2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

PLSNet: A Simple Network Using Partial Least Squares Regression for Image Classification

Ryoma Hasegawa
Department of Electrical and Electronic Engineering
Meijo University
Nagoya, Japan
153433027@ccalumni.meijo-u.ac.jp

Kazuhiro Hotta
Department of Electrical and Electronic Engineering
Meijo University
Nagoya, Japan
kazuhotta@meijo-u.ac.jp

***Abstract*—PCANet is a simple network using Principal Component Analysis (PCA) for image classification and obtained high accuracies on a variety of datasets. PCA projects explanatory variables on a subspace that the first component has the largest variance. On the other hand, Partial Least Squares (PLS) regression projects explanatory variables on a subspace that the first component has the largest covariance between explanatory and objective variables, and the objective variables are predicted from the subspace. If class labels are used as objective variables for PLS, the subspace is suitable for classification. Stacked PLS is a simple network using PLS for image classification and obtained high accuracy on the MNIST database. However, the performance of Stacked PLS was inferior to PCANet on the others. One of differences between Stacked PLS and PCANet is network architecture. In this paper, we combine the network architecture of PCANet with PLS and propose a new image classification method called PLSNet. It obtained higher accuracies than PCANet on the MNIST and the CIFAR-10 datasets. Furthermore, we change how to make filters for extracting features at the second convolution layer, and we call it Improved PLSNet. It obtained higher accuracies than PLSNet. In addition, we give it deeper network architecture, and we call it Deep Improved PLSNet. It obtained higher accuracies than Improved PLSNet.**

Keywords—Deep Learning; Convolutional Neural Network; PCANet; Partial Least Squares Regression; Stacked PLS; PLSNet;

I. INTRODUCTION

In recent years, we are easily able to treat a lot of images because of spread of internet, smart phone, and digital camera, etc. For example, we use surrounding words around images or tags annotated by human to find them. However, we sometimes are not able to find them correctly if the words and tags are mismatched. Furthermore, there are images without tags. Therefore, automatic image classification by computer is required to find images without those metadata.

In recent years, researches based on Convolutional Neural Network (CNN) have been widely done in computer vision after the success on ImageNet Large Scale Visual Recognition Challenge 2012 [1]. They obtained high accuracies on image classification [1, 2, 3, 4], fine-grained image classification [5], video classification [6], object detection [3, 7] and other tasks [8, 9]. Furthermore, CNN which pre-trained large-scale dataset as ImageNet [10] is useful as a powerful feature descriptor [11]. One of reasons why CNN gives the state-of-the-art accuracy is hierarchical feature extraction.

PCANet is a simple network using PCA for image classification [12]. First it crops a lot of local regions from training images and applies PCA to the regions. It considers weights obtained by PCA to be filters for extracting features and convolutes the filters on images. Almost the same processes are repeated, and it obtains some feature maps. Finally, it encodes the feature maps and classifies images by classifiers such as nearest neighbor [13] and Support Vector Machine (SVM) [14]. It obtained high accuracies on a variety of datasets such as the MNIST database [15] and the CIFAR-10 dataset [16].

PLS regression is widely used in chemo-metrics [17]. PCA projects explanatory variables on a subspace that the first component has the largest variance. On the other hand, PLS regression projects explanatory variables on a subspace that the first component has the largest covariance between explanatory and objective variables, and the objective variables are predicted from the subspace. If class labels are used as objective variables for PLS, the subspace is suitable for classification. Consequently, the subspace is more suitable for classification than PCA [18]. In recent years, PLS regression was also used in computer vision and obtained high accuracy on pedestrian detection [18].

Stacked PLS is a simple network using PLS for image classification [19]. First it crops a lot of local regions from training images and applies PLS to the regions. We consider weights obtained by PLS to be filters for extracting features and convolute the filters on images. Next, it carries out pooling and normalization in reference to CNN. Almost the same processes are repeated, and it obtains some feature maps. Finally, it concatenates the feature maps and classifies images by SVM. It obtained high accuracy on the MNIST database. However, the performance of Stacked PLS was inferior to PCANet on the others. One of differences between Stacked PLS and PCANet is network architecture.

