Small sample learning with high order contractive auto-encoders and application in SAR images

Appendix A

A Background: Regularized auto-encoders

When AEs are used as a tool for dimension reduction, AEs perform as a dimensionality reduction method, and the solution for the minimum loss problem is unique. However, the basic AEs are always over-complete form in deep neural networks, thus we need to provide more hidden layer representation than input layer, which makes the problem an ill-posed one. The only way to turn it into a posed problem is regularization. Thus, in mathematical analysis, the auto-encoders should be trained and learned under regularized constraints.

Studies show that the auto-encoders are easy to apply in the deep structure, especially because of the over-complete hidden layer used in the algorithm. Once the input data is completely copied to the hidden layer and directly produce an output, which will be the complete factorization of the input, so called over-fitting. It is necessary to limit the "encoding" of the auto-encoders, but the restrictions cannot be applied to the dimension of hidden layers, by which the training becomes questioning its original issue. This series of methods are called "regularization".

Regularization is widely used in statistical learning(Mohammadi H et al, 2013). Regularization of the auto-encoders is to constrain the hidden layer representation, and the auto-encoders is regularized by applying penalty to the final loss function. But then the hidden layer will reconstruct to its former state, and based on the decoder's inverse of the encoder. Particularity, regularization is an irreversible algorithm that does not affect energy minimization.

Similar as Kamyshanska et. al.(Kamyshanska et. al. 2015), we can also establish the regularization constraint by the energy model of the auto-encoders. In this paper, we study the factor of regularization constraints from the energy perspective, and further study the validity of these constraints in the sample scarcity model.

The most frequently studied regularized automatic encoders include Sparse Auto-Encoder, Denoising Auto-Encoder, and Contractive Auto-Encoder are still useful and is our foundation in studying data scarcity problem.

A.1 Sparse auto-encoders (SAE)

Sparse auto-encoders are the simplest encoders whose rules punish the deviation of the hidden layer cells by moving them in a smaller (even negative) direction(Bengio Y et al. 2006; Schölkopf B et al. 2006; Lee H et al. 2007; Goodfellow I J et al. 2009). Or apply the penalty directly to the hidden layer activation function, so that the activation function is more likely to be saturated (i.e. Sigmoid function of the largest and smallest) (Ranzato M et al. 2007). The current SAE is the most commonly used as L1 or Student-t constraint, where the average value is obtained by constraining the hidden layer output, and the amplitude of the Kullback-Liebler Divergence of the obtained average is limited.

A.2 Denoising auto-encoders (DAE)

DAE(Vincent P et al, 2008) is based on the corruption of input assumption, and the training data is added to the noise, so the auto-encoders is designed to learn reconstruct raw input that is not corrupted by noise. Thus, this forces the encoder to learn the more robust to the input noise, which is why its generalization is stronger than the average AEs.

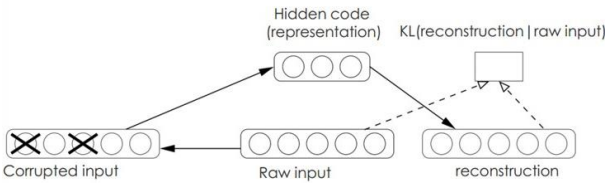


Fig. 1. Denoising Auto-encoders

Where E is the mean of the contaminated sample points and the corruption of the samples are generated using q (x ~ | x (t)). DAE is trained by Stochastic Gradient Descending(SGD) algorithm. The introduction of noise is very useful to learn the true innate representation of the samples. Thus DAE is proved as the energy distribution implementation(Swersky K et al. 2011) as Score Matching method, whose strategy is equivalent to energy minimization in DAE. And score matching is very similar to the energy distribution of the GRBM, while the GRBM training process is associated with the additive noise to traditional auto-encoders, so the method is called denoising regularization factor.

