2018 IEEE Third International Conference on Data Science in Cyberspace

The Feature Representation Ability of Variational AutoEncoder

Chenxi Dong[1,2],Tengfei Xue[2,3],Cong Wang[2,3]
[1]School of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education, Beijing, China
[3]School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
dcx1770@bupt.edu.cn,tffeiba@126.com,wangc@bupt.edu.cn

***Abstract*—As an important generation model, variational autoencoder plays an important role in image feature extraction, text generation, and text compression. In this paper, from the perspective of feature expression, we mainly study the representation ability and stability of variational autoencoder for image features. We extract the features from the original pixels and the normalized pixels of the image respectively. Through the performance of the image classification task, we evaluate the representation ability of the variational autoencoder and compared with the traditional methods of dimensionality reduction principal components analysis, autoencoder. The experiments on multiple datasets prove that variational autoencoder is a new nonlinear dimensionality reduction method, which can represent the data effectively and stably.**

***Keywords*—Variational Autoencoder, Feature Representation, Dimensionality Reduction**

I. Introduction

In recent years, deep learning has made breakthroughs in the field of computer vision, including image classification (AlexNet [1], ZFNet [2], VGG [3], GooLeNet [4], ResNet [5]) and segmentation (R-CNN [6], SPPNet [7], Fast-R-CNN [8], Faster-R-CNN [9], YOLO [10]). The success of deep learning in vision can be attributed to: (a) models with high capacity; (b) increased computational power; and (c) availability of large-scale labeled data [11].

However, compared with the high-cost large-scale labeled dataset, there is all unlabeled data or a lot of unlabeled data with a small part of labeled data in the real world, which leads more and more people to pay attention to unsupervised learning. Therefore, deep generation model for unsupervised learning, which task is learning representation, demonstrates its understanding of the data by generating a new sample, such as variational autoencoder [12, 13] and generative adversarial networks [14]. More and more people focus on the research progress of deep generation model.

As the main task of unsupervised learning, the purpose of dimensionality reduction is to reduce the data complexity as much as possible while preserving the related structural information and the features that contribute to the variance of the data. Autoencoder (AE) uses a 3-layer neural network to reduce the dimension of data and compress data, which can improve the performance of classification, retrieval and other tasks, such as Zhu Z et al. project 3D shapes into 2D space and use autoencoder for feature learning on the 2D images. High accuracy 3D shape retrieval performance is obtained by aggregating the features learned on 2D images [15]; Sun Wenjun et al. realize a deep neural network algorithm by the sparse autoencoder combined with the denoising autoencoder, to achieve unsupervised feature learning for fault diagnosis of induction motors [16]. With an organic combination of variational inference and probability graph models, variational autoencoder has also achieved excellent results in dimensionality reduction and feature extraction. For example, Hjelm R D et al. demonstrate one approach to training Helmholtz machines, variational autoencoders (VAE), as a viable approach toward feature extraction with magnetic resonance imaging (MRI) data [17]; Xu W et al. propose the Semi-supervised Sequential Variational Autoencoder (SSVAE), which improves the classification accuracy compared with pure-supervised classifiers and achieves competitive performance against previous advanced methods [18]; Yunchen Pu et al. use a deep Convolutional Neural Network as an image encoder and a deep Generative Deconvolutional Network as a decoder, what's more, the latent code is linked to generative models for labels (Bayesian support vector machine) or captions (recurrent neural network) [19].

In this paper, we mainly study the representation ability of variational autoencoder for image features. We extract the features from the original pixels and the normalized pixels of the image respectively. Through the performance of the image classification task, we compare with the traditional methods of dimensionality reduction principal components analysis (PCA), autoencoder and evaluate the representation ability and stability of variational autoencoder. Finally, we demonstrate and argue it based on experiments with multiple datasets.

II. Related Work

A. Principal Components Analysis

Principal components analysis is a common method for data analysis. The PCA transforms the original data into a set of linearly independent representations of each dimension through linear transformation, which is used to extract the main feature components of the data and retain the original data information as much as possible. PCA is often used in the dimensionality reduction of high-dimensional data.