In this paper, we combine the network architecture of PCANet with PLS and propose a new image classification method called PLSNet. First it applies PLS to a lot of local regions cropped from training images and convolutes weights obtained by PLS on images in reference to Stacked PLS. It repeats almost the same processes to obtain some feature maps and encodes the feature maps in reference to PCANet. Finally,

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

it classifies images by SVM. Furthermore, we change how to make filters for extracting features at the second convolution layer, and we call it Improved PLSNet. In addition, we give it deeper network architecture, and we call it Deep Improved PLSNet.

We evaluated respective PLSNets on the MNIST database and the CIFAR-10 dataset. PLSNet, Improved PLSNet and Deep Improved PLSNet obtained 99.41%, 99.46% and 99.48% acuuracy on the MNIST database. Those accuracies is slightly higher than PCANet. On the other hand, PLSNet and Improved PLSNet obtained 79.15% and 80.73% accuracy on the CIFAR-10 dataset. Although Deep Improved PLSNet has lower dimension than PCANet, the accuracy is higher than PCANet on the CIFAR-10 dataset.

This paper is organized as follows. In section 2, we explain the algorithm of PLS regression briefly. In section 3, we describe the details of our new image classification method called PLSNet. In section 4, we show experimental results on the MNIST and the CIFAR-10 datasets. Finally, we state conclusion and future works in section 5.

II. Partial Least Squares Regression

PLS regression projects explanatory variables on a subspace that the first component has the largest covariance between explanatory and objective variables, and the objective variables are predicted from the subspace. In this paper, a matrix for explanatory variables is denoted as $\boldsymbol{X} \in R^{n\times d}$ where n denotes the number of samples and d denotes the number of dimensions per sample. A vector for objective variables is denoted as $\boldsymbol{y} \in \boldsymbol{R}^{n\times 1}$ where n denotes the number of samples. The most famous algorithm of PLS called Nonlinear Iterative Partial Least Squares (NIPALS) is as follows.

```
$Centering\ \boldsymbol{X}\ and\ \boldsymbol{y}$
$for\ i = 1\ to\ component$
    $\boldsymbol{w}_i = \frac{\boldsymbol{X}^T\boldsymbol{y}}{\|\boldsymbol{X}^T\boldsymbol{y}\|_2}$
    $\boldsymbol{t}_i = \boldsymbol{X}\boldsymbol{w}_i$
    $\boldsymbol{p}_i = \frac{\boldsymbol{X}^T\boldsymbol{t}_i}{\boldsymbol{t}_i^T\boldsymbol{t}_i}$
    $c_i = \frac{\boldsymbol{y}^T\boldsymbol{t}_i}{\boldsymbol{t}_i^T\boldsymbol{t}_i}$
    $\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{t}_i\boldsymbol{p}_i^T$
    $\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{t}_i c_i$
$end\ for$
$\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{component}]$
$\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_{component}]$
$\boldsymbol{W}^* = \boldsymbol{W}(\boldsymbol{P}^T\boldsymbol{W})^{-1}$
```

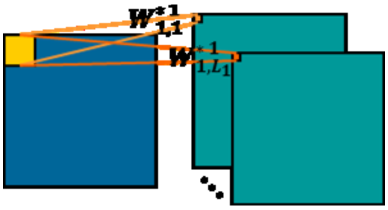

Fig. 1. The network architecture at the first convolution layer of respective PLSNets

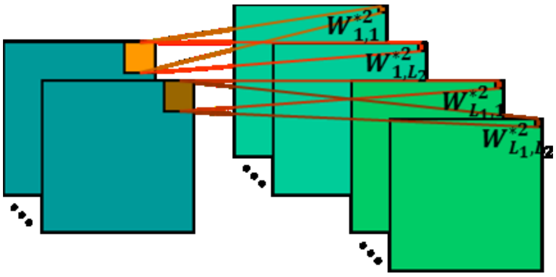

Fig. 2. The network architecture at the second convolution layer of Improved PLSNet

