A Novel Approach for Regularization of Convolutional Neural Network

Yuan Zhang[1,2*], BiMing Shi[1]
[1]Safety Technology and Engineering Specialty, Anhui University of Science and Technology
Huainan 23200, China

[2]Department of Basic Courses, Anhui Medical College
Hefei, 230001, China
*Corresponding Author, email: Cheungyuan1107@163.com

Journal of Digital Information Management

ABSTRACT: *At present, the traditional convolutional neural network (CNN) can easily cause overfitting in the training process, thereby resulting in an invalid training model. Thus, this study proposed a novel CNN regularization method to avoid overfitting in the training process and to increase the image classification accuracy of CNN. The proposed method uses failure probability as the theoretical basis. First, the failure probability density (FPD) function of image pixel points was deduced, and the FPD prediction model of pixel points in image neighborhood was established. Second, this method evaluated the FPD of image eigenvalues, selected the pixel points with the minimum FPD in the image neighborhood as the retention characteristics, and extracted excellent features. Finally, this method conducted a labeling classification experiment on three image classification datasets (MNIST, CIFAR-10, and CIFAR-100) and compared the image classification accuracy of the three existing popular pooling methods (Max, Average, and Stochastic pooling). Results demonstrated that the proposed regularization method based on FPD can increase the classification accuracy of the image data. The classification accuracies gained by the training datasets are all higher than 90%, and the classification accuracies gained by the testing datasets are all higher than 85%. The proposed CNN regularization method possesses good generalization capability and engineering applicability.*

Subject Categories and Descriptors
I.4.8 [Scene Analysis]: Object recognition; **I.5.1 [Models]:** Neural nets

General Terms
Neural Network, Image Classification, Probability Theory, Image Processing, Pattern Recognition.

Keywords: Failure Probability Density, Pooling, Regularization, Convolutional Neural Network, Overfitting.

Received: 1 November 2016, Revised: 27 December 2016, Accepted: 11 January 2017

1. Introduction

Convolutional neural network (CNN) is widely applied in machine learning fields such as computer vision, pattern recognition, and scene classification, and they have become hot research topics nowadays. Given its weight sharing and local receptive field characteristics [1], CNN exhibits a strong learning ability and fast image processing. Traditional CNN is composed of two layers, namely, convolutional and pooling. Generally, in CNN, all units in the same layer share equal weights and have the same detected features, thereby enabling translation invariant feature detection when these detected features are delivered to high levels. Next, CNN is always insensitive to distor-

tion through the pooling layer. However, due to weight sharing and the translation-invariant feature of image features, CNN can easily train image noises or errors together in the image feature prediction model. The image labeling classification accuracy of training data is high and even reaches 100% during enough training iterations if no adequate training dataset exists. However, the image labeling classification accuracy in practical test is low, which is known as the overfitting phenomenon and finally results in an invalid prediction model and causes the model to lose generalization ability [2].

Some studies on the overfitting problems of CNN have reported that overfitting occurs when the training model has high complexity and inadequate training samples [3]. Therefore, the training model can reduce its complexity and interferences from noises or wrong information through regularizing network [4, 5]. Conversely, the convolutional layer in the CNN structure has features of one image neighborhood through convolution, while the pooling layer acquires new characteristics by integrating feature points in a small neighborhood through the pooling technology. Pooling results can reduce features and parameters, thus decreasing the calculated amount for image feature classification. Therefore, the pooling technology is the most common non-parameter regularization CNN approach [6]. Thus, studying and optimizing the pooling method of CNN are important. A novel pooling method was proposed by combining traditional CNN regularization method and failure probability theory, which was used as the regularization strategy of CNN.

