DEEP MULTI-VIEW ROBUST REPRESENTATION LEARNING

Zhenyu Jiao, Chao Xu

School of Electronics Engineering and Computer Science
Peking University,
Beijing 100871, China

ABSTRACT

Multi-view representations are widely existed in practical applications, the quality of latent representation learned from multi-view observations often suffer from noise and outliers in original data. In this work, we propose an auto encoder based deep multi-view robust representation learning (DMRRL) algorithm, which can learn a shared representation from multi-view observations and the algorithm is robust to noise and outliers by using Cauchy estimator as loss function. When the label of the original data is available, the algorithm can gain better representation by adding an auxiliary loss. We validate our methods on the CMU PIE dataset (noise applied) with face recognition task and UCF101 dataset with human motion recognition task, demonstrating that DMRRL is an effective algorithm for practical applications.

Index Terms— Multi-view learning, Cauchy Estimators, auto encoder, robust algorithm

1. INTRODUCTION

Multi-view representation exists in many real world applications, such as the multiple kinds of features in classification, simultaneously recorded audio and video, parallel text in multi-languages. The information obtained from one view is not sufficient to describe the full example. The goal of multi-view learning is to learn a latent representation utilized the connection and difference between multiple views, such that the original data can be accurately modeled.

During the past decades, many algorithms have been proposed. These algorithms can be roughly divided into 3 classes, co-training, multiple kernel learning(MKL), subspace learning. This paper focus on the subspace learning.

Subspace learning obtains a latent subspace shared by multiple views by assuming that the input views are generated from this subspace. Canonical correlation analysis (CCA) [1] is a standard statistical technique for finding linear projections of two random vectors that are maximally correlated. Kernel canonical correlation analysis (KCCA) [2] used the kernel trick to get a nonlinear form of CCA. DCCA [3] was a DNN-based extension of CCA and DCCAE [4] added an auto encoder regularization term to DCCA. [5] showed the ability of DNN-based multi-view learning method in real challenge. The shortcoming of CCA based method is that they can only be applied with the two-view scene, and is helpless to the multi-view scene.

Some approaches can also deal with multi-view scene. Shared Gaussian Latent Variable Model (sGPLVM) [6] used Gaussian process to learn common latent structure by multi-view observations. FLSSS [7] [8] factorized the information into shared parts and effectively account for the dependencies and independencies between the different input views. MSL [9] presented a convex formulation of multi-view subspace learning that enforces conditional independence while reducing dimensionality. MISL [10] integrated the encoded complementary information in multiple views to discover a latent intact representation.

While many multi-view learning algorithms have been presented in recent years, an algorithm which is robust to noise and outliers in original data and has strong learning ability is still scare. In this work, we keep the ability to discover a robust latent presentation from the data in [10] and utilize the learning ability provided by deep learning framework [11].

The contributions of this paper is as follows. We propose an auto encoder based deep multi-view robust representation learning algorithm which uses Cauchy estimator as loss function. This algorithm takes advantage of the modelling capability of deep framework and is robust to noise and outliers in original data. When the label of original data is available, the algorithm can gain better representation by adding an auxiliary loss. Experiment shows that our approach can get the latent representation which accurately models the original data even if the multi-view observations are contaminated by noise.

2. THE DEEP ROBUST MODEL

In this section, we first formulate the problem in subsection 1. Subsection 2 introduces the Cauchy estimator which is robust to outliers. Subsection 3 presents the DMRRL algorithm. Finally we extend the DMMRL algorithm by making use of label information in subsection 4.

2.1. Problem Formulation

In multi-view learning, an example e is represented by multi-view features $x^v (1 \leq v \leq N_v)$, where N_v is the number of views. Each view feature captures partial information of the example and the view generation can be formulated by:

$$x^v = g^v(e) + \varepsilon^v$$

where ε^v is view-dependent noise, $g^v(e)$ is the view generation function. The view generation function $g^v(e)$ is non-invertible, because some information miss when the example e is reflected to the particular view representation x^v. With multi-view observations x^v, the recover of the example e become possible, benefited from the connections and difference between multiple observations.