A.3 Contractive auto-encoders (CAE)

The contractive factor(Alain G et al, 2014) is based on the first-order derivatives of encoder function in auto-encoders, and the extended form of CAE algorithm can operate as both sparsity and denoising regularization. The CAE algorithm mentioned in this article is specific to the Jacobian matrix of the encoder as a regularized penalty factor for loss function, usually in the form of the Jacobian Frobenius norm.

$$
\begin{aligned}
J_{CAE} &= \sum_{t} L\left(x^{(t)}, g_{\theta}\left(f_{\theta}\left(x^{(t)}\right)\right)\right) + \lambda \left\| J\left(x^{(t)}\right)\right\|_{F}^{2} \\
J_{j}(x) &= f_{\theta}\left(x^{(t)}\right)_{j}\left(1 - f_{\theta}(x)_{j}\right) W_{j} \\
\left\| J\left(x^{(t)}\right)\right\|^{2} &= \sum_{j}\left(f_{\theta}(x)\right)_{j}\left(1 - f_{\theta}(x)_{j}\right)^{2} \left\| W_{j}\right\|^{2}
\end{aligned}
\qquad (1)
$$

Where λ is a parameter that controls the intensity of the regularization term (which is independent of the training process and is only relevant to the system properties).

CAE is advantaged over other regularization methods, simply because the sensitivity of the feature rather than the reconstruction of the regularization; it is calculated of the penalty factor by numerical analysis rather than random generated regularization; λ parameters ensure that the regularization can be balanced between reconstruction error and robustness. However, one of the disadvantages of CAE's advantage relies on its analytic penalty factor, mainly due to its robustness to input with small perturbations. When the input is widely separated and the neighborhood classification margin is large, such input error will result in a greater probability of positive false error. The CAE differs from other regularized auto-encoders for preserving the input value by increasing saturation rate of the encoder, while maintaining the saturation of the samples. The higher sensitivity of perturbations in specific directions is also known as a neighborhood characteristic for CAE, which makes it maintain the ability to learn a density manifold embedded in the higher data space.

In summary, regularized auto-encoders has made more attempts for the general data processing, and for the regularization of the method of theoretical verification, experimental analysis, to prove the feasibility of the method, but for the existing sample scarcity model , We need to consider the validity of the rule constraints and the limited content of the sample.

Appendix B

B Our proposed generalized regularization

B.1 Energy form requirements

According to (Kamyshanska et. al. 2015), they established the auto-encoders' energy form without regularization as

$$E(x) = \int s(Wx + b_h)dv - \frac{1}{2}\| x - b_r\|_2^2 + const \qquad (2)$$

From where the regularization term is works as a sink of Laplacian onto the energy function, thus, we deduced from the second order expansion that the Hessian of Energy should be negative in any observed point x_0. Thus, the generalized regularization rules must first satisfy the conditions for energy analysis, i.e.

$$\sum_k \left(s'\left(w_k^T x\right)\right)^2 \left\|w_k\right\|_2^2 < D \qquad (3)$$

From the numerical calculation, the main characteristic of w_k is when CAE is optimal in the hidden layer space; D is the dimension of the input samples. We propose the spatial-sensitive rules for regularization w.r.t. the discussion above.

B.2 Spatial gradient of contraction

In this paper, a spatial gradient based on the higher-order derivative is proposed. For the sample scarcity problems, new added term is

$$H \propto E\left(\frac{\left\|J\left(x\right) - J\left(x+\varepsilon\right)\right\|}{\left|f\left(x\right)+\varepsilon\right|}\right) \qquad (4)$$

