ARTIFICIAL NEURAL NETWORK
HOMEWORK 3 REPORT

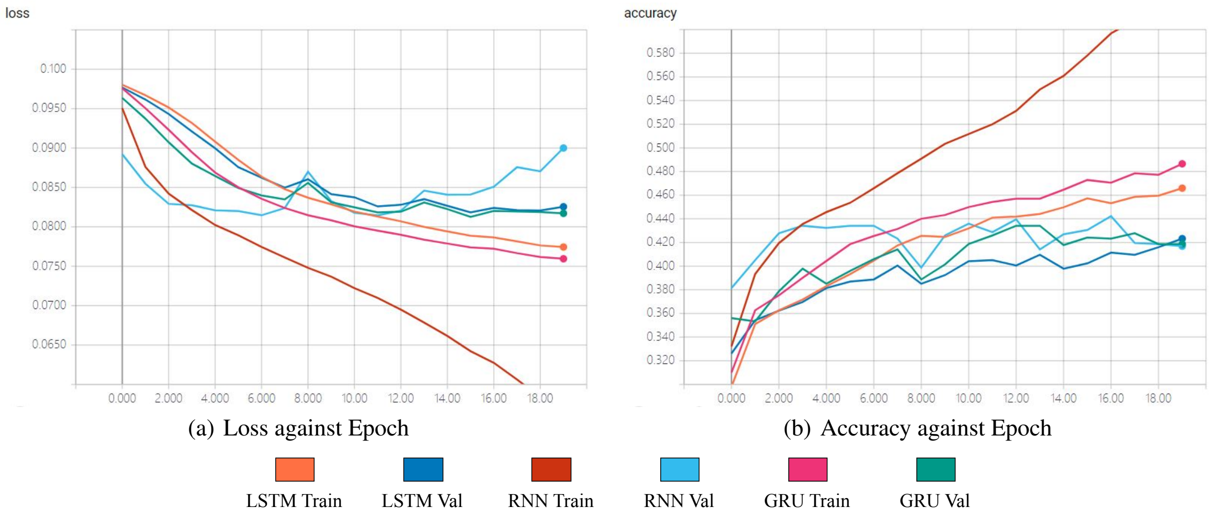
Jiahao Li
Department of Computer Science
Tsinghua University
lijiahao17@mails.tsinghua.edu.cn

1 Experimental Settings

Unless otherwise specified, all experiments are conducted with the same settings stated as follows. Every model is with one recurrent layer which consists of 512 hidden units. It is trained by a gradient descent optimizer with default learning rate of 0.005 for 20 epochs with batch size of 16.

2 Loss and Accuracy of Three RNN Models

As shown in Figure 1, the loss and accuracy values against every epoch are plotted with `tensorboard`, including training and validation process of 3 kinds of RNN cells. The experimental results are also provided in Table 1.



(a) Loss against Epoch

(b) Accuracy against Epoch

Figure 1: Loss and accuracy values against every epoch.

Model	Train Loss	Train Accuracy (%)	Validation Loss	Validation Accuracy (%)
RNN	**0.0564**	**64.75**	0.0851	**44.23**
GRU	0.0759	48.65	**0.0819**	43.42
LSTM	0.0774	46.59	0.0825	42.33

Table 1: Experimental results of three kinds of RNN cells.

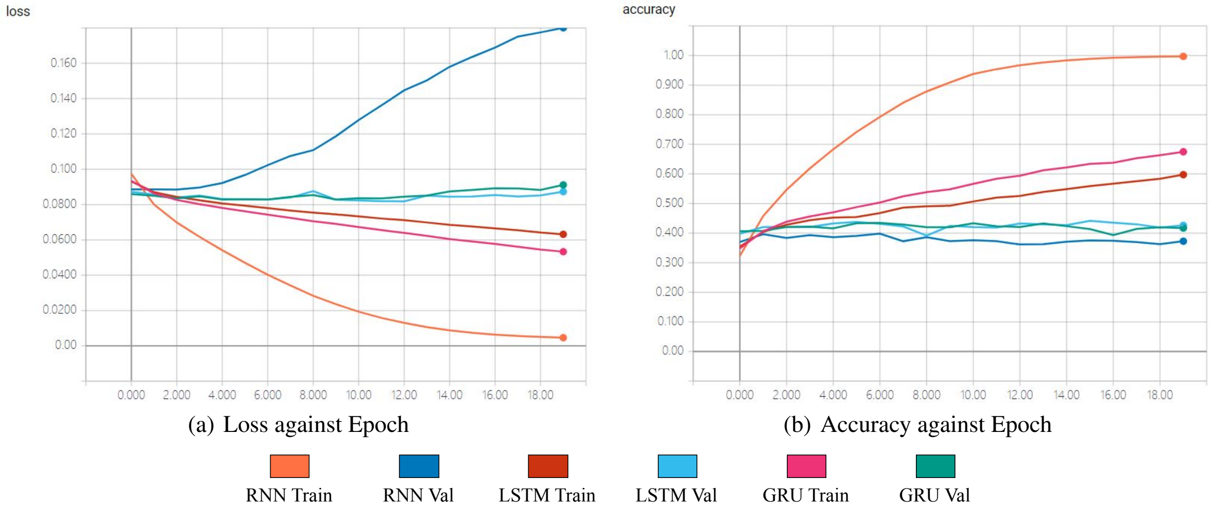
The basic RNN cell exhibits serious overfitting problems. It achieves highest accuracy and lowest loss on training set but obtains relatively poor results on validation set. Despite the overfitting issue, it achieves the best performance among these three kinds of cells.