The goal of multi-view subspace learning is to infer, for each multi-view observation, a shared latent representation e' of dimension d_l such that the original data can be accurately modeled. A straightforward way to evaluate the latent representation is to apply view-generation function to e', such that the result $x^{v\prime} = g^v(e')$ should be similar to x^v. This process correspond to the design of auto encoder, so we use auto encoder as a base modal. The traditional AE minimize the empirical risk $\{x^v - x^{v\prime}\}$ with the L2 loss, but the L2 loss is not robust to outliers which are induced by the noise, and the performance will be seriously degraded. A robust estimators is needed.

2.2. Cauchy Estimators

Cauchy estimators is first used by [10] in multi-view learning, and preforms strong robustness to outliers.

$$\rho(x) = log(1 + (\frac{x}{c})^2)$$

Its derived function:

$$\varphi(x) = \frac{2x}{x^2 + c^2}$$

has a clear bound $[-\frac{1}{c}, \frac{1}{c}]$. This is different from the least-squares loss function whose derived function is unbounded such that the outliers only have a small effect on parameters estimation. Theoretically, Cauchy estimators is proven to have a breakdown point of nearly 50 percent [12], any single observation is insufficient to yield significant influence.

2.3. Deep Multi-View Robust Latent Space Learning

We now describe the AE-based multi-view robust representation learning algorithm considered here, with corresponding schematic diagram given in Fig 1.

Let $X = \{x_i | 1 \leq i \leq N\}$ be the training set which contains N examples. Each example $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector which is represented by N_v views with concatenation $[x_i^1, x_i^2, \cdots, x_i^{N_v}]$.

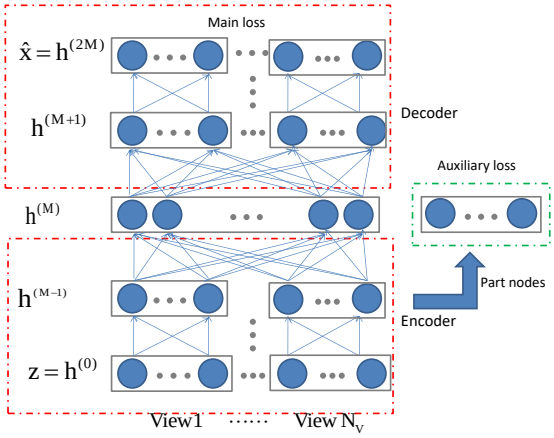


Fig. 1. Schematic diagram of AE-based multi-view robust representation learning model

Assume there are $2M + 1$ layers in the designed network and u^m units in the mth layer, where m = $1, 2, \cdots, 2M$. The output of x at the mth layer is computed as:

$$z^{(m)} = W^{(m)} a^{(m-1)} + b^{(m)} \tag{1}$$

$$h^{(m)}(x) = a^{(m)} = f(z^{(m)}) \in \mathbb{R}^{u^{(m)}} \tag{2}$$

where $W^m \in \mathbb{R}^{u^m \times u^{m-1}}$ is the parameter matrix and $b^m \in \mathbb{R}^{u^{(m)}}$ is the bias unit in the mth layer. We let $z^{(m)}$ to denote the total weighted sum vector of inputs in layer m and $a^{(m)}$ to denote the activation vector in layer m. f is a activation function which applies component wisely, such as widely used sigmoid, tanh or relu functions. We write $h^{(m)}$ to denote the output of the neural network which contains first m layers. For the first layer, we assume $a^{(0)} = x$ and $u^{(0)} = d$.

Each input x can be represented as latent representation $h^M(x)$ at the Mth layer through the encoding part (first M layers) and finally be reconstructed as $h^{2M}(x)$ at the $2M$th layer through the decoding part (last M layers). The reconstruction error can be measured using Cauthy loss:

$$\begin{aligned} \min_x J = \frac{1}{N} \sum_{i=1}^{N} \log(1 + \frac{\|x_i - h^{(2M)}\|_2^2}{c^2}) + C_3 r \\ + C_1 \sum_{m=1}^{2M} \|W^{(m)}\|_F^2 + C_2 \sum_{i=1}^{N} \|h^{(M)}\|_2^2 \end{aligned} \tag{3}$$

where $\|W^{(m)}\|_F^2$ and $\|h^{(M)}\|_2^2$ are regularization terms which aim to avoid overfitting. C_1 and C_2 are tunable non-negative regularization parameters that can be determined using cross validation.