This satisfies the energy requirements because it changes where Hessian of higher order is infinitesimal. In Bishop pattern recognition algorithms(Bishop et al. 2006), the square sum of the diagonal elements of the Hessian matrix is used to penalize the input data, and the computational superior is described. However, to ensure the calculation functions well even in sample scarcity problems, we exploit the second term as the CAE + H algorithm(Rifai S et al. 2011), which also satisfies sensitivity in a certain range to the second-order characteristics of data generating density. According to this idea, this paper extracts the high-order contractive information to exploit the auto-encoders' expressiveness, and the density of the sample space is estimated. In this article, estimation for Hessian strategy is adopted for data scarcity problem, i.e., by randomly selecting of ε to estimate the expectation of Hessian value in the sample neighborhood. The former solution, adapting square sum of diagonal elements of Hessian matrix as regularization factor, however, was used more in unsupervised learning; and, the problem got complexed with calculating for Hessian. On the contrary, in our paper, the value of the higher order contractive factor can be obtained by statistically average over x neighborhood. On the other hand, from a statistical point of view, the penalty factor needs to be estimated by means of existing sample points, and the sample generating distribution is more-or-less implicit in this process, and this factor can no longer be regarded as an independent "a priori " but as a new strategy to exploit the data information. Yet by adding the Hessian regularization term, we cannot simply take the influence on density estimation of the auto-encoders to prove the validity; on the contrary, we still needs further study of the small sample to verify its effectiveness.

For ill-posed problem with limited data, we introduce the gradient feature of the second order spatial Hessian. Compared with the simpler generalized algorithm, the sample space of the second order spatial matrix will be adequate for a certain sample size, and the result of the calculation tends to be optimal in contractive ratio.

Then, we solve the ill-posed problem by studying the geometric meaning of the second order derivative, also known

as Hessian. As the contractive first-order derivatives, it is the same with Hessian, which can accelerate its loss descending form through the first-order hidden layer derivatives, and with the second-order derivatives, which will lead the energy moving towards the sinks of sampled data. With the scarcity of the target sample, the descriptiveness of the hidden layers in general is determined by whether $f(x)$ is sufficiently sensitive to the tangential changes along the manifold, which will ensure better performance even in the limited samples. So that GAE algorithm we propose, with certain conditions, can be a very effective solution for density estimation of the samples.

B.3 Generalized regularization auto-encoders model

By studying the above regularization factors, we propose a generalized regularization auto-encoders model, the minimization error is:

$$J_{GAE} = \sum_t L\left(x^{(t)}, g_\theta\left(f_\theta\left(x^{(t)}\right)\right)\right) + \lambda\left\|J\left(x^{(t)}\right)\right\|_F^2 - \mu E\left[\frac{\left\|J(x) - J(x+\varepsilon)\right\|}{f(x)+\varepsilon}\right] \tag{5}$$

The model is an improvement and estimation of the Contracted Regularizer (CR). The model of the single layer structure is shown in Fig.2.

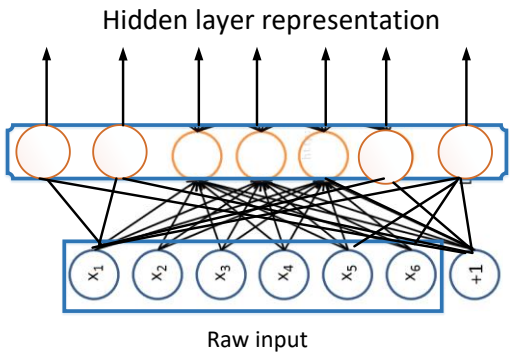


Fig. 2 GAE model structure

Complexity analysis: For CAE, the complexity of the algorithm depends on the dimension of the input sample and the hidden layer space, expressed as o ($d_x d_h$). When adding the second order Hessian factor as the penalty factor, the computational complexity is also depends on randomized sampling within the neighborhood of x for n_c samples to estimate Hessian, thus o ($N_c d_x d_h$) is the complexity, in which the Hessian matrix is calculated by assuming that n_c samples are obtained by Gaussian white noise sampling. In GAE, similarly with CAE calculation, the Jocobian's complexity is o($d_x d_h$) and then the most important to note we choose nc as 5 throught the samples we used, thus the complexity is also o($C d_x d_h$)it does not adds greatly than CAE, which is acceptable.

