SCIENCE CHINA
Information Sciences

• LETTER •

September 2018, Vol. 61 099101:1–099101:3
https://doi.org/10.1007/s11432-017-9214-8

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

Qianwen YANG[1,2,3] & Fuchun SUN[1,2,3*]

[1] *Department of Computer Science and Technology, Tsinghua University, Beijing* 100084*, China;*
[2] *State Key Lab of Intelligent Technology and Systems, Tsinghua University, Beijing* 100084*, China;*
[3] *Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing* 100084*, China*

Received 25 April 2017/Accepted 16 August 2017/Published online 3 May 2018

Citation Yang Q W, Sun F C. Small sample learning with high order contractive auto-encoders and application in SAR images. Sci China Inf Sci, 2018, 61(9): 099101, https://doi.org/10.1007/s11432-017-9214-8

Dear editor,
Recently auto-encoders (AEs) are used as intermediate layers or unsupervised learning stages in deep learning networks [1]. However, unlike other deep learning algorithms, which can extract higher-order abstract features using deep structures [2,3], AEs are typically more easily decoded than convolutional networks for their physical meanings. Applications are generally limited to encoding training data to hidden space.

In this study, we consider the case of remote sensing images. We propose a new theory on how regularized AEs can learn the appropriate features for target recognition even with extremely limited examples and prove its effectiveness. This study primarily aims at solving the problem of sample scarcity for remote sensing images, particularly synthetic aperture radar (SAR) images. SAR image samples are frequently difficult to obtain, and even for known data, transfer learning of deep convolutional neural networks cannot solve problems because of the differences between data source. We tried to solve the scarcity of samples in unsupervised AEs as representation learning effectively.

Generalized regularized auto-encoders. Why use regularized methods? AEs act as a dimensionality reduction method and provide a unique solution for the minimum loss problem. However, AEs have overcomplete hidden layers in deep neural networks; this makes the problem ill posed. The problem can be converted to a solvable optimization problem using regularization, which implies that AEs should be trained and learned under regularized constraints.

Ref. [2] shows that AEs are representations of local statistics, particularly for overcomplete hidden layers used in AEs; however, with limited samples, AEs will easily become overfit.

Regularization is widely used in statistical learning [4] and in sparse representation [5, 6]. Particularly, regularization is an irreversible algorithm that does not affect real reconstruction error, which could be viewed as an additive rectifier.

The most frequently studied regularized AEs include sparse AEs (SAE), denoising AEs (DAE), and contractive AEs (CAE). These AEs are useful, and they are the basis for investigating the data scarcity problem in this study.

The current SAE is the most commonly used as the L1 or Student's t constraint, where average loss is obtained by constraining hidden layer output. DAE [7] is based on the corruption of input assumptions; it is considered as an alternative implementation [2, 3] of the score matching method. The contractive factor [2] is based on the first-order derivatives of the encoder function in AEs. It is advantageous over other regularization methods in that it provides a balance between reconstruc-

[*] Corresponding author (email: thusunfc@163.com)

tion error and robustness. The disadvantage of CAE is that it relies on analytic penalty factor: if training set is disproportionally separated and the neighborhood classification margin is large, such an input error will result in a higher probability of positive false error.

We proposed regularized AEs as a solution to existing sample scarcity. However, contractive, denoising, and sparse regularizers are not fully applicable in small sample problems. Similar to former study [8], we established regularization constraints using the energy model of AEs. Based on the study of regularization constraints from the energy perspective, we tested the validity of these constraints in a small sample model.

Energy form requirements. Kamyshanska et al. [8] established a regularization term condition based on energy analysis as follows:

$$\sum_k s'(w_k^{\mathrm{T}}x)\,\|w_k\|_2^2 - D < 0, \tag{1}$$

where s is the sigmoid activation function, s' is its derivative function, x is input data, w_k is optimal weights of the hidden layer, and D is the dimension of the input samples. We propose spatial-sensitive rules for regularization according to the abovementioned discussion.

Our proposed spatial gradient is based on a higher-order derivative. The regularized additive term is as follows:

$$H \;\propto\; \mathrm{E}\left(\frac{\|J\,(x) - J\,(x+\varepsilon)\|}{|f\,(x)+\varepsilon|}\right), \tag{2}$$

where H is the Hessian term, $\mathrm{E}(\cdot)$ is expectation, f is the function of the encoders, and $J(x)$ is the Jacobian of $f(x)$.

However, to ensure that the regularizer functions well even in sample scarcity problems, we utilize the second term as the CAE+H algorithm [9], where we use high-order contractive information to capture the data generating density.

Generalized auto-encoders (GAE). The higher-order contractive factor is the statistical average over x neighborhoods, and the penalty factor estimated using existing samples ensures implicit expression of sample generating distribution. In addition, with limited training samples, the representativeness of the hidden layers is determined by whether $f(x)$ is sufficiently sensitive to the tangential changes along the manifold. Thus, under certain conditions, the loss function of the GAE is as follows:

$$\begin{aligned} J_{\mathrm{GAE}} = &\sum_t L(x, g(f(x))) + \lambda\|J(x)\|_F^2 \\ &- \mu\mathrm{E}\left[\frac{\|J(x) - J(x+\varepsilon)\|}{f(x)+\varepsilon}\right]. \end{aligned} \tag{3}$$

The GAE algorithm can be an effective solution for the density estimation of the samples. Model complexity is related to only numbers of neighbors C used to estimate Hessian, input data dimension d_x and hidden layer size d_h. GAE's complexity is $o(Cd_xd_h)$; compared to former study [2,3], it does not add much to CAE, which is acceptable.