978-1-5386-4210-8/18/$31.00 ©2018 IEEE
DOI 10.1109/DSC.2018.00108

B. Autoencoder

Autoencoder is an unsupervised learning algorithm that uses backpropagation to train. The autoencoder encodes the input into a new expression and then decodes it back into the input. the goal of autoencoder is to make the output of the model equal to the input.

In general, the dimension of the hidden layer is much less than the data dimensions. Whats more, there are some restrictions can be added to the network of autoencoder, such as sparse restriction and noise restriction, which can obtain low dimensional expression and sparse coding of raw data.

C. Variational Autoencoder

The variational autoencoder is a directed graph structure, and its graph model is shown in Figure 1. The unobserved continuous variable z is generated from some prior normal distribution $p_\theta = N\left(\mu, \sigma^2\right)$. Let us consider some dataset $X = \left\{x^{(i)}\right\}_{i=1}^{N}$ consisting of N i.i.d. samples of some continuous or discrete variable x. The true posterior density $p_\theta\left(z|x\right)$ is intractable. Therefore, we use a recognition model $q_\phi\left(z|x\right)$, which is an approximation to the intractable true posterior $p_\theta\left(z|x\right)$. We will minimize the KL divergence of approximation from the true posterior. While this KL divergence is zero, $q_\phi\left(z|x\right)$ is equal to $p_\theta\left(z|x\right)$, that is $p_\theta\left(z|x\right) = q_\phi\left(z|x\right)$. We will get the true posterior distribution.

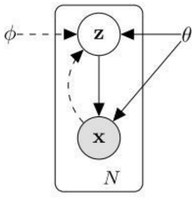

Fig. 1: The Graph Structure of Variational Autoencoder

The KL divergence of approximation from the true posterior $D_{KL}\left(q_\phi\left(z|x\right)||\,p_\theta\left(z|x\right)\right)$ can be written as:

$$\begin{aligned} D_{KL}\left(q_\phi\left(z|x\right)||\,p_\theta\left(z|x\right)\right) &= \int_{-\infty}^{\infty} q_\phi\left(z|x\right)\log\frac{q_\phi\left(z|x\right)}{p_\theta\left(z|x\right)}\mathrm{d}z \\ &= \log p_\theta\left(x\right) \\ &\quad + D_{KL}\left(q_\phi\left(z|x\right)||\,p_\theta\left(z\right)\right) \\ &\quad - E_{q_\phi(z|x)}\left[\log p_\theta\left(x|z\right)\right] \\ &\geq 0 \end{aligned}$$

That is,

$$\log p_\theta\left(x\right) \geq -D_{KL}\left(q_\phi\left(z|x\right)||\,p_\theta\left(z\right)\right) + E_{q_\phi(z|x)}\left[\log p_\theta\left(x|z\right)\right]$$

The right half of the inequality is called the variational lower bound on the marginal likelihood of data x.

$$\begin{aligned} L\left(\theta, \phi; x\right) = &- D_{KL}\left(q_\phi\left(z|x\right)||\,p_\theta\left(z\right)\right) \\ &+ E_{q_\phi(z|x)}\left[\log p_\theta\left(x|z\right)\right] \end{aligned} \quad (2\text{-}1)$$

The first term $D_{KL}\left(q_\phi\left(z|x\right)||\,p_\theta\left(z\right)\right)$ of Eq. (2-1) can be integrated analytically, and the second term $E_{q_\phi(z|x)}\left[\log p_\theta\left(x|z\right)\right]$ requires estimation by sampling: Firstly, we reparameterize the approximation $q_\phi\left(z|x\right)$ using a differentiable transformation $g_\phi\left(x, \epsilon\right)$ of an auxiliary noise variable ϵ; Secondly, we estimator $E_{q_\phi(z|x)}\left[\log p_\theta\left(x|z\right)\right]$:

$$E_{q_\phi(z|x)}\left[\log p_\theta\left(x|z\right)\right] = \frac{1}{M}\sum_{m=1}^{M}\log p_\theta\left(x|z^m\right)$$

where $z^m = g_\phi\left(x, \epsilon^m\right), \epsilon^m \sim N\left(0, I\right)$.