Appendix C

C Experiments and analyses

The main purpose of this experiment is to verify the reliability of proposed regularized auto-encoders, and by the design of single-layer generalized AEs (GAE) model, we first study data density estimation in our image dataset, through the handwritten digits and SAR image data experiments we will verify the following questions:

(1) GAE is effective in handwritten digits and SAR image target identification and classification;

(2) GAE's property as regularized auto-encoders vs. the number of input samples, and then study the model of its hyper parameters;

(3) GAE's effectiveness on the sample scarcity problem of the SAR image applications;

C.1 Experiments design

Comparative experimental protocols include regularized automatic encoders such as DAE, CAE, CAE + H, GAE (three automation factors), and RBM network models.

DAE-Denoising auto-encoders

CAE-Contractive auto-encoders

CAE + H - Higher-order contractive auto-encoders, regularization of higher - order derivatives

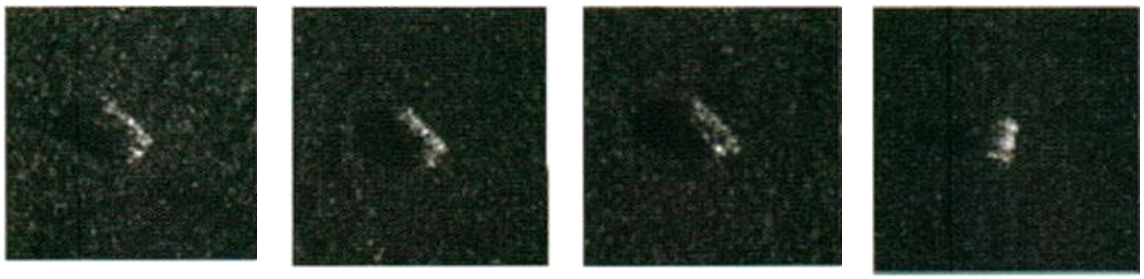
GAE - Regularized Auto-encoders based on higher-order contractive gradient

RBM-CD-Restricted Boltzmann Machine

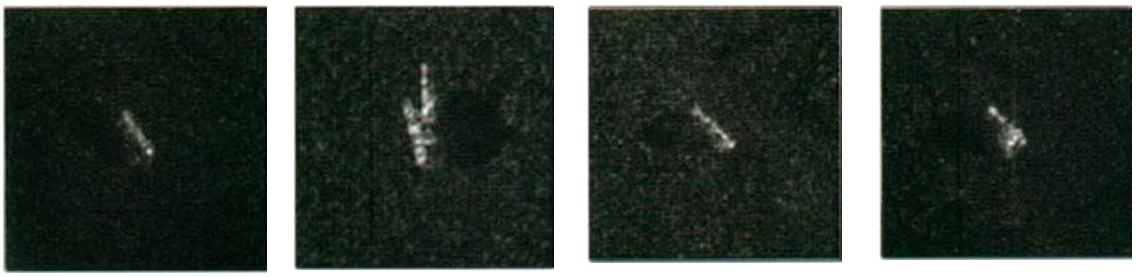
The data used in the experiment are mainly MNIST handwritten database and MSTAR database, for their difference between the natural image data and the remote sensing images. This experiment is designed in order to verify the applicability of the algorithm to the sample scarcity problem in SAR images.

MNIST database is an established database of handwritten digits from 0 to 9, which contains 60,000 training samples and 10,000 test samples. Each sample image is a gray-scale picture with resolution of 28 * 28, and we randomly select a number of samples in each category, in order to re-establish a series of sample set from 50, 100 up to 10,000 for different training samples. Assume that the sample series are all sufficient representative, that is, the collected samples are representative enough for each category of the digits. With a number of randomly selected training sample sets, in the test set, we also randomly select 1000 test samples for example to test the 10000 sample learned algorithm as test set.

