E-commerce Item Identification Based on Improved SqueezeNet

Kai Fan1*, Lianqiang Niu2 and Shengnan Zhang2,

1School of Information Science and Engineering, Shenyang University of Technology, Shenyang, 110870, China
2School of Software, Shenyang University of Technology, Shenyang, 110870, China

*Corresponding author email: fankai@sefonsoft.com

Abstract. In order to improve the recognition rate of product images in e-commerce recommendation scenes, we proposed a high-performance improved SqueezeNet convolutional neural network, which uses a Fire Module structure containing two large convolution kernels to reduce the complexity of the model while fully extracting features, and adding a pooling layer behind each Fire Module to effectively filter key features of the classification. Experiments show that the algorithm in this paper has a better application in image recognition in e-commerce scenes, and the recognition accuracy is improved by 1%.

1. Introduction

Most of the commodities in the e-commerce system have both text and image features, and they are mainly based on image features. Therefore, the efficient use of image features has become an important method for implementing recommendations in e-commerce systems. At present, the extraction of image features is mainly based on neural network methods. How to choose a suitable network structure and take into account both complexity and accuracy is a common problem in neural network design. In recent years, some simplified convolutional neural network models have appeared to try to resolve the contradiction between network complexity and accuracy, such as ShuffleNet, SkipNet and SqueezeNet. These models proposed a new network structure to minimize the model volume while minimizing the impact on recognition accuracy. Among them, as an improved model of the Alexnet network (2012 ImageNet competition champion), SqueezeNet not only maintains the same accuracy as AlexNet, but also reduces the size by 50 times by proposing the Fire Module structure. However, in the previous experiments of the e-commerce recommendation, we found that the SqueezeNet network has a low recognition rate for image parts. The reason is that due to the large differences of products in the same classification, the Fire Module structure in SqueezeNet and the setting of the pooling layer cannot effectively extract unified features that can represent the classification, resulting in poor recommendation results. This article optimizes the hyper parameters such as the number of Fire Modules and the size of the convolution kernel, and adds a pooling layer after each Fire Module, so that the network can increase the recognition rate and faster convergence speed is applied to the identification of e-commerce items.
2. SqueezeNet Network Model Overview

2.1. The Basic Structure of SqueezeNet

SqueezeNet [1] is a streamlined neural network model proposed by UC Berkeley and others in 2016. As shown in figure 1, this network is an improvement on AlexNet proposed in 2012, which solves the problem of slow training speed caused by too many model parameters in AlexNet. This network significantly reduces the model complexity and improves the convergence speed of the network. UC Berkeley and others have proved through a large number of experiments that SqueezeNet is similar to AlexNet in classification accuracy, but the amount of parameters is much less than AlexNet. Among them, SqueezeNet was 1.24 MB, AlexNet was 61.10 MB, and the number of parameters varied by 50 times. The main reason for this is that SqueezeNet used the following three strategies to adjust the network structure:

- The convolution kernel of 3×3 is adjusted to 1×1;
- Reduce the number of feature maps input into 3×3 convolution kernel;
- Delay the sampling operation required in the network backward.

Among the three strategies, the first two aims are to reduce the number of parameters in the model, so that the experiment can improve the training speed under the same hardware conditions. The third strategic goal is to improve model classification accuracy.

![Figure 1. Structure of SqueezeNet.](image1.png)

The innovation of SqueezeNet is proposed Fire Module structure, as shown in figure 2. The Fire Module contains Squeeze and Expand structures. Where the Squeeze is composed of S 1×1 convolution kernels and the Expand contains E1 1×1 and E2 3×3 convolution kernels. The advantage of this is that, assuming that the input image of the Fire Module is H×W×M and the output is H×W×(E1+E2), the number of channels output is changed to make E1+E2 less than M to achieve the purpose of compression model, where E1=E2=4S. SqueezeNet reduced the convolutional kernel’s size of the network, reducing the amount of parameters without decreasing the accuracy, thus improving the model convergence speed under the same conditions.

![Figure 2. Structure of fire module.](image2.png)

2.2. SqueezeNet Existing Improved Algorithm

Although use the above three improved strategies, the recognition results differ greatly in different application scenarios. Therefore, scholars complete the recognition task in specific scenarios by adjusting the network structure of SqueezeNet or combining with other networks.

Yiwei Dong [2] et al. used packet bottlenecks and singular bottlenecks to reduce the parameter amount of SqueezeNet, making the network more suitable for mobile terminals. Junying Gan [3] and others proposed a lightweight Fast-SqueezeNet model to improve the accuracy of fingerprint classification. Jun Wu [4] et al. optimized the convolution layer setting method after SqueezeNet pooling, and
enhanced the accuracy of face recognition. Baohua Qiang [5] and others combined the human bone keypoint detection model SqueezeNet to reduce training time and significantly improve accuracy. Qiang Guo [6] et al. proposed FastSqueezeNet, which improved the model's convergence speed while improving accuracy. Yinhui Yan [7] combined SqueezeNet to optimize the original SSD and applied it to the management of ETC entrances and on-street parking spaces. Qianqian Ma [8] and others proposed the SqueezeNet-RM model, which showed faster detection speed and higher generalization ability in robot grab detection technology. Aili Wang [9] and others combined SqueezeNet and OctConv to improve the accuracy and efficiency of lidar data classification. Jixiao Wang [10] and others used SqueezeNet to extract infrared and visible image features for weighted fusion, and finally obtained the fused image.

