Noname manuscript No.
(will be inserted by the editor)

Deep Regularized Discriminative Network

Nazneen N. Sultana · Bappaditya Mandal* · N. B. Puhan.

Received: date / Accepted: date

Abstract Traditional linear discriminant analysis (LDA) approach discards the eigenvalues which are very small or equivalent to zero, but quite often eigenvectors corresponding to zero eigenvalues are the important dimensions for discriminant analysis. We propose an objective function which would utilize both the principal as well as nullspace eigenvalues and simultaneously inherit the class separability information onto its latent space representation. The idea is to build a convolutional neural network (CNN) and perform the regularized discriminant analysis on top of this and train it in an end-to-end fashion. The backpropagation is performed with a suitable optimizer to update the parameters so that the whole CNN approach minimizes the within class variance and maximizes the total class variance information suitable for both multi-class and binary class classification problems. Experimental results on four databases for multiple computer vision classification tasks show the efficacy of our proposed approach as compared to other popular methods.

Keywords Convolutional neural network · latent space representation · regularization; subspace learning.

Nazneen N. Sultana
School of Electrical Sciences
Indian Institute of Technology, Bhubaneswar, India
E-mail: nns11@iitbbs.ac.in

*Bappaditya Mandal (corresponding author)
School of Computing and Mathematics
Keele University, United Kingdom
E-mail: b.mandal@keele.ac.uk

N. B. Puhan
School of Electrical Sciences
Indian Institute of Technology, Bhubaneswar, India
E-mail: nbpuhan@iitbbs.ac.in

1 Introduction

Linear discriminant analysis (LDA) is a method from multivariate statistics which attempts to find a linear projection of high-dimensional observations onto a lower-dimensional space [10]. It finds the optimal decision boundaries in the resulting lower dimensional subspace. LDA is an efficient way to separate the features on the basis of class information, but since it requires inverse operation it often becomes problematic if the dimension becomes very high as compared to the number of available training samples. Thereby it ignores the eigenvectors corresponding to zero eigenvalues so as to have the within class scatter matrix non-singular. In Sharma *et al.* [29] an improved regularized LDA is proposed which is carried out by adding a perturbation term α to the diagonal elements of within class matrix to make it non-singular and invertible. However, the eigenvectors corresponding to zero eigenvalues also contain the important class discriminatory information as reported in [27,17, 19,6]. Thus, we aim to utilize both the principal as well as nullspace eigenvalues and extend the beneficial properties of the proposed regularized fisher method (low intra-class variability, high total-class variability, optimal decision boundaries). This is done by reformulating its objective to learn linearly separable representations based on a deep neural network (DNN) for both binary as well as multi-class problem.

LDA is used widely as a supervised dimensionality reduction method in computer vision and pattern recognition. Its recent generalization to non-Euclidean Grassmann manifolds can be found in [33]. This aims to impose the highest possible variance among classes, by maximizing the between-class distances, whilst minimizing the within-class scattering. Recently, deep learning combined with various multivariate statistics methods have achieved great success [12]. Andrew *et al.* [4] introduced a deep canonical correlation analysis (DCCA) which can be viewed as a non-linear extension of CCA . In their evaluations, they argued that DCCA learns representations with significantly higher correlation than those learned by CCA and Kernel (non-linear) CCA. They experimented using the MNIST handwritten data and simultaneous recording of articulatory and acoustic data. Ghassabeh [13]*et al.* presents new adaptive algorithms for online feature extraction using principal component analysis (PCA) and LDA for classification purpose. In Al-Waisy *et al.* [2], they have merged the advantages of local handcrafted feature descriptors with the Deep Belief Networks for the face recognition problem in unconstrained conditions and have obtained better performances.

PCANet proposed by Chan *et al.* [5] which includes cascading of PCA, binary hashing and block histogram computations. This can be seen as an unsupervised convolutional deep learning approach. Due to computational complexity these multi-stage filter banks are limited to two stages but can be extended to any number. They also experimented further modifications on PCANet as RandNet and LDANet. RandNet and LDANet share the same methodology like PCANet, but their cascaded filters are either selected randomly as in RandNet or learned from LDA in case of LDANet. Lifkooee *et*

al.[24] combines regular deep convolutional neural network with the Laplacian of Gaussian filter (LoG) right before fully connected layer and they have shown that the proposed feature descriptor along with LoG introduced in CNN further improves the performance of deep learning.

