A novel CNN+LSTM classification model based on fashion-MNIST

Yaran Ji

Yaran Ji, "A novel CNN+LSTM classification model based on fashion-MNIST," Proc. SPIE 12258, International Conference on Neural Networks, Information, and Communication Engineering (NNICE 2022), 122580S (15 July 2022); doi: 10.1117/12.2639667

SPIE.

Event: International Conference on Neural Networks, Information, and Communication Engineering (NNICE 2022), 2022, Qingdao, China

A Novel CNN+LSTM Classification Model Based on Fashion-MNIST

Yaran Ji*
Xiamen University Malaysia, Jalan Sunsuria, Bandar Sunsuria,43900 Sepang, Selangor
*Corresponding Author's E-mail: MAT1909405@xmu.edu.my

ABSTRACT

Nowadays, Convolutional Neural Network (CNN) based image recognition is a popular research direction. This study uses the Fashion-Mnist dataset, which is more challenging than the Mnist dataset. aims to add Long short-term memory (LSTM) to the structure of CNN to create a hybrid model of CNN and LSTM, called CNN+LSTM model. This model is used to complete and optimize the image classification problem on Fashion-Mnist dataset. The final image classification accuracy of the obtained model is 91.36%, which still needs to be improved, but the accuracy results are better compared to the accuracy of other models.

Keywords: Image Classification; Convolutional Neural Network; Long Short-Term Memory; Fashion-MNIST Dataset.

1. INTRODUCTION

1.1. Convolutional Neural Network

In deep learning, a convolutional neural network (CNN) is a class of artificial neural networks, most commonly applied to analyze visual imagery (Yamashita, Nishio, Do, & Togashi, 2018). Nowadays, computer vision develops rapidly, such as image classification, target detection, neural style transfer, medical image analysis, and so on, which are widely used (Wang, Yang, Wang, Wang, & Li, 2019). One of the big differences between machines and human beings is that human beings can learn gradually, which depends on the connection and structure of human brain neurons. For machines to have the ability to learn like human beings in some aspects, we need to imitate the structure of human brains. It can be said that the main structure of CNN is similar to the structure of the nerve conduction process of human brain neurons (Sahinbas & Catak, 2021). But the original CNN study was inspired by a study of the tissue of the visual cortex in mammals. CNN was first proposed in the 1960s when Hubel and Wiesel first observed that visual cortical neurons were sensitive to moving edges during experiments on cat visual cortex cells (Hernandez-Garcia, 2020). in the 1980s, Fukushima and Miyake proposed a receptive domain based on the "receptive field" of neocognitron, which can be considered as the first implementation of CNN (Chen, Jiang, Li, Jia, & Ghamisi, 2016).

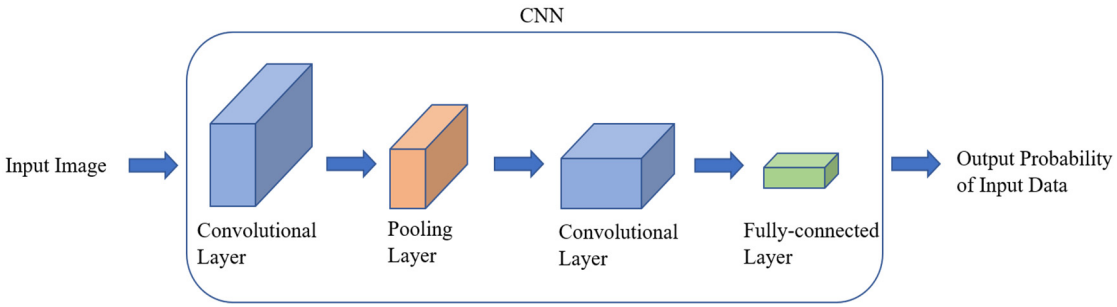


Figure 1. The Architecture of The Typical CNN Model in Image Classification

CNN is a breakthrough in image recognition and is often used in image classification. Compared with other algorithms used for image classification, CNN is more convenient and effective, because CNN can look at groups of pixels in an area of an image and learn the filters that have to be set manually in other algorithms (Sharma, Jain, & Mishra, 2018). In this way, it is not necessary to manually extract features from images during experiments. Instead, model learning can extract features by itself by training the model. Figure1 shows the schematic of a typical CNN model. Each layer of CNN can increase the complexity of model learning and improve the accuracy of image classification.