3. Improved SqueezeNet Model
The improved model in this paper starts with the size of the convolution kernel in the Fire Module structure and the setting of the pooling layer in the network. The reasonable Fire Module structure can fully extract the image features. The multi-layer pooling method helps to further filter the key features, so as to effectively match the text data features and improve the accuracy of the recommendation results. The Fire Module of the SqueezeNet network is composed of 1 × 1 and 3 × 3 convolution kernels. The convolution size is small and the feature extraction granularity is fine. It is suitable for the same type of data with little difference. However, the data of e-commerce are usually of the same type with larger features, such as "pants" in the data set. As shown in figure 3 below, although all belong to the same classification, due to the large feature gap, the fine-grained convolutional kernel in the Fire Module of SqueezeNet network is not conducive to obtaining the composite features representing categories. Therefore, in this case, a coarse-grained approach is more conducive to better classification results.

![Figure 3. Example of data classified as "pants".](image)

The setting of the convolution kernel size in the Fire module of this experiment needs to ensure the following two points:

- The number of output feature maps of the two parts of the Expand structure should be consistent. The purpose is to ensure that the number of feature maps formed after the Expand convolves the image to extract the features should be the same.
- Because the convolution kernels of the two parts of the Expand structure are different in size, it is necessary to add 0 to the output feature map so that the output of the feature maps of the two parts can be completely stitched, which is convenient for the pooling operation.

In addition to the setting of the convolution kernel granularity, after the Fire Module extracts the features of the image, how to set the pooling structure to properly filter the features is also a key factor affecting the accuracy of the recognition. Therefore, on the basis of full experiments, this paper makes two improvements to SqueezeNet for this e-commerce scene dataset:

- Modify the original Fire Module containing 1 × 1 and 3 × 3 to 3 × 3 and 5 × 5. By increasing the size of the convolution kernel, prevent over-fitting caused by the too fine granularity of the original network;
- Adjust the number of Fire Modules in the original SqueezeNet network, and add a pooling layer after each Fire Module. The purpose of this is to reduce the amount of parameters while extracting key features. The structure of a single Fire Module is shown in figure 4.
4. Experiment and Analysis

This experiment uses the deep learning framework Caffe to train the network. The Caffe framework was developed by Dr. Jia Yangqing of the University of California, Berkeley, and is a widely used deep learning framework.

The experimental device is a personal computer with Intel i5 and NVIDIA GeForce 840M and 8G memory. The validity of the model proposed in this paper is verified on this device.

The experimental data uses a self-built e-commerce data set, which contains 100 categories of 100 product images in each classification. We separated training data from test data at a ratio of 3:1.

Some of the hyperparameters in SqueezeNet are set as follows: the learning rate is 0.001, the learning strategy is "poly", the maximum number of iterations is 10,000, the momentum is set to 0.9, and the weight parameter is 0.0002.

4.1. Experiment of Fire Module Convolution Kernel Size

Optimization was performed on the Fire Module that originally contained two types of convolution kernels of 1×1 and 3×3. Under sufficient experimental conditions, it was determined that 3×3 and 5×5 performed better on this experimental data set. The experimental results are shown in Table 1.

| No. | Expand1 | Expand2 | top5-accuracy |
|-----|---------|---------|---------------|
| 1   | 1×1     | 3×3     | 97.0%         |
| 2   | 1×1     | 5×5     | 97.0%         |
| 3   | 3×3     | 5×5     | 98.0%         |
| 4   | 5×5     | 7×7     | 87.0%         |

Experimental results show that when the convolution kernel sizes are 3×3 and 5×5, although there may still be some problems that reduce the recognition accuracy due to too few pooling layers and insufficient feature screening granularity, which lead to overfitting. So the following experiments on the pooling layer.

4.2. Experiment of Pooling Layer

In the structure of the original Fire Module, the pooling layer was added only after Fire Module3 and Fire Module5. This experiment wants to verify the impact of adding a pooling layer after each Fire Module on the recognition rate of this dataset. Results of the experiment are shown in Table 2.

|               | top5-accuracy |
|---------------|---------------|
| only Fire Module3 and Fire Module5 have pooling layer (SqueezeNet) | 97% |
| each Fire module are followed by a pooling layer | 98% |
The experimental results show that, compared with only Fire Module3 and Fire Module5, when the pooling layer is added after each Fire Module, the key features are extracted effectively, and the image features are fully utilized to improve the accuracy.

