Advamced Machine Learning Homework 1

Kolmakova Elizaveta
Innopolis University
Kazan, Russia
e.kolmakova@innopolis.university

***Index Terms*—AML, LSTM, params tuning, comparison, sequence modeling**

I. Introduction and Motivation

This is a comprehensive report-comparison of 3 machine deep learning models: Simple LSTM, LSTM with the best parameters and simple neural network without any recurrent connections. The motivation of this work is ho investigate whether LSTM network is capable to get significantly better performance in comparison with simple non-recurrent network and find out the importance of parameter selection, rather it is important part. The bigger ultimate question is complication worth to compute in order to get better performance.

The work is constructed from the following sections: section 2 briefly describes theoretical aspects and architecture of every model, section 3 introduce experiment settings, section 4 shows graphical results and section 5 represent conclusions with discussion.

1) Section 1. Introduction and Motivation.
2) Section 2. Theoretical aspects and architectures.
3) Section 3. Experiments settings.
4) Section 4. Results.
5) Section 5. Conclusion and discussion.

II. Theoretical aspects and architectures.

Three models were used in this work: basic LSTM, upgraded LSTM with best parameters, simple neural network with linear layers.

A. Basic LSTM

LSTM network consists from lstm-cells with two subcells: memory and state. Every cell has three gates: remember gate, forget gate and focus gate. Remember gate analyse the income information whether it is important to add to long-term memory. Forget gate is capable to say what information is no longer relevant, while focus gate define the information which is supposed to be considered at the moment. Due to this architecture the network is capable to detect long-term dependencies within embedding representation of some sequence information.

The basic model contains simply input layer for embedding representation, LSTM layer and output layer for final results.

B. Upgraded LSTM

This model has passed parameter tuning process which had defined the most appropriate parameters within which the performance of the network is the best in comparison with other configurations. The main parameters were: optimizer, amount of training epochs, loss function computation, learning rate, etc.

C. Simple model

This model simply consists from two linear layers with different dimensions.

III. Experiments settings

The task is to train a neural network to distinguish female name from male name. The given dataset contains list of different names and its gender.

The initial data is important to analyze since its representation highly impacts on the network results. Therefore every name were translated into embedding representation: every letter in word were encoded according to its char number. This is the most naive representation.

In order to conduct a validation there was a split on training and test sets. From definition test set is not supposed to be used for validation, but due to fact we obtain an additional set with labeled data gives us an opportunity to use it for validation, while typical test set should be applied for prediction.

To get more stable training the normalization of data is needed. In this work MinMaxScaler was applied to tranform values to particular interval from -1 to 1. Output results were embedded through one-hot-encoder: female name was encoded as [0,1], male name - [1,0]. The following stage for data was creation of tensors, tensor datasets and tensor loaders in order to work with Pytourch since it is a requirement of pytourch.

A. LSTM

Technically, LSTM network for this task contains from three main part: LSTM layer with hidden cells and final linear layer.

B. Upgraded LSTM

In order to obtain the best parameters which would provide with the highest performance measured by accuracy, the simple version of GridSearch was applied. GridSearch is simple algorithm of brute force of all parameters. There were 5 parameter's groups: learning rate, amount of epochs,

optimizers, loss functions, amount of hidden layers, - 240 unique combination.

Due to computational complication some short amount of datapoints (1280 out of 83288) were used in GreedSearch. Further there is a presentation of different parameters group and combinations with the best accuracy and the best convergence rate.

1) Learning rate: 0.001, 0.01, 0.1
2) Amount of epochs: 1, 11, 21, 31
3) Optimizers: Adam and SGD
4) Loss functions: BCE and MSE
5) Amount of hidden layers: 100, 150, 200, 250, 300

The table I represent combinations with the best accuracy:

Based on the result presented in table **??** other parameters were suggested to be checked.

1) Learning rate: 0.01, 0.1
2) Amount of epochs: 30, 50, 70, 90
3) Optimizers: Adam and SGD
4) Loss functions: BCE and MSE
5) Amount of hidden layers: 150, 200, 250, 300

This new parameters invoke the following results in Table II:

After exploration of results of GridSearch the following parameters were selected as preferable: learning rate = 0.01, SGD optimizer, BCE loss function, amount of epochs = 70, amount of hidden layers is 200.

C. Simple model

Simple model does not contain any recurrent or other complicated layers, while there are only two linear layers.

After training different architectures of neural networks the weights were saved and then reloaded to apply it on test dataset in order to understand the performance of proposed model. The performance of proposed models are described in the following section.