2. State of The Art

Overfitting can easily make features that were extracted during NN training insignificant. To address this problem, many scholars proposed new data training methods continuously. Some methods improved the overfitting problem significantly, but they often have poor generalization capability on different image datasets. The pooling layer in a CNN structure aims to prevent overfitting of training through dimensionality reduction. Traditional pooling technologies have max pooling and average pooling [7]. The max pooling technology extracts maximum pixels in the image neighborhood and abandons the remaining pixels. This process achieves the goal of zooming image features and reducing image dimensions, finally preventing overfitting. The max pooling technology retains the textural features in images. However, the limitation on the size of the image neighborhood will increase the variance of the estimated value and may fail to load the maximum layer [8]. In short, the max pooling technology is simple and feasible, but the image feature extraction is excessively rough. The average pooling technology calculates the pixel mean of feature points in an image neighborhood, which retains the image background characteristics [9]. However, the parameter error of the convolutional layer causes an offset of the estimation mean, and as a result, the features of the image foreground are unclear and not sufficiently explicit. Hilton recently proposed an improved regularization method called dropout [10,11], which reduces weight calculation by randomly deleting half weights of connections in one layer of the neural network, thus preventing overfitting. However, the position of random deletion in the network layer is crucial. Inappropriate position selection may delete important image features. However, no universal applicable standards for selection of position exist, and positions can be selected only according to experience [10,11,12]. Moreover, dropout has weak generalization ability. The stochastic-pooling regularization method [12] proposed by Matthew D. Zeiler does not establish the prediction model of image pixels in a simple manner; rather, it provides the probability of pixel points according to numerical values and then collects samples randomly according to the probability. In other words, feature points with high pixel values have a high probability of being retained. The ability of the method to extract the textural features of images is between that of max and average. The method has a good generalization ability and can be viewed as the product of model averaging. Nevertheless, the experiment demonstrated that stochastic pooling can obtain an excellent effect only by cooperating with the dropout model [12,13], thereby increasing the complexity of the training model. Moreover, the approach still requires empirical parameter adjustment, and its generalization ability does not reach the optimum level [14]. The above studies show that the pooling technology is the main technological method for regularization of CNN but has poor generalization ability in different experiments on image classification data. It has a low accuracy of image classification and depends on empirical judgment for parameter selection. In this study, a prediction model based on FPD was constructed through scientific deduction of the probability theory by combining the failure probability theory and traditional CNN pooling technology. Therefore, a novel pooling method that can enhance the model generalization ability and accuracy was proposed.

The remainder of this study is organized as follows.Section 2 introduces three common pooling technologies and their characteristics. Section 3 constructs the FPD function by combining the failure probability theory and proposes the prediction model based on FPD. This model was applied to the pooling layer of the traditional CNN. Section 4 compares the proposed regularization method with three traditional pooling technologies and verifies its feasibility and accuracy with three image classification datasets. Section 5 concludes the study.

3. Methodology

In the process of traditional pooling method of CNN, we can choose the overlapping or non-overlapping unit[15], the proposed method in this study is better than that of the non-overlapping unit. We set the size of a $z \times z$ pooling kernel in a size of $M \times M$ convolutional mapping, Extracted the feature of invariant local transformations in an $M \times M$ image, and summarized the best performance of the $z \times z$ neighborhood as an output of pooling [16,17,18,19]. The

probability density analysis of the basic eigenvalue was obtained by analyzing the probability density of the variable of $M \times M$ convolutional eigenvalue,and understood the importance of each feature,thus, the lower value of the failure probability was regarded as a deterministic feature[20]. The probability density represented the influence of on the uncertainty of image feature, obtained the importance order of eigenvalue of FPD. We can prioritized the most important feature variables or ignore the less important variables in the process of design and optimization with this order. The calculation of FPD had shift invariance and similarity invariance[21],it meant that the original image feature variables had the same FPD in the neighborhood [22].

A hidden layer is generated by each of the convolutional layer, which contains several neural units. The pooling procedure has been performed on the feature value of the hidden layer, where the computational complexity can be reduced and overfitting can be alleviated. In an $M \times M \times H$ image, $N \times H$ image feature mappings were obtained with N $z \times z$ convolution kernels. Each feature mapping was independent and can be expressed as follows,

$$y = g(\omega_i x_i + b) \tag{1}$$

Where g(**.**) is the output function of the convolution layer, y is the eigenvalue, and $x_i \in X$ is the image input of the convolution layer. An $n \times n$ pooling unit was used to perform non-overlapped translation pooling of the mapping of $(M - z + 1) \times (M - z + 1)$ after pooling. In this way, one pooling area can be obtained by each convolution mapping. The failure range of the output eigenvalue for every pooling area was set as $F = \{x: G(x) \leq 0$. Therefore, the function was used to describe eigenvalue of the image. The failure probability of the eigenvalue can be expressed as,

