Fall 2018 NLP Assignment 2

adc563

October 31, 2018

1 Code

The code can be found on the following **page on GitHub**.

2 Data & the Problem

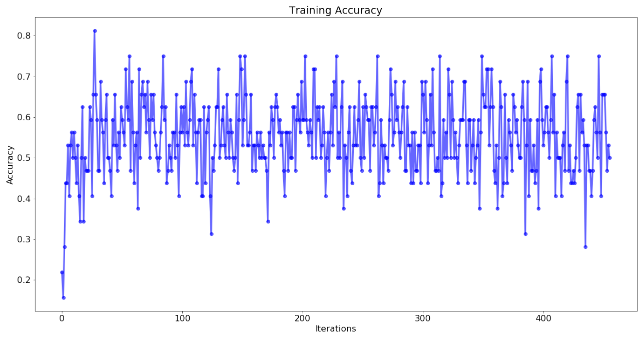
The models are trained and validated on the provided SNLI dataset that consists of hypothesis-premise pairs with entailment, neutrality, or contradiction labels. The performance is evaluated on the MultiNLI dataset. In all models, pretrained word embeddings from **https://fasttext.cc/docs/en/english-vectors.html** were used. The task is a 3-class classification problem, based on predicting whether the premise entails, is neutral toward or contradicts the hypothesis.

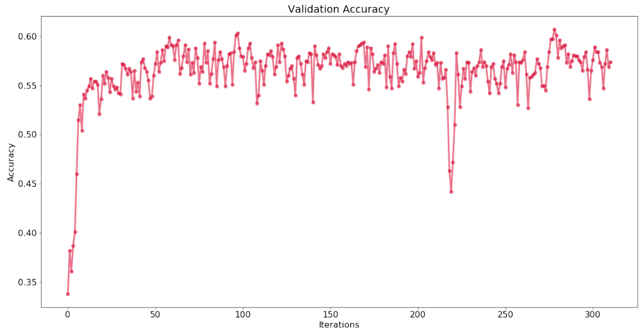
3 The Models and Training

3.1 CNN Encoder

In the CNN encoder, two convolutional layers with ReLU nonlinearity and max-pooling are used. The embeddings of the two sentences -the hypothesis and the premise- are fed into the encoder separately. Both go through the two convolutions, each followed by ReLUs, and then max-pooled. Once the hidden representations come out of the max-pools, they are interacted (concatentation, subtraction or element-wise multiplication were tried), and the single vector that is the result of this interaction goes through two consecutive fully connected layers. Finally, the output is sent to softmax to generate the probability distribution.

The evaluation metric of the model is cross entropy, and the optimizer used is Adam. Below are the initial 10-epoch training (left) and validation (right) accuracy plots of the model.





The left axis shows the initial training (80%) and validation (60%) accuracy scores reached. Please see the next page for hyperparameter search results.

3.1.1 Hyperparameter Search for CNN

The hyperparameter search is done on the SNLI validation set. Please see the sets for each hyperparameter included in Table 1 below.