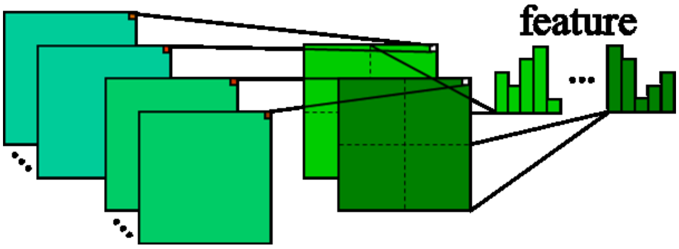

Fig. 3. The network architecture at the output layer of respective PLSNets

The number of components used for PLS is determined by using accuracy of validation samples. $\boldsymbol{W}^*$ is a matrix which projects $\boldsymbol{X}$ on a subspace that the first component has the largest covariance between $\boldsymbol{X}$ and $\boldsymbol{y}$.

III. Proposed Method

We explain the overview of our new image classification method called PLSNet. First it crops a lot of local regions from training images. We define them and their class labels as $\boldsymbol{X}$ and $\boldsymbol{y}$ respectively and apply PLS to them. Next, it convolutes $\boldsymbol{W}^*$ obtained by PLS as filters for extracting features on images. The subspace projected by $\boldsymbol{W}^*$ is suitable for image classification because the first component has the largest covariance between the local regions and their class labels. Therefore, it obtains some feature maps for image classification by convoluting $\boldsymbol{W}^*$ on images. It uses zero padding when it carry out the convolution. Almost the same processes are repeated, it extracts features hierarchically. Finally, it encodes them in reference to PCANet and classifies images by SVM. We describe the details of PLSNet as follows.

A. The First Convolution Layer

PLSNet crops a lot of local regions from training images and applies PLS to the regions. It convolutes $\boldsymbol{W}^*$ obtained by PLS as filters for extracting features on images. If the number of components used for PLS is L_1, it obtains L_1 feature maps per image. The network architecture at the first convolution layer of respective PLSNets is shown in Fig. 1. In Fig. 1, W_{1,L_1}^{*1} means the L_1-th filter for extracting features at the first convolution layer learned from the local regions which are cropped from training images.

B. The Second Convolution Layer

PLSNet crops a lot of local regions from L_1 feature maps of training images and applies PLS to the regions. It convolutes $\boldsymbol{W}^*$ obtained by PLS as filters for extracting features on images. If the number of components used for PLS is L_2, it obtains $L_1 L_2$ feature maps per image.

We change how to make filters at the second convolution layer of PCANet. PCANet applies PCA to $\boldsymbol{X} \in R^{nL_1 \times d}$ where n denotes the number of local regions per feature map obtained at the first convolution layer and d denotes the number of dimensions per local region. This means that PCANet does not distinguish L_1 feature maps and it applies PCA to all local regions which are cropped from L_1 feature maps. However, every filter is orthogonal each other. Therefore, we consider that it should distinguish L_1 feature maps and make filters for each feature map. In this paper, PLSNet which uses how to make filters of PCANet is called as PLSNet, and PLSNet which uses how to make filters for each feature map is called as Improved PLSNet. The network architecture at the second convolution layer of improved PLSNet is shown in Fig. 2. In Fig. 2, W_{L_1,L_2}^{*2} means the L_2-th filter for extracting features at the second convolution layer learned from the local regions which are cropped from the L_1-th feature maps of training images at the first convolution layer.

C. The N-th convolution Layer

PLSNet crops a lot of local regions from $\prod_{i=1}^{N-1} L_i$ feature maps of training images and applies PLS to the regions. It convolutes $\boldsymbol{W}^*$ obtained by PLS as filters for extracting features on images. If the number of components used for PLS is L_N, it obtains $\prod_{i=1}^{N} L_i$ feature maps per image. In this paper, PLSNet which has more than three convolution layers is called as Deep PLSNet.