$$P_{G(X)} = \int_{-\infty}^{0} f(y)\, dy \tag{2}$$

Where $f(y)$ is the FPD function[27]. However, the distribution of the pixel variable of the original image,X, is often unknown. In most existing research, the non-parametric estimation e.g., Kernel Density Estimation are used to efficiently calculate FPD $f(y)$. with unknown distribution. Therefore, we also use the kernel density estimation algorithm to estimate the FPD . Every convolutional mapping of the pooling layer was assumed to contain N samples, y_i $(i = 1,...,N)$, which is assumed to be sampled from the FPD function, $f(y)$ to construct the failure value matrix. This matrix can be expressed as follows,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ y_n \end{bmatrix} \tag{3}$$

Moreover, K out of N sample data points were assumed to be inside of the failure area F, where a binomial distribution can be expressed as follows[25],

$$Bin(K \mid N, P_{g(x)} = \frac{N!}{K!\,(N-K)!} P^K (1-P)^{1-K} \tag{4}$$

When the sample size N becomes large, the approximation $K \approx N \times P$ can be justified. In contrast, when the sample size N becomes small, the following equation can be derived,

$$P_{G(x)} = \frac{K}{NV} \tag{5}$$

Where V is the set of the failure range, F. The failure range F was assumed to be a very small cube centered at y and with h as its side length. The number of 3D data points falling inside this area was K. The kernel function was defined as follows,

$$K = \sum_{1}^{N} k\ \frac{(y - y_n)}{h} \tag{6}$$

Substituting Eq. (6) to Eq. (5) and assuming the estimated value of the kernel density $f(y)$ was $f(y) \approx f^*(y)$after derivation, the following equation can be obtained:

$$f^*(y) = \frac{1}{Nh} \sum_{1}^{N} K(y - y_i) \tag{7}$$

Where K(**.**) denotes the kernel density function and h denotes the bandwidth. The key ingredient in this algorithm was how to determine a suitable K(**.**) and choose an optimal bandwidth. In this study, Gaussian function was used as kernel function and the bandwidth was set as h=2.

Another important factor in this algorithm is to rank the numerical value of the failure probability density for the image convolutional feature. These are based on the geometric position of the response value $e_{ij} = f^*(y)$ of the FPD for the image and selection of the minimum FPD value, $\min(e_{ij})$ can also be obtained. The minimum probability density value indicated that the FPD of the image eigenvalue was the smallest, the sample failure probability was low, and the image eigenvalue was also the most reliable one. Therefore, in pooling neighborhood of size $z \times z$, the image convolutional eigenvalue is for the position of the reserved failure probability density, where all the remaining eigenvalue $g(x_{ij})$ can be simply abandoned. A $z \times z$ pooling area can be transformed to an image eigenvalue after pooling. In Figure.1, the chart of the process is shown.

FPD-pooling model can be considered as a new type of model represented by a set of locations in the new regions. In training phase, new positions can be obtained to produce a new model for sampling. Subsequently, a model can be effectively constructed based on the use of failure probability [23] to estimate the feature values. In

