J Electr Eng Technol.2018; 13(5): 2099-2106
http://doi.org/10.5370/JEET.2018.13.5.2099

ISSN(Print) 1975-0102
ISSN(Online) 2093-7423

Discriminative Manifold Learning Network using Adversarial Examples for Image Classification

Yuan Zhang† and Biming Shi*

Abstract – This study presents a novel approach of discriminative feature vectors based on manifold learning using nonlinear dimension reduction (DR) technique to improve loss function, and combine with the Adversarial examples to regularize the object function for image classification. The traditional convolutional neural networks (CNN) with many new regularization approach has been successfully used for image classification tasks, and it achieved good results, hence it costs a lot of Calculated spacing and timing. Significantly, distrinct from traditional CNN, we discriminate the feature vectors for objects without empirically-tuned parameter, these Discriminative features intend to remain the lower-dimensional relationship corresponding high-dimension manifold after projecting the image feature vectors from high-dimension to lower-dimension, and we optimize the constrains of the preserving local features based on manifold, which narrow the mapped feature information from the same class and push different class away. Using Adversarial examples, improved loss function with additional regularization term intends to boost the Robustness and generalization of neural network. experimental results indicate that the approach based on discriminative feature of manifold learning is not only valid, but also more efficient in image classification tasks. Furthermore, the proposed approach achieves competitive classification performances for three benchmark datasets : MNIST, CIFAR-10, SVHN.

Keywords: Manifold learning, Discriminative feature, CNN, Adversarial examples, t-SNE, Dimensionality reduction.

1. Introduction

Manifold learning, as one method of machine learning and pattern recognition, has been widely used in dimension reduction [1]. Its main idea is to map high dimensional data into low dimension, ensuring the low dimensional data reflect the essential structural features of the original high dimensional data [2]. The premise of manifold learning exists an assumption, that is, some high dimensional data is actually a manifold structure of low dimension embedded in the high dimensional space. The purpose of manifold learning is to map high dimensional data back into low dimensional space, which reveals its essence. The learning thought of discriminating image feature based on manifold learning is assuming the feature built on low dimensional image is actually embedded in high dimensional image, so the feature that high dimensional image map is mapped back into low dimension remains a maximal approximation [3]. Given this, the learning of discriminating feature is to constrain the feature representation algorithm by featuring the local relations of these manifold vectors, and then to optimize the discriminating feature on basis of relations, thus formulating the optimization criterion [4].

DNN(Deep Neural Network) has successfully learned meaningful image presentation in a variety of tasks [2]. We can master more complex image feature relations, by adding more hidden layers and hidden neurons. However, increasing complexity of the model leads to more computational space. Therefore, many regularization techniques have been very common into the Neural Network for more solutions, such as Dropout [5], Drop Connect [6], Batch Normalization, Maxout [7], Stachostic-pooling [8] and Inception [9]. In the past years, DCNN, as one of the important network structure for image classification, has been adopted by many researchers, which could find more image presentation in the shallow level feature and the high level feature. It obtains more valuable image information to resolve the task of image classification, through the construction of convolution operation, showing an advanced achievement [10]. Whereas, CNN not only requires a number of learning parameters and learning space, but also abundant adjustable work to gain superior performance. For this reason, many researchers have joined the GPU operation.

The performance decreased when facing certain intentional or unintentional interference, although good results have been achieved by means of these methods. These interferences, rarely seen in the human world, created

† Corresponding Author: Safety Technology and Engineering Specialty, Anhui University of Science and Technology, Huainan, 232000, China; Dept. of Basic Courses, AnHui Medical College, Hefei, 230001, China. (yuanzhang@aust.edu.cn)
* Safety Technology and Engineering Specialty, Anhui University of Science and Technology, Huainan, 232000, China. (bmshi@aust.edu.cn)
Received: June 17, 2017; Accepted: May 23, 2018

lack of confidence of Neural Network, which was called Adversarial Examples [11]. Adversarial examples have become one effective solution for the security and robustness of DNN. Existing classification methods have reached a high level, but it shows a much higher error in the face of intentional or unintentional interference. At present, the feature extraction algorithm exhibits a good performance in the training data formed naturally, but among the features whose probability are not high, the classification illusion occurs. The CNN [12], popular in the current computer vision, using the convolution method as the approximative distance perception of Euclidean space, would be a little disappointing in case of the occurence of these classification errors. There exists a clear defect in the Euclidean approximative distance. If Euclidean approximative distance of the image is infinitesimal, the completely different classification results can be obtained in the network performance. Therefore, it is fair to say this conclusion will become the defect of DNN, especially DCNN, and such phenomenon also exists in the linear classifier. Hoping to solve this problem, [3,11,13] has begun to carry out experiments with Adversarial Examples. Although there is not a successful model yet, the most advanced accuracy remains in the original training set [2].

