DOI: 10.7641/CTA.2017.60614

“智能科学与大数据工程”专题报告

编者按: 由华南理工大学主办, 中山大学协办的“2016年智能科学与大数据工程国际会议”(The 2016 International Conference on Intelligence Science and Big Data Engineering (IScIDE 2016))于2016年5月13–15日在广州举行. 会议收到大量高水平学术论文, 集中反映了中国学者在智能科学和大数据工程方面的研究现状和前沿成果. 为了更好地推进这些成果的应用, 组委会选取了其中部分优秀论文集中刊登在《控制理论与应用》2017年第34卷第6期, 以期达到加强宣传和技术推广的效果.

《控制理论与应用》编辑部

图像分类卷积神经网络的特征选择模型压缩方法

邹月娴†, 余嘉胜, 陈泽晗, 陈 锦, 王 毅

(北京大学 信息工程学院 现代信号与数据处理实验室, 广东 深圳 518055)

摘要: 深度卷积神经网络(convolutional neural networks, CNN)作为特征提取器(feature extractor, CNN–FE)已被广泛应用于许多领域并获得显著成功. 根据研究评测可知CNN–FE具有大量参数, 这大大限制了CNN–FE在如智能手机这样的内存有限的设备上的应用. 本文以AlexNet卷积神经网络特征提取器为研究对象, 面向图像分类问题, 在保持图像分类性能几乎不变的情况下减少CNN–FE模型参数量. 通过对AlexNet各层参数分布的详细分析, 作者发现其全连接层包含了大约99%的模型参数, 在图像分类类别较少的情况, AlexNet提取的特征存在冗余. 因此, 将CNN–FE模型压缩问题转化为深度特征选择问题, 联合考虑分类准确率和压缩率, 本文提出了一种新的基于互信息量的特征选择方法, 实现CNN–FE模型压缩. 在公开场景分类数据库以及自建的无线胶囊内窥镜(wireless capsule endoscope, WCE)气泡图片数据库上进行图像分类实验. 结果表明本文提出的CNN–FE模型压缩方法减少了约83%的AlexNet模型参数且其分类准确率几乎保持不变.

关键词: 卷积神经网络; 图像分类; 特征提取; 特征选择; 模型压缩

中图分类号: TN912.35 **文献标识码**: A

Convolutional neural networks model compression based on feature selection for image classification

ZOU Yue-xian†, YU Jia-sheng, CHEN Ze-han, CHEN Jin, WANG Yi

(ADSPLAB/ELIP, School of Electronic and Computer Engineering, Peking University, Shenzhen Guangdong 518055, China)

Abstract: Deep convolutional neural networks (CNN) feature extractor (CNN–FE) has been widely applied in many applications and achieved great success. However, evaluating shows that the CNN–FE holds abundant parameters which largely limits its applications on memory-limited platforms, such as smartphones. This study makes an effort to trim the well-known CNN–FEs, AlexNet, to reduce its parameters meanwhile the image classification performance almost remains unchanged. This task is considered as a CNN–FE model compression problem. Through carefully analyzing the parameter distribution of AlexNet, we find about 99% of parameters are in its fully connected layer but the deep features are redundant for image classification tasks with small number of categories. Moreover, we propose to convert the CNN–FE model compression problem into a feature selection problem. Specifically, a feature selection method, which is based on mutual information and a novel criterion related to the classification accuracy and the compression ratio, has been proposed. Image classification experiments on a public scene categories database and our self-built wireless capsule endoscope (WCE) bubble dataset show that our proposed CNN–FE model compression method reduces more than 83% size of the AlexNet while almost maintaining the classification accuracy.

Key words: convolutional neural networks; image classification; feature extractor; feature selection; model compression

Received 15 August 2016; accepted 20 June 2017.

†Corresponding author. E-mail: zouyx@pkusz.edu.cn; Tel: +86 755-26032016.

Recommended by Associate Editor: YU Zhu-liang.

Supported by Shenzhen Science & Technology Fundamental Research Program (JCYJ20150430162332418).

1 Introduction

Undoubtedly, the outstanding performance of convolutional neural networks (CNN) in ImageNet2012 has brought revolutions to the computer vision[1], which drew broad attention in both academic and industrial areas. A huge amount of research showed that deep CNN has been continuously advancing the image classification accuracy[1–3], meanwhile it can be also treated as a generic feature extractor for various tasks such as object detection[4–5], semantic segmentation[4, 6], image retrieval[7] and etc.