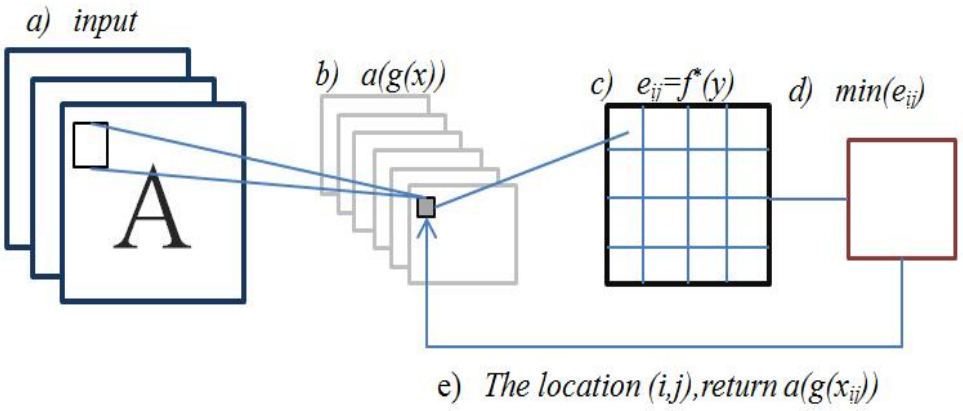

Figure 1. The chart of the FPD-pooling. a) 32×32 input image. b) 5×5 convolution kernel activation function $a(g(x))$ of each convolution layer. c) The estimate value of FPD $e_{ij} = f * (y)$. d) The minimum value of FPD $\min(e_{ij})$, record the corresponding position (i,j). e) Return the features value of the corresponding position $a(g(x_{ij}))$

CNN, it is assumed that L pooling layers are present, the spatial size of each pooling layers are *2×2, 3×3, 4×4* of pooling region, respectively, where n^l pooling models can be produced. We utilize FPD-pooling to achieve similar performance as in Max-pooling, Average-pooling, Stochastic-pooling in training phase of the dataset. Therefore, using FPD-pooling for weighting the activations with test datasets is modified as demonstrated in the latter part of the study.

The training algorithm is described as follows:

Algorithm: Train dataset using the value of FPD

Input: *sample variable x, Stochastic parameter, Stride u, convolution kernel size z×z, bandwidth h*

1: Output: *Feature value max(f(y))*

2: Feedforward propagation:

3: Computing the feature value of convolution:

4: $Y = g(\sum \omega_i x_i + b) =$ *relu* $g(\sum \omega_i x_i + b)$

5: Extracting failure field samples $G(X) \leq 0$

6: Computing FPD: $e_{ij} = f * (y)$

7: Extracting the position of the minimum FPD: $min(e_{ij})$

8: Return the corresponded (i,j) convolution feature valu $g(x_{ij})$ *from image*

9: Drop the feature value of the rest position

4. Result Analysis And Discussion

In this study, all the experiments were conducted on *MATLAB 2015B*[24]. The *matconvnet-1.0-beta23* toolbox was used. The network structure of *simplynn* in the *matconvnet* toolbox[14] package was employed for testing, which included a *5×5* convolution kernel and 64 feature images. The *relu* function[25] was used as the output of the activation function. The proposed pooling algorithm was then used to extract eigenvalues of the feature. Finally, the *softmax* function was utilized as the classification loss function for the fully connected layer. Moreover, mini-batch stochastic gradient descent was used with a stepsize of 100. The Backpropagation algorithm was employed to update the weights. For each of the datasets, the same network architecture was used as shown in Figure.2, for fair comparison of the results. All of the hyperparameters remained to be same and 300 iterations were conducted for three different datasets (i.e., CIFAR-10[26], CIFAR-100[26], and MNIST[27]).

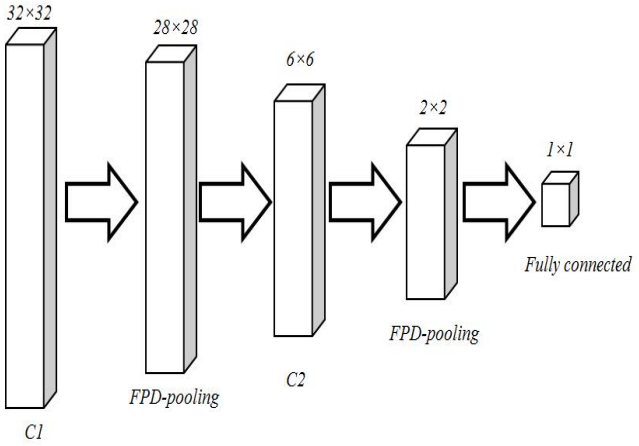

Figure 2. The architecture of network. Convolution layer and pooling layer,with convolutional kernel size of 5×5 pooling size of 3×3 and stride 2

4.1 Cifar-10

In this dataset, it consisted of 50,000 training samples of natural images, where 10 categories are included. For a total of 5,000 training samples, the 10-class classification was carried out. In this dataset, each image was a three-layer RGB image with an image size of 32×32. All images were manually labeled. The convolution kernel size was set to be 5×5, the stepsize was set to be 2 and pooling size was set to be 3×3, where the non-overlapped pooling was adopted. Therefore, eigenvalue of the image could be

decreased by 75% after each FPD-pooling. Training data based on the previously described network structure were used for training four groups of models, i.e., the Max-pooling, Average-pooling, Stochastic-pooling and FPD-pooling models, respectively. In Figure.3, results of error comparison after 300 iterations are given. The FPD-pooling model can desirably avoid overfitting in this dataset, which was very different from all the compared three other models. The training results only show slight improvements with FPD-pooling model as given in Figure 3.