D. Output Layer

PLSNet binarizes the feature maps obtained at the N-th convolution layer by positive or negative values. Next, it converts $\prod_{i=1}^{N} L_i$ feature maps to $\prod_{i=1}^{N-1} L_i$ feature maps by equation (1).

$$F^O = \sum_{i=1}^{L_N} 2^{i-1} H\left(F_{N-1,i}^{N}\right), \qquad (1)$$

In equation (1), $F_{L_{N-1},i}^{N}$ denotes the i-th feature map at the N-th convolution layer obtained from the feature map at the N-1-th convolution layer, H(·) denotes binarization function and F^O denotes converted feature map at the output layer. We convert the feature maps to real number by equation (1). Then, we consider the real number to be a bin of histogram. We consider a case $L_1 = L_2 = 2$. In this case, the minimum value of the real number in the feature map is $2^0 \times 0 + 2^1 \times 0 = 0$, and the maximum value of the real number in the feature map is $2^0 \times 1 + 2^1 \times 1 = 3$. We consider all possible value to be a bin of histogram, and the number of bin is $2^2 = 4$. Since the number of feature map at the first convolution layer is $L_1 = 2$, the dimension of the feature is $2 \times 4 = 8$. When we expand PLSNet into N convolution layers, the number of bin B_d of histogram is equation (2).

$$B_d = \prod_{i=1}^{N-1} L_i \cdot 2^{L_N} \qquad (2)$$

PLSNet divides the feature maps obtained by equation (1) into some blocks with overlap and makes histograms for every block. It obtains position invariance within every block by carrying out the process. Finally, we consider the concatenation of the histograms to be a feature vector and classify them by linear SVM. The network architecture at the output layer of respective PLSNets is shown in Fig. 3.

IV. Experiments

We show experimental results on the MNIST database and the CIFAR-10 dataset. This section is organized as follows. In section A, we explain the datasets for experiments and implementation details. In section B, we show the accuracies on the MNIST database. In section C, we visualize the filters obtained by respective PLSNets on the MNIST database. In section D, we show the accuracies on the CIFAR-10 dataset. In section E, we visualize the filters obtained by respective PLSNets on the CIFAR-10 dataset.

A. Datasets

The MNIST is a dataset for handwritten digits recognition. It consists of 10 classes from 0 to 9. Each image is gray-scale with 28 × 28 pixels, and the dataset contains 60,000 training images and 10,000 test images. In experiments, we use the last 10,000 training images as validation samples and the remaining training images as training samples. We choose the optimal cost of SVM by the accuracy of validation samples.

The CIFAR-10 is a dataset for general object recognition. It consists of 10 classes such as airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Each image is natural RGB with 32 × 32 pixels, and the dataset contains 50,000 training images and 10,000 test images. In experiments, we use the last 10,000 training images as validation samples and the remaining training images as training samples. We choose the optimal cost of SVM by the accuracy of validation samples.

B. Evaluation on the MNIST database

We evaluate PLSNet on the MNIST database. In this experiment, the sizes of filters at the first and second convolution layers are set to 7 × 7, the number of components used for PLS at the both convolution layers are set to 8, the size of block at the output layer is set to 7 × 7 and the stride of block at the output layer is set to 3 × 3 in reference to PCANet. Furthermore, we evaluate Improved PLSNet and Deep Improved PLSNet. In experiment, we use the same hyper-parameters with the PLSNet. The difference between PLSNet and Improved PLSNet is how to make filters at the second and third convolution layers. If the number of components used for PLS at the third convolution layer is set to 5, the number of bin of Deep PLSNet is the same with the number of bin of PLSNet whose number of components used for PLS at the second convolution layer is set to 8 by equation (2). Therefore, the number of components used for PLS

TABLE I. EVALUATION ON THE MNIST DATABASE

Method	Accuracy [%]
PCANet	99.38
Stacked PLS	99.42
PLSNet	99.41
Improved PLSNet	99.46
Deep Improved PLSNet	**99.48**

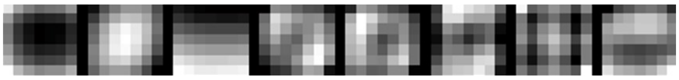

Fig. 4. Filters at the first convolution layer by respective PLSNets on the MNIST database