MSTAR dataset consist of 11 categories of samples, including eight classes of military vehicles, which we take as our research target. Each class has around 170-220 training samples, and there are about 2000 samples for training, also test set has about 1000 samples; and each of the sample is 128 by 128 image clip from larger scene SAR image. As shown in Fig.3, The dataset consists of three major categories, B(BRDM_2, BTR-60, 2S1), T(T62, T72) and others(D7, ZIL131, ZSU_23_4). Similarly, we sample the image as a dataset from 80, 160 up to 2000 as new subsets for training the models. And also test set from the 1000 samples.



2S1 BRDM_2 BTR70 D7



T62 T72 ZIL131 ZSU_23_4

Fig. 3 MSTAR dataset samples

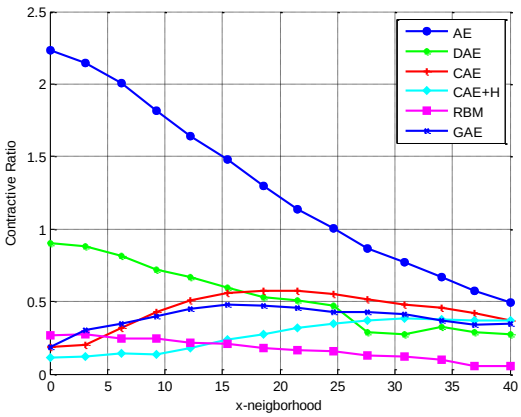
The reconstructed error is measured by the cross entropy information in the experiment and is calculated by the stochastic gradient descent. RBM model is trained by Contrastive Divergence (CD) algorithm, implemented as a comparison on limited selection of MNIST and MSTAR dataset.

As a study on the single-layer encoder structure, this article only employ the traditional single-layer network model as comparison for effectiveness in information extraction and small sample processing performance. All of the single-layer network outputs will be entered into the linearly encoded Logistic Regression (Logistic regression model) for the target classification, but the learning process in GAE model is mainly two-way information representation via Stochastic Gradient Descending algorithm along the loss function, which now is embodied as the higher order contractive gradient.

C.2 GAE target recognition analysis

On the analyses of the results, for MNIST handwritten data recognition in the experiment, regularization factors can extract the higher dimension information; whereas MSTAR database not only make use of higher order dataset information, also noise suppression performs very good. As a view of the Gaussian process in estimation Hessian, the feature GAE learned is very close to the Gabor filter, which serves as the band-pass filtering as the primitive processing in images. Therefore, the GAE shows more captive ability of dealing with the visual information than any of the regularized AEs. In MSTAR dataset, CAE algorithm shows a strong noise suppression, and it shows not only denoising effect, but also its ability to extract the smoother features with the manifold of data density; as for RBM network, it is also very competitive in extracting useful features even though its lack of competence in lower dimensional data inputs. Anyway, from the view of feature extraction, GAE and RBM are both quite effective.

In the study of CAE based methods, it is also called contractive ratio for the penalty term in the range of $U\left(x_0,\ \varepsilon\right)$. Compared with CAE algorithm and CAE + H algorithm, the maximum and minimum values of the generalized penalty factor in the sample neighborhood space. However, at the same time, the change result is converged in the nearest neighborhood. The Hessian penalty factor ensures the smoothness of the manifold in the process of moving away from the sample point.