International Conference on Neural Networks, Information, and Communication Engineering (NNICE 2022),
edited by Rajeev Tiwari, Proc. of SPIE Vol. 12258, 122580S · © 2022 SPIE
0277-786X · doi: 10.1117/12.2639667

1.2. Long Short-Term Memory.

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture that was used for model temporal sequences in the field of deep learning (Sak, Senior, & Beaufays, 2014). One of the advantages of recurrent neural networks is that they attempt to connect prior information to the present task, but when the information is too far apart from the task, the RNN may no longer be able to connect long-term information to the task. However, LSTM is able to learn long-term dependency and can avoid long-term dependency problems well (Salehinejad, Sankar, Barfett, Colak, & Valaee, 2017), and is good at dealing with long-term correlation. Therefore, LSTM overcomes the limitation of the short-term memory of RNN very well. In addition, LSTM has feedback connections, and many backpropagations lead to exponential explosion or exponential decay (Kag, Zhang, & Saligrama, 2019), while LSTM can capture long-term temporal dependencies, and it was originally introduced to solve the gradient disappearance and gradient explosion problems (J. Zhang, Zeng, & Starly, 2021). LSTM has a wide range of applications, such as machine translation, speech recognition, etc., thanks to its ability to learn sequential dependencies in sequence prediction problems.

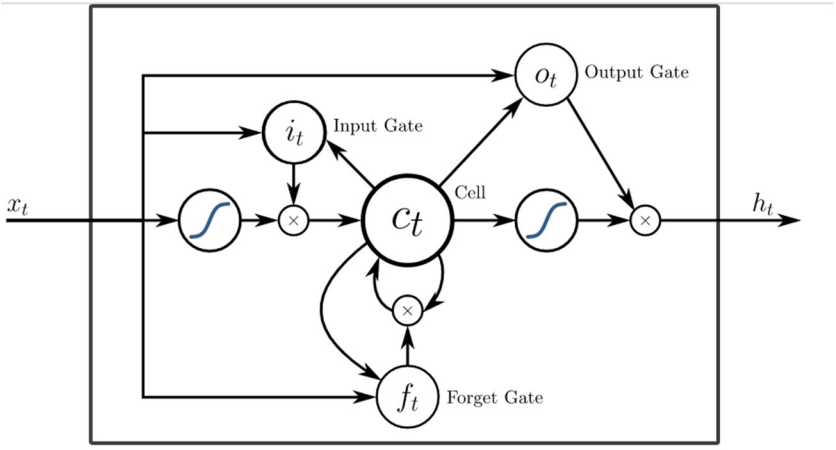


Figure 2. The Architecture of The Typical LSTM Cells

As shown in Figure 2 (Jenkins, Gee, Knauss, Yin, & Schroeder, 2018), the basic structure inside the LSTM cell is a forget gate, an input gate, an output gate and a cell state. The forget gate determines how much data is to be "forgotten" and how much data is to be used next. By combining LSTM with CNN, errors are propagated back to the CNN model through LSTM across multiple input images, so as to better train the CNN model. Therefore, the model with CNN+LSTM structure is suitable for sequence prediction problems with spatial inputs.

1.3. Literature Review.

It can be said that the field of computer vision made a breakthrough in 2012 when AlexNet won ImageNet and has been growing rapidly ever since (LeCun, Bottou, Bengio, & Haffner, 1998). Image classification is a very important part of the computer vision field, which can recognize images and classify them in one of the pre-defined different categories.

The present study is based on a convolutional neural network for image classification. Convolutional neural networks are now an important pillar of many computer vision systems. The first work on modern convolutional neural networks appeared in the paper "Gradient-Based Learning Applied to Document Recognition" by Yann LeCun et They worked out a CNN model that could aggregate simple features into more complex ones and showed that the CNN model could be used for handwritten character recognition (Jiang & Song, 2020). At that time, their experiments were based on the MNIST dataset, and now, the accuracy of the advanced CNN model for image classification on the MNIST dataset is almost perfect.