It is quite clear that deep CNN, such as AlexNet[1] and, is extremely effective in various computer vision task as a feature extractor (CNN–FE), but they require a big dataset for training and tuning their huge number of parameters[8–10]. Due to the large quantity of model parameters, it is difficult to apply the well-trained CNN--FE on a memory-limited platform, such as smartphones and portable devices. Therefore, there are some efforts have been made to trim the CNN–FE model. Gong et al. tackled the CNN–FE model storage problem by investigating information theoretical vector quantization methods for compressing the parameters of CNNs[11]. Denton speeded up the bottleneck convolution operations in the first layers of a CNN by a factor $2-3\times$ using low-rank projection approach, while compressing the fully connected layers by using SVD[12]. Han et al. proposed a three-step method to learn only the important connections in fully connected layer to compress the CNN[13]. However, it is noted that the methods discussed above all modified the original CNN architecture, need to retrain the CNN by using the original training set and evaluating the performance by the original testing set. Therefore, we can see that these methods treat the CNN as an end-to-end trainable classifier instead of a generic feature extractor. Essentially, these methods can not be considered as CNN-model compression methods and they are not suitable for transferring the well-trained CNN as a generic feature extractor to dataset with small number of categories and memory-limited applications.

In this paper, we strive to propose a novel CNN–model compression method to transfer the well-trained AlexNet feature extractor (AlexNet–FE) to the image classification task with small number of categories. Firstly, the distribution of parameters in AlexNet–FE is carefully analyzed, and it is found that 99% of the parameters are in the fully connected layer. Secondly, when the response of every unit in the fully connected layer is treated as an independent feature value, the redundancy of the features can be calculated through mutual information. Thirdly, the CNN–model compression problem is converted to a feature selection problem. Therefore, aiming at compressing the CNN–model while maintaining its classification accuracy, we propose a feature selection based on mutual information and a novel selection criterion, which is directly related to the classification accuracy and compression ratio. Experiment results of scene categories database[14] and WCE bubble dataset show that our proposed method reduces more than 83% parameters of AlexNet–FE while almost retaining the classification accuracy.

The rest part of this paper is organized as follows: In Section 2, the distribution of parameters of CNN–FE and the redundancy of deep features are presented. The proposed CNN–model compression method based on feature selection and a novel selection criterion is introduced in Section 3. The experimental results are given in Section 4 and conclusions are drawn in Section 5.

2 Architecture analysis of AlexNet–FE

Just to prove our research motivation and consider the page limitation, in this subsection, we take AlexNet as example. According to [7, 10], the responses from the higher-level layers of AlexNet have been proven to be effective generic features with benchmark image classification performance on various image datasets. To make the presentation clear and differ the original AlexNet from the compressed AlexNet, the original AlexNet and the compressed AlexNet are respectively termed as AlexNet and C–AlexNet in the following paper.

Generally, the first six layers of the AlexNet are taken as a generic feature extractor (termed as AlexNet–FE), the architecture of which is shown in Fig.1. The parameters about each layer of AlexNet–FE are described in Table 1. The details about AlexNet are described in [1].

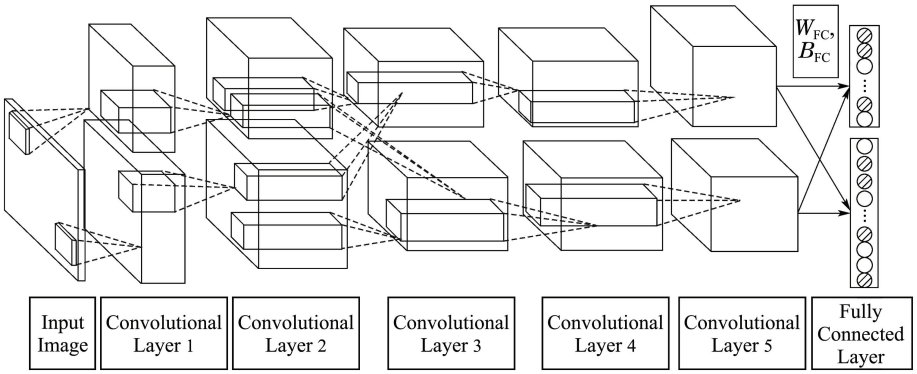


Fig. 1 The architecture of the AlexNet feature extractor

Table 1 The parameters of AlexNet–FE

Layer (k)	Type	Number of feature map n	Number of neurons	Width of convolutional kernel s
0	input	3	—	—
1	convolution	96	—	11
2	convolution	256	—	5
3	convolution	384	—	3
4	convolution	384	—	3
5	convolution	256	9126	3
6	full connection	—	4096	—

2.1 Distribution of parameters in AlexNet–FE

Let's define $N(k)$ as the number of parameters in kth layer of AlexNet–FE.

Firstly, the number of parameters in kth convolutional layer is calculated as bellow:

$$N(k) = n_{k-1}s_k^2 n_k + n_k,\ k = 1, 2, \cdots, 5, \tag{1}$$

where n_k is the number of the output feature maps in the k-th convolutional layer, n_{k-1} is the number of the output feature maps in the $(k-1)$th convolutional layer, s_k is the width of the convolutional kernel. And $n_0 = 3$ is the number of channel of the input image.