In this experiment, the value of stepsize was significantly adjusted, since it was too small to jump out from the local optimum. Otherwise, the model will hover near the global optimum, where the stepsize was chosen to be $steosize = 100$. In Figure.3, the overfitting phenomenon can be clearly observed in the Max-pooling. And the errors obtained by Average-pooling and Stochastic-pooling are relatively high, the proposed FPD-pooling not only can avoid overfitting, but also can obtain low errors for both training and testing.

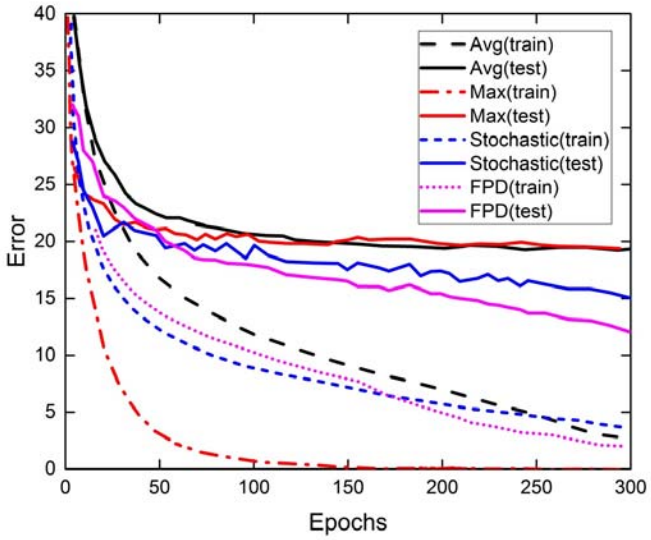

Figure 3. The FPD-pooling method had a learning rate of 0.01 and error performance of the iteration was 300 times, where the bandwidth of FPD is $h = 2$. A comparison of FPD-pooling of the training and testing datasets in the iteration after 150 times indicated that the training data had a 0.8% error rate reduction as compared to Stochastic-pooling. The test data had less than 1.5% of Stochastic-pooling error

4.2 Mnist

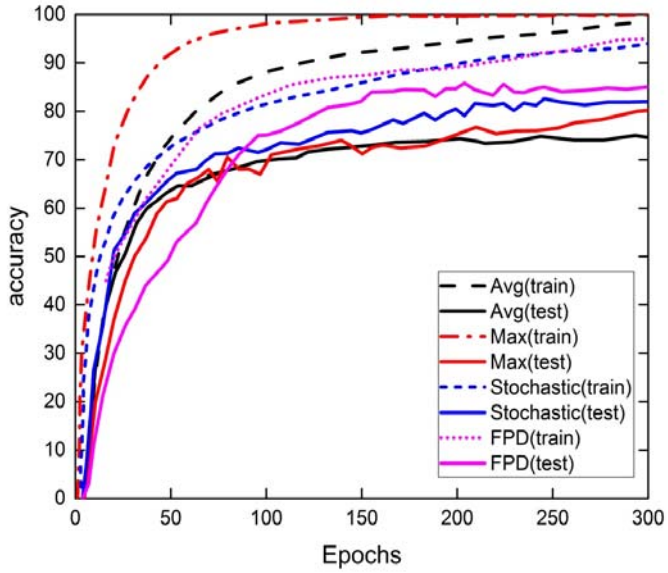

Figure.4 The proposed FPD-pooling was 87% accuracy at training time,and 80% accuracy in test dataset. Moreover,Max-pooling presents overfitting, Avg-pooling indicates the best accuracy at training time,while FPD-pooling indicates the highest accuracy compared with other methods