Based on the described previously, this study combines manifold learning techniques to discriminate feature learning, and constraints manifold feature of the image projection with a design to learn relation weight of the projection feature image, so as to optimize the discriminant feature. In this research, the training of countering samples are also included, and the regularization are added to train the feature weight of the neural network, in order to achieve the purpose of maximum optimizing efficiency of CNN training and improvement of the network structure robustness. In the following, the manifold learning method and the regularization term with adversarial examples, as well as the CNN structure are described in detail. Finally, the MINST handwritten digit, CIFAR-10 and SVHN data are tested, compared with visual effects for feature discrimination of common manifold learning techniques, and compared with the performance of the convolutional network using other regularization methods.

2. Motivation

2.1 Manifold learning-discriminative feature learning

The feature extraction is one critical step in the image classification. Great efforts have begun in the pre-training of the image data. Dimension reduction involves mapping high dimensional data to low dimension, and retains original image features as much as possible. Classical nonlinear dimension reduction algorithms of manifold learning, like principal component analysis (PCA), multidimensional scaling (MDS), or the self-organizing map (SOM), have been successfully applied in the social sciences and microbiology [14]. In recent years, advanced algorithms such as LLE, Isomap, Laplacian, Eigenmap, stochastic neighbor embedding (SNE), t-distributed SNE, have been born. All these methods are nonlinear dimension reduction, which makes the data originally dependent on the curvature or mixed cluster of complex shape visualize correctly, like the case in real life [10]. Therefore, it provides a new idea for the visualization method of given nonlinear data samples [14].

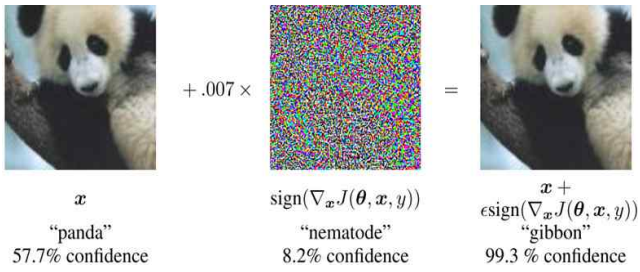

Fig. 1. A demonstration of fast adversarial example generation applied to GoogLeNet on ImageNet [11]

2.2 Adversarial examples

In our study, in order to ensure the relationship between adjacent convolution image in convolution layer and the convolution image of adversarial examples, the missing parameters are included[16]. As mentioned above, only the adjacent convolution image information of the training samples will cause improper embedding. These samples themselves have been sufficiently dense, so the adversarial examples that approximate the original samples are generated to prevent overfitting and improve the robustness of the training [17]. For instance, a misclassification resulted from interference sample mentioned in [11], creates a panda image that human vision can not distinguish, through the confidence level of minimum convolution space to calculate the space, but the networking learning is mistaken for a gibbon. As shown in Fig. 1[11]:

The interference terms of adversarial examples are obtained through updating samples instead of updating parameters. Set θ is the weight parameter of the model; x is the input of the model; y is the output mapping target that corresponds to x. In the maching learning task, $J(\theta; x, y)$ is used to train the neural network. We focus on the linear cost function of weight θ, to get an interference term of optimal and maximum norm constraint. The form is as follows [3]:

$$\eta = \varepsilon sign(\nabla_x J(\theta; x, y)) \tag{1}$$

This formula is referred to as "Fast Gradient Symbolic Method" that generates adversarial examples, and this gradient can be calculated fast and efficient with back-propagation algorithm. Thus, the sample update can be achieved through the following formula.

$$x' := x + \eta \nabla_x J(\theta; x, y) \tag{2}$$

Actually, η is the learning rate.

3. Relative Work

t-SNE is a nonlinear and unsupervised manifold learning technique. Assuming that We have a high dimensional data set $X=\{x_1, x_2, x_3 \ldots x_n\}$; *t-SNE* tries to find low dimensional performance $Y=\{y_1, y_2, y_3 \ldots y_n\}$, where $y_n \in R^d$; $x_n \in R^D$; $d<D$, y_n can show the feature of high dimensional data, $p_{m|n}$ represents x_m is the probability of x_n neighbors, expressed by the following formula [15]:

$$p_{m|n} = \frac{\exp(-\| x_n - x_m \|^2)}{\sum_{t \neq n} \exp(-\| x_n - x_t \|^2)} \tag{3}$$