Experiments and analyses. Experiments are performed to verify the reliability of the proposed GAE algorithm. First, data density estimation for handwritten digits and SAR image data is examined in the following aspects.

(1) Effectiveness on digits and SAR images for a small sample problem.

(2) Regularization property vs. the number of input samples, and the effect of hyperparameters on the regularizer.

The data used in the experiment is primarily the MNIST handwritten digit database and moving-stationary target automatic recognition (MSTAR) database.

Reconstruction error is measured using cross entropy information in the experiment and is calculated using stochastic gradient descent. This study employs only a conventional single-layer network model to compare the effectiveness of the restricted Boltzmann machine (RBM), DAE, CAE, and CAE+H algorithms in a small sample.

Regularization property. In the study of CAE-based methods, the contractive ratio is an indicator of the contractiveness of the regularizer. The Hessian penalty factor ensures the smoothness of the manifold captured during sampling from a given data neighborhood. The geometric effect of contractive penalty improves tangent space estimation in the neighborhood of x0. In these experiments, we analyzed the spectrum of Jacobian singular values. With a higher-order contractive factor, the GAE exhibits considerably promising results in terms of producing a closer estimation on the manifold.

Selection of hyperparameters. Hyperparameters represent the regularization intensity in the model. With respect to the abovementioned analysis, the effects of the two regularization factors are different. In the experiments, two considerably different categories must be studied, i.e., whether to contract (increase λ) or improve the convergence rate (increase μ). These categories are discussed separately. We performed experiments using several sets of hyperparameters and selected the set with the best performance for further study.

Recognition analysis. For MNIST handwritten data recognition, regularization factors can extract higher-dimension information; MSTAR data exhibits noise suppression performance. In the Gaussian process for Hessian estimation, the feature learned by the GAE is considerably close to the Gabor filter, which acts as a band-pass filter for primitive processing of images. Therefore, the GAE is more capable of processing visual information than other methods. In the MSTAR dataset, the CAE algorithm exhibits strong noise suppression. In addition to this denoising effect, it extracts the smoother features with the data density manifold. Moreover, the RBM network is extremely competitive in extracting useful features, even though it lacks accuracy for small samples.

The experiments carried out using the GAE algorithm for different sample sizes are shown in Table 1. In the MNIST database, the increase in sample size is generally stable after it crosses 1000. Among all algorithms that are compared, the GAE provides the highest test accuracy. However, the RBM exhibits significant increase when sample size increase to above 150 as shown in Table 1. In both MNIST and MSTAR database, the GAE shows the best performance in lower dimensions for less than 500 samples.

Table 1 Test error for MNIST and MSTAR databases[a)]

MNIST	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	**7.31**±0.33	—	—	**20.22**	**14.45**	**12.01**	**9.34**
MSTAR	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	15.27±2.03	—	—	**30.49**	**22.14**	**19.84**	**15.45**

a) Bold digits show best error rates.

The reasons that the GAE achieves better low-sample-size performance in SAR images may be as follows.

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

(2) The GAE algorithm is advantageous in the case of limited samples. The accuracy of target recognition depends on the generalization of hidden layer representations, in which the Jacobian contractive factor is affected. The first-order regularization term ensures the reliability of GAE algorithm, while the second-order factor constrains its distribution onto a smooth manifold.

Conclusion. The proposed regularized AEs provide the best results in limited samples. This is primarily because of the effectiveness of the learned algorithm, particularly for SAR images.

In future, we will study the performance of GAE algorithm by adding more layers; this is difficult in small samples because deep models are more easily overfit. In addition, automatic recognition on a single sample is the next problem that we will attempt to solve using the GAE algorithm.

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell, 2013, 35: 1798–1828

2 Alain G, Bengio Y. What regularized auto-encoders learn from the data-generating distribution. J Mach Learn Res, 2014, 15: 3563–3593

3 Rifai S, Vincent P, Muller X, et al. Contractive auto-encoders: explicit invariance during feature extraction. In: Proceedings of International Conference on Machine Learning (ICML), Bellevue, 2011

4 Evgeniou T, Poggio T, Pontil M, et al. Regularization and statistical learning theory for data analysis. Comput Stat Data Anal, 2002, 38: 421–432

5 Liu H, Yu Y, Sun F, et al. Visual-tactile fusion for object recognition. IEEE Trans Automat Sci Eng, 2017, 14: 996–1008

6 Liu H, Liu Y, Sun F. Robust exemplar extraction using structured sparse coding. IEEE Trans Neural Netw Learn Syst, 2015, 26: 1816–1821

7 Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning (ICML), Helsinki, 2008. 1096–1103

8 Kamyshanska H, Memisevic R. The potential energy of an autoencoder. IEEE Trans Pattern Anal Mach Intell, 2015, 37: 1261–1273

9 Rifai S, Mesnil G, Vincent P, et al. Higher order contractive auto-encoder. In: Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Athens, 2011. 645–660