In this dataset, hand-written numbers ranging from 0 to 9 are considered, which includes 60,000 hand-written images of size 28×28 for training and 10,000 test data. The data were normalized to [0, 1], where the spatial size of convolution kernel was 5×5, the stride was 2, and the size of pooling was 3×3. No pre-training and pre-processing procedures were performed, but a simple four-layer convolutional network structures, i.e., convolution layer-pooling layer-convolution layer-fully connected layer, were used. The results of testing are given in Figure.3.In Figure.4, the accuracy of Max-pooling was 100% in training, which has obviously indicated overfitting phenomenon. Moreover, the performance obtained by the proposed method was as bad as the Stochastic-pooling, but not as good as Average-pooling. In the testing phase, the proposed FPD-pooling can achieve the best performance.

4.3 Cifar-100

Similar to CIFAR-10, the dataset consisted tiny images of 100 classes, including 5,000 images for training and 10,000 images for testing. In particular, there are 500 images in each category. The pre-processing procedures are carried out similar to those for CIFAR-10, and data value was normalized to [0, 1].

In Table.1, the error rate of training data is given, which can be expressed by Average-pooling and Stochastic-pooling coinciding with the expectation. The overfitting problem could be observed in the Max-pooling. Notably, both top1err and top5err by the proposed algorithm has minimum error rates. As can be seen in Table 1, the proposed algorithm can achieve the highest accuracy rate.

Pooling method	***Train error %***	***Test top 1error %***	***Test top5error %***
Max-pooling	*0%*	*30.2%*	*20.8%*
Average-pooling	***2.3%***	*25.6%*	*17.6%*
Stochastic-pooling	*1.9%*	*17.3%*	*15.4%*
Ours	*2.1%*	***15.5%***	***13.2%***

Table 1. Three pooling methods compared to FPD-pooling method based on *top1error* and *top5error*

5. Conclusion

To increase the image classification accuracy of CNN learning and prevent overfitting in the training process, this study deduces the prediction function of FPD by combining the failure probability theory and traditional CNN pooling technology. Thus, a pooling technology based on FPD is proposed, and its image labeling classification accuracies on three popular image datasets are tested. The conclusions are as follows:

The conclusions are as follows:

(1) The proposed regularization method based on failure probability theory can be used in the pooling layer of traditional CNN and predict image labeling classification successfully.

(2) According to image classification accuracy, a smaller FPD of an image feature corresponds to a greater importance of the image feature. Research also demonstrates that the proposed regularization method based on FPD is obviously superior to the traditional CNN pooling technology in terms of classification accuracy.

(3) Results of the proposed regularization method on different quantities and types of datasets reflect that the proposed method has a stronger generalization ability than the traditional pooling technology.

In this study, a novel CNN regularization method is proposed by combining theoretical deduction and experiments on datasets. The established pooling model based on FPD has stronger generalization ability and higher classification accuracy. Given the inadequate number of experimental datasets and inadequate structural depth of the used CNN, future studies will increase the sample data of experiments and combine and amend the depth learning network and the proposed model, which will increase image recognition accuracy.

References

[1] Darrell T, Huang C, Jia Y(2012).Beyond spatial pyramids: Receptive field learning for pooled image features.In: *Computer Vision and Pattern Recognition,* p. 3370-3377. Rhode Island,USA: IEEE, June. 2012.

[2] Wan, L., Zeiler, M., Zhang, S., Cun, Y. L., Fergus, R. (2013). Regularization of neural networks using dropconnect. In: *Proceedings of the 30th International Conference on Machine Learning (ICML-13),* p. 1058-1066. Atlanta, USA: International Machine Learning Society, June. 2013.

[3] Foresee, F. D., Hagan, M. T.(1997). Gauss-Newton approximation to Bayesian learning. In: *IEEE International Conference on Neural Networks*, p.1930-1935.Houston,USA: IEEE, June.1997.

[4] Bishop, C. M. (1995). *Neural networks for pattern recognition.* Oxford, England:Oxford university press.400-450.

[5] Jin, Y., Okabe, T., Sendhoff, B.(2004). Neural network regularization and ensembling using multi-objective evolutionary algorithms. *In:* Congress on Evolutionary Computation(CEC2004), p. 1-8. Portland, USA: IEEE, June. 2004.