Compared to to the basic RNN cell, the basic LSTM cell requires greater amount of computation and thus more time to train. After being trained for the same number of epoch, it results in an inferior performance. However, it can be observed that the accuracy on validation set keeps going up, indicating that its potential is not fully developed due to the limited training epochs.

The GRU cell performs slightly better than the basic LSTM cell, achieving higher accuracy and lower loss in both training and validation phases. Unlike basic RNN cell, both GRU and basic LSTM cells do not have noticeable overfit problem.

3 Ablation Study of Self-Attention Mechanism

To evaluate the effect of self-attention mechanism, the self-attention module is replaced by a linear layer, while the other parts of the model remain the same. The only variable in this experiment is the self-attention module. The experimental results are shown in Figure 2 and Table 2.



(a) Loss against Epoch

(b) Accuracy against Epoch

Figure 2: Loss and accuracy values against epoch without applying self-attention mechanism.

Model	Train Loss	Train Accuracy (%)	Validation Loss	Validation Accuracy (%)
RNN-with-Attention	0.0564	64.75	**0.0851**	**44.23**
RNN-without-Attention	**0.0046**	**99.70**	0.1023	39.78
GRU-with-Attention	0.0759	48.65	**0.0819**	**43.42**
GRU-without-Attention	**0.0533**	**67.46**	0.0830	43.32
LSTM-with-Attention	0.0774	46.59	**0.0825**	42.33
LSTM-without-Attention	**0.0632**	**59.73**	0.0845	**44.14**

Table 2: Experimental results of three kinds of RNN cells with and without self-attention mechanism.

Without self-attention module, the overfitting issues of all three models become more conspicuous. Compared to the self-attention models, the validation accuracies of GRU and basic RNN cells without attention mechanism have dropped 4.5% and 0.1%, respectively. Besides, the slight improvement of LSTM cells may be largely due to the reduction of total parameters, which enables the LSTM to learn more rapidly. In general, the self-attention module plays an important role in preventing overfitting and improving the performance because it provides proper context information of a sequence.

4 Hyperparameter Finetuning

4.1 Comparison of Different Layers

In this section, the performances of models with 1 and 2 layers are evaluated with the same settings. The experimental results are shown in Table 3. In general, models with 2 layers obtain worse results on validation set than the 1-layer models do. A possible reason for this phenomenon is that the 2-layer networks have more parameters and require more time to train.

Model	Train Loss	Train Accuracy	Validation Loss	Validation Accuracy
1-layer-RNN	0.0564	64.75	0.0851	**44.23**
2-layer-RNN	**0.0488**	**70.33**	**0.0825**	44.05
1-layer-GRU	**0.0759**	**48.65**	**0.0819**	**43.42**
2-layer-GRU	0.0768	47.41	0.0838	42.51
1-layer-LSTM	**0.0774**	**46.59**	**0.0825**	**42.33**
2-layer-LSTM	0.0806	43.70	0.0842	40.60

Table 3: Experimental results of 3 kinds of models with 1 and 2 layers.

4.2 Comparison of Different Units

To evaluate the effect of hidden units, the number of hidden units of 3 kinds of RNN cells is adjusted from $\{256, 512, 1024\}$. The experimental results are shown in Table 4. It is shown that the overfitting problem of basic RNN cell is significantly relieved by the hidden units reduction, which also improves its performance on the validation set.