Stuhlsatz *et al.* [31] initially proposed the idea of combining LDA with neural networks. In their proposed approach they pre-train a stack of restricted Boltzmann machines and this pre-trained model is finetuned with respect to a linear discriminant criterion. LDA has the disadvantage that it overemphasizes large distances at the cost of confusing neighbouring classes. Thus, to tackle this problem they introduced a heuristic weighing scheme for computing the within-class scatter matrix required for LDA optimization. The LDA based objective function proposed by Dorfer *et al.* [9] is a non-linear extension of classic LDA where the objective function is obtained from the general LDA eigenvalue problem while still allowing to train the CNN architecture with stochastic gradient descent and back-propagation.

In this paper, we propose to modify the LDA based objective function which would utilize both the principal as well as nullspace eigenvalues onto its latent space representation for both multi-class as well as binary class problem. Extensive experimental results on multiple computer vision classification tasks illustrates the superiority of our proposed approach as compared to other popular methods. Below we describe our proposed method in details.

2 Proposed Approach

The approaches mentioned so far are based on the study of multi-variate statistics. In our work, we propose to train a CNN architecture in an end-to-end fashion with a new objective function which would enable the network to inherit the property of maximizing the total variation and minimizing the within class variation.

Deep Learning has become state-of-the-art for many image based applications of classification, object recognition, segmentation, image captioning and natural language processing [26,14]. The mathematical model of Convolutional Neural Network (CNN) is explained by Kuo *et al.*[22] where the fundamental questions about the structure of the convolutional neural networks is explained. There are many variations of deep convolutional neural networks for various vision tasks. The intuition behind our approach is to use the proposed regularized Fisher method as the objective function on top of a powerful feature learning model. The optimization of parameters is carried by back-propagating the error of the proposed objective function through the entire network. One of our objectives in this work is to come up with a CNN architecture that can be generically applied to many computer vision classification tasks. For experimental evaluation, we evaluated our proposed objective function on various benchmark databases like MNIST (handwritten digit recognition), CIFAR-10 (natural image classification) and ISBI (skin cancer detection into melanoma and non-melnoma cases) to show that the objective

function is effective for both multi-class as well as binary class classification problems.

2.1 Deep Regularized Discriminative Network over simple ConvNet

Deep learning networks are different from the simple single-hidden-layer neural networks by their depth. Deep-learning networks effectively learn the features automatically without human intervention, unlike most traditional machine-learning algorithms. A neural network with P hidden layers is represented as a non-linear function $f(\Theta)$, where $\Theta = \{\Theta_1, ..., \Theta_P\}$. In supervised learning for N number of samples, we have $x = \{x_1, ...x_N\}$ as training data and $y = \{y_1, ...y_N\} \in 1, ..., C$, where C is the number of classes. In the last layer, we have softmax as the classifier which gives the normalized probability of the data that belongs to a particular class. The output, $o_i = \{o_{i1}, ...o_{iC}\}$ is a function of $f(x_i, \Theta)$. The network is optimized using stochastic gradient descent or any other optimizer like Adam with the goal of finding optimal model parameters Θ by minimizing the objective function $l_i(\Theta)$.

$$\Theta = \underset{\Theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i}^{N} l_i(\Theta) \tag{1}$$

where $l_i(\Theta) = f((x_i, \Theta), y_i)$. For categorical cross entropy (CCE), the loss function is defined as

$$l_i(\Theta) = -\sum_{i}^{C} y_{i,j} \log(p_{i,j}) \tag{2}$$

where $p_{i,j}$ is the network output probability and $y_{i,j}$ is 1 if observation x_i belongs to class y_i for $(j = y_i)$ and 0 otherwise. Figure 1 shows the deep regularized network where the objective is different from the CCE in maximizing the total scatter matrix eigenvalues and minimizing the within class scatter matrix eigenvalues. In the following subsections, detail description of the proposed objective function and the related analysis are discussed.

2.2 Proposed objective function

Linear discriminant analysis tries to find out the axes which maximize the between-class scatter matrix S_b, while minimizing the within-class scatter matrix S_w in the projective subspace $A \in \mathbb{R}^{l \times d}$. The projective subspace is a lower dimensional subspace, i.e., $l = C - 1$ where C is the number of classes. The resulting projection matrix onto this subspace $x_i A^T$ are maximally separated in this space [10]. Fisher criterion is defined as the ratio of between-class and within-class variances, given by:

$$J(W) = \frac{|W^T S_b W|}{|W^T S_w W|}. \tag{3}$$

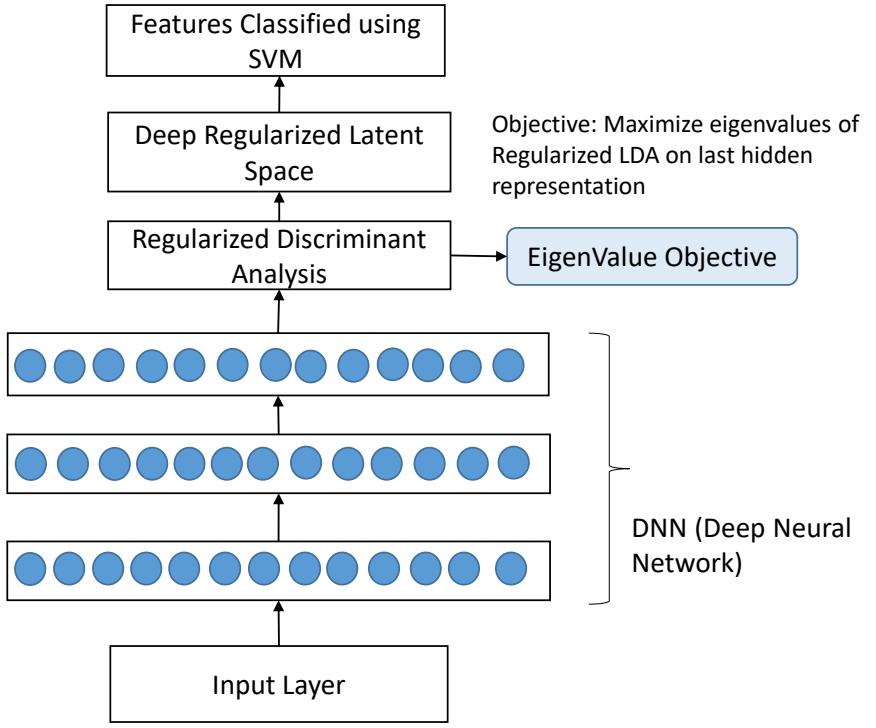


Fig. 1: Schematic sketch of deep regularized discriminative network which learns the linear separability property in the latent representation. Here the objective is to maximize the eigenvalues so that the class separability also increases.

Here W is the weight vector. To compute the within class scatter matrix,

$$S_c = \frac{1}{(N_c - 1)} \sum \bar{X_c}^T \bar{X_c} \tag{4}$$

$$S_w = \frac{1}{(C)} \sum S_c \tag{5}$$

The total scatter matrix is computed using,

$$S_t = \frac{1}{N - 1} \bar{X}^T \bar{X} \tag{6}$$

where X is the input data matrix; in our case it would be the output of the CNN model and N_c is the sample numbers in that particular class. N is the total samples and $\bar{X_c} = X_c - m_c$, m_c is the mean of that class, $\bar{X} = X - m$ where m is the total mean of the samples. The output predicted values from

the CNN model (`y_pred`) is used as X values for the computation of S_w, as in (5). To extract discriminative features, at first we perform eigen decomposition of the within-class scatter matrix S_w, given by:

$$S_w = \Phi \Lambda \Phi^T. \tag{7}$$

Here, Φ contains the eigenvectors and Λ are the eigenvalues of S_w. Then the eigenvectors are sorted according to the eigenvalues in descending order. Matrix Φ is then split into W_1 and W_2, where W_1 is the matrix which contains the eigenvectors corresponding to those eigenvalues which are greater than a certain minimum variance. For our experimentation, we took minimum variance value as $1e-2$. W_2 matrix are the eigenvectors corresponding to those eigenvalues whose variance are less than the minimum variance. W_1 matrix is divided with the square root of the corresponding eigenvalues and W_2 matrix is divided with the square root of the minimum eigenvalues. These two matrices are concatenated to form Ψ as shown in (8) and it is multiplied with the `y_pred` to form the model output y.

$$\Psi = [\Phi_i(\Lambda_i)^{-\frac{1}{2}} \qquad \Phi_i(\Lambda_{smallest})^{-\frac{1}{2}}]. \tag{8}$$

$$y = \Psi^T \texttt{y\_pred} \tag{9}$$

Then, we compute the total scatter matrix S_t using (6). After computing the covariance matrix, the projection matrix Ω is selected by eigen decomposition of S_t and selecting the eigenvectors in Φ_{wy} according to the most significant eigenvalues Λ_{wy}. Eigen decomposition of S_t is given by:

$$S_t = \Phi_{wy} \Lambda_{wy} \Phi_{wy}^T. \tag{10}$$

Using the eigenvalues of S_t matrix, we formulate the objective as,

$$\underset{\Theta}{\operatorname{argmax}} \frac{1}{C-1} \sum_{i}^{C-1} \Lambda_{wy} \tag{11}$$

The objective of combining this with the deep neural net is that of maximization of the individual eigenvalues of S_t and minimization of the eigenvalues of S_w. In particular we expect maximization (minimization) of the eigenvalues of S_t (S_w) leads to maximizing (minimizing) separation in the respective eigenvector direction. Thus we would achieve the target of minimizing the within-class variation and maximizing the total variation. Deep neural network with categorical cross entropy (CCE) or binary cross entropy loss function does not take into account this aspect of discriminatory power. CCE main objective is to maximize the likelihood of the class labels according to the target labels.