2. DATA AND METHOD

This section is divided into two sections. The first section introduces the data set used in the study, and the second section explains the main structure and flow of the model with the highest accuracy.

2.1. Dataset

Some The dataset I chose is the Fashion-MNIST dataset. Fashion-MNIST is a dataset of Zalando's article images, it consists of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated

with a label from 10 classes (K. Zhang). Table1 represents the names of the 10 classes and their corresponding labels and image examples.

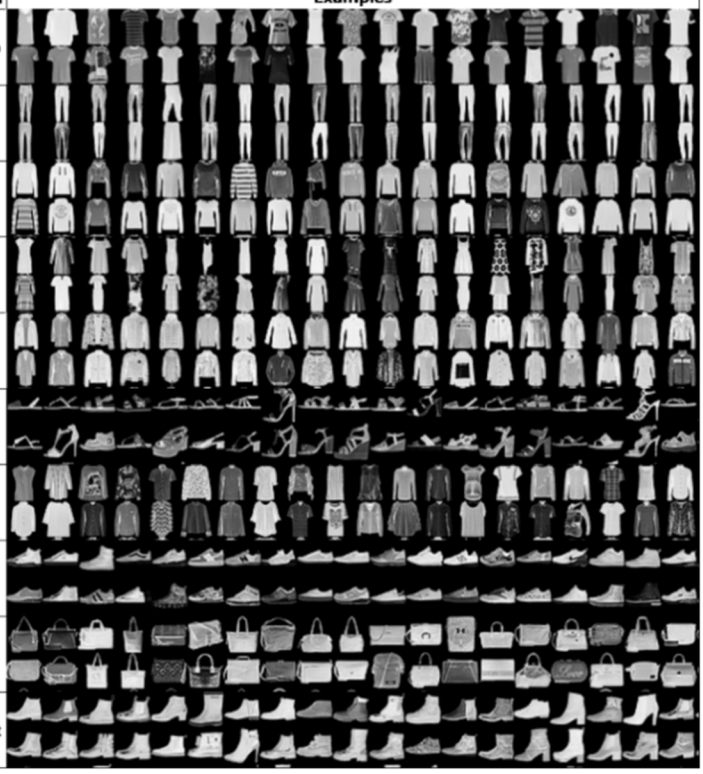
Label	Description	Examples
0	T-shirt/top	
1	Trouser	
2	Pullover	
3	Dress	
4	Coat	
5	Sandal	
6	Shirt	
7	Sneaker	
8	Bag	
9	Ankle boot	

Figure 3. Fashion-MNIST Dataset

The reasons that I chose the Fashion-MNIST dataset as the data source are: first, the original MNIST dataset is simple and the models generated on that basis may not be usable in other datasets, while the Fashion-MNIST dataset is more challenging and complex than the MNIST dataset, and therefore has more optimization possibilities. Secondly, the original MNIST dataset is overused, while the Fashion-MNIST dataset is used for a shorter period time, which is not the case.

2.2. Model Construction

Optimization requires trial and error. Figure3 is the architecture of the model with the highest accuracy obtained after several attempts. Here is the definition of accuracy. Accuracy: The ratio of true positive and true negative in all rated cases. It is defined as $\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$ where TP is the true positive cases, TN is the true negative cases and FP is false-positive cases.