Secondly, the number of parameters in the fully connected layer ($k = 6$) is calculated as

$$N(6) = n_{\text{conv5}} \times n_{\text{fc}} + n_{\text{fc}}, \tag{2}$$

where n_{conv5} is the number of the output neurons of the 5th convolutional layer, n_{fc} is the number of the neurons in the fully connected layer.

Moreover, we can define the following ratios:

$$R_N(k) = N(k)/(\sum_{a=1}^{6} N(a)),\ k = 1, 2, \cdots, 6. \tag{3}$$

According to (3) and Table 1, the ratio about the number of parameters of every layer in AlexNet is computed and shown in Fig.2. From Fig.2, it is obvious that the number of parameters in all the convolutional layers are much less than the parameters in the fully connected layer where about 99% of the parameters are in the fully connected layer. From Table 1, we can clearly see that 99% of the parameters are in the fully connected layer. Therefore, in order to reduce the number of parameters in AlexNet–FE, it is a straightforward idea that we should focus on reducing the number of parameters in the fully connected layer. Moreover, from (2), we can see that the number of the parameters of the fully connected layer can be reduced by reducing n_{conv5} or n_{fc}. In this paper, we make an effort to reduce n_{fc} to compress the O–AlexNet–FE model.

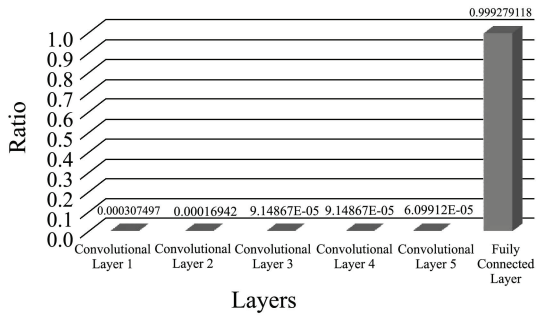


Fig. 2. The ratio of number of parameters of every layer

To differentiate the notation of the number of output neurons in the fully connected layer (n_{fc}), we denote $n_{\text{fc-s}}$ as the number of the remaining output neurons in the fully connected layer of the C–AlexNet. Accordingly, the compression ratio of the C–AlexNet–FE is defined asp

$$\text{Com}(n_{\text{fc-s}}) = \\ (\sum_{k=1}^{5} N(k) + n_{\text{conv5}} \times n_{\text{fc-s}} + n_{\text{fc-s}})/(\sum_{k=1}^{6} N(k)). \tag{4}$$

From Fig.2, It is noted that the number of parameters in all the convolutional layers are much less than the parameters in the fully connected layer. Therefore, Eq.(4) can be approximated as

$$\text{Com}(n_{\text{fc-s}}) \approx \\ \frac{n_{\text{conv5}} \times n_{\text{fc-s}} + n_{\text{fc-s}}}{n_{\text{conv5}} \times n_{\text{fc}} + n_{\text{fc}}} = \\ \frac{n_{\text{fc-s}}}{n_{\text{fc}}},\ n_{\text{fc}} >> n_{\text{conv5}}. \tag{5}$$

2.2 Analysis of redundancy in deep features

Supposed by [7], the responses of the fully connected layer in AlexNet are treated as the deep features which are termed as $\boldsymbol{F}_{\text{v}} \in \mathbb{R}^{n_{\text{fc}}}$, and the ith entry $\boldsymbol{F}_{\text{v}}^{(i)}$ can be viewed as an independent feature. In this section, by calculating the mutual information between $\boldsymbol{F}_{\text{v}}^{(i)}$ and the ground truth label $y \in \mathbb{R}$, we observe that the deep features have redundancy. These observation reveals that the dimension of $\boldsymbol{F}_{\text{v}}$ can be further reduced, which means we can use less number of neurons in the fully connected layer to remove redundant features ($n_{\text{fc}} < n_{\text{fc-s}}$). The problem becomes how to select the $n_{\text{fc-s}}$ neurons from n_{fc} neurons. Our derivation details are described below.

In probability theory and information theory, it is well-known that the mutual information (MI) of two random variables is a measure of the statistical dependency between these two variables. The smaller MI is, the more independent the variables will be. When two variables are independent, MI becomes zero.

Hence the mutual information between $\boldsymbol{F}_{\text{v}}^{(i)}$ and

y can be computed as

$$\mathrm{MI}(\boldsymbol{F}_{\mathrm{v}}^{(i)}, y) = H(\boldsymbol{F}_{\mathrm{v}}^{(i)}) + H(y) - H(\boldsymbol{F}_{\mathrm{v}}^{(i)}, y), \tag{6}$$

where $\boldsymbol{H}(\boldsymbol{F}_{\mathrm{v}}^{(i)})$ and $H(y)$ are the marginal entropies, $H(\boldsymbol{F}_{\mathrm{v}}^{(i)}, y)$ is the joint entropy of $\boldsymbol{F}_{\mathrm{v}}^{(i)}$ and y, and $\mathrm{MI}(\boldsymbol{F}_{\mathrm{v}}^{(i)}, y) \in [0, 1]$.