Here the objective function is designed to consider only the k eigenvalues that do not exceed a certain threshold for variance maximization:

$$\underset{\Theta}{\text{argmax}} \frac{1}{k} \sum_{i}^{k} v_i \quad with(v_i, ...v_k) \;=\; \{v_j | v_j \;<\; min\{v_i, ...v_{n-1}\} + \epsilon\} \quad (12)$$

where for symbol easiness we have considered Λ_{wy} as v and n is the rank of the covariance matrix which is equal to one less than the number of samples $(n-1)$. This formulation of objective function allows to train the deep networks with backpropagation in end-to-end fashion. This is similar to the classic LDA but it lifts the constraint that generally occurs for binary classification where C (number of classes) is 2 and the l-dimensional projection matrix with classic LDA method will be $l = C - 1$ i.e, $2 - 1 = 1$. The above proposed objective function can be used for both multi-class as well as binary class classification problems.

3 Experimental Results

One of the key objectives of our work is to propose a CNN architecture that can be generically applied to many vision tasks. For our experimental evaluation we considered four publicly available databases, namely MNIST (hand written digits recognition), CIFAR-10 (natural scenes classification), ISBI 2016 (skin cancer classification) and ISBI 2017 (skin cancer classification). We compare our results with various other similar approaches available for vision classification.

3.1 Databases

- MNIST [23]: The MNIST or handwritten digits database consists of a 60,000 training set examples, and 10,000 testing set examples. The images have been size-normalized and centered to a defined size of 28×28 gray scale images. The database is freely available to public under a Creative Commons Attribution-Share Alike 3.0 license.
- CIFAR-10 [21]: The CIFAR-10 database is freely obtained under MIT licensing (MIT), used for object recognition application is an established computer-vision database which consists of 60000 32×32 colour images in 10 classes, with 6000 images per class. There are a total of 50000 training images and 10000 test images.
- ISBI 2016 [16]: The ISIC archive, containing training database of 900 images of dermoscopic lesion and 369 in testing database in JPEG format, obtained under CC0 licensing. From leading clinical centers internationally, these images have been collected that are acquired from various devices used at each center. It has both natural (skin hairs, veins) as well as man-made artifacts which becomes difficult to classify without pre-processing.
- ISBI 2017 [7]: International skin imaging collaboration (ISIC) is an international effort to improve melanoma diagnosis. In 2017 challenge, the database consists of more images in number as compared to 2016 including Seborrheic keratosis, a benign skin tumor derived from keratinocytes

(non-melanocytic) along with benign nevus (melanocytic) and melanoma (melanocytic). The training data consists of 2000 images (374 melanoma, 254 seborrheic keratosis and 1626 benign nevus) and testing data consists of 600 images (117 melanoma images), all obtained under CC0 licensing. This is the largest among all state-of-the-art melanoma databases.

3.2 Experimental Setup

The general structure of the CNN model is based on VGG model using 3×3 convolutions [30]. We experimented with and without including the Batch-Normalization layer after each convolutional layer [18]. This layer helps in increasing the convergence speed and also the performance of the model. For non-linearity RELU is used, since it greatly accelerate the convergence rate of stochastic gradient descent or any other optimizer as compared to the *sigmoid*/*tanh* functions [20]. All the networks are trained using Adam optimizer, but the learning rate is decreased to half after every 200 epochs. The batch size for MNIST data and CIFAR-10 is 1000 and for ISBI 2016 and ISBI 2017, the batch size is 400, as the training data is quite small in case of ISBI databases.

Related methods show that mini-batch learning on distribution parameters (in this case covariance matrices) is feasible if the batch-size is sufficiently large to be representative for the entire population [32]. Even though a large batch size is required to have stable estimates, it is limited by the data availability, image size and memory available on the GPU. Table 1 shows detail CNN model specifications for the CIFAR-10 and MNIST databases. The total number of trainable parameters for CIFAR-10 model is 5,752,414 and MNIST is 467,486. In all our experiments, the proposed method is validated with the existing ones using the same corresponding datasets and protocols. They are implemented on a system with Intel Core i7 processor, 16GB RAM, and NVIDIA GeForce GTX-1050Ti GPU card.