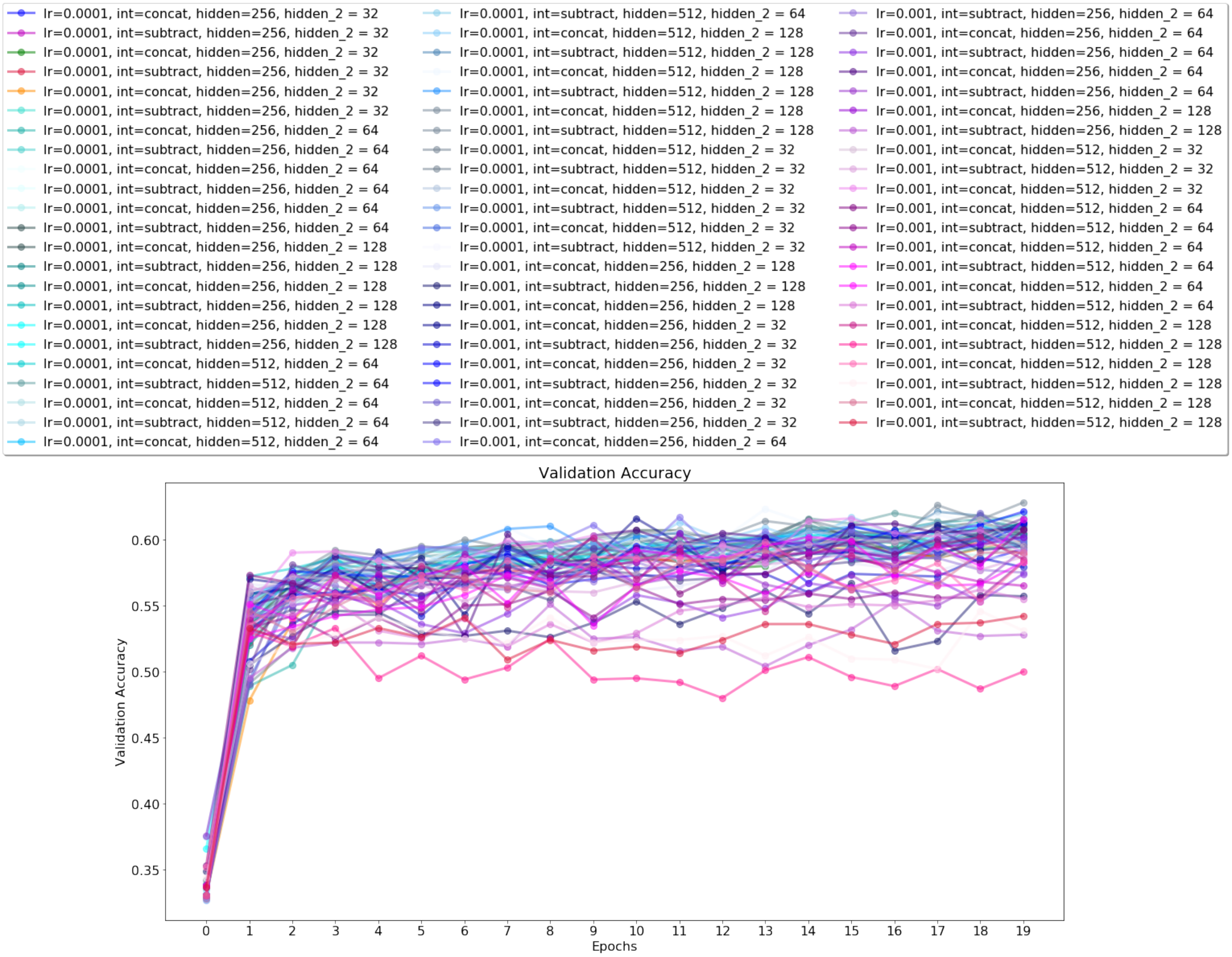
Table 1: Hyperparameter Space

Parameter	**Value Set**
Learning Rate	$\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$
Hidden Size (1)	{128, 256, 512}
Hidden Size (2)	{128,64,32}
Kernel Size	{3, 5, 11}
Interaction	{concat., mult., subt.}
Dropout	{0.1, 0.3, 0.5}

The highest validation accuracy score reached by a **CNN** - with configuration:lr=10^{-3},hidden=512, hidden_2=64, dropout=0.1 - is **63%** on the SNLI validation set. In general, it was observed that,

- Concatenation and subtraction learn and generalize better than multiplication,
- Relatively smaller hidden sizes (¡512) learn slower however generalize better than relatively large hidden sizes,
- Kernel size 3 performs better than kernel size 5 in validation.

On the below figure, the validation accuracy scores of the models included in the CNN hyperparameter search are plotted. The legend is large due to the large size of parameter space, however, it can be seen that models with larger hidden sizes (e.g. the pink lines) have shown inferior performance, even though they had been performing very well in terms of training accuracy. On the other hand, the validation scores of the models with smaller (e.g. 128) hidden size are much better (around 63%).