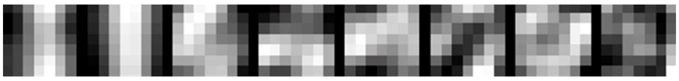

Fig. 5. Filters at the second convolution layer by PLSNet on the MNIST database

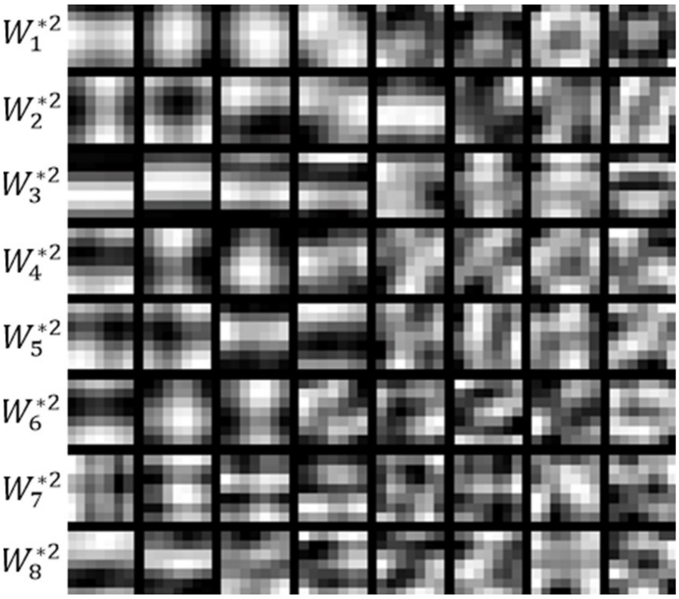

Fig. 6. Filters at the second convolution layer by Improved PLSNet on the MNIST database

at the third convolution layer is set to 5 in Deep Improved PLSNet.

The results are shown in Table I. Table I shows that PLSNet obtained 99.41% accuracy. Improved PLSNet and Deep Improved PLSNet obtained 99.46% and 99.48% respectively, and these accuracies are higher than Stacked PLS. Since the accuracy of PCANet is high, the improvement of our method is not so large. However, we see that our proposed method obtained better accuracy than PCANet.

TABLE II. EVALUATION ON THE CIFAR-10 DATABASE

Method (the size of filters, the component of PLS)	Accuracy [%]
PCANet (5 × 5, 40 - 8)	77.14
PCANet (combined)	78.67
PLSNet (3 × 3, 12 - 8)	76.78
PLSNet (5 × 5, 28 - 8)	76.62
PLSNet (combined)	79.15
Improved PLSNet (3 × 3, 12 - 8)	78.3
Improved PLSNet (5 × 5, 28 - 8)	79.07
Improved PLSNet (combined)	**80.73**
Deep Improved PLSNet (3 × 3, 12 - 8 - 5)	79.58

C. Visualization of filters for the MNIST database

We visualize filters obtained by respective PLSNets on the MNIST database. Fig. 4 shows the filters at the first convolution layer in respective PLSNets. From left to right shows the filters for the first component to the eighth components. Fig. 4 demonstrates that respective PLSNets extracts a variety of edges.

Fig. 5 and 6 show the filters at the second convolution layer obtained by PLSNet and Improved PLSNet respectively. From left to right show the filters for the first component to the eighth component in these figures. In Fig. 6, from up to down shows the filters made for the first feature map to the eighth feature map at the first convolution layer. Fig. 5 shows that PLSNet obtains only a kind of filters. On the other hand, Fig. 6 shows that Improved PLSNet obtains eight types of filters. We consider that this is the reason why Improved PLSNet obtained higher accuracy than PLSNet.