3.3 Results and Discussion

3.3.1 MNIST

The MNIST database consists of 28×28 gray scale image with labels as 0 to 9. The data structure consists of 60,000 samples of which 50,000 is training data and 10,000 is validation data. The test sample consists of 10,000 images, same protocol as that in [9]. Since the proposed method requires large batch size, thus for MNIST we took 1000 as the batch size. The optimizer is the Adam optimizer and the initial learning rate is reduced to half for every 200 epochs. For final classification, we use the linear support vector machine (SVM) classifier.

Table 2 shows the comparison of our proposed approach as compared to various relevant methods on MNIST database. From the results it can be seen

Table 1: Our proposed CNN model specifications for CIFAR-10 and MNIST databases.

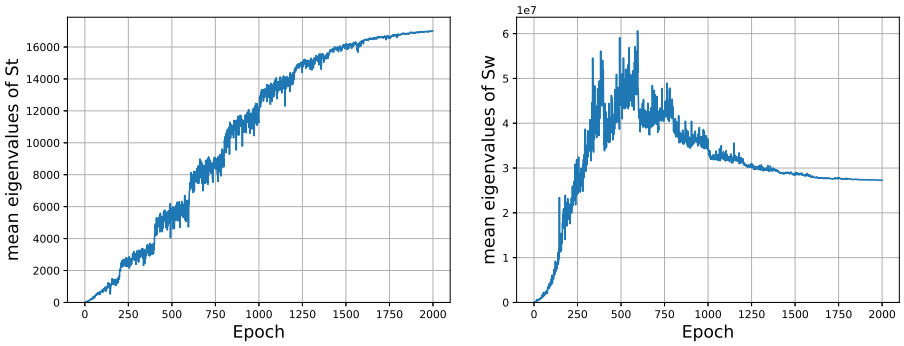
CIFAR-10	MNIST
Input $3 \times 32 \times 32$	Input $1 \times 28 \times 28$
3×3 Conv (pad-1)-64-BN-ReLu	
3×3 Conv (pad-1)-64-BN-ReLu	
2×2 Max-Pooling + Drop-Out (0.25)	
3×3 Conv (pad-1)-128-BN-ReLu	3×3 Conv (pad-1)-96-BN-ReLu
3×3 Conv (pad-1)-128-BN-ReLu	3×3 Conv (pad-1)-96-BN-ReLu
2×2 Max-Pooling + Drop-Out (0.25)	2×2 Max-Pooling + Drop-Out (0.25)
3×3 Conv (pad-1)-256-BN-ReLu	
3×3 Conv (pad-1)-256-BN-ReLu	
3×3 Conv (pad-1)-256-BN-ReLu	
3×3 Conv (pad-1)-256-BN-ReLu	
2×2 Max-Pooling + Drop-Out (0.25)	
3×3 Conv (pad-0)-1024-BN-ReLu	3×3 Conv (pad-0)-256-BN-ReLu
Drop-Out (0.5)	Drop-Out (0.5)
1×1 Conv (pad-0)-1024-BN-ReLu	1×1 Conv (pad-0)-256-BN-ReLu
Drop-Out (0.5)	Drop-Out (0.5)
1×1 Conv (pad-0)-10-BN-ReLu	1×1 Conv (pad-0)-10-BN-ReLu
2×2 Global Average Pooling	5×5 Global Average Pooling
Regularized LDA Layer	

BN: Batch Normalization, ReLu: Rectified Linear Activation Function, Conv: Convolutional layer.

Table 2: Comparison of test errors (%) on MNIST database using our proposed approach and other relevant methodologies.

Method	Test Error (in %)
NIN [25]	0.47
Conv. Maxout + Dropout [15]	0.45
ScatNet-2 [3]	0.43
PCANet-1 [5]	0.62
DeepLDA [9]	0.29
Proposed method	0.35

that our proposed method with new cost function is second best and comparable with the other state of-the-art reported performances. So it is evident that adding the latent space representation into the cost function, by maximizing the between-class and minimizing the within-class eigen representation efficiently learns the features required for classification. Thus the training is done in an unsupervised manner and using linear SVM, we do the final classification using the testing data.



(a) (b)

Fig. 2: (a) shows the evolution of mean eigenvalues of S_t with respect to epoch number, (b) depicts the minimization of within class scatter matrix S_w with respect to epoch, on MNIST database.

Table 3: Comparison of test accuracy (%) on CIFAR-10 database using our proposed approach and other relevant methodologies.