t represents t elements neighboring x_n. Considering the distance between the outlier x_n and all other nodes is overlarge, whether where y_n, the mapping point of the outlier in the low dimensional space, is, the penalty value is not too high, so the more simple and intuitive way are adopted to define:

$$p_{n|m} = \frac{p_{n|m} + p_{m|n}}{2N} \tag{4}$$

N is the amount of data points, and such definition not only meets the symmetry, but also ensures the x_i penalty value not too small. The similarity and distance relation of high dimensional space should be also reflected in the low dimensional space. The mapping data y_i in the low dimensional space should satisfy the probability $q_{n|m}=p_{n|m}$. With respect to *t-SNE* method, its low dimensional representation probability can be written as:

$$q_{nm} = \frac{(1+\| y_n - y_m \|)^k}{\sum_n \sum_{t,t \neq n} (1+\| y_n - y_t \|)^k} \tag{5}$$

$q_{nn}=0$, where k is the degree of freedom for t-distribution. The smaller the k, the longer the tail of the distribution

In order to monitor the learning, supposing that there is a high dimensional data set X and its corresponding label $T=\{t_1, t_2 \ldots t_j\}$, and j represents j_{th} categories. We make x_n and x_m from the same label, then the mapping probability should be equal and equal to one. We define $p_{nn}=0$ and $p_{nm}=1$, *as* $t_n=t_m$, while $p_{nm}=0$, as $t_n \neq t_m$ Hence, the pre-treatment probability of high dimensional space is seen as the the demand probability performance of low dimensional space through given samples.

We redefine the joint probability formula in low dimensional space:

$$q_{nm} = \exp(-\| y_n - y_m \|^k) \tag{6}$$

Assuming low dimensional data point Y; figure $G=\{Y, \Omega\}$, the manifold used to represent the space relationship of low dimensional mapping data; Ω is the weight matrix including the boundary connected node, which is believed to be closely related to the weight matrix. The weight ω_{nm} is the proximity connecting y_n, y_m. The weight controls various features of the image, including structure, connectivity and tightness, which is also the eigenvalue weight of the original image after reduction. The image based on the relation is often characterized with the use of Euclidean distance on the basis of Gauss Kernel. It is believed that the formula is:

$$\omega_{nm} = \begin{cases} exp(\frac{-\| y_n - y_m \|^k}{\rho}) & ; \quad e(y_n, y_m) = 1 \\ 0 & ; \quad otherwise \end{cases} \tag{7}$$

Where ρ is the nuclear scale parameter, $e(y_n, y_m)$ represents whether it is the nearest neighbor data point. For a map of $G=\{y, \Omega\}$, its dispersion can be expressed as:

$$F(Y) = \sum_{n,m} \| y_n - y_m \|^k \omega_{nm} \tag{8}$$

This shows the tightness between the nodes. At the same time, minimization denotes the weight of close relationship, which can constraint the mapping output y_n and reserve the maximum manifold features based on the local relations of image input. But this method does not better take into account the issue of the feature vector of projection [16]. In this study, the issue of discriminant features class based on manifold learning projection is to be discussed in the following. Besides, a new non parametric dimension reduction, like *t-SNE* in the *t*-distribution, can obtain an effective and flexible visualized dimension reduction of high dimensional data. However, a drawback of non-parametric is a weakly generalization ability to extended data outside of samples. For this reason, in order to enhance the *t-SNE* generalization ability of manifold learning, adversarial examples are increased into the original samples. And the features of interference contained are studied, to produce the new description of anti-interference. It mantains the powerful performance of *t-SNE*, while avoiding the issue of inaccurate classification for extended data.

4. Networks Architecture

In this study, we use a traditional CNN, in which a convolutional layer learns the feature weight from the results of one pooling-layer and a full-connection layer. In addition, the weight of the feature vectors for the filters have been studied through the discriminant feature approach of manifold learning, and the training weight of adversarial examples are also included. Since the pooling

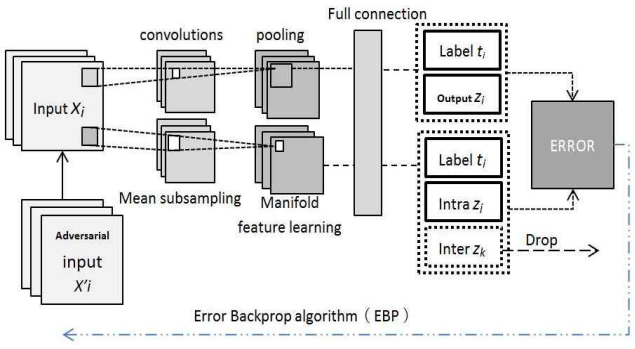

Fig 2. Illustration of the structure of improved CNN. Two groups of sample x_i is training for the feature weight with the addition of adversarial samples x'_i. One group is the traditional CNN, which is classifyed through convolution-layer and pooling-layer and the full-connection layer; the other group is the weight learned through the relation of manifold discriminant feature vectors, which retains the learned ω^{intra}, namely, in this class maintaining the local feature $z_{j,}$ meanwhile, casting away the feature vector z_k that the learned ω^{inter} from corresponded map, and categorizing the trained feature vector $z_{i,}$ z_j. The error is calculated through corresponding sample label, and the weights are derivated by *EBP* algorithm (Error Back Propagation algorithm)