For the parameters ϕ and θ of Eq. (2-1), we can use a fully connected neural network or a convolutional neural network to estimate them. What's more, parameters can be updated using SGD, AdaDelta [20], Adagrad [21], Adam [22] and so on.

III. The Representation Ability of Variational Autoencoder

In this paper, we mainly study the representation ability and stability of variational autoencoder for image features. Through the performance of the image classification task, we evaluate the effectiveness of feature extraction and compare with the traditional methods of dimensionality reduction PCA, AE.

A. The Variational Constraint

Traditional autoencoder reconstructs the model input through the decoder network to learn the representation of the data. Therefore, the most work of autoencoder is to ”remember” or ”copy” the input of the model without generating a new sample.

In order to establish a generation model better, compared with autoencoder, variational autoencoder adds a variational constraint to latent variable z, that is z obeys gaussian distribution. then samples from the distribution, and generates a new sample through the decoder network.

After adding the variational constraints, the variational autoencoder maps the model input to a latent variable which obeys gaussian distribution through the encoder network. Compared with the autoencoder, the parameters of gaussian distribution about the same type of input after mapping are approximately the same, there are quite different parameters of gaussian distribution if the model input belongs to different types. So, the variational autoencoder sample from gaussian distributions of different parameters and generate different new data through the decoder network.

Since variational autoencoders can get gaussian distribution with different parameters from different types of inputs, the latent variables have obvious differences between different categories. They often achieve better results in other tasks such as classification and retrieval.

B. The Feature Representation of Variational Autoencoder

In general, as for variational autoencoder, the dimensions of the latent variable are much smaller than the dimensions of the original data. Whats more, if the latent variable has different dimensions, the expression ability of variational autoencoder will be different. In this paper, we realize three kinds of representation methods (PCA, AE, VAE) in the different dimensions of the latent variable, and extract the feature of the original data respectively. Through the performance of the image classification task with the new feature, we compare the ability of each method to extract features under latent variable of different dimensions and evaluate the representation ability of variational autoencoder.

As for variational autoencoder that has been trained completely, $q_\phi(z|x)$ is very close to the true posterior distribution $p_\theta(z|x)$, that is $D_{KL}(q_\phi(z|x)\,||\,p_\theta(z|x))$ closes to zero.

In this way, we can use the encoder network $q_\phi(z|x)$ to reduce the dimension of the real dataset $X = \left\{x^i\right\}_{i=1}^{N}$ that is close to the true low-dimensional distribution.

Since the posterior distribution $q_\phi(z|x)$ is close to the true posterior distribution $p_\theta(z|x)$, the latent variable maybe gets better results than traditional dimensionality reduction method. At present, variational autoencoder has achieved good results in extracting features from a variety of tasks, including medical image feature extraction and text classification.

IV. Experiments

In order to explore the effectiveness and stability of representation ability of variational autoencoder, we use variational autoencoder, principal component analysis, and autoencoder to extract features of the same dimension from multiple datasets (MNIST, SVHN, CIFAR10), and then evaluate their performance on the nearest neighbor classifier. Not only that, we visualize the features extracted by the three methods on the MNIST dataset and compare the effects of the extracted features more intuitively and clearly.

A. Visualization of MNIST Feature

The original pixel of the MNIST dataset has 784 dimensions. We extract its 20-dimensional feature by three methods and randomly select 5000 samples of the 10,000 test sets for visual analysis using the t-SNE [23] method (Figure 2).

Comparing the three images, we can see that among the features extracted by principal component analysis, there are many data that overlap with 9 and 5, which may lead to misclassification. The boundaries of feature categories extracted by the autoencoder are more clear, but there are still some categories that overlap, features extracted by autoencoder are more discriminating than principal component analysis. However, compared with the former two, the features extracted by variational autoencoder have the clearest boundary among the categories, and there is almost no overlap between different categories, indicating that variational autoencoder extracts the best low-dimensional features. That confirms that variational autoencoder can effectively extract the features of the data.