Therefore, the smaller $\mathrm{MI}(\boldsymbol{F}_{\mathrm{v}}^{(i)}, y)$ means that $\boldsymbol{F}_{\mathrm{v}}^{(i)}$ and y is less correlated. For the image classification task, the smaller $\mathrm{MI}(\boldsymbol{F}_{\mathrm{v}}^{(i)}, y)$ also indicates that there are less contribution of $\boldsymbol{F}_{\mathrm{v}}^{(i)}$ in classifying the image to y category and $\boldsymbol{F}_{\mathrm{v}}^{(i)}$ is redundant.

Based on the basic principles discussed above, it is possible to visualize the redundancy of the deep features. We calculate the histograms of the mutual information of deep features with scene categories database[14] and WCE bubble dataset and plot them in Fig.3 and Fig.4, respectively. Since Fig.3 and Fig.4 are computed from two different datasets, we can observe they have clearly different distributions. Moreover, we observe that majority values of the mutual information between deep features and category labels are in the range from 0 to 0.4, which means that many entries of the deep features are less correlated with the image categories. In other word, for image classification with these two datasets, the deep features extracted by the AlexNet–FE are of redundancy. Motivated by these observations, we make an effort to reduce the dimensions of deep features by selecting the most category-related features.

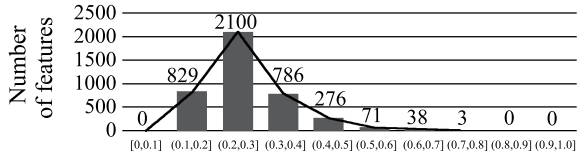


Fig. 3. The histogram of mutual information of deep features with scene categories database[14]

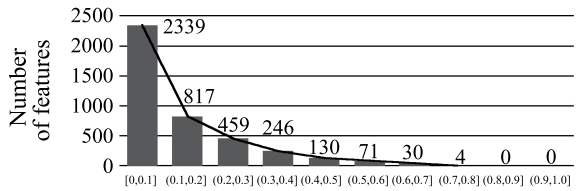


Fig. 4. The histogram of mutual information of deep features with WCE bubble dataset

3 Proposed CNN-model compression method

In this section, we firstly introduce the relationship between the feature selection and model compression. Secondly, we propose a feature selection method based on the mutual information and a novel selection criterion.

3.1 The relationship between the feature selection and model compression

As described in Subection 2.3, the deep features extracted by AlexNet–FE are redundant for image classification task with small number of categories, such as with fifteen scene categories and WCE bubble data.

As shown in Fig.1, $\boldsymbol{W}_{\mathrm{FC}}$ and $\boldsymbol{B}_{\mathrm{FC}}$ are the parameters of the fully connected layer, where $\boldsymbol{W}_{\mathrm{FC}} \in \mathbb{R}^{n_{\mathrm{fc}} \times n_{\mathrm{conv5}}}$ is the weight matrix between the 5th convolutional layer and the fully connected layer, $\boldsymbol{B}_{\mathrm{FC}} \in \mathbb{R}^{n_{\mathrm{fc}}}$ is the bias vector of the fully connected layer.

Let us denote $\boldsymbol{M}_{\mathrm{S}} \in \mathbb{R}^{n_{\mathrm{fc-s}} \times n_{\mathrm{fc}}}$ as the selecting matrix for the feature selection. The selected feature $\boldsymbol{F}_{\mathrm{V}}{}' \in \mathbb{R}^{n_{\mathrm{fc-s}}}$ is given by

$$\boldsymbol{F}_{\mathrm{v}}' = \boldsymbol{M}_{\mathrm{s}} \boldsymbol{F}_{\mathrm{v}}. \tag{7}$$

Besides, we denote $\boldsymbol{F}_5 \in \mathbb{R}^{n_{\mathrm{conv5}}}$ as the response of the 5th convolution layer in AlexNet–FE. Therefore, we have

$$\boldsymbol{F}_{\mathrm{v}} = \boldsymbol{W}_{\mathrm{FC}} \boldsymbol{F}_5 + \boldsymbol{B}_{\mathrm{FC}}. \tag{8}$$

Substituting (8) into (7), we get

$$\boldsymbol{F}_{\mathrm{v}}' = \boldsymbol{M}_{\mathrm{s}} \boldsymbol{W}_{\mathrm{FC}} \boldsymbol{F}_5 + \boldsymbol{M}_{\mathrm{s}} \boldsymbol{B}_{\mathrm{FC}}. \tag{9}$$