Model	Train Loss	Train Accuracy	Validation Loss	Validation Accuracy
256-unit-RNN	0.0652	56.91	**0.0792**	**46.41**
512-unit-RNN	0.0564	64.75	0.0851	44.23
1024-unit-RNN	**0.0440**	**73.93**	0.0824	44.78
256-unit-GRU	0.0768	47.72	0.0818	42.87
512-unit-GRU	0.0759	48.65	0.0819	**43.42**
1024-unit-GRU	**0.0754**	**49.08**	**0.0817**	43.32
256-unit-LSTM	0.0777	46.43	**0.0822**	41.87
512-unit-LSTM	0.0774	46.59	0.0825	42.33
1024-unit-LSTM	**0.0773**	**46.77**	0.0823	**43.32**

Table 4: Experimental results of models with different sizes of hidden units.

4.3 Comparison of Different Epochs

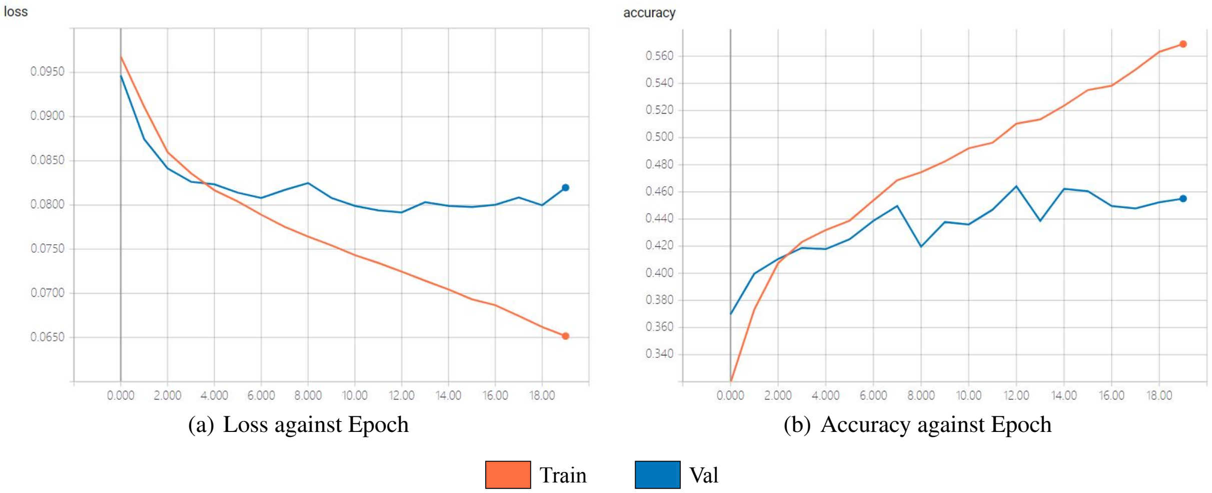
A larger epoch of 100 is also applied in this experiment to fully exploit the potential of models. The experimental results are shown in Table 5. A large epoch enables the model to converge completely. In spite of the overfitting problem, the best validation results are far better than that of a small epoch.

Model	Train Loss	Train Accuracy	Validation Loss	Validation Accuracy
20-epoch-RNN	0.0564	64.75	0.0851	44.23
100-epoch-RNN	**0.0002**	**100.00**	**0.0818**	**44.35**
20-epoch-GRU	0.0759	48.65	0.0819	43.42
100-epoch-GRU	**0.0476**	**70.12**	**0.0815**	**44.60**
20-epoch-LSTM	0.0774	46.59	0.0825	42.33
100-epoch-LSTM	**0.0504**	**67.66**	**0.0819**	**44.12**

Table 5: Experimental results of models trained for different epochs.

5 Loss and Accuracy of Final Model

The loss and accuracy of the final model is shown in Figure 3.



(a) Loss against Epoch

(b) Accuracy against Epoch

Figure 3: Loss and accuracy values against every epoch.

6 Implementation Details of Final Model

The final network is an RNN model with 1 layer and 256 units. The model is trained for 20 epochs with batch size of 16. A Gradient Descent Optimizer with learning rate of 0.005 is used for training. The other hyperparameters remain unchanged. The best checkpoint on validation set achieves an accuracy of 46.41%, and is selected as the final model. The results of the final model on test set are generated and submitted.

The basic RNN cell is a simple and effective model for sentence-level sentiment classification. Compared to the GRU cell and the basic LSTM cell, it requires less computational cost and converges more quickly. Its biggest weakness is the overfitting problem. To address this issue, reducing the hidden units is an alternative. It cuts down the total parameters, and thus alleviates the overfitting problem. Despite the low accuracy on the training set, it effectively improves the performance on the validation set.