[6] Graham, B. (2014). Fractional max-pooling. arXiv preprint arXiv 1412.6071.

[7] Swietojanski, P., Li, J., Huang, J. T. (2014). Investi-

Investigation of maxout networks for speech recognition. *In: I*nternational Conference on Acoustics, Speech and Signal Processing (ICASSP), p.7649-7653. Florence, Italy: IEEE, May. 2014.

[8] Yu, Dingjun.Wang, Hanli.Chen, Peiqiu.Wei, Zhihua(2014).Mixed Pooling for Convolutional Neural Networks.In: *The 9th international conference on rough sets and knowledge technology*,pages 364-375.Shanghai,China:Springer International Publishing, October. 2014.

[9] Xiong, W., Du, B., Zhang, L., Hu, R., Tao, D. (2016). Regularizing Deep Convolutional Neural Networks with a Structured Decorrelation Constraint. *In:* The IEEE International Conference on Data Mining (ICDM), p. 519-528. Barcelona ,Spain:IEEE, December. 2016.

[10] Hinton G E, Srivastava N, Krizhevsky A(2012). Improving neural networks by preventing co-adaptation of feature detectors. *Computer Science*,3 (4) 212-223.

[11] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov(2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15 (1) 1929-1958.

[12] Zeiler, M., Fergus, R. (2013). Stochastic pooling for regularization of deep convolutional neural networks. *In:* Proceedings of the International Conference on Learning Representation (ICLR), p. 105-116. Arizona,USA: ICLR, May 2013.

[13] Iosifidis, Alexandros,Tefas, Anastasios,Pitas, Ioannis(2015). DropELM: Fast Neural Network Regularization with Dropout and DropConnect. *Neurocomputing,*162: 57-66.

[14] Zeiler M D, Taylor G W, Fergus R(2011). Adaptive Deconvolutional Networks for Mid and High Level Feature Learning.In: *International Conference on Computer Vision*,pages 2018-2025. Barcelona, Spain: IEEE, November. 2011.

[15] Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in cognitive sciences*, 11(10) 428-434.

[16] Goh, A., Lenglet, C., Thompson, P. M., Vidal, R. (2009). Estimating orientation distribution functions with probability density constraints and spatial regularity. *In: International Conference on Medical Image Computing and Computer-Assisted Intervention*,pages 877-885. London, England: Springer Berlin Heidelberg, September.2009.

[17] Traven, H. G. (1991). A neural network approach to statistical pattern calssification by 'semiparametric' of probability density functions. *IEEE Transactions on Neural Networks*, 2(3) 366-377.

[18] Luo X, Lu Z, Xu X(2014). Non-parametric kernel estimation for the ANOVA decomposition and sensitivity analysis. *Reliability Engineering & System Safety*, 130(3) 140-148.

[19] Rashki,Miri, Mahmoud,Moghaddam, Mehdi Azhdary(2012). A new efficient simulation method to approximate the probability of failure and most probable point. *Structural Safety*,39 22-29.

[20] Dehnad, K.(1986). Density Estimation for Statistics and Data Analysis. *Technometric*, 29 (4) 296-297.

[21] Lv Zhaoyan and Lv Zhenzhou(2014).Reliability Sensitivity Analysis Method Based on Weight Index of Density. *Acta Aeronautica et Astronautica Sinica*,35 (1) 179-186.

[22] Ciresan, D. C., Meier, U., Masci, J., Maria Gambardella, L., Schmidhuber, J (2011). Flexible, high performance convolutional neural networks for image classification. *In:* Proceedings-International Joint Conference on Artificial Intelligence(IJCAI), p. 1237-1242. Barcelona,Spain: IJCAI, July. 2011.

[23] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Computer Science*,12 (1) 107-115.

[24] MatConvnet toolbox.*http://www.vlfeat.org/matconvnet/*

[25] Nair, V., Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *In:* Proceedings of the 27th international conference on machine learning,p. 807-814. Haifa, Israel: International Machine Learning Society, June. 2010.

[26] CIFAR-10,CIFAR-100. *http://www.cs.toronto.edu/~kriz/cifar.html*

[27] MNIST DATASET. *http://yann.lecun.com/exdb/mnist/*