To solve the optimization problem in (3), we use the stochastic gradient descent method to obtain the parameters

$W^{(m)}$ and $b^{(m)}$. $\|h^{(M)}\|_2^2$ only influenced by first M layers, so it is better to separate it out. Then the loss function can be divided into two parts as:

$$J_x = C_2 \sum_{i=1}^{N} \|h^M\|_2^2 + C_3 r \tag{4}$$

$$J_{AE} = \frac{1}{N} \sum_{i=1}^{N} \log(1 + \frac{\|x_i - h^{2M}\|_2^2}{c^2}) + C_1 \sum_{m=1}^{2M} \|W^{(m)}\|_F^2 \tag{5}$$

The gradients of function J on $W^{(m)}$ and $b^{(m)}$ can be calculated as follows:

$$\frac{\partial J}{\partial W^{(m)}} = s(m-M)\frac{\partial J_x}{\partial W^{(m)}} + \frac{\partial J_{AE}}{\partial W^{(m)}} \tag{6}$$

$$\frac{\partial J}{\partial b^{(m)}} = s(m-M)\frac{\partial J_x}{\partial b^{(m)}} + \frac{\partial J_{AE}}{\partial b^{(m)}} \tag{7}$$

where $s(m-M)$ is the unit step function which satisfies that $s(m-M) = 1$ for $1 \leq m \leq M$ and $s(m-M) = 0$ for $M < m \leq 2M$.

Now we can get $\frac{\partial J_{AE}}{\partial W^{(m)}}$ and $\frac{\partial J_{AE}}{\partial b^{(m)}}$ as follow

$$\frac{\partial J_{AE}}{\partial W^{(m)}} = \frac{2}{N} \sum_{i=1}^{N} \delta^{(m)} (a^{(m-1)})^T + 2C_1 W^{(m)} \tag{8}$$

$$\frac{\partial J_{AE}}{\partial b^{(m)}} = \frac{2}{N} \sum_{i=1}^{N} \delta^{(m)} \tag{9}$$

in which $\delta_{AE}^{(m)}$ is computed by

for the output layer,set

$$\delta_{AE}^{(2M)} = (\frac{a^{(2M)} - x}{c^2 + \|x - a^{(2M)}\|^2}) \bullet f'(z^{(2M)}) \tag{10}$$

for m = $1, 2, \cdots, 2M-1$,set

$$\delta_{AE}^{(m)} = (W^{(m+1)^T} \delta^{(m+1)}) \bullet f'(z^{(m)}) \tag{11}$$

The $\frac{\partial J_x}{\partial W^{(m)}}$ and $\frac{\partial J_x}{\partial b^{(m)}}$ for $m = 1, 2, \cdots, M$ can be computed with similar formulas.

Then $W^{(m)}$ and $b^{(m)}$ can be updated repeatedly by the gradient decent algorithm as follows to train the neural network until loss convergence:

$$W^{(m)} = W^{(m)} - \lambda \frac{\partial J}{\partial W^{(m)}} \tag{12}$$

$$b^{(m)} = b^{(m)} - \lambda \frac{\partial J}{\partial b^{(m)}} \tag{13}$$

where λ is the learning rate.

After $\{W^{(m)}\}_{m=1}^{2M}$ and $\{b^{(m)}\}_{m=1}^{2M}$ is obtained, then we can exploit first M layers of the neural network to infer latent representation of multi-view observations.

2.4. Using Label Information

Sometimes, the multi-view observations have class labels which are common in practical problems where multi-view learning is used as an approach to infer a shared latent p-resentation from multi-view observations. This class label constraint condition can be applied to the loss function as an auxiliary loss and we found this measure can significantly improve the quality of latent presentation. This is mainly due to the auxiliary loss and main loss can play a role as an regularization term alternately to avod loss function traps into a local minimum. We added an auxiliary output fully connected layer behind the Mth layer, as shown in Fig 1. Let $h_{aux}(x)$ denote the output of this layer using a softmax activation function and N_c is the num of class. We convert the labels to a N_c dimensional binary vector $\{y_1, y_2, \cdots, y_N\}$, the conversion is as known as "one-hot" encoding).