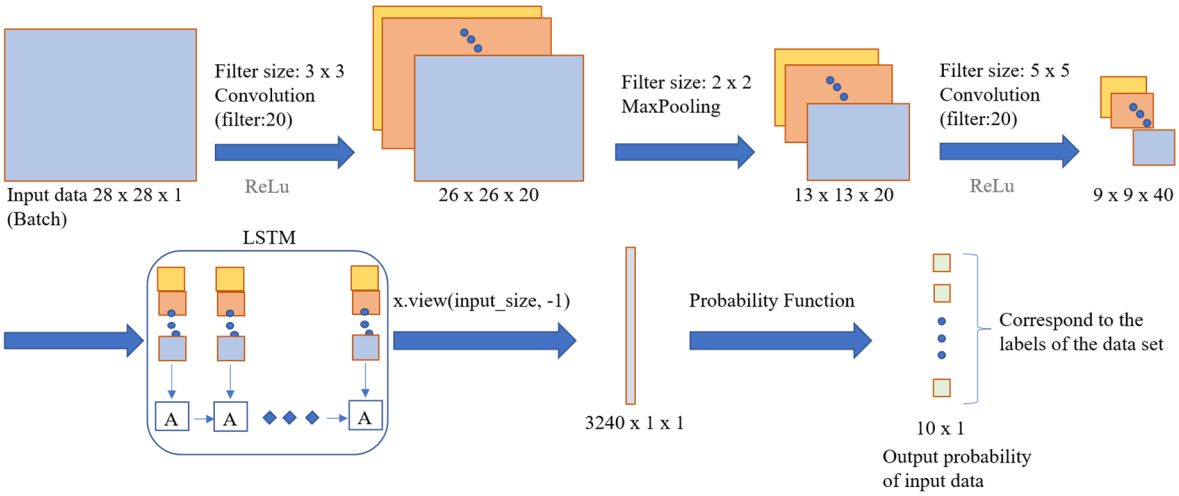


Figure 4. The Architecture of The Model 1 --with The Highest Accuracy.

It can be seen from Figure3 that the input data in each round are 28×28×1 images. Through the first layer, the convolution layer is equipped with 20 3×3 filters, and the convolution layer performs the convolution operation on the data. Each convolutional neuron processes data only for its receptive field. Convolutional layers convolve the input and pass its result to the next layer, namely the pooling layer. The pooling layer reduces the dimension of data and merges the outputs of clusters of neurons into a single neuron in the next layer. The pooling layer can help reduce the number of parameters required by the model and also effectively avoid over-fitting. Then pass the convolution layer with 20 5×5 filters. LSTM was then added, and the error was calculated and then propagated back. The weights and feature detectors were adjusted each time during the repeat process to help optimize the performance of the model. It can improve the accuracy of classification. After that, the image becomes a long vector that passes through the fully connected layer. In the fully connected layer, each neuron in the previous layer is connected to each neuron in the next layer. The fully connected layer can make use of the output value of the previous layer according to the training data, and finally obtain the required probability of image classification.

3. RESULTS

Table2 is the structure of these five models. Model1, Model2, and Model3 use the structure of CNN+LSTM, and Model4 and Model5 use the structure of CNN. The experiment hopes to extract features from data from the Fashion-MNIST dataset through layers of Convolutional Neural Network and improve spatial learning ability. Sequence prediction can be supported by adding LSTM to improve the time learning ability, to improve the accuracy of image classification.

Table 1. Structure of The 5 Models

Model1	Structure	Model2	Structure	Model3	Structure	Model4	Structure	Model5	Structure
Conv1	(1,20,3)	Conv1	(1,20,3)	Conv1	(1,20,3)	Conv1	(1,10,5)	Conv1	(1,5,5)
Pool1	(4,4,0)	Pool1	(4,4,0)	Pool1	(4,4,0)	Pool1	(4,4,0)	Pool1	(4,4,0)
Conv2	(20,40,5)	Conv2	(20,30,3)	Conv2	(20,40,5)	Conv2	(10,20,5)	Conv2	(5,10,3)
Fc1	(3240,500)	Fc1	(3630,300)	Pool2	(4,4,0)	Pool2	(4,4,0)	Fc1	(1000,100)
Fc2	(500,10)	Fc2	(300,10)	Fc1	(640,500)	Fc1	(320,100)	Fc2	(100,10)
LSTM	(20,10,10)	LSTM	(10,10,1)	Fc2	(500,10)	Fc2	(100,10)	Softmax	(10,1)
Softmax	(10,1)	Softmax	(10,1)	LSTM	(20,20,1)	Softmax	(10,1)		
				Softmax	(10,1)				

The accuracy of the five models was compared. Table3 is the code name of the five models mentioned above and the final accuracy value of image recognition. It can be found that the model using CNN+LSTM structure can obtain better accuracy of image classification in the Fashion-MNIST dataset.