layer of convolution network lacks the ability of anti-interference [18], and a number of weight parameters need learning, the dimension reduction techniques of manifold learning and the way of adding adversarial examples are used to pretreat the processing image, with regularization of the network, reduction of the overfitting, increase of the network robustness. The structure of CNN is in Fig. 2:

Given $\{X\}$ *is* N training image recognition samples of size $m \times n$, and each example is annotated the label T=1,2,3…k. K is the number of categories. First, each trained image is separated into $l \times l$ blocks, which is then generate a data matrix vector $P=\{p_{i,1}, p_{i,2}, p_{i,3} \cdots p_{i,j}\}$, $p_{i,j}$ is the j_{th} vector of the i_{th} image. For each image forming each average normalization block, a whole data matrix of normalization is conducted as $p'=\{p'_{i,1}, p'_{i,2} \cdots p'_{i,mn}\}$

Hence, we have a dimension of the number of $N \times m \times n$ and the size of $l \times l$, and we prefer building a mapping matrix Y of low dimensional space, for retaining the feature of high dimensional data space. The depth learning both has the ability to learn complex nonlinear relations in the feature vector and to discriminate feature in the image classification. We hope that the depth learning network can discriminate the feature vector to realize the image classification by learning the relation-weight of mapping in the low dimensional space matrix Y. Accordingly, the mapping low dimensional matrix reached the goal of dimensional reduction, while achieving the purpose of discriminating the feature vectors in the process of learning the weight of mapping vectors.

The vector block P_{ij} corresponding to label t_i, is seperated into inter-class and intra-class. k_1 is the number of intra-class vector block; k_2 is the number of inter-class vector block. P_{ij} can be divided into two classes of examples, intra-class $p_{i_1}, p_{i_2}, p_{i_3} \cdots p_{i_{k_1}}$ and inter-class $p_{i^1}, p_{i^2}, p_{i^3} \cdots p_{i^{k_2}}$. The two classes of samples form a new low dimensional performance, creating a new feature matrix:

$$z = \{z_i, \overbrace{z_{i_1}, z_{i_2} \ldots z_{i_{k_1}}}^{k_1}, \overbrace{z_{i^1}, z_{i^2} \ldots z_{i^{k_2}}}^{k_2}\}$$

The Euclidean distance between Z_i and K_1 samples is indicated as:

$$A_1(z_i) = \sum_1^{k_1} \| z_i - z_{i_j} \|^k \tag{9}$$

It shows that the Euclidean distance between Z_i and K_2 *is*:

$$A_2(z_i) = \sum_1^{k_2} \| z_i - z_{i^k} \|^k \tag{10}$$

The weights of the feature connection with the relationship between distance are classified into two classes: ω^{intra} *and* ω^{inter} denotes the connection weights of intra-class and inter-class, respectively, so k_1 and k_2 *is* the distance of the mapping connection image:

$$\omega^{intra} = \begin{cases} exp\dfrac{(-\| p_i - p_j \|^k)}{\rho}; t(p_i) = t(p_j), e(p_i, p_j) = 1 \\ 0 \qquad\qquad ; otherwise \end{cases} \tag{11}$$

and

$$\omega^{inter} = \begin{cases} exp\dfrac{(-\| p_i - p_j \|^k)}{\rho}; \ t(p_i) \neq t(p_j), e(p_i, p_j) = 1 \\ 0 \qquad\qquad ; \ otherwise \end{cases} \tag{12}$$

$F_G(Z)$ is combined with formulas (9), (10), in which retains local features and divided class with relation graph, it can be expressed as:

$$F_G(Z) = A_1(z_i)\omega^{intra} + A_2(z_i)\omega^{inter} \tag{13}$$

Where the first term displays the information of intra-class with only local relationship among feature vectors of the same class based on discriminant manifold, and the second term, which is also called penalty term, punishes relationship among feature vectors of different classes with the process shown in the Fig. 3. This approach highlights good robustness of manifold learning technology. On basis of this, we apply this approach into CNN.

