# Convolutional LSTM for hand-written letter image sorting

Bae Byung Uk
Department of Electrical and Computer Engineering
Seoul National University
peardanny@snu.ac.kr

## Abstract

*The goal of hand-written letter image sorting is to classify EMNIST alphabet data set. Most of classification method use Convolutional Nerual Net(CNN) to solve this problem. In this paper, CNN and Long Short-Term Memory(LSTM) are used to perform classification. For better performance residual nets(ResNET) and ensemble method are used.*

## 1. Introduction

Classification of EMNIST alphabet datset is one of the criteria to evaluate model's performance for image classification. EMNIST dataset has 4 times more data than MNIST, which is a set of handwritten digits with a 28 x 28 format. As there are 26 alphabets, model has to perform 26 class classification. Most of the image classification are well performed by using CNN, however in this paper, we will use Convolutional LSTM(Conv LSTM) for this task.[1] ConvLSTM has two main phase of CNN and LSTM. Teacher forcing, ResNET, ensemble methods are used.

### 1.1. Convolutional Nerual Network (CNN)

First phase is CNN. CNN is the deep neural network, which uses convolution. CNN exploit local connectivity, parameter sharing,so it's useful to handle with 2D images. For ConvLSTM, CNN will take 10 sequential images. These 10 images will be taken to one CNN Net, so this CNN will be shared CNN. Every images will be 28 x 28 size of gray scale, and has class index from 0 t0 25 with alphabetical order. From these sequential images CNN will produce hidden representations for LSTM. [2]

### 1.2. Long Short-Term Memory (LSTM)

LSTM is one of the Recurrent Nerual Networks(RNN). LSTM makes the weight of a self-loop gated, integration of sequential data can be handled dynamically. This feature makes LSTM stronger than plain RNN. In this model, by using hidden representations of CNN, LSTM will produce 10 class predictions of 26 labels logits. [3]

### 1.3. Residual net (ResNET)

There are many CNN architectures and ResNET is one of them. When network becomes deeper, the model is much harder to trained due to gradient vanishing problem. By using residual block with skip connection, problem can be alleviated. For our ConvLSTM, ResNET will be used for CNN architecture.

### 1.4. Teacher Forcing

Teacher Forcing is a method for RNN. Teacher forcing gives ground truth to model. Instead of giving model's output of prior time step this will make model to converge faster. By using teacher forcing training will become faster and more efficient. To prevent overfitting, Teacher forcing will be regulated by teacher forcing ratio(0.5).

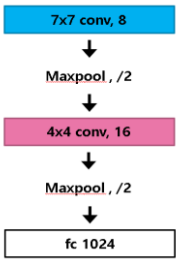
### 1.5. Ensemble

Ensemble method is a machine learning paradigm. Ensemble method trains multiple models simultaneously and integrate them together for better performance. Ensemble method is easy way to improve performance and robustness of model. For ConvLSTM model, voting classifiers is used for ensemble.

## 2. Experiments

### 2.1. Plain CNN

For comparative analysis, simple plain CNN was used for criteria. See figure 1 and table 1 for architecture of CNN and performance. As one can see performance is better than pure guess, but not good enough.




| | **Plain** |
|---|---|
| **Validation accuracy** | 31.24 |

Figure 1: Architecture of simple plain CNN (left)
Table 1: Validation accuracy of simple plain CNN(right)

## 2.2. ResNET

Now we use ResNET instead of plain CNN. Structure of ResNET is as below.

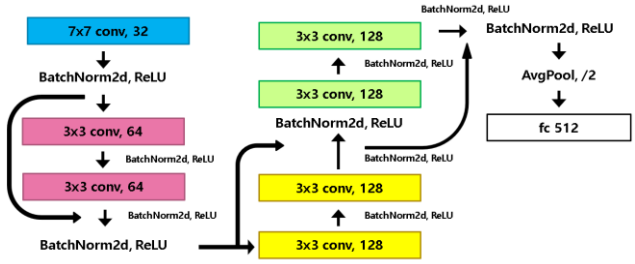



Figure 1: Architecture of simple plain CNN (left)
Comparison of plain CNN and ResNET model is as below. Validation accuracy becomes dramatically increased.

| | Plain CNN | ResNET |
|---|---|---|
| **Validation accuracy(%)** | 31.24 | 87.40 |

Table 2: Comparison of validation accuracy between Plain CNN & ResNET

## 2.3. Hyper parameter

For ConvLSTM there are hyper parameters such as filter size, RNN hidden size, number of layers of RNN, drop out rate of RNN. Table below show validation accuracy according to change of these hyper parameters.

| RNN hidden size | Validation accuracy(%) | RNN num layers | Validation accuracy |
|---|---|---|---|
| **8** | 45.67 | 1 | 84.88 |
| **50** | 60.34 | 2 | 87.40 |
| **64** | 72.77 | 3 | 86.23 |
| **100** | 87.40 | 4 | 86.66 |

Table 3: Validation accuracy according to hyper parameters

## 2.4. Teacher forcing

Teacher forcing is powerful method for RNN. However, when teacher forcing ratio is too high model becomes overfitted. Table below show comparison of validation accuracy between non teacher forcing model, highly teacher forced model, and model with teacher forcing ration 0.5. When teacher forcing ratio is 1.0, model's test accuracy becomes approximately 100%. However, validation accuracy is decreased compared to model with ratio of 0.5.

| Teacher forcing ratio | 0 | 0.4 | 0.5 | 0.6 | 1.0 |
|---|---|---|---|---|---|
| **Validation accuracy(%)** | 85.76 | 87.24 | 87.40 | 86.57 | 86.88 |

Table 3: Comparison of Validation accuracy according to teacher forcing ratio.

## 2.5. Ensemble

Ensemble is powerful method for machine learning. As table below shows ensemble method increase validation accuracy when voting classifier is used.

| | Validation accuracy(%) |
|---|---|
| **Non-Ensemble Model** | 87.40 |
| **Voting classifier ('Adam', lr = 1e-03, weight decay = 0.005)** | 89.746 |
| **Voting classifier ('Adam', lr = 2e-03, weight decay = 0.007)** | 88.23 |
| **Fusion classifier ('Adam', lr = 1e-03, weight decay = 0.005)** | 69.78 |

Table 4: Comparison of non-ensemble and ensemble model with various hyper parameter

# 3. Conclusion

By using ConvLSTM model, model achieved validation accuracy of 89.746%. Resnet was used for CNN, and teacher forcing method and LSTM was used for RNN, finally for overall model ensemble method was used. Every method has effective influence and change of hyper parameter was valid too. For better performance, more powerful ensemble method or powerful CNN network can be valid. However, due to google Colab runtime restriction, using either model was impossible.

References


[1] S. Xingjian, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Neural Information Processing Systems, 2015.

[2] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:1603.07285, 2016.Actual Author Name. The frobnicatable foo filter, 2014. Face and Gesture (to appear ID 324).

[3] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. AutoAugment: Learning augmentation strategies from data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of CVPR, pages 770–778, 2016. pp. 770-778