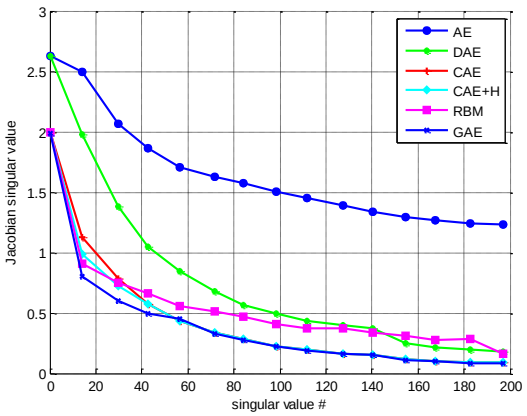


Fig. 4 J(x) characteristics in sample neigborhood and its eigenvalue distribution

From the previous study of contractive regularization, the geometric effect of contractive penalty makes the neighborhood of x0 have a better tangent space estimation. With higher order contractive factor, GAE is much promising in producing an closer estimation on the manifold. Thus, the singular value of Jacobian will be dropping faster and the constraint is tighter to acquire the requirements fitting for the accuracy of manifold.

By studying the contractive characteristics and eigenvalue of other penalty factors, it can be concluded that in order to make the GAE algorithm a better approximation to the manifold, it is necessary to increase the penalty of $J(x)$ and Hessian matrix. GAE also performs better density estimation of the manifold, at the same time drives the training towards convergence faster. From the view of information theory, the higher-order contractive factor is to make sure the encoder can learn as much as possible without loss of useful information, and it will ensure that the equivalent of high-level information in the limited sample. Then we study the optimization with penalty factor, and get the optimal composition as regularization for SAR images by fine-tuning. Then we discuss the sample capacity of the regularization factors for different data, i.e. different regularization methods could only process a different minimum size of the sample space for density estimation.

We analyze on how GAE is capable to achieve better low-dimensional performance in SAR images, the following may be reasonable explanations:

1. GAE improves the dimension and converges rapidly during the training process because of the addition of the manifold descending factor of second order gradient to achieve better optimization in manifold approximation.

2. The GAE algorithm is advantaged not only in convergence rate, but also in limited samples. The accuracy of target recognition depends on the generalization of the hidden layer representations, in which the Jacobian contractive factor. The first-order regularization term guarantees its reliability (CAE), while the second-order factor constrains its distribution onto a smooth manifold.

C.3 Choice of hyperparameters

Hyper-parameters represents of the regularization intensity in the model, w.r.t. the above analysis, the effects of the two regularization factors are different. In our experiments, two very different categories need to be studied, whether to contract(increase λ) or improve the convergence rate(increase μ) thus will need to be discussed separately. We experimented several sets of hyper parameters, Fig.5 shows the reconstruction accuracy vs. parameters in MNIST dataset, but the axis here is in range of the valid values of each hyperparameters. The *y* axis only represent the magnitude of recognition rate change.

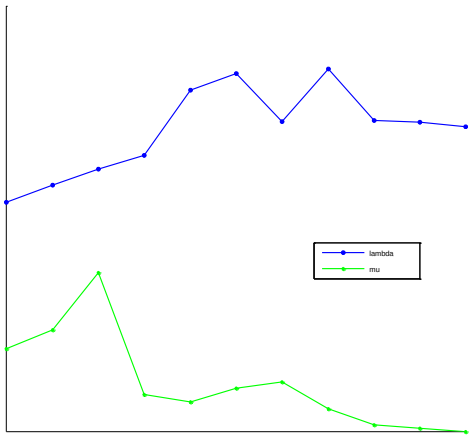
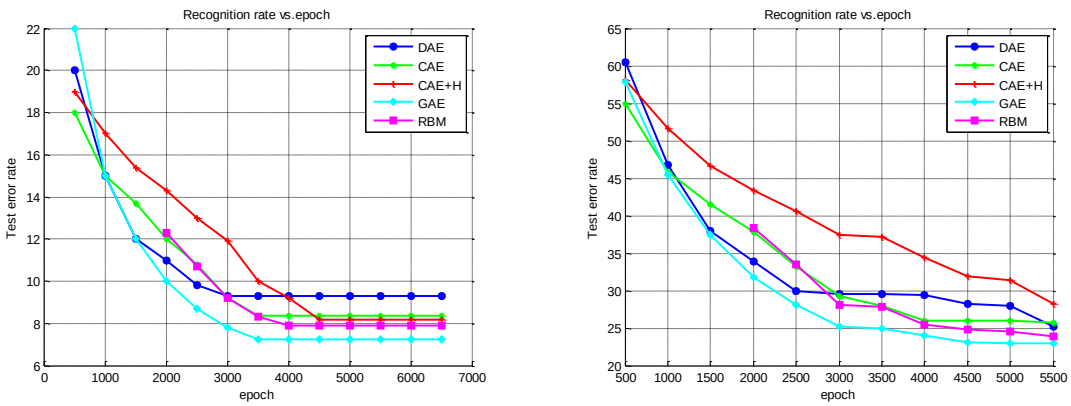