CNN is a feed-forward neural network, which can

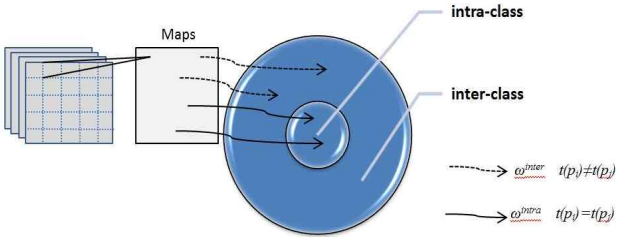

Fig. 3. Illustration of learning discriminative feature from neighbor patchs.when p_i and p_j belongs to the same label,the feature patchs distribute to intra-class, otherwise, distribute to inter-classe

respond to surrounding units with multiple hidden layers. The map of image sample x_i is inputted into hidden layers, then, an output z_i is obtained by correspondent label. The network optimizes weight ω to minimize the loss function through trainning. Therefore, the loss function of weight is expressed as:

$$\Gamma(W)=\frac{1}{N}\sum_{i=1}^{N}\Phi(t_i,f(x_i)) \tag{14}$$

where $f(x_i)$ is the activation function of output feature.

This study adds manifold discriminant feature learning and learns the close weight of potential embedding feature vector by CNN, the regularization term is added in object recognition function on basis of discriminant feature learning, and the object function can be redefined by combining with formula (13) as:

$$\Gamma(W;Z)=\frac{1}{N}\sum_{i=1}^{N}\left\{\Phi(t_i,f(x_i))+\frac{1}{k}\lambda\sum_{1}^{2k}F_G(Z)\right\} \tag{15}$$

Where N stands for all sample images, k is the nearest neighbor number, and λ is used to balance regularization parameter. The object recognition function described in Formula (15) is the neural network output based on manifold constraints, and $2k$ is output of all mapped feature vectors, where k is the number of relationsin intra-class feature, and k is the number of relations in inter-class feature, also known as penalty term forcing the constraints of recognition features.

The above-mentioned neutral network will encounter intentional interference or blind spots in images during training, therefore, training on counter samples is added, and with a regularization term added in image classifacation function, the error induced by intentional interference and regularization can be reduced. A new object loss function can be established by combining with formula (15) as:

$$\Gamma(W;Z)=\frac{1}{N}\sum_{i=1}^{N}\left\{\Phi(t_i,f(x_i))+\frac{1}{k}\lambda\sum_{1}^{2k}F_G(Z)\right\}+\gamma\mathrm{H}(\chi,\chi') \tag{16}$$

$$\mathrm{H}(x,x')=\frac{1}{n}\sum_{n}h(x_n,x_n') \tag{17}$$

Where x' is generated from formula (8). Object recognition formula (17) learns the training weight of input image x_i through the first term, constraints the relation graph weight of manifold discriminant feature learning through the second term, and increases the regularization weight of adversarial samples training through the third term. As a result, the robustness and generalization ability of the function increases.

In general, the weights of CNN is updated by multiple iterations using SGD(Stochastic Gradient Descent) algorithm on the training sets, and EBP algorithm is used to improve the convergence, so the formula is expressed as follows:

$$\nabla_w\Gamma(W;Z)=\sum_{i}^{N}\frac{\partial\Gamma(W;Z)}{\partial z_i}\frac{\partial z_i}{\partial W} \tag{18}$$

Where $\nabla_{\mathrm{w}}\Gamma(\mathrm{W};\mathrm{Z})$ is the gradient of object recognition function and the expected weight matrix W.

5. Experiement

In this study, we will use the proposed approach to test in MNIST, CIFAR-10, SVHN Three benchmark dateset. First, we will also compare our approach with the traditional nonlinear manifold learning theory, *ISOMAP*, *LLE*, *t-SNE* and contrast the regular training with the currently popular architecture of CNN, CNN+*dropout*, CNN+*Maxout*, CNN+ *Stochastic-pooling*. Second, we will compare different dataset with proposed approach. Our labs are equipped with Tensor Flow [19] deep learning framework, 4 GPUS GTX1080ti and 64G memory. OS is Ubuntu 16.04.We use the encapsulated network architecture from TensorFlow. Finally, We use convolution kernel size of 10 ×10, strike is 2 and 1000 iterations with each dateset.