IV. Results

The table III represent result of every model based on shown accuracy. The following graphics represent:

1) Dependency between training accuracy at the moment of different training epochs
2) Dependency between training loss at the moment of different training epochs
3) Dependency between training f-measure at the moment of different training epochs
4) Dependency between training computational time at the moment of different training epochs

Figure 1 shows plots with dependencies, then distribution of weights and biases for every layer and, finally, histogram of these weights and biases for layers for basic NN with two linear layers.

Figure 2 shows plots with dependencies, then distribution of weights and biases for every layer and, finally, histogram of these weights and biases for layers for basic LSTM model.

Figure 3 shows plots with dependencies, then distribution of weights and biases for every layer and, finally, histogram of these weights and biases for layers for LSTM model after parameter tuning.

After exploration of results from Figure 3 the next issue was detected: due to too small dataset for parameter grid search not all essential information was detected therefore suggested best parameters were selected not precisely. That is why another configuration of parameters was also considered: learning rate = 0.001, Adam optimizer, BCE loss function, amount of epochs = 50, amount of hidden layers is 200. The table III can be updated then into Table IV and graphical representation can be updated info Figure 4 as well. In Figure 4 orange color represent previous parameters, blue color represents new relevant parameters.

The following plots in Figure 6 and distribution graphics in Figure 7 would present a difference in model performance or weight distribution. Color representation is presented in Figure **??**.

V. Conclusions

Based on the results the following was concluded:

1) Performance of LSTM model is better than performance of linear neural network. This is due to sequence nature of data, since it is simply the sequence of encoded letters.
2) But the level of complexity of LSTM model is an important point. For such easy task as short sequence binary classification, having the basic LSTM model was enough for obtaining a results with the probability of correct answer more than 70%. The model with better parameters has not managed to overperform significantly. Moreover, the computational time had grown twice for such complex model with best parameters.
3) LSTM model with better parameters which has managed to show significant performance in f-measure in comparison with others and one of the best performance in accuracy turned out to be overfitted. The accuracy on test set is significantly less than accuracy of training. This indicate high variance of the model, while the variance and biases as well should be in balance.
4) The best choice in terms of efficiency is basic LSTM model since it showed the lowest loss, good performance in accuracy and f-measure and was relatively fast to be computed.

Number	Accuracy, %	Learning rate	Optimizer	Loss	Epochs	Hidden layers
Best result accuracy						
1	69	0.01	SGD	MSE	31	250
2	70	0.1	SGD	BCE	31	150
3	71	0.01	SGD	MSE	31	200,150,100
4	72	0.01	SGD	BCE	31	100
5	73	0.01	SGD	BCE	31	150
Best convergence						
1	from 48 to 61	0.1	Adam	MSE	21	200
2	from 52 to 67	0.01	Adam	MSE	31	250
3	from 53 to 69	0.1	Adam	BCE	31	150
4	from 58 to 70	0.01	Adam	MSE	31	250,200,150
5	from 50 to 67	0.1	Adam	BCE	21	200

TABLE I: Best parameters according to GridSearch

Number	Accuracy Initial, %	Best Accuracy, %	Learning rate	Optimizer	Loss	Epochs	Hidden layers
1	71	71	0.01	Adam	BCE	30	150
2	70	71	0.01	SGD	MSE	30	200
3	69	71	0.1	SGD	BCE	30	150
4	61	75	0.01	Adam	BCE	50	150
5	76	77	0.01	SGD	BCE	50	150
6	60	72	0.01	Adam	BCE	50	20
7	72	73	0.01	SGD	BCE	50	200
8	59	73	0.01	Adam	Mse	50	150
9	72	73	0.01	SGD	BCE	50	200
10	73	74	0.01	SGD	MSE	50	150
11	61	73	0.01	Adam	MSE	50	200
12	73	74	0.01	SGD	MSE	50	200
13	71	72	0.01	SGD	MSE	50	200
14	72	73	0.01	SGD	BCE	50	300
15	70	72	0.1	SGD	BCE	70	200
16	58	78	0.1	Adam	BCE	70	150
17	70	72	0.1	SGD	BCE	70	200
18	79	80	0.01	SGD	BCE	70	150
19	59	78	0.01	Adam	BCE	70	200
20	77	80	0.01	SGD	BCE	70	200
21	55	75	0.01	Adam	MSE	70	150
22	75	78	0.01	SGD	MSE	70	150
23	53	77	0.01	Adam	BCE	70	200
24	77	78	0.01	SGD	MSE	70	200