Method	Accuracy (in %)
NIN+ Dropout [25]	89.59
Conv. Maxout + Dropout [15]	88.32
PCANet-2 [5]	78.67
DeepLDA [9]	92.42
Proposed method	90.04

Figure 2 (a) shows the evolution of mean eigenvalues of the total scatter matrix with varying epochs during the training. Figure 2 (b) shows the eigenvalues of within class scatter matrix with respect to varying epochs, which initially increases but later decreases; thus achieving our objective of minimizing the within class and maximizing the total variation among different classes as shown in Figures 2 (a) and (b).

3.3.2 CIFAR-10

The CIFAR-10 database consists of 32×32 size image containing 10 different classes. The database structure consists of 50,000 training samples and 10,000 testing samples, same as that in [9]. We normalize the pixel values between 0-1. Table 1 describes the network structure, and similar to MNIST approach described above the initial learning rate is reduced to half for every 200 epochs. Table 3 summarizes the comparison of our proposed approach and various relevant methods on this database. It can be seen that our proposed methodology has achieved second best accuracy for this natural image classification task.

3.3.3 ISBI 2016 and ISBI 2017

To show the efficacy of the proposed objective function, we have conducted experimentation on both multi-class (MNIST and CIFAR-10) and binary class classification databases (ISBI 2016 and 2017). ISBI databases consist of dermoscopic lesion images for the diagnosis of skin cancer melanoma from the non-melanoma cases. ISBI 2016 database consists of 900 training set and 379 testing set. The database is unbalanced with 727 benign images and 173 melanoma images. Similarly, ISBI 2017 database consists of 2000 training samples and 600 testing samples. As stated by Wang *et al.* [32], minibatch learning with covariance estimates requires large batch size such that it could represent the entire population. Thus to overcome the batch size problem due to limited availability of ISBI training and testing data as well as due to large size of these images (224×224) and limited amount of memory available in GPU, we first performed fine-tuning of pretrained ResNet-50 model which has 25,636,712 parameters and then extracted the features from the last convolutional layer. We used these 2-dimensional features as inputs to train MLP (multi-layer perceptron) or fully connected layers. The fully connected layers used for training with the proposed objective function can be represented as,

$$\begin{aligned} \Theta_{MLP} = Input(900, 2048) \rightarrow \\ Dense(2048) - Sigmoid - l2regularizer \rightarrow \\ Dense(1024) - Sigmoid - l2regularizer \rightarrow \\ Dense(1024) - Sigmoid - l2regularizer \rightarrow \\ Dense(100) - Sigmoid - l2regularizer \end{aligned} \tag{13}$$

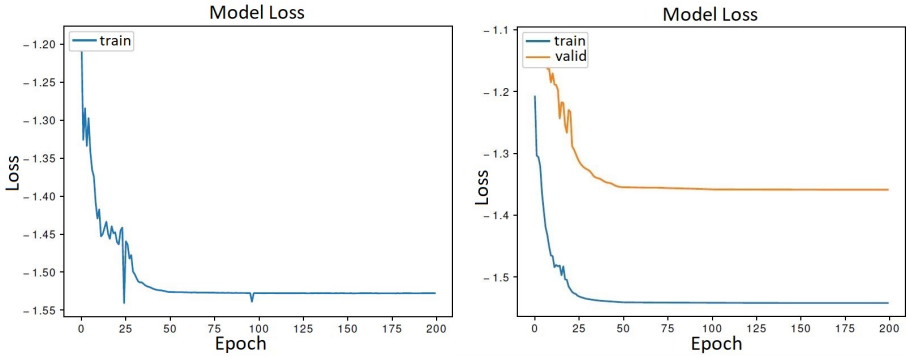
Sigmoid activation function is the most favoured activation function for shallow networks. We experimented using RELU and *tanh* as well, but there was no significant improvement using them. Activation function adds non-linearity to the existing nodes of the network. For deeper networks, RELU is the best activation function since RELU increases the convergence rate. Disadvantage of RELU is that ReLU units can be fragile during training and can erode easily [1]. The following performance criteria are used for comparison of the proposed approach with the existing methodologies:

- Accuracy: The ratio of correct prediction to that of total predictions, mathematical formulation as,

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}, \tag{14}$$

  where TP = True Positive, TN = True Negative, FP = False Positive, FN = Flase Negative.
- Sensitivity: The ability of the algorithm to correctly predict the diseased cases (*i.e.* malignant),

$$SE = \frac{TP}{TP + FN} \tag{15}$$



(a) (b)

Fig. 3: Loss with respect to number of epochs during training (a) loss vs epochs on ISBI 2016 database (with training data only as validation database is unavailable) (b) loss vs epochs on ISBI 2017 database (for both training and validation datasets).