5.1 MNIST

The data set is the 10 categories of fully handwritten digital data sets, which is made up by 10 numbers from 0 to 9 and contains 60000 handwriting image training data of 28×28 and 1000 testing data. The data format of *MATLAB* has been normalized to [0,1], network structure as shown in Fig. 2. We take the method referred in this study to conduct the Dimension Reduction (DR) operation to the data. In order to visualized the dimension reduction, we map the dimension of principle sample to the 2-dimension, topological structure to D-500-500-2000-d, of which $d = 2$. In order to contrast the results on our experiments of dimension reduction, we conduct the comparing in the method of nonlinear dimensionality reduction *ISOMAP*, *T-SNE* with different k and the method of the study. The

Table 1. Accuracy of different regularization methods on MNIST

Approach	Accuracy
CNN+Dropout	89.097 ± 0.219
CNN+Maxout	90.613 ± 0.129
CNN+Stochastic-pooling	91.670 ± 0.213
OURES	91.910 ± 0.215

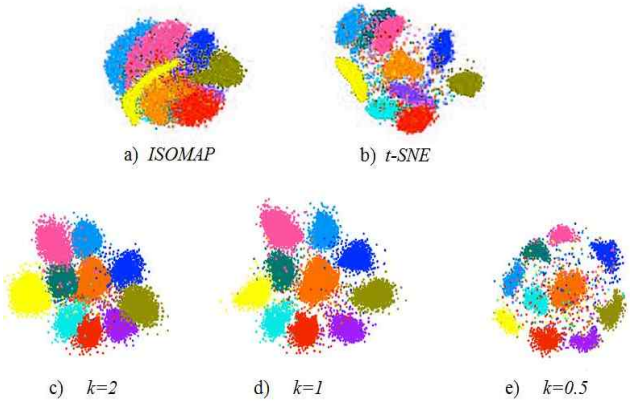

Fig. 4. 2-Dimension test results for MNIST samples, a) *ISOMAP*, (b) *t-SNE*, (c) ours *k=2*, (d) ours *k=1*, (e) ours *k=0.5*.This graph indicates The effect of our DR is significantly better than that of *t-SNE* or *Isomap*, and the visualization effect of DR is getting better with smaller *k*

following figure 4 shows the contrast result.

With the network topology puts forward in this study conducting training relation weight, the learning rate is 0.01. We add the original sample into the network training to contrast the popular regularization method *dropout*, *maxout*, *stoachastic-pooling*. And we get the accuracy of the test data using our approach as the following Table 1.

5.2 CIFAR-10

The dataset contains 60000 color images of size 32×32, 10 categories and 60000 samples for each category. 50000 of samples are used for training, which consist of 5 batches, 10000 samples for each batch, and 10000 for testing. We use 4 regularization methods with the network structure described earlier, including *dropout, maxout, stochastic-pooling* and the method based on manifold learning in this study, to compare the accuracy rate feature learning. The result is shown in Fig. 5. The accuracy of our proposed regularization method based on manifold learning on the different feature kernel increases steadily with the increase of the characteristic kernel, and the accuracy of *dropout* performance is lowest on the same characteristic kernel, and the accuracy of *maxout* is lower when the number of characteristic kernel is less, and the accuracy of *stochastic-pooling* performance is steady. Our method has relatively stable accuracy, which is the best in 35 characteristic kernels.

Table 2. Time cost in with distinct regularization methods in one category of CIFAR-10

Methods	Time cost(s)
CNN+*Maxout*	830
CNN+*Dropout*	927
Ours	**600**
CNN+*Stochastic-Pooling*	650

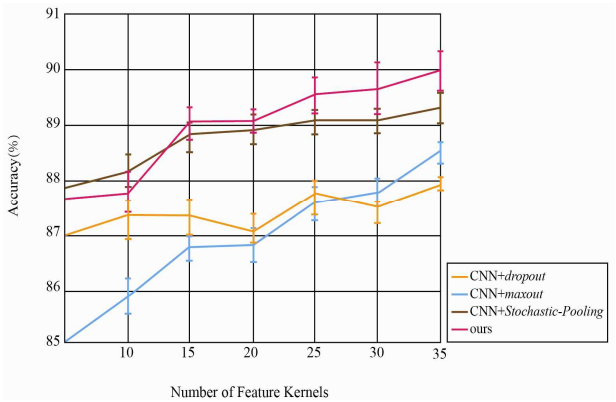

Fig 5. Classification accuracy on CIFAR-10 of different regularation methods with varying the number of feature filters. Where 10 feature kernels are trained, the accuracy of *maxout* is about 85.6%-86.2%, and the accuracy of *dropout* is in 86.9%-87.5%. Our proposed method is at 87.5%-88.2%, and the accuracy of *stochastic-pooling* is 87.9%-88.5%. With the increase in the number of feature kernels, the other four methods are steadily increasing except *dropout*. Where 35 feature kernels are trained, the accuracy of our method is about 89.5%-90.3%, and the accuracy of *stochastic-pooling*, *maxout* and *dropout* are 89.1%-89.6%, 88.3%-88.7%, 87.8%-88.1%, respectively

For comparing the time consumption of the four methods, we randomly select 30 images of one category from CIFAR-10 to calculate the training time, which is shown in Table 2 as follows. It can be seen that the training time based on manifold learning is the shortest. Although the number of weights of discriminative feature learning is more in the process of reducing dimension in this study, the discriminative feature learning is dependent on weight relations, and the fine-tune will not produce more weight calculation time, so the time is reduced.