3.2 RNN Encoder

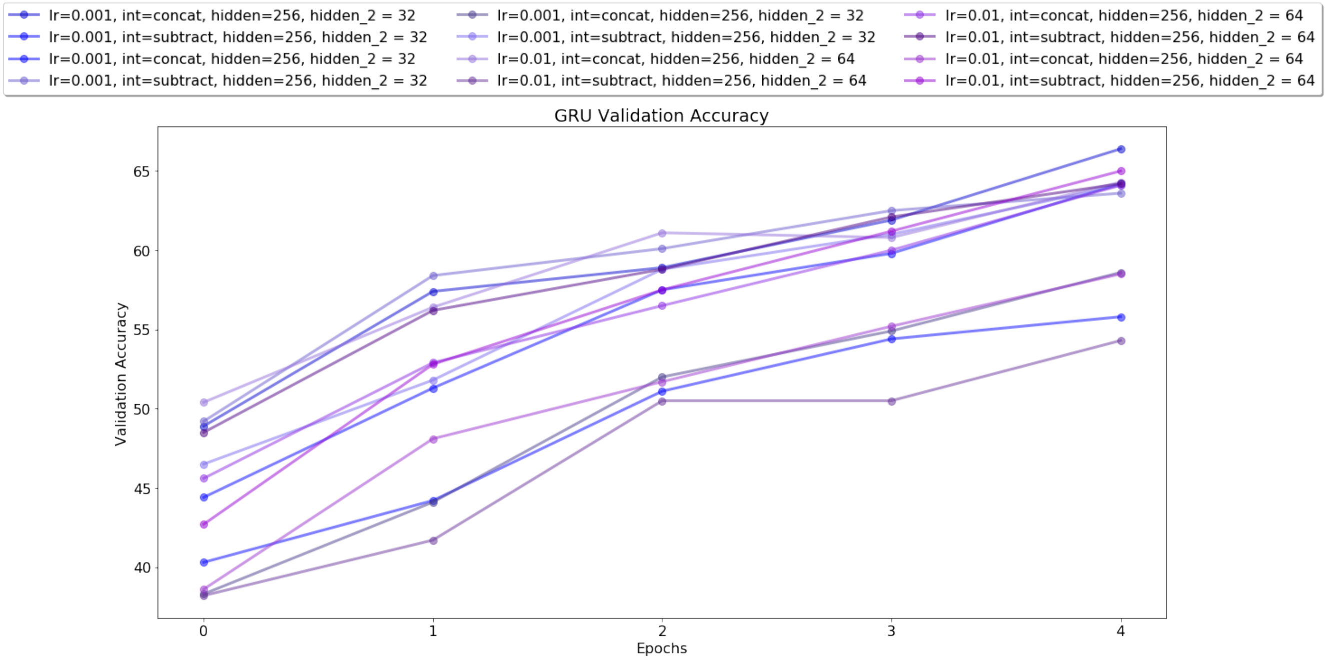
The RNN encoder network (class biGRU) is a single-layer bidirectional GRU. The hidden states of each sentence is obtained by summing up bi-GRU outputs across time. Here, I should note that I changed the strategy and used a single data loader and separated the GRU and linear forward passes into two networks since the first network I tried (both encoder and linear in the same class) did not learn. As a consequence, the encoder model is trained twice; once to embed the first sentence and once to embed the second. Then, these two embeddings are fed into the linear layer network.

3.3 Hyperparameter Search

The hyperparameter search space for the bidirectional GRU encoder network is relatively smaller than that of CNN, since it was training training slower and CNN and it was implemented later. However, even with the smaller search space the GRU encoder reached better reults on the validation set.

Parameter	**Value Set**
Learning Rate	$\{10^{-4}, 10^{-3}, 10^{-2}\}$
Hidden Size (1)	{512}
Hidden Size (2)	{128,64}
Interaction	{concat., mult., subt.}
Dropout	{0.1}

The maximum validation accuracy reached by a GRU is **66.4**%. The configuration is: lr=10^{-3}, hidden size=512, linear input size=64, interaction=concatenation, and dropout=0.1. Please see the validation accuracy plots from the first 5 epochs below.