D. Evaluation on the CIFAR-10 dataset

First we evaluate PLSNet on the CIFAR-10 dataset. In this experiment, we train two PLSNets with different hyper-parameters in reference to PCANet and evaluate three PLSNets which are the two PLSNets and combination of the two PLSNets. The sizes of filters at the first and second convolution layer are set to 3 × 3 for one PLSNet and 5 × 5 for the other PLSNet, the number of components used for PLS at the convolution layer are set to 12 - 8 for one PLSNet and 28 – 8 for the other PLSNet where the former numbers (12, 28) denote the number of components at the first convolution layer and the later numbers (8, 8) denote the number of components at the second convolution layer, the sizes of block at the both output layers are set to 8 × 8, and the strides of block at the both output layers are set to 4 × 4 in reference to PCANet. For the convenience of time, we do not use Spatial Pyramid Pooling [3] which is used in PCANet. Furthermore, we evaluate Improved PLSNet and Deep Improved PLSNet. In these experiments, we use the same hyper-parameters with the PLSNet. For the same reason as the experiments on the MNIST database, the number of components used for PLS at the third convolution layer is set to 5 in Deep Improved PLSNet.

The results are shown in Table II. The numbers in each method denote the size of filters at the both convolution layers and the number of filters at the first and second convolution

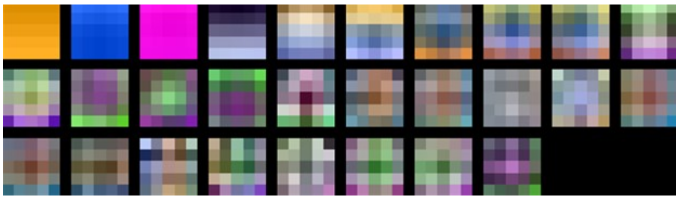

Fig. 7. Filters at the first convolution layer by respective PLSNets (5 × 5, 28 - 8) on the CIFAR10 dataset

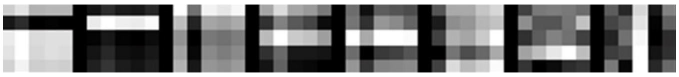

Fig. 8. Filters at the second convolution layer by PLSNet (5 × 5, 28 - 8) on the CIFAR-10 dataset

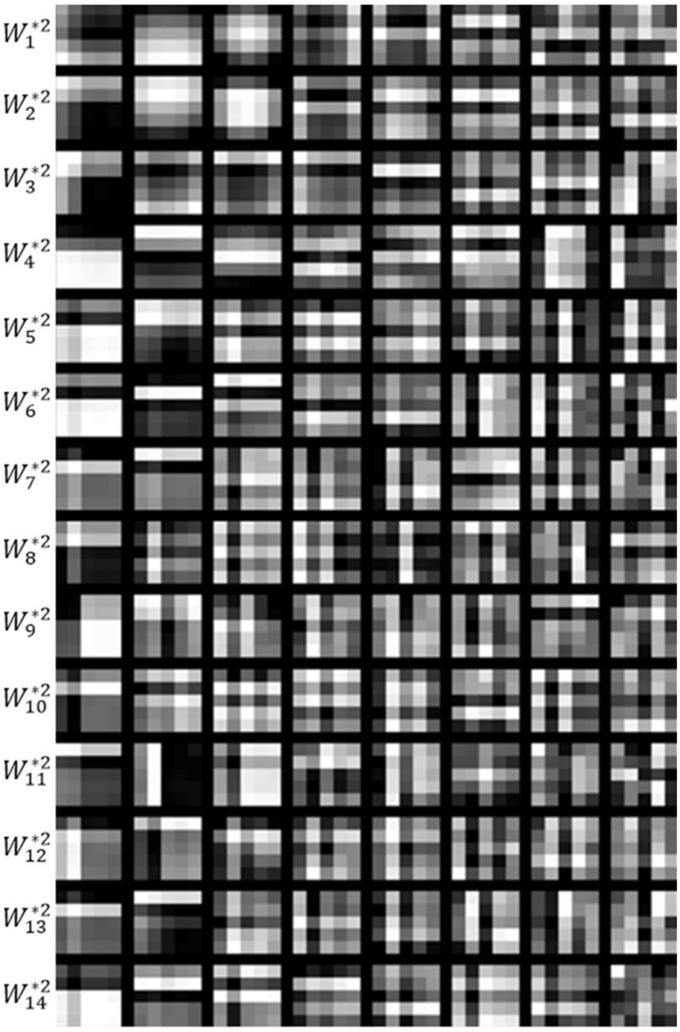

Fig. 9. Filters at the second convolution layer by Improved PLSNet (5 × 5, 28 - 8) on the CIFAR-10 dataset