Table 2. A Comparison of Accuracy between 5 Models on Fashion-MNIST Dataset

Model	Accuracy
Model1	91.36%
Model2	90.98%
Model3	89.97%
Model4	88.72%
Model5	88.43%

It can be seen that the accuracy of the first three models is better than the other two models, and the accuracy of Model1 is the highest. The roles of the layers that all five models have are as follows: the convolution layer is equivalent to a filter that extracts features from the image. The pooling layer shrinks the large image while preserving the important features. The fully connected layer converts the filtered images into categories. However, the first three models add the LSTM, which, as we mentioned before, avoids the drawbacks of traditional RNNs and is better adapted to the models used for image classification, and then the models increase the accuracy.

Neural networks are trained by the stochastic gradient descent method, usually, we need to minimize the error, so we need to design a likelihood loss function to calculate the error of the model. loss function can summarize all aspects of the model into a single number to evaluate the performance of the algorithm that solves the task. Therefore, the choice of the loss function is also very important, only by choosing the appropriate loss function can we estimate the accuracy of the model more accurately. For any given problem, a lower log loss value means better predictions. in this study, we have chosen Maximum likelihood seeks to find the optimum values for the parameters by maximizing a likelihood function derived from the training data.

As seen in Figure4, the likelihood loss during training appears to be a convex function of converges to 0. However, the fluctuation of the lines in the figure shows that the performance of the model is not very stable.

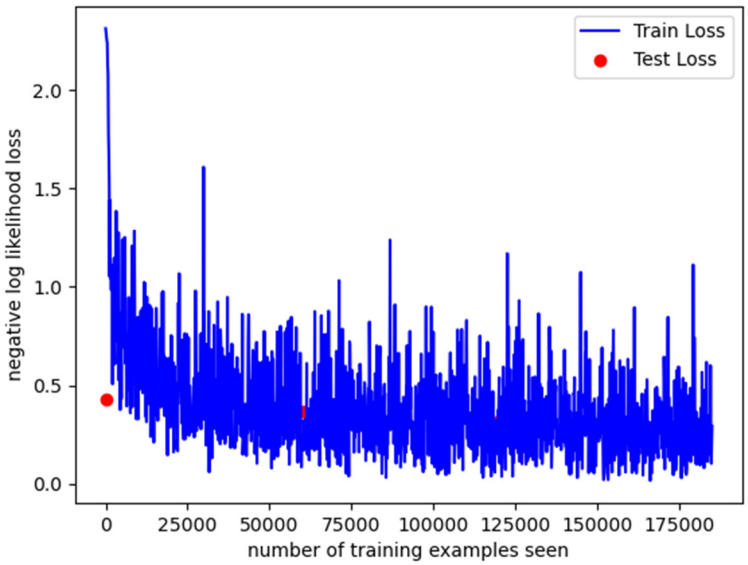


Figure 5. Likelihood loss of training process with epoch k

Figure5 shows that the likelihood loss during testing appears to be a convex function of converges to 0. Compared with Figure4, the fluctuations in Figure5 look much smaller and the performance is more stable.