Now the auxiliary loss can be given by the categorical crossentropy:

$$J_s = -\frac{C_3}{N} \sum_{i=1}^{N} < y_i, log(h_{aux}(x_i)) > \tag{14}$$

where C_3 is tunable non-negative parameters that determine the weight of the auxiliary loss. The objective function of the supervised DMRRL method can be formulated as:

$$\min J = J_s + J_{AE} + J_x \tag{15}$$

Now the gradient of the objective function J with respect to $W^{(m)}$ and $b^{(m)}$ are calculated by:

$$\frac{\partial J}{\partial W^{(m)}} = \frac{\partial J_{AE}}{\partial W^{(m)}} + s(m-M)(\frac{\partial J_x}{\partial W^{(m)}} + \frac{\partial J_s}{\partial W^{(m)}}) \tag{16}$$

$$\frac{\partial J}{\partial b^{(m)}} = \frac{\partial J_{AE}}{\partial b^{(m)}} + s(m-M)(\frac{\partial J_x}{\partial b^{(m)}} + \frac{\partial J_s}{\partial b^{(m)}}) \tag{17}$$

The calculation of $\frac{\partial J_s}{\partial W^{(m)}}$ and $\frac{\partial J_s}{\partial b^{(m)}}$ can be inferred from Equation (8) (9) (10) (11). Then we update $\{W^{(m)}\}_{m=1}^{2M}$ and $\{b^{(m)}\}_{m=1}^{2M}$ with Equation (12) (13).

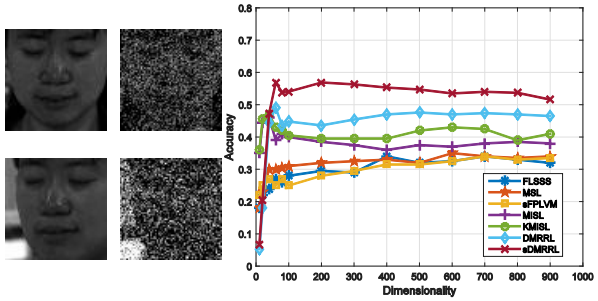
3. EXPERIMENT

3.1. Face Recognition

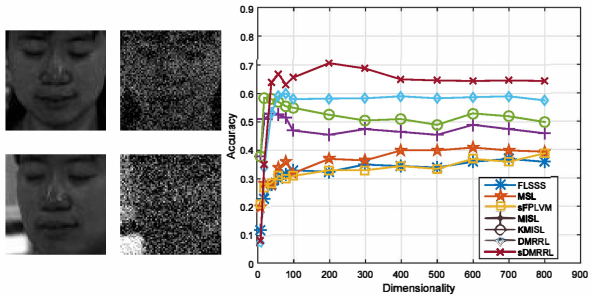
PIE face database [13] of CMU is a database of 41368 images of 68 people, each person under 13 different poses, 43 different illumination conditions, and with 4 different expressions. In order to facilitate comparison, we used identical settings to [10]. We selected two near frontal poses (C9 and C29) as two views, then each pair of images from the 2 poses with same illumination and expression of one person can be seen as two-views observation. The selected pair of images was separated into train set (random 50 percent images) and test set(other images). Different algorithms were employed to

project the input to low dimensional latent presentation and then a k-nearest neighbor classifier based on Euclidean distance was trained for recognition.

We set the num of latent layers $M = 2$, the num of neighbors $k = 3$ and experimented on different dimensional latent presentation. To show the robust of the DMRRL and sDMRRL, we applied noise to the images. Fig 2 show the images after applying noise and the result under 2 kind of noise (SNR=2 and SNR=5). We find DMRRL and sDMRRL can get a effective representation which is robust to noise in the multi-view observation, the recognition accuracy significantly outperform other algorithms when feature space dimension $d_l > 40$. The success of the DMRRL is due to the nonlinear learning ability of the deep network, and the robustness of Cauchy estimator.



(a) SNR = 2



(b) SNR = 5

Fig. 2. Face recognition accuracies of different algorithms on different dimensional spaces.

3.2. Human Action Recognition

UCF101 [14] is an action recognition data set of realistic action videos, collected from YouTube. It consists of 101 action classes over 13k clips and 27 hours of video data. The videos in 101 action categories are grouped into 25 groups, where each group can consist of 4-7 videos of an action. As the experiment settings in [10], we split the entire datasets to training and test samples three times, each split randomly selecting two-thirds of the data for training and the rest for testing. The split keep the videos belonging to the same group seperate in training and testing.