Fig. 5 Hyperparameters vs. Test accuracy

According to the curve, we can see that the accuracy is significantly influenced by contractive factor. Firstly the contraction factor increases with the penalty intensity, but when it reaches a certain degree The Frobenius norm of $J(x)$ it sharply drops in certain parameters, but remain relatively high until too large to constrain the output generality. The second order information is similar to the contractive factor, but its influence is not so significant compared to the first-order penalty .

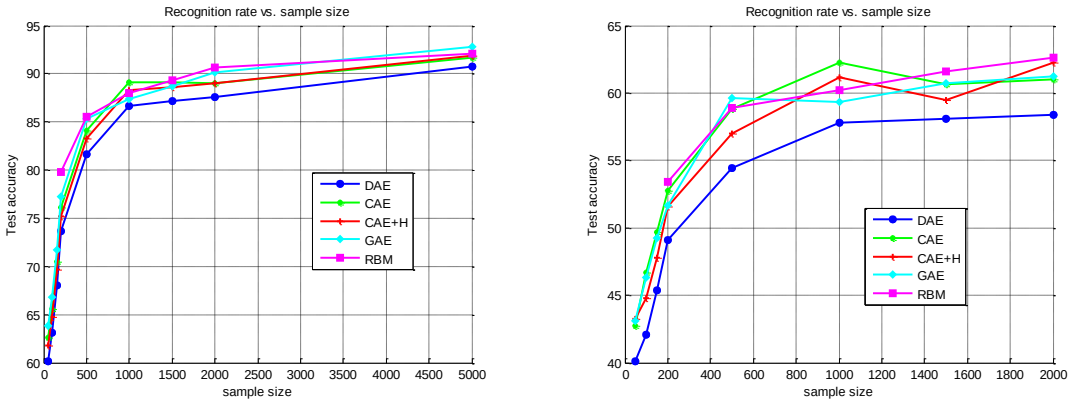
C.4 Small samples analyses

In GAE algorithm, experiments based on different sample sizes shown in Table 1, and Fig. 6. In MNIST database, The sample size raise is generally stable when it reaches 1000, and among all the algorithms we compared, GAE is the best in test accuracy, but also, RBM shows great increase when sample size reach 150 and more, so the best methods for MNIST are GAE and RBM. Then in MSTAR dataset, it differs greatly for GAE shows best performance in lower dimensions less than 500 samples. And the explanation is in MSTAR dataset, we calculated only the central sample's direction, geometric shape etc., for which the GAE shows best representation in hidden layer vectors. Thus, in SAR dataset, GAE shows great lower-size sample tolerance. This shows our proposed method as competitive as a SAR image representation learning machine.