– Specificity: It is the ability of the algorithm to correctly predict the non-diseased cases (*i.e.* benign),

$$SP = \frac{TN}{TN + FP} \tag{16}$$

– AUC: Area under receiver operating characteristic curve. It is the graph between true positive rate against the false positive rate.
– Average Precision: Average precision (AP) is the area under the precision-recall curve. The detailed explanation can be found in [16].

Since DeepLDA approach uses the traditional LDA where we could get at most, number of classes minus one as the principal eigenvalues which in this database would be $(2-1) = 1$. Thus, at the end there would be only one eigenvalue to maximize so as to have maximum inter class separation and minimum within class separation. In our approach, we use total class scatter matrix variance information to find the optimal projection among all the training data samples. This has enabled us to select up to $n-1$, where n is the total number of training samples. The model loss plot with respect to varying epochs are shown in Figure 3(a) for ISBI 2016 and 3(b) for ISBI 2017 databases respectively. The plot shows that in both the cases the loss decreases evenly with increase in number of epochs and finally converges.

Tables 4 and 5 show the various comparison of this approach with the existing ones on ISBI 2016 and 2017 databases, respectively. The results obtained on these databases do not exceed the best accuracy so far obtained but show a new approach to proceed by inheriting the class separability into the deep neural net as a result of changing the objective function. We implemented DeepLDA method [9] and experimented on ISBI databases. These

Table 4: Comparison of the proposed approach with the existing state-of-the-art methodologies on ISBI 2016 database.

Methods	Accuracy	AUC	AP	SE	SP
LDF-FV (fusion)[35]	0.868	0.852	0.684	0.426	0.977
CNN-FV (fusion)[36]	0.831	0.796	0.535	-	-
FCRN+deep ResNet[34]	0.855	0.804	0.637	0.507	0.941
Ensemble model [8]	0.805	0.838	0.645	0.693	0.832
Deep Bayesian Active Learning [11]	-	0.750	-	-	-
ResNet features+SVM	0.738	0.620	0.313	0.347	0.835
DeepLDA [9]	0.839	0.807	0.595	0.546	0.911
Proposed method	0.849	0.818	0.629	0.640	0.901

Table 5: Comparison of the proposed approach with the existing state-of-the-art methodologies on ISBI 2017 database.

Methods	Accuracy	AUC	AP	SE	SP
RECOD-TITANS [28]	0.872	0.874	0.715	0.547	0.950
ResNet features+SVM	0.783	0.689	0.379	0.350	0.888
DeepLDA [9]	0.831	0.791	0.544	0.470	0.919
Proposed method	0.833	0.793	0.566	0.555	0.901

tables show that our proposed approach achieves third best in its accuracy and AUC on ISBI 2016 database and second best for these metrics on ISBI 2017 database. Fisher vector based methods [35] and [36] use 32,768 (even after dimensionality reduction using principal component analysis) and 12,800 feature dimensions, respectively, for final feature matching, which are very high as compared to ours that uses only 2048 feature dimensions and 899 (number of samples -1) for final classification purpose. For ISBI 2017 [28], the authors used two pretrained CNN models ResNet-101 and Inception-v4. Experimentation using large number of data requires huge computational resources such as large memory CUDA-compatible GPUs. The training time and complexity are huge as compared to our approach, which uses only 2048 features and still achieve competitive accuracy performances. Our method is simple, efficient, requires less computing time and complexity that can be generically applied to many computer vision classification tasks.

4 Conclusions

In this paper, we have proposed an objective function which would work for both binary as well as multi-class classification problems. The proposed loss function minimizes the within class variance and maximizes the total class variance. We experimented our method on popular databases for various applications like MNIST (hand written digit recognition) and CIFAR-10 (natural image classification), and we have shown that the proposed approach achieves competitive performances on these databases as compared to other methods. For the application of melanoma detection (skin cancer detection

into melanoma and non-melanoma cases), since the number of images are few we trained the network using multi-layer perceptron and are able to achieve an accuracy of 84.9% on ISBI 2016 and 83.3% on ISBI 2017 databases. These experimental results show the efficacy of our proposed approach as compared to other methods for many computer vision classification tasks.

