NLP HW1

Jianzhi Li, Jing Qian, Xin Wu, Yuhong Zhu

1 Training on SNLI

1.1 Model Implementation

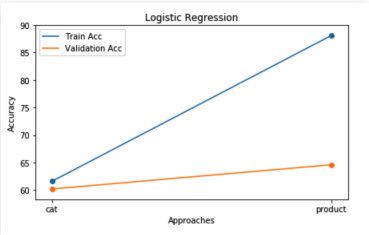
We implemented the logistic regression and neural networks models in PyTorch – as nn.Module. The logistic regression has two layers, an embedding layer that maps indexed data into vectors, and a linear layer that transfers embedded vectors to 3-class output. In neural networks, we have two more hidden layers which are both linear layers, so in total we have to add three layers and two activation functions. We use ReLu as our activation function. To be more specific, our first linear classifier maps our embedded data into vectors (from embedding dimension(we chose 100) to dimension 20), and then we apply an activation function to the result to get the non-linearity. Then we apply another linear classifier to the result data(from dimension 20 to dimension 10), and here again we apply the ReLu to the result to get the non-linearity. Lastly, we apply the third linear classifier to the result data (from dimension 10 to dimension 3).

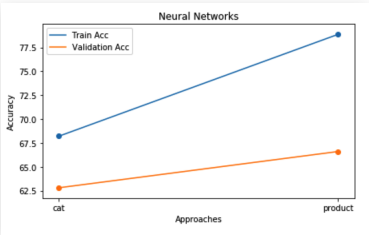
1.2 Hyper-parameters Tuning

For both models, we

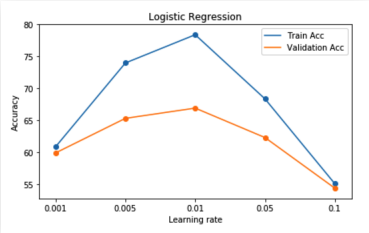
1. changed the interacting ways of the two encoded sentences: concatenation and element-wise multiplication.
2. tuned the learning rate from [0.001,0.005,0.01,0.05,0.1] as optimization hyperparameters
3. varied the size of the vocabulary [2000,5000,10000,15000]
4. varied the embedding dimension [50,100,200,300]

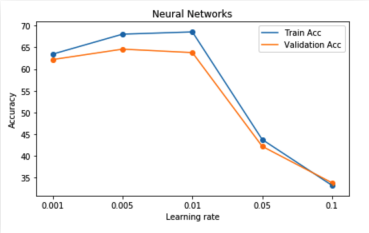
The plot below shows that element-wise multiplication performs better. In Logistic regression model, element-wise multiplication's validation accuracy = 64.6% and concatenation's validation accuracy = 60.2%. In Neural Network, element-wise multiplication's validation accuracy = 66.6% and concatenation's validation accuracy = 62.8%



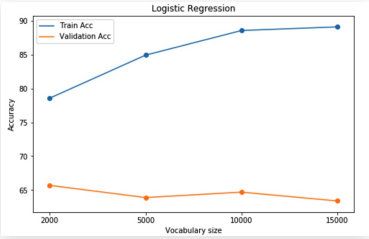


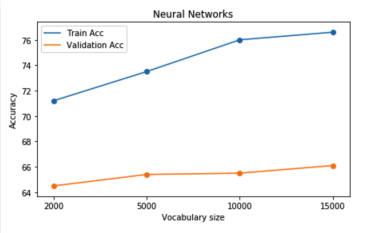
The plot below indicates that for logistic regression, when learning rate = 0.01, the validation accuracy = 66.9% is the highest, and that for neural network, the when learning rate = 0.005, the validation accuracy = 64.6% is the highest.



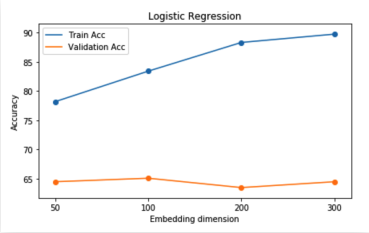


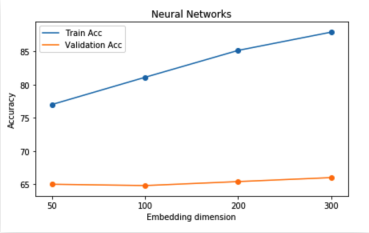
The plot below shows that for Logistic Regression when size of vocabulary = 2000, the validation accuracy =65.7%. And for neural Networks, when learning rate is 15000, the validation accuracy =66.1% is the highest.