To construct the multi-view setting, we chose Histograms of Oriented Gradients (HOG) [15] and Motion Boundary Histograms (MBH) [16] descriptors as two views of feature extracted from video clip. For feature extraction, we used identical settings to [17]. We train a codebook for each descriptors type using 100,000 randomly sampled features with k-means. The size of the code book is set to 4000. Multi-view learning algorithm is followed to get the latent representation which helps to improve the recognition accuracy. For DMRRL and sDMRRL, we set num of layers $M = 2$ and the feature space dimension $d_l = 1500$. A linear SVM is used for classification, and the multiclass support is handled according to a one-vs-one scheme. The performance is measured by the average classification accuracy over all classes on three splits. The experiment result comparison between different algorithm is shown in Table 1. DMRRL and sDMRRL can utilize the complementary information across multiple views to form a better hidden representation and classification accuracy is improved with the merged features.

Algorithm	MBH+HOG	MBH	HOG
sGPLVM	58.43%	52.80%	42.04%
FLSSS	60.75%	53.56%	43.12%
MSL	60.49%	54.70%	42.74%
MISL	62.57%	55.75%	43.81%
KMISL	64.23%	56.23%	44.62%
DMRRL	64.82%	56.77%	45.01%

Table 1. Average accuracy of different algorithm on UCF101

4. CONCLUSIONS

This work presents the deep multi-view robust representation learning to get effective latent representation from multi-view observations which may contain noise. The DMRRL combines the deep auto encoder with Cauchy estimator such that the feature representation generated by DMRRL is effective and robust to noise. When label information is available, category constraints can be added to loss function to further improve the quality of latent representation. The experiment results suggest that the effective representation based on DMRRL or sDMRRL is robust to noise and is promising for practical applications.

5. REFERENCES

[1] Harold Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 28, pp. 321–377, 1935.

[2] Shotaro Akaho, "A kernel method for canonical correlation analysis," *In Proceedings of the International*

Meeting of the Psychometric Society (IMPS2001, vol. 40, no. 2, pp. 263–269, 2001.

[3] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *ICML*, 2013, pp. 1247–1255.

[4] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes, "On deep multi-view representation learning," in *Proc. of the 32st Int. Conf. Machine Learning (ICML 2015)*, 2015, pp. 1083–1092.

[5] Jun Qi and Javier Tejedor, "Deep multi-view representation learning for multi-modal features of the schizophrenia and schizo-affective disorder," in *ICASSP*, 2016.

[6] Aaron P. Shon, Keith Grochow, Aaron Hertzmann, and Rajesh P. N. Rao, "Learning shared latent structure for image synthesis and robotic imitation.," *Proc Nips*, pp. 1233–1240, 2005.

[7] Mathieu Salzmann, Carl Henrik Ek, Raquel Urtasun, and Trevor Darrell, "Factorized orthogonal latent spaces," *Journal of Machine Learning Research*, vol. 9, pp. 701–708, 2010.

[8] Yangqing Jia, Trevor Darrell, and Mathieu Salzmann, "Factorized latent spaces with structured sparsity," *Advances in Neural Information Processing Systems*, pp. 982–990, 2010.

[9] Martha White, Yaoliang Yu, Xinhua Zhang, and Dale Schuurmans, "Convex multi-view subspace learning," *Nips*, pp. 1673–1681, 2012.

[10] C. Xu, D. Tao, and C. Xu, "Multi-view intact space learning.," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 37, no. 12, pp. 2531–2544, 2015.

[11] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng, "Multimodal deep learning," in *International Conference on Machine Learning, ICML 2011, Bellevue, Washington, Usa, June 28 - July*, 2011, pp. 689–696.

[12] Ivan Mizera and Christine H. Mller, "Breakdown points of cauchy regression-scale estimators," *Statistics & Probability Letters*, vol. 57, no. 1, pp. 79–89, 2002.

[13] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database," in *IEEE International Conference on Automatic Face and Gesture Recognition, 2002. Proceedings*, 2002, pp. 46 – 51.

[14] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *Computer Science*, 2012.

[15] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, no. 12, pp. 886–893, 2005.

[16] Navneet Dalal, Bill Triggs, and Cordelia Schmid, *Human Detection Using Oriented Histograms of Flow and Appearance*, Springer Berlin Heidelberg, 2006.

[17] Heng Wang, Alexander Klser, Cordelia Schmid, and Cheng Lin Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.