Then the after selection, $\boldsymbol{W}_{\mathrm{FC-s}} \in \mathbb{R}^{n_{\mathrm{fc-s}} \times n_{\mathrm{conv5}}}$ and $\boldsymbol{B}_{\mathrm{FC-s}} \in \mathbb{R}^{n_{\mathrm{fc-s}}}$ can be computed as

$$\begin{aligned} \boldsymbol{W}_{\mathrm{FC-s}} &= \boldsymbol{M}_{\mathrm{s}} \boldsymbol{W}_{\mathrm{FC}}, \\ \boldsymbol{B}_{\mathrm{FC-s}} &= \boldsymbol{M}_{\mathrm{s}} \boldsymbol{B}_{\mathrm{FC}}. \end{aligned} \tag{10}$$

As we designed that $n_{\mathrm{fc-s}}$ is smaller than n_{fc} after feature selection operated, $\boldsymbol{W}_{\mathrm{FC-s}}$ and $\boldsymbol{B}_{\mathrm{FC-s}}$ are essentially the compressed weight matrix and bias vector of C–AlexNet–FE, respectively. Therefore, with the feature selection, we can not only reduce the dimension of $\boldsymbol{F}_{\mathrm{v}}$, but also reduce the parameters of the fully connected layer by compressing $\boldsymbol{W}_{\mathrm{FC}}$ and $\boldsymbol{B}_{\mathrm{FC}}$.

3.2 Feature selection approach

In this subsection, a feature selection approach using mutual information is proposed to determine the feature selection matrix $\boldsymbol{M}_{\mathrm{S}}$.

In order to preferentially select the features that are most correlated to the ground truth label (category), the features are firstly sorted according to the mutual information between the feature vector $\boldsymbol{F}_{\mathrm{v}}$ and the label y in descending order, which is denoted as

$$\boldsymbol{X}_{\mathrm{S}} = \{\boldsymbol{F}_{\mathrm{v}}^{(S_1)}, \boldsymbol{F}_{\mathrm{v}}^{(S_2)}, \cdots, \boldsymbol{F}_{\mathrm{v}}^{(S_{4096})}\}, \tag{11}$$

where $\boldsymbol{X}_{\mathrm{S}}$ is the sorted feature set, which satisfies the conditions $\mathrm{MI}(\boldsymbol{F}_{\mathrm{v}}^{(S_1)}, y) > \mathrm{MI}(\boldsymbol{F}_{\mathrm{v}}^{(S_2)}, y) > \cdots >$

$\mathrm{MI}(\boldsymbol{F}_{\mathrm{v}}^{(S_{4096})}, y)$.

Then, the feature subsets denoted as $\boldsymbol{X}_{\mathrm{S}}^{j}$, which contains the first j element of $\boldsymbol{X}_{\mathrm{S}}$, are created as candidate feature subsets, denoted as

$$\boldsymbol{X}_{\mathrm{S}}^{j} = \{\boldsymbol{F}_{\mathrm{v}}^{(S_1)}, \boldsymbol{F}_{\mathrm{v}}^{(S_2)}, \cdots, \boldsymbol{F}_{\mathrm{v}}^{(S_j)}\}, \quad j = 1, \cdots, 4096. \tag{12}$$

The optimal feature subset $\boldsymbol{X}_{\mathrm{S}}^{j*}$ is determined according to the following selection criterion

$$j^* = \arg\max_{j}(\mathrm{Sco}(\boldsymbol{X}_{\mathrm{S}}^{j})), \tag{13}$$

where $\mathrm{Sco}(\boldsymbol{X}_{\mathrm{S}}^{j})$ is a score function for the feature subset $\boldsymbol{X}_{\mathrm{S}}^{j}$.

There are different methods to design the score function. Most commonly, it is defined by the classification accuracy. In this study, our goal is to compress the AlexNet–FE meanwhile to maintain the classification accuracy. From this point, we design the $\mathrm{Sco}(\boldsymbol{X}_{\mathrm{S}}^{j})$ as

$$\mathrm{Sco}(\boldsymbol{X}_{\mathrm{S}}^{j}) = \mathrm{Sco}_{\mathrm{drop}}(X_{\mathrm{S}}^{j}) \times \mathrm{Sco}_{\mathrm{comp}}(X_{\mathrm{S}}^{j}), \tag{14}$$

where

$$\begin{cases} \mathrm{Sco}_{\mathrm{drop}}(\boldsymbol{X}_{\mathrm{S}}^{j}) = \mathrm{e}^{a}[\mathrm{Acc}(X_{\mathrm{S}}^{j}) - \mathrm{Acc}(X_{\mathrm{S}}^{4096})], \\ \mathrm{Sco}_{\mathrm{comp}}(\boldsymbol{X}_{\mathrm{S}}^{j}) = 1 - \mathrm{Com}(j). \end{cases} \tag{15}$$

In Eq.(15), $\mathrm{Acc}(\boldsymbol{X}_{\mathrm{S}}^{j})$ is the classification accuracy for the feature subset $\boldsymbol{X}_{\mathrm{S}}^{j}$. $\mathrm{Acc}(\boldsymbol{X}_{\mathrm{S}}^{4096})$ is the classification accuracy of the AlexNet–FE.