The plot below shows that for logistic regression when embedding size = 100 , the validation accuracy = 65.1%. For neural networks when embedding size = 300, the validation accuracy = 66.0%





1.3 Result Analysis

False predictions:

1. True label: (Contradiction)

Predicted label: (Entailment)

Three women on a stage , one wearing red shoes , black pants , and a gray shirt is sitting on a prop , another is sitting on the floor , and the third wearing a black shirt and pants is standing , as a gentleman in the back tunes an instrument .

There are two women standing on the stage

2. True label: (Entailment)

Predicted label: (Contradiction)

Four people sit on a subway two read books , one looks at a cellphone and is wearing knee high boots .

Multiple people are on a subway together , with each of them doing their own thing .

3. True label: (Entailment)
Predicted label: (Contradiction)
Two people are in a green forest .
The forest is not dead .
Reasoning: It is likely that the proportion of overlapping words has an effect on the final prediction. The three examples with wrong predicted results are all long sentences and have similar lengths for premise and hypothesis. The first example with many word overlaps is predicted as entailment while the other two have few overlaps. However, the true labels are opposite. The model may learn sentences as highly correlated when they contain similar information, and therefore the predicted label turns to be entailment and vice versa. The three correctly predicted examples can further support our hypothesis: all three are short and are entailment. They all have high proportions of same words in the sentences and thus are predicted correctly as entailment.

Correct Predictions:
1. True Label:(Entailment)
Predicted Label: (Entailment)
Man observes a wavelength given off by an electronic device .
The man is examining what wavelength is given off by the device .

2. True Label:(Entailment)
Predicted Label:(Entailment)
bicycles stationed while a group of people socialize .
People get together near a stand of bicycles .

3. True Label: (Entailment)
Predicted Label: (Entailment)
Man in overalls with two horses .
a man in overalls with two horses.

2 Evaluating on MultiNLI

Validation Accuracy		
	Logistic Regression	Neural Networks
fiction	32.26%	34.57%
telephone	32.64%	32.64%
slate	34.43%	34.83%
government	31.29%	32.19%
travel	32.99%	29.12%

By comparing the validation accuracy across genres, we found out that slate has the highest accuracy in both models, with validation accuracy in logistic regression is 34.43%, and with validation accuracy in neural

networks is 34.83%, and government has the lowest accuracy with validation accuracy in logistic regression, which is 31.29% and travel has the lowest validation accuracy in neural networks, which is 29.12%. The difference means that our trained model is best fitted to the slate genre, and least fitted to the government genre and travel genre.

It is reasonable that the SNLI validation accuracy is much higher than those genres, because the model is trained based on the SNLI data. Thus, this model has a much higher accuracy when evaluating on the validation set of SNLI; where as the accuracy on the MNLI validation dataset is much lower.

3 Fine-tuning on MultiNLI

Validation Accuracy		
	Logistic Regression	Neural Networks
fiction	32.16%	34.57%
telephone	35.82%	31.34%
slate	33.13%	35.03%
government	33.46%	34.84%
travel	34.52%	31.57%

4 Pre-Trained Word Embedding

We downloaded the 1 million word vectors trained on Wikipedia 2017 and took the first 50,000 words. We then added unknown $< UNK >$ and padding $< PAD >$ vector into the weight matrix which is of dimension (50002,300). We froze all the embedding vectors except for the $< UNK >$ word embedding and trained the model on SNLI data. Afterwards, we tested the model on SNLI validation data and MNLI validation data by genre.

We trained the Logistic regression model and used the pretrained weight matrix and the validation accuracy is 41.00%.

We trained the Neural Networks model and used the pretrained weight matrix and the validation accuracy is 35.44%

Validation Accuracy		
	Logistic Regression	Neural Networks
fiction	31.46%	34.87%
telephone	29.45%	36.52%
slate	35.13%	34.83%
government	29.04%	36.71%
travel	30.86%	35.44%