MNIST MSTAR

(a) Test error convergence



MNIST MSTAR

(b) Test Accuracy vs number of samples

Fig. 6 Test results of MNIST and MSTAR with different methods

Table 1 Test error of MNIST and MSTAR dataset

MNIST	**Training samples**	**Size of training set**					
	10000	**50**	**100**	**500**	**1000**	**2000**	**5000**
DAE	9.23±0.81	39.84	36.86	26.38	18.36	13.33	12.37
CAE	8.34±0.32	37.39	34.41	23.93	15.91	11.88	10.97
CAE+H	8.17±0.45	38.23	35.25	24.77	16.75	11.72	11.01
GAE	**7.23**±0.38	**36.17**	**33.19**	**22.71**	**14.69**	**12.66**	**9.89**
RBM-CD	**7.31**±0.33	—	—	**20.22**	**14.45**	**12.01**	**9.34**
MTAR	**Training samples**	**Size of training set**					
	2000	**50**	**100**	**150**	**200**	**500**	**1000**
DAE	17.52±2.13	60.60	56.79	38.76	29.51	23.25	19.86
CAE	15.05±2.01	57.59	51.23	37.06	26.34	17.46	17.29
CAE+H	**14.61**±1.32	57.03	54.69	36.98	25.56	20.93	19.67
GAE	**15.15**±1.89	**54.10**	**51.85**	**36.20**	**23.47**	**20.90**	**15.43**
RBM-CD	15.27±2.03	—	—	**30.49**	**22.14**	**19.84**	**15.45**

REFERENCES

Alain G, Bengio Y. 2014. What regularized auto-encoders learn from the data-generating distribution. Journal of Machine Learning Research, 15(1):3563-3593

Bengio Y, Courville A, Vincent P. 2013. Representation learning: a review and new perspectives. IEEE Transactions on Pattern Analysis & Machine Intelligence, 35(8):1798-1828

Bengio Y, Lamblin P, Popovici D, et al. 2006. Greedy layer-wise training of deep networks. International Conference on Neural Information Processing Systems. MIT Press: pp.153-160

Bishop C M. 2006. Pattern Recognition and Machine Learning,Christopher M..Bishop,.Springer: pp.249-254

Bishop C M. Curvature-driven smoothing: a learning algorithm for feedforward networks. IEEE Transactions on Neural Networks, 1993, 4(5):882

Goodfellow I J, Le Q V, Saxe A M, et al. 2009. Measuring invariances in deep networks. International Conference on Neural Information Processing Systems. Curran Associates Inc.: pp.646-654

Kamyshanska H, Memisevic R. 2015. The Potential Energy of an Autoencoder. IEEE Transactions on Pattern Analysis & Machine Intelligence, 37(6): pp.1261-73

LeCunn Y, 2016. Predictive Learning, NIPS

Lee H, Ekanadham C, Ng A Y. 2007. Sparse deep belief net model for visual area V2. Advances in Neural Information Processing

Systems, Vol 20: pp.873-880

Mohammadi H, Riche R L, Touboul E, et al. 2013. On regularization techniques in statistical learning by Gaussian processe. Journal of Virology, 88(6): pp.3161

Ranzato M, Boureau Y L, Lecun Y. 2007. Sparse feature learning for deep belief networks. Advances in Neural Information Processing Systems: pp.1185-1192

Rifai S, Mesnil G, Goire, et al. 2011. Higher order contractive auto-encoder: pp.645-660

Rifai S, Vincent P, Muller X, et al. 2011. Contractive Auto-Encoders: Explicit Invariance During Feature Extraction. International Conference on Machine Learning, ICML

Schölkopf B, Platt J, Hofmann T. 2006. Efficient Learning of Sparse Representations with an Energy-Based Model. Advances in Neural Information Processing Systems: pp.1137 - 1144

Swersky K, Ranzato M, Buchman D, et al. 2011. On Autoencoders and Score Matching for Energy Based Models. International Conference on Machine Learning, ICML 2011, Bellevue, Washington, Usa, June 28 - July: pp.1201-1208

Vincent P. 2011. A connection between score matching and denoising autoencoders. Neural Computation, 23(7):pp.1661-1674

Vincent P., H. Larochelle Y. Bengio and P.A. 2008. Manzagol, Extracting and Composing Robust Features with Denoising Autoencoders, Proceedings of ACM Twenty-fifth International Conference on Machine Learning, ICML: pp. 1096 – 1103