4 Multi-NLI Test Performance

4.1 CNN

The best CNN model is selected based on the validation accuracy on the SNLI dataset adn evaluated on the Multi-NLI genres. The results are in the below table. The validation score of the same CNN model was 64% on the SNLI validation set, however it can be seen that the best test accuracy score attained is only 47.66%.

There are several possible causes of the lower accuracy on Multi NLI dataset.

- Since the sentence pairs in the MNLI dataset are genre-specific, I expect the classifier to embed many more **unknown** token then it did for the training and validation sets (SNLI).

- Even if the tokens are known to the classifier, they might have been misinterpreted since the **context** they are used is different then the language context they have been used in the training and validation sets.

Please see Section 5 for misclassification examples.

Table: Best CNN Test Results

Genre	Test Accuracy
Telephone	47.66%
Slate	40.12%
Travel	44.60%
Government	47.34%
Fiction	45.43%

4.2 RNN

The best RNN model selected based on the hyperparameter search is evaluated on Multi-NLI genre validation sets.

Table: Best RNN Test Results

Genre	Test Accuracy
Telephone	45.12%
Slate	42.06%
Travel	43.40%
Government	45.08%
Fiction	41.91%

5 Misclassification Examples

An example from each genre was selected.

Government
Sentence 1: [’IDPA’, ”’s”, ’OIG’, ”’s”, ’mission’, ’is’, ’to’, ’prevent’, ’detect’, ’and’, ’eliminate’, ’fraud’, ’waste’, ’abuse’, ’and’]
Sentence 2: [’IDPA’, ”’s”, ’OIG’, ”’s”, ’mission’, ’is’, ’clear’, ’and’, ’cares’, ’about’, ’payment’, ’programs’]
Explanation: This example was labeled as entailment, while it’s true label is neutral. Since this is the government genre, there are many words that are used in different contexts than their daily or more informal uses. This might be the cause of misclassification, even though there are not any unknown tokens.

Telephone
Sentence 1: [’because’, ’like’, ’Tech’, ’is’, ’known’, ’to’, ’be’, ’a’, ’good’, ’engineering’, ’school’, ’and’, ’A’, ’and’, ’M’, ’maybe’, ’is’, ’known’, ’more’, ’for’]
Sentence 2: [’A’, ’and’, ’M’, ”’s”, ’computer’, ’department’, ’is’, ’unk’, ’very’, ’well’, ’regarded’, ’.’]
Explanation: This example was labeled as neutral, but in fact it’s contradiction. The language is very informal.

Slate
Sentence 1:[’A’, ’button’, ’on’, ’the’, ’Chatterbox’, ’page’, ’will’, ’make’, ’this’, ’easy’, ’,’, ’so’, ’please’, ’do’, ’join’, ’in’, ’.’]
Sentence 2: [’They’, ’had’, ’to’, ’submit’, ’a’, ’written’, ’request’, ’before’, ’being’, ’accepted’, ’.’]
Explanation: This pair is labeled as entailment, but it is a contradiction. However, it is very hard to understand without taking into account the context.

Fiction
Sentence 1: [’And’, ’if’, ’they’, ’did’, ’come’, ’,’, ’as’, ’remote’, ’as’, ’that’, ’is’, ’,’, ’you’, ’and’, ’your’, ’men’, ’look’, ’strong’, ’enough’, ’to’, ’handle’, ’anything’, ’.’]
Sentence 2: [’The’, ’men’, ’were’, ’warriors’, ’.’]
Explanation: This pair is labeled entailment, but its true label is neutral.

Travel
Sentence 1:['In', 'this', 'enclosed', 'but', 'airy', 'building', ',', 'you', "'ll", 'find', 'ladies', 'with', 'large', 'machetes', 'expertly', 'chopping', 'off', 'hunks', 'of', 'kingfish', ',', 'tuna', ',', 'or', 'shark', 'for', 'eager', 'buyers', '.']
Sentence 2: ['You', "'ll", 'find', 'small', 'lepers', 'chopping', 'of', 'chunks', 'of', 'tuna', ',', 'its', 'the', 'only', 'place', 'they', 'can', 'work', '.']
Explanation: This pair is a contradiction example that is labeled as neutral. There are not any unknown tokens however it is a very unusual context, even for travel.