5 Declarations

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Cs231n: convolutional neural networks for visual recognition. http://cs231n.stanford.edu (2019)
2. Al-Waisy, A.S., Qahwaji, R., Ipson, S., Al-Fahdawi, S.: A multimodal deep learning framework using local feature representations for face recognition. Machine Vision and Applications **29**(1), 35–54 (2018)
3. Andén, J., Sifre, L., Mallat, S., Kapoko, M., Lostanlen, V., Oyallon, E.: Scatnet. Computer Software. Available: http://www. di. ens. fr/data/software/scatnet/.[Accessed: December 10, 2013] **2** (2014)
4. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on Machine Learning, pp. 1247–1255 (2013)
5. Chan, T.H., Jia, K., Gao, S., Lu, J., Zeng, Z., Ma, Y.: Pcanet: A simple deep learning baseline for image classification? IEEE Transactions on Image Processing **24**(12), 5017–5032 (2015)
6. Cheng, D., Zhang, S., Liu, X., Sun, K., Zong, M.: Feature selection by combining subspace learning with sparse representation. Multimedia Systems **23**, 285–291 (2017)
7. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1710.05006 (2017)
8. Codella, N.C., Nguyen, Q.B., Pankanti, S., Gutman, D., Helba, B., Halpern, A., Smith, J.R.: Deep learning ensembles for melanoma recognition in dermoscopy images. IBM Journal of Research and Development **61**(4), 5–1 (2017)
9. Dorfer, M., Kelz, R., Widmer, G.: Deep linear discriminant analysis. arXiv preprint arXiv:1511.04707 (2015)
10. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of human genetics **7**(2), 179–188 (1936)
11. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. arXiv preprint arXiv:1703.02910 (2017)
12. Gao, G., Liu, L., Wang, L., Zhang, Y.: Fashion clothes matching scheme based on siamese network and autoencoder. Multimedia Systems **25**, 593–6028 (2019)
13. Ghassabeh, Y.A., Moghaddam, H.A.: Adaptive linear discriminant analysis for online feature extraction. Machine vision and applications **24**(4), 777–794 (2013)
14. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 1. MIT press Cambridge (2016)
15. Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y.: Maxout networks. arXiv preprint arXiv:1302.4389 (2013)
16. Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1605.01397 (2016)

17. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J.: Face recognition using laplacianfaces. IEEE PAMI **27**(3), 328–340 (2005)
18. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
19. Jiang, X.D., Mandal, B., Kot, A.: Eigenfeature regularization and extraction in face recognition. IEEE PAMI **30**(3), 383–394 (2008)
20. Karpathy, A.: Cs231n convolutional neural networks for visual recognition. Neural networks **1** (2016)
21. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images (2009)
22. Kuo, C.C.J.: Understanding convolutional neural networks with a mathematical model. Journal of Visual Communication and Image Representation **41**, 406–413 (2016)
23. LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database. AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist **2** (2010)
24. Lifkooee, M.Z., Soysal, Ö.M., Sekeroglu, K.: Video mining for facial action unit classification using statistical spatial–temporal feature image and log deep convolutional neural network. Machine Vision and Applications pp. 1–17 (2018)
25. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
26. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
27. Martinez, A.M., Kak, A.C.: Pca versus lda. IEEE PAMI **23**(2), 228–233 (2001)
28. Menegola, A., Tavares, J., Fornaciali, M., Li, L.T., Avila, S., Valle, E.: Recod titans at isic challenge 2017. arXiv preprint arXiv:1703.04819 (2017)
29. Sharma, A., Paliwal, K.K., Imoto, S., Miyano, S.: A feature selection method using improved regularized linear discriminant analysis. Machine vision and applications **25**(3), 775–786 (2014)
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
31. Stuhlsatz, A., Lippel, J., Zielke, T.: Feature extraction with deep neural networks by a generalized discriminant analysis. IEEE transactions on neural networks and learning systems **23**(4), 596–608 (2012)
32. Wang, W., Arora, R., Livescu, K., Bilmes, J.: On deep multi-view representation learning. In: International Conference on Machine Learning, pp. 1083–1092 (2015)
33. Yu, H., Xia, K., Jiang, Y., Qian, P.: Fréchet mean-based grassmann discriminant analysis. Multimedia Systems (2019)
34. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE transactions on medical imaging **36**(4), 994–1004 (2017)
35. Yu, Z., Jiang, X., Wang, T., Lei, B.: Aggregating deep convolutional features for melanoma recognition in dermoscopy images. In: International Workshop on Machine Learning in Medical Imaging, pp. 238–246. Springer (2017)
36. Yu, Z., Ni, D., Chen, S., Qin, J., Li, S., Wang, T., Lei, B.: Hybrid dermoscopy image classification framework based on deep convolutional neural network and fisher vector. In: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on, pp. 301–304. IEEE (2017)