$\mathrm{Com}(j)$ is the compression ratio for the feature subset $\boldsymbol{X}_{\mathrm{S}}^{j}$, which is defined in Eq.(5).

Super parameter a is used to trade off the impact of the accuracy and the compression ratio. Specifically, in Eq.(15), the $\mathrm{Sco}_{\mathrm{drop}}(\boldsymbol{X}_{\mathrm{S}}^{j})$ is a measurement of the drop from $\mathrm{Acc}(\boldsymbol{X}_{\mathrm{S}}^{4096})$ to $\mathrm{Acc}(\boldsymbol{X}_{\mathrm{S}}^{j})$. Considering the drop is usually very small, we formulate $\mathrm{Sco}_{\mathrm{drop}}(\boldsymbol{X}_{\mathrm{S}}^{j})$ in the exponential form to magnify its influence on the $\mathrm{Sco}(\boldsymbol{X}_{\mathrm{S}}^{j})$. In addition, the $\mathrm{Sco}_{\mathrm{comp}}(\boldsymbol{X}_{\mathrm{S}}^{j})$ is a measurement of the compression performance. With these two terms, (13) favors the result with very small drop of accuracy and high compression ratio.

After the optimal feature subset $\boldsymbol{X}_{\mathrm{S}}^{j*}$ is obtained, the entry $\boldsymbol{M}_{\mathrm{s}}(p, q)$ in pth row and qth column of $\boldsymbol{M}_{\mathrm{s}}$ can be calculated as follows

$$\boldsymbol{M}_{\mathrm{S}}(p, q) = \begin{cases} 1, \text{ if } S_p = q, \ p = 1, 2, \cdots, j^*, \\ 0, \text{ otherwise}. \end{cases} \tag{16}$$

For presentation completeness, we summarize the proposed CNN–model Compression algorithm in Table 2.

Table 2 The algorithm of CNN–model compression (feature selection method)

1. Calculate the mutual information between image features $\boldsymbol{F}_{\mathrm{v}}^{(i)}$ and categories y, then sort them in descending order $\boldsymbol{X}_{\mathrm{S}} = \{\boldsymbol{F}_{\mathrm{v}}^{(S_1)}, \boldsymbol{F}_{\mathrm{v}}^{(S_2)}, \cdots, \boldsymbol{F}_{\mathrm{v}}^{(S_{4096})}\}$, s.t. $\mathrm{MI}(\boldsymbol{F}_{\mathrm{v}}^{(S_1)}, y) > \mathrm{MI}(\boldsymbol{F}_{\mathrm{v}}^{(S_2)}, y) > \cdots > \mathrm{MI}(\boldsymbol{F}_{\mathrm{v}}^{(S_{4096})}, y)$.
2. Define the subset of features that contains j elements which in front of $\boldsymbol{X}_{\mathrm{S}}^{j}$ as the candidate subset $\boldsymbol{X}_{\mathrm{S}}^{j} = \{\boldsymbol{F}_{\mathrm{v}}^{(S_1)}, \boldsymbol{F}_{\mathrm{v}}^{(S_2)}, \cdots, \boldsymbol{F}_{\mathrm{v}}^{(S_j)}\}$, $j = 1, \cdots, 4096$.
3. Set BestNumber $= 0$, MaxValue $= 0$, $j = 1$.
4. Calculate $\mathrm{Sco}(\boldsymbol{X}_{\mathrm{S}}^{j})$, if $\mathrm{Sco}(\boldsymbol{X}_{\mathrm{S}}^{j}) >$ MaxValue: MaxValue $= \mathrm{Sco}(\boldsymbol{X}_{\mathrm{S}}^{j})$, BestNumber $= j$, $j = j + 1$.
5. If $j \leqslant 4096$, jump to step 4, otherwise, jump to Step 6.
6. Get the optimal subset of the features: $\boldsymbol{X}_{\mathrm{S}}^{\mathrm{BestNumber}}$

4 Experiments results

We evaluate our proposed CNN–model compression method as well as the SqueezeNet compression method proposed by Han et al in 2016[15] on two visual classification datasets: scene categories database[14] and self-built WCE bubble dataset. The former is widely used in image classification task with 4485 images of 15 categories. The latter is established for medical diagnosis analysis. It includes 4000 images of 2 categories (bubble and normal) from 10 individuals, while there are 200 bubble images and 200 normal images in each individual.