layers respectively. Table II shows that PLSNet (combined) obtained 79.15% accuracy. It is 0.48% higher than PCANet (combined). Furthermore, Improved PLSNet (combined) obtained 80.73% accuracy which is 2.06% higher than PCANet (combined). Deep Improved PLSNet (3 × 3, 12 - 8 - 5) obtained 79.58% accuracy which is 1.28% higher than Improved PLSNet (3 × 3, 12 - 8). For the convenience of time, we do not evaluate Deep Improved PLSNet with the other parameters such as Deep Improved PLSNet (5 × 5, 28 - 8 - 5) and (combined). However, we will obtain higher accuracy if we select adequate hyper-parameters.

E. *Visualization of filters for the CIFAR-10 dataset*

We visualize filters obtained by respective PLSNets on the CIFAR-10 dataset. Fig. 7 shows the filters at the first convolution layer in respective PLSNets (5 × 5, 28 - 8). From left to right show the filters from the first component to the twenty-eighth components. Fig. 7 demonstrates that respective PLSNets extracts a variety of features including simple and complex patterns.

Fig. 8 and 9 show the filters at the second convolution layer obtained by PLSNet and Improved PLSNet respectively. From left to right show the filters for the first component to the eighth component in these figures. In Fig.9, the filters at top row are made for the first feature map at the first convolution layer. Filters at bottom row are made for the fourteen feature map at the first convolution layer. Fig. 8 shows that PLSNet obtains only a kind of filters. On the other hand, Fig. 9 shows that Improved PLSNet obtains fourteen types of filters. We consider that this is the reason why Improved PLSNet obtained higher accuracy than PLSNet.

V. Conclusion

In this paper, we proposed a new image classification method called PLSNet, Improved PLSNet and Deep Improved PLSNet. PLSNet obtained higher accuracy than PCANet and Stacked PLS on the MNIST database and the CIFAR-10 dataset.

In experiments, hyper-parameters used for PLSNet are the same as PCANet for fair comparison. Therefore, we will obtain higher accuracy if we select hyper-parameters properly. This is a subject for future works.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems 25, pp.1097-1105, 2012.

[2] M. Lin, Q. Chen, and S. Yan, “Network in network”, Proc. International Conference on Learning Representations, 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition”, Proc. European Conference on Computer Vision, pp.346-361, 2014.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions”, Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1-9, 2015.

[5] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, “The application of two-level attention models in deep convolutional neural network for fine-grained image classification”, Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.842-850, 2015.

[6] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural

networks", Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1725-1732, 2014.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.580-587, 2014.

[8] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: closing the gap to human-level performance in face verification", Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1701-1708, 2014.

[9] V. Badrinarayanan, A. Kendall, R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation", Proc. International Conference on Computer Vision, 2015.

[10] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database", Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.248-255, 2009.

[11] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks", Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1717-1724, 2014.

[12] T. H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcanet: a simple deep learning baseline for image classification?", IEEE Transactions on Image Processing, Vol.24, No.12, pp.5017-5032, 2014.

[13] S. Dudani, "The distance-weighted k-nearest-neighbor rule", IEEE Transactions on Systems, Man and Cybernetics, Issue.4, pp.325-327, 1976.

[14] V. Vapnik, "Statistical learning theory", Wiley, New York, 1998.

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", Proceedings of the IEEE, Vol.86, No.11, pp.2278–2324, 1998.

[16] The CIFAR-10 dataset, https://www.cs.toronto.edu/~kriz/cifar.html.

[17] H. Wold, "Soft modeling by latent variables: the nonlinear iterative partial least squares approach", Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett, 1975.

[18] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis", Proc. International Conference on Computer Vision, pp.24-31, 2009.

[19] R. Hasegawa and K. Hotta, "Stacked partial least squares regression for image classification", Proc. Asian Conference on Pattern Recognition, 2015.