neutral, but the true label is contradiction
"A waitress is serving customers at a restaurant . "
"The waitress is sitting in a chair ignoring the customers around her . "
Analysis: Probably this is because the train data seldom contain both "serve" and "sit" at the same time, and thus the sentences are considered to be unrelated.

**Evaluating on MultiNLI**

Table 1: Validation accuracy for different genres on MNLI data

| model | government | telephone | slate | fiction | travel |
|---|---|---|---|---|---|
| 3 (CNN) | 39.86 | 38.91 | 40.32 | 40.0 | 40.0 |
| 6 (RNN) | 39.96 | 39.50 | 37.72 | 41.21 | 39.20 |

The validation accuracy on mnli data is much lower than validation accuracy on snli data. This reflects that both models are not good enough to handle with the data that they have never seen. Both models have similar results on all the genres, which shows that their encoded methods have similar effects on the data. Fiction has relatively high accuracy among all the genres, and this is probably because the sentences in fictions are well organized, and relatively formal. In comparison, telephone data has low accuracy when using CNN model, and maybe this is due to the broad topics and variety of contents in the telephone data.

Link to the Github page: `https://github.com/cdsherry/DS-GA1011-NLP-hw2`