5.3 SVHN

SVHN is a real-image dataset, which is used to develop machine learning and object recognition algorithms, and requires minimum data preprocessing and formatting. It has 10 categories, 73257 for training, 26032 for testing, and 531131 extra non- annotated images. We use our proposed discriminative feature learning with adversarial examples on SVHN and contrasting other manifold

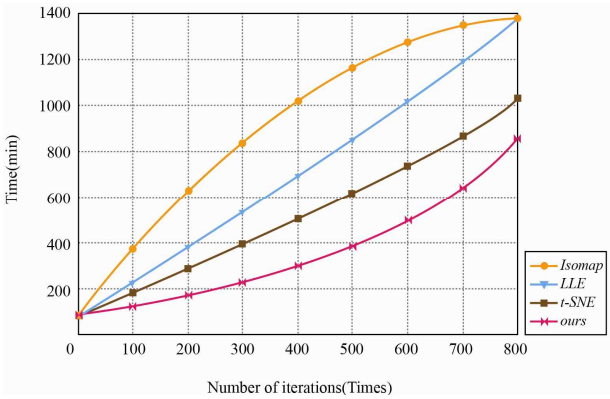

Fig. 6. Illustration of the time cost by varying iterations with different methods of dimension reduction . The time cost of the Isomap algorithm from 100-800 iterations was about 100 to 1380 minutes, and the LLE algorithm took about 100 to 1380 minutes, while the time spent in the process of iterations are reduced, and the time cost of the t-SNE algorithm in 800 iteration took only about 1080 minutes, but our proposed manifold learning based on the nonlinear dimensionality reduction only took about 820 minutes, which perform the most efficient

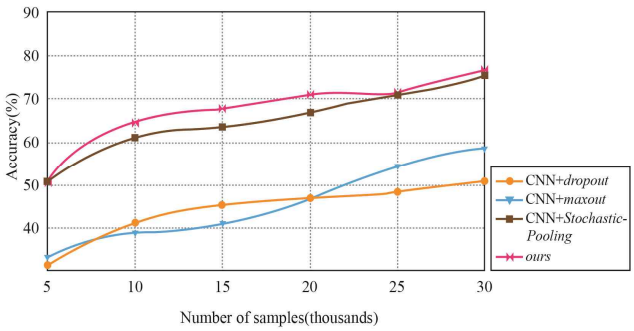

Fig. 7. Illustration of accuracy in the number of image samples with different regularization methods. While 5,000 sample images are trained, the accuracy of CNN with dropout and maxout regularization method was only more than 30%, and the accuracy of our method and stochastic-pooling method was almost same. However, when 30,000 samples was trained, the accuracy of our proposed method had exceeded 75%, which is obviously better than the other three regularization methods

learning method, *IOSMAP*, *LLE*, *t-sne*. In the 1000 iterations, we recorded training times based on four manifold learning methods, by contrast with the other three methods, we found that we spent much less time. The time cost of training is shown in the Fig. 6.

Moreover, we used four varying regularization methods for training in SVHN, and Fig. 7 showed our proposed method had the highest accuracy of classification prediction at different times of iteration.

6. Conclusion

In the study, we address a method on image classification based on the discriminant feature learning of manifolds and conducting the experiment with the structure of CNN. We add adversarial examples with anthropogenic disturbances for increasing the number of samples to improve the accuracy. By improved technology of manifold learning, we conduct the supervised discriminant feature training. We learn the weight relation of feature relation graph to classify the mapped samples as intra-class and inter-class samples and then drop the inter-class samples, conduct EBP derivation on the results of the image classifying and calculate the error combining with the traditional CNN training and the regularization term to constrain classification. We evaluate this approach by contrasting the time and accuracy rate with the advanced regularization method like dropout, maxout, stochastic-pooling and traditional nonlinear manifold learning method like Isomap, LLE, t-SNE. In the process of learning the discrimination feature, the speed increased obviously and the accuracy rate improved a lot. Especially, when adding the interrupt training of adversarial examples, the approach's generalization ability is strengthened and a new regularization method is constructed with strong robustness. As a result, we find that the adversarial examples training and manifold based discriminant feature learning relation weights have achieved state-of-the-art accuracy.