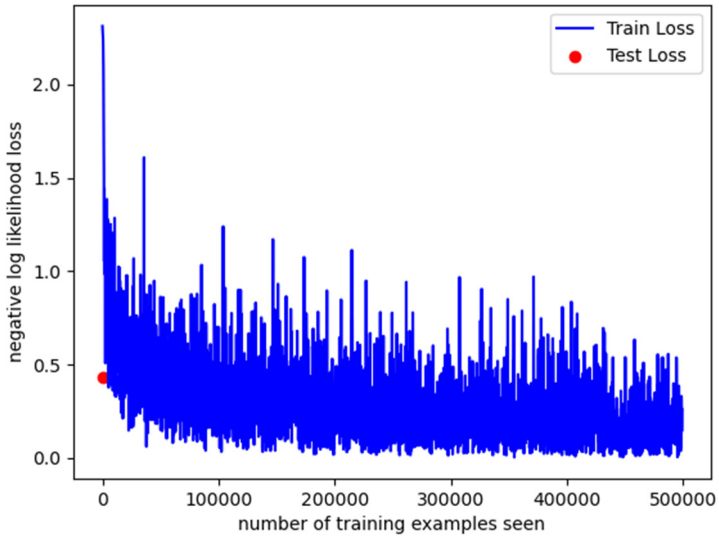


Figure 6. Likelihood loss of testing process with epoch k

4. CONCLUSION

Nowadays, computer vision is developing rapidly, where the image classification problem is a very fundamental and important part, which is applied in several aspects of daily life, such as automotive industry, medical industry, retail industry, security industry, etc. In this paper, the features of convolutional neural network and long and short-term memory are first outlined separately, and then the generative learning capability of CNN in the feature extraction phase is utilized to improve the representational capability of the model; the LSTM network is utilized to avoid the failure of the recurrent neural networks through a pattern that is similar to that of Back Propagation Through Time. Eventually, both of them are combined to build models for image classification on the Fashion-MNIST dataset. As technology continues to evolve, image recognition models based on CNN structures have been overused, and the accuracy of image recognition can be improved by adding LSTM appropriately. In this study, a model of CNN+LSTM for image classification of Fashion-MNIST dataset was developed, several different models were tested, the model structure was improved, and the highest accuracy rate finally obtained was 91.36%. This accuracy result still has room for improvement, but the accuracy of the model with CNN+LSTM structure is superior compared to the CNN model. Many studies have shown that the combination of multiple model approaches can make the models complement each other and optimize the overall model, so more combinations of different models can be tried in the future to improve the accuracy of image classification.

REFERENCES

[1] Chen, Y., Jiang, H., Li, C., Jia, X., & Ghamisi, P. (2016). Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing, 54(10), 6232-6251.
[2] Hernandez-Garcia, A. (2020). Data augmentation and image understanding. arXiv preprint arXiv:2012.14185.
[3] Jenkins, I. R., Gee, L. O., Knauss, A., Yin, H., & Schroeder, J. (2018). Accident scenario generation with recurrent neural networks. Paper presented at the 2018 21st International Conference on Intelligent Transportation Systems (ITSC).
[4] Jiang, Y., & Song, Y. (2020). High-Accuracy Offline Handwritten Chinese Characters Recognition Using Convolutional Neural Network. Journal of Computers, 31(6), 12-23.
[5] Kag, A., Zhang, Z., & Saligrama, V. (2019). Rnns evolving on an equilibrium manifold: A panacea for vanishing and exploding gradients? arXiv preprint arXiv:1908.08574.

[6] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
[7] Sahinbas, K., & Catak, F. O. (2021). Transfer learning-based convolutional neural network for COVID-19 detection with X-ray images. In Data Science for COVID-19 (pp. 451-466): Elsevier.
[8] Sak, H., Senior, A. W., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling.
[9] Salehinejad, H., Sankar, S., Barfett, J., Colak, E., & Valaee, S. (2017). Recent advances in recurrent neural networks. arXiv preprint arXiv:1801.01078.
[10] Sharma, N., Jain, V., & Mishra, A. (2018). An analysis of convolutional neural networks for image classification. Procedia computer science, 132, 377-384.
[11] Wang, W., Yang, Y., Wang, X., Wang, W., & Li, J. (2019). Development of convolutional neural network and its application in image classification: a survey. Optical Engineering, 58(4), 040901.
[12] Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. Insights into imaging, 9(4), 611-629.
[13] Zhang, J., Zeng, Y., & Starly, B. (2021). Recurrent neural networks with long term temporal dependencies in machine tool wear diagnosis and prognosis. SN Applied Sciences, 3(4), 1-13.
[14] Zhang, K. LSTM: An Image Classification Model Based on Fashion-MNIST Dataset.