TABLE II: Best parameters according to GridSearch

Model	Initial Acc, %	Final Acc, %	Epochs	Test Acc, %	Overfitted?	Underfitted?
Simple	62	63	50	63	No	No
LSTM	63	73	50	73	No	No
Param LSTM	62	63	70	63	No	No

TABLE III: Test evaluation

Model	Initial Acc, %	Final Acc, %	Epochs	Test Acc, %	Overfitted?	Underfitted?
Simple	62	63	50	63	No	No
LSTM	63	73	50	73	No	No
Param1 LSTM	62	63	70	63	No	No
Param2 LSTM	62	73	50	63	Yes	No

TABLE IV: Test evaluation

Fig. 1: Results for basic linear neural network

Fig. 2: Results for basic LSTM neural network

Fig. 3: Results for LSTM neural network

Fig. 4: Results for LSTM neural network with different parameters in comparison manner

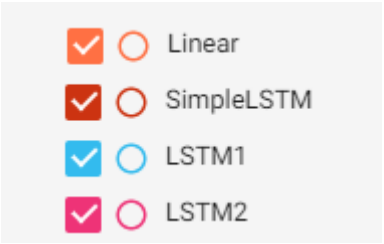

Fig. 5: Graphs colors

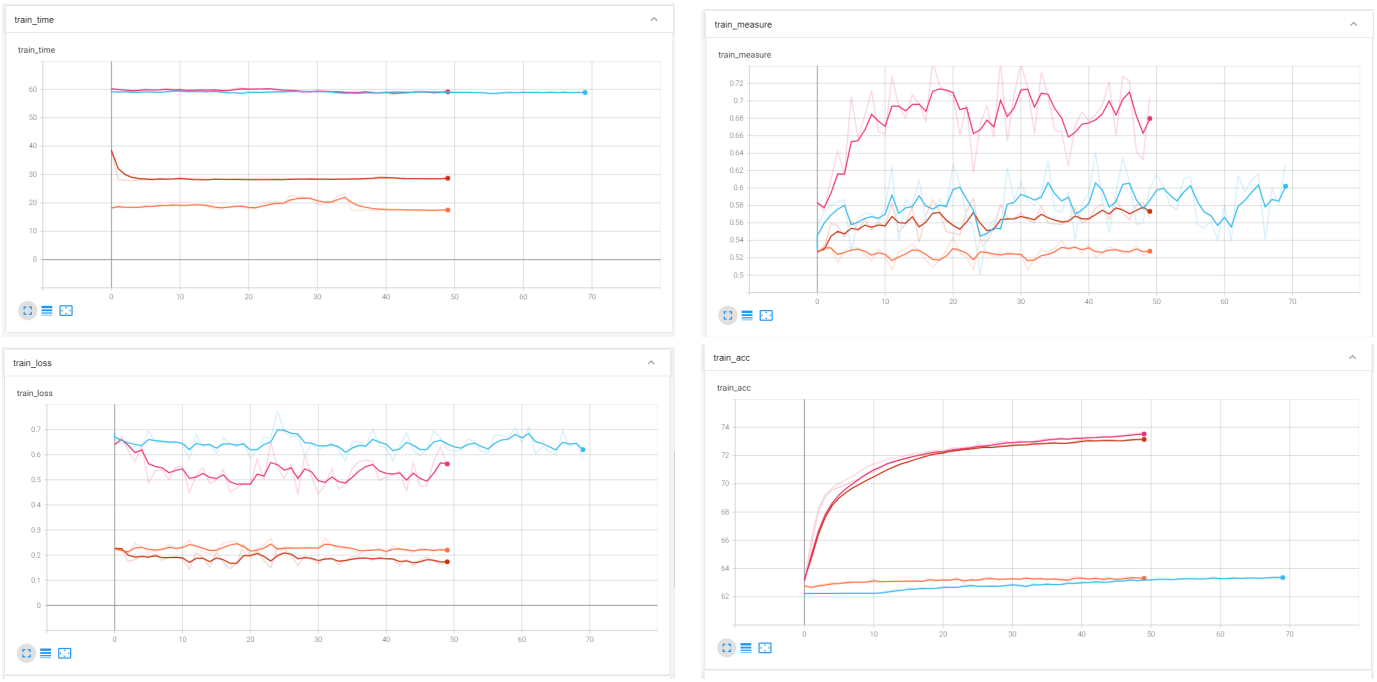

Fig. 6: Comparison of models in plots

Fig. 7: Comparison of weights distribution in models