Acknowledgements

This work was supported by funds from Major Project of Natural Science Research of Department of Education Anhui Province of China under Grant KJ2018ZD062 and Key Project of Natural Science Research of Department of Education Anhui Province of China under Grant KJ2017A466 and Proje-ct of support program for outstanding young people of Department of Education Anhui Province of China under Grant GXYQ2018170.

References

[1] Tomar, V. S. and Rose, R. C., "Manifold Regularized Deep Neural Networks," *INTER-SPEECH.,* 2014.

[2] Feng, Z., Jin, L., Tao, D. and Huang, S., "Dlanet: a manifold-learning-based discriminative feature learning network for scene classification,". *Neuro-computing*, *157*, pp. 11-21, 2015.

[3] Lee, T., Choi, M. and Yoon, S., "Manifold regularized deep neural networks using adversarial examples," *Computer Science*, 2015.

[4] Masci, J., Boscaini, D., Bronstein, M. M. and Vandergheynst, P., "Shapenet: convolutional neural

networks on non-euclidean manifolds," *Epfl*, pp. 832-840, 2015.
[5] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. R., "Improving neural networks by preventing co-adaptation of feature detectors," *Computer Science,* vol. 3, no. 4, pp. 212-223, 2012.
[6] Wan, L., Zeiler, M., Zhang, S., Cun, Y. L. and Fergus, R., "Regularization of neural networks using dropconnect," *International Conference on Machine Learning*, pp. 1058-1066, 2013.
[7] Goodfellow, I. J., Wardefarley, D., Mirza, M., Courville, A. and Bengio, Y., "Maxout net-works," *Computer Science*, pp. 1319-1327, 2013.
[8] Zeiler, M. D. and Fergus, R., "Stochastic pooling for regularization of deep convolutional neural networks," *Eprint Arxiv,* 2013.
[9] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. and Anguelov, D., et al., "Going deeper with convolutions," P. *Computer Vision and Pattern Recognition*. IEEE, pp. 1-9, 2015.
[10] Yang, W., Sun, C. and Lei, Z., "A multi-manifold discriminant analysis method for image feature extraction," *Pattern Recognition*, vol. 44, no. 8, pp. 1649-1657, 2011.
[11] Goodfellow, I. J., Shlens, J. and Szegedy, C., "Explaining and harnessing adversarial examples," *Computer Science,* 2014.
[12] Lécun, Y., Bottou, L., Bengio, Y. and Haffner, P., "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998.
[13] Tabacof, P. and Valle, E., "Exploring the space of adversarial images," *Computer Science,* 2015.
[14] Gisbrecht, A., Schulz, A. and Hammer, B., "Parametric nonlinear dimensionality reduction using kernel t-sne," *Neurocomputing*, vol. 147, no. 1, pp. 71-82, 2015.
[15] Hinton, G. E., "Visualizing high-dimensional data using t-sne," *Vigiliae Christianae*, vol. 9, no. 2, pp. 2579-2605, 2008.
[16] Han, Y., Xu, Z., Ma, Z. and Huang, Z., "Image classification with manifold learning for out-of-sample data," *Signal Processing*, vol. 93, no. 8, pp. 2169-2177, 2013.
[17] Vural E, Guillemot C., "Out-of-Sample Generalizations for Supervised Manifold Learning for Classification," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 25, no. 3, p. 1410, 2016.
[18] Papernot N, Mcdaniel P, Goodfellow I, et al. "Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples", 2016.
[19] TensorFlow. http://www.tensorflow.org/

Yuan Zhang was born in HeFei, China. He received his B.S. degree from Anhui Polytechnic University, Wuhu, China, in 2005. He received his M.S. degree from Hefei University of Technology and Ph.D. Candidates from Anhui University of Science and Technology, Huainan, China, in 2010 and 2016, respectively. Since 2005, he has been engaged in Anhui Medical College. He is presently an Associate Professor in Department of Basic Courses at AnHui Medical College. His current research interests include AI, machine learning, image recognition, image processing.

Biming Shi was born in Taihu, China. He received his B.S. degree from Anhui University of Science and Technology, Huainan, China, in 1987. He received his M.S. degree from Anhui University of Science and Technology and Ph.D. from China University of Mining and Technology, Xuzhou, China, in 1995 and 2004, respectively. Since 2004, he has been engaged in Anhui University of Science and Technology. He is presently an Professor and doctor advisor in School of energy and safety at Anhui University of Science and Technology. His current research interests include Theory and technology of mine ventilation, prevention and control of mine gas.