4.1 Experimental settings

The AlexNet–FE used in our experiments is implemented by a well-known deep learning framework Caffe[16]. The input color image of AlexNet–FE is set to the size of 227×227. The specific architecture is presented in Table 1. For our proposed CNN–model compression method, super parameter a is set to an empiric value of 20. In the experiment of scene categories database, the training and testing strategy is the same as those used in [14]. For WCE bubble dataset, the images from 5 individuals are randomly selected as the training images and the remaining images are used for testing. For SqueezeNet, the parameter setting follows the protocol used in [15].

4.2 Performance compression and classification results on two datasets

Since we take AlexNet as a generic feature extractor. Therefore, we take the output of the AlexNet as the input vector of a linear SVM classifier. To make the presentation clarify, let us denote AlexNet–FE–SVM[10] indicating the image classification system using AlexNet as feature extractor and linear SVM as a classifier. Similarly, C–AlexNet–FE–SVM

represents the image classification system using C–AlexNet as feature extractor and linear SVM as a classifier (our proposed method).

The experiment results are shown in Table 3. For scene categories database, more than 83% parameter in AlexNet–FE can be reduced by using our proposed model compression method while the drop of classification accuracy is smaller than 1%. For WCE bubble dataset, more than 95% parameters in AlexNet–FE are reduced with a little drop of classification accuracy. It is clear that our proposed method is able to compress the AlexNet–FE for image classification task with small number of categories, while maintaining the classification accuracy.

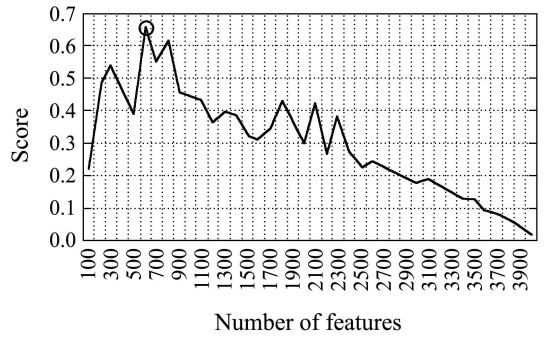
Table 3 The CNN–model compression results

Database	Method	Original→ Compressed	Average Accuracy
Fifteen scene categories	AlexNet–FE–SVM	—	83.67%
	Ours	240 MB→ 40 MB	82.68% (**0.99%** ↓)
WCE Bubble	AlexNet–FE–SVM	—	98.51%
	Ours	240 MB→ 12 MB	98.50% (**0.01%** ↓)

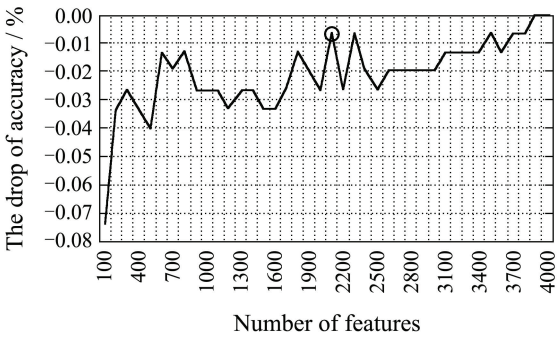
4.3 The impact on compression and classification with different selection criterions

As described in Subsection 3.2, this paper proposed a novel feature selection criterion which takes both the classification accuracy and the compression ratio into account. However, in [17], Peng only takes the classification accuracy as the feature selection criterion.

In this section, we evaluate these two different feature selection criterions with the scene categories database for AlexNet–FE. Fig.5 shows the correlation between the feature number and the score. Although using the traditional criterion would achieve the goal of maintaining accuracy, it is clear that our method can reduce more feature dimension than it. Specifically, our proposed criterion (shown in Fig.5(a)) gives the optimal number of features is 600 (reaches the highest score), which means that more than 83% of the parameters is reduced at the price of 0.01% drop of the classification accuracy in validation set. However, we also can see that the approach by [17] (shown in Fig.5(b)) only reduces 50% of the parameters by selecting the first 2100 features without losing the classification accuracy. Therefore, by considering the classification accuracy and the compression ratio, our proposed feature selection criterion is more suitable for compressing the AlexNet–FE. The same conclusion is made on WCE bubble dataset.



(a) Our proposed feature selection method.



(b) feature selection method proposed in [17]

Fig. 5. Score versus number of features with fifteen scene categories

4.4 Performance comparison of different compression models

Very recently, a novel DNN compression method was proposed by Han et al. in 2016[15] and it is termed as SqueezeNet. For evaluating purpose, one experiment has been conducted on two datasets and the results are given in Table 4. From Table 4, it is clear that SqueezeNet[15] achieves higher performance than ours in terms of model size and classification accuracy. It is worthwhile pointing out that our model compression method is much flexible to extend to different CNN model since we did not squeeze the convolutional layers. However, SqueezeNet is more complex and much difficult to apply to other CNN models since it is a delicate designed compression model.