4.3. Experiment on the Number of Network Layers
The purpose of this experiment was to determine the required number of layers of the network. The structure of Fire Module2 to Fire Module8 in SqueezeNet network was similar. Under the premise of retaining the structure of Fire Module5 to Fire Module8, this experiment added and deleted the number of layers of Fire Module2 to Fire Module4 to determine the final number of network layers. Results of the experiment are shown in Table 3.

| No. | Fire Module2 | Fire Module3 | Fire Module4 | top5-accuracy |
|-----|--------------|--------------|--------------|---------------|
|     | Squeeze2     | pool2        | Squeeze3     | pool3         | Squeeze4      | pool4         |               |
| 1   | 3×3          | 3×3          | 1×1          | 3×3           | 1×1           | 3×3           | 96.0%         |
| 2   |              |              |              | 3×3           | 97.0%         |
| 3   |              |              | 3×3          | 3×3           | 95.0%         |

The results showed that when the Fire Module5 to Fire Module8 network structure was kept basically unchanged, the recognition rate was the highest when the Fire Module2 structure was removed. Therefore, it can be concluded that when a Fire Module structure is removed, the feature extraction effect of this experimental data set is the best, and the final recognition rate is improved by 1%.

4.4. Result Analysis

4.4.1. Effectiveness analysis
This experiment has been iterated for 10,000 times, and the results are shown in Figure 5 to Figure 8. Figure 5 shows the variation curve of the error value with iterations. Figure 6 shows the change curve of iteration times with the accuracy rate. Figure 7 and Figure 8 show the change curve of learning rate with the number of iterations and the iteration curve of learning rate with the training time, respectively. It can be concluded that the experimental parameters are set reasonably and the results are reliable.

4.4.2. Error analysis
Using this network to identify the test set, we found that the following data cannot be properly classified. Partial results are shown in Figure 9 and Figure 10.
In the above results, figure 9 is the classification of "Sofa", and the left side is the wrongly recognized data. It can be seen that the misidentified data differs greatly from most of the data in this classification in the shape of the outline, there is a large difference in the image, which leads to recognition errors, figure 10 is the classification of "fanned fans", the left side is the incorrectly identified data. It can be seen that it is also “fanned fans”, but if the contour shapes differ greatly, it will still lead to the error.

5. Conclusion
Taking advantage of characteristics of convolutional neural network in weight sharing and less learning parameters, this paper optimizes based on SqueezeNet network by increasing the size of the convolution kernel in the Fire Module and adding a pooling layer after each Fire Module. It effectively extracts key features in the image, so that the recognition rate of the image is improved to a certain extent compared with the original network. It should be pointed out that, in the e-commerce scene, when the contour shapes of the images are too different, they still cannot be effectively identified. If you continue to subdivide the categories, it will greatly expand the number of classifications, resulting in slower convergence speed, if the extracted features are filtered multiple times, the amount of features will be too small and affect the recognition accuracy. Finding a balance between this factors to identify or add other characteristic information to improve the rate of recognition is a future development direction.

References
[1] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: Alexnet-level accuracy with 50x. fewer parameters and< 0.5 mb model size, J. arXiv preprint arXiv:1602.07360, 2016.
[2] Yiwei Dong, Jin Yu. SqueezeNet-based lightweight convolutional neural network SlimNet, J. Computer applications and software, 35(2018)226-232 (in Chinese).
[3] Junying Gan, Ling Qi, Chuanbo Qin, et al, Lightweight fingerprint classification model combined with transfer learning, J. Journal of Image and Graphics. 24(2019) 1086-1095(in Chinese).
[4] Wu Jun, Qiu Yang, Lu Zhongliang. Face recognition based on improved squeezeNet, J. Science Technology and Engineering, 19(2019)218-223(in Chinese).
[5] Baohua Qiang, Yijie Zhai, Jinlong Chen, et al, Lightweight human bone keypoint detection model based on improved CPMs and SqueezeNet, J. Computer application. 1-7(2020)(in Chinese).
[6] Qiang Guo, Quanwei Liu, Wei Wang. Fast SqueezeNet Algorithm and Its Application in Estimating Metro Population Density, J. Control Theory and Application, 1-11(2019)(in Chinese).
[7] Yinhui Yan, Zhengjin Zhou, Tao Xiang. Improved SSD model based on SqueezeNet and its application, J. China Transportation Information, S1(2018) 189-191(in Chinese).
[8] Qianqian Ma, Xiaojuan Li, Zhiping Shi. Research on Robot Grab Detection of Lightweight Convolutional Neural Network, J. Computer Engineering and Applications, 1-11(2020)(in Chinese).
[9] Wang Aili,Wang Minhui,Jiang Kaiyuan, Cao Mengqing,Iwahori Yuji. A Dual Neural Architecture Combined SqueezeNet with OctConv for LiDAR Data Classification, J. Sensors (Basel, Switzerland), 19(2019)22.
[10] Jixiao Wang, Yang Li, Jiabao Wang, et al. Lightweight image fusion method based on SqueezeNet, J. Computer application, 1-7(2020)(in Chinese).