Table 4 Performance comparison on two datasets

Database	Method	Original→ Compressed model size	Average accuracy
Fifteen scene categories	SqueezeNet (2016)	240 MB→ 5 MB	89.58%
	Ours (2015)	240 MB→ 40 MB	82.68%
WCE bubble	SqueezeNet (2016)	240 MB→ 5 MB	99.44%
	Ours(2015)	240 MB→ 12 MB	98.50%

5 Conclusion

In this paper, by carefully analyzing the AlexNet--FE, we found out that the majority network parameters are occupied by the fully connected layer. Moreover, investigating the mutual information between the fully-connected layer in CNNs and the corresponding class labels, we learnt that, for image classification tasks, the feature vectors at the fully connected layer are of redundancy. Motivated by these observations, we propose an effective model compression method based on a novel feature selection criterion which takes the compression ratio and the classification accuracy into account simultaneously. Experiment results with scene categories database and WCE bubble dataset show that our proposed method reduces more than 83% parameters of AlexNet–FE while almost retaining the classification accuracy. It is worthwhile to point out that our proposed compression method based on feature selection is more general and easily applied to different CNN models.

References:

[1] KRIZHEVSKY A, STUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [C] //*Advances in Neural Information Processing Systems*. Lake Tahoe, USA: MIT Press, 2012, 25: 1097 – 1105.

[2] SERMANET P, EIGEN D, ZHANG X, et al. Overfeat: integrated recognition, localization and detection using convolutional networks [J]. *Eprint ArXiv: 1312.6229*, 2013.

[3] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks [C] //*Computer Vision–ECCV 2014*. Zurich: Springer, 2014: 818 – 833.

[4] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] //*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Columbus: IEEE, 2014: 580 – 587.

[5] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [C] //*Computer Vision–ECCV 2014*. Zurich: Springer, 2014: 346 – 361.

[6] HARIHARAN B, ARBELÁEZ P, ARBELÁEZ R, et al. Simultaneous detection and segmentation [C] //*Computer Vision–ECCV 2014*. Zurich: Springer, 2014: 297 – 312.

[7] RAZAVIAN A S, AZIZPOUR H, SULLIVAN J, et al. CNN features off-the-shelf: an astounding baseline for recognition [C] //*IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Columbus: IEEE, 2014: 512 – 519.

[8] YOSINSKI J, CLUNE J, BENGIO Y, et al. How transferable are features in deep neural networks? [C] //*Advances in Neural Information Processing Systems*. Canada: MIT Press, 2014: 3320 – 3328.

[9] SUNDERHAUF N, SUNDERHAUF C, UPCROFT B, et al. Fine-grained plant classification using convolutional neural networks for feature extraction [C] //*Working Notes of CLEF 2014 Conference*. Sheffield, UK: CEUR Workshio, 2014.

[10] ZHOU B, ZHOU A, XIAO J, et al. Learning deep features for scene recognition using places database [C] //*Advances in Neural Information Processing Systems*. Canada: MIT Press, 2014: 487 – 495.

[11] GONG Y, LIU L, YANG M, et al. Compressing deep convolutional networks using vector quantization [J]. *Eprint ArXiv: 1412.6115*, 2014.

[12] DENTON E L, ZAREMBA W, BRUNA J, et al. Exploiting linear structure within convolutional networks for efficient evaluation [C] //*Advances in Neural Information Processing Systems*. Canada: MIT Press, 2014: 1269 – 1277.

[13] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural networks [J]. *Eprint arXiv: 1506.02626*, 2015.

[14] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories [C] //*IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. New York: IEEE, 2006: 2169 – 2178.

[15] 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]. *Eprint arXiv: 1602.07360*, 2016.

[16] JIA Y, SHELHAMER E, DONAHUE J, et al. Caffe: convolutional architecture for fast feature embedding [C] //*Proceedings of the 22nd ACM International Conference on Multimedia*. Orlando: ACM, 2014: 675 – 678.

[17] PENG H, LONG F, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(8): 1226 – 1238.

作者简介:

邹月娴 (1964–), 女, 博士, 教授, 博士生导师, 目前研究方向为机器视觉、模式识别与机器学习, E-mail: zouyx@pkusz.edu.cn;

余嘉胜 (1989–), 男, 硕士, 目前研究方向为计算机视觉, E-mail: 178142718@qq.com;

陈泽晗 (1992–), 男, 硕士, 目前研究方向为计算机视觉, E-mail: zehanchen@pku.edu.cn;

陈 锦 (1992–), 男, 硕士, 目前研究方向为计算机视觉, E-mail: chenjin@pku.edu.cn;

王 毅 (1992–), 男, 硕士, 目前研究方向为计算机视觉, E-mail: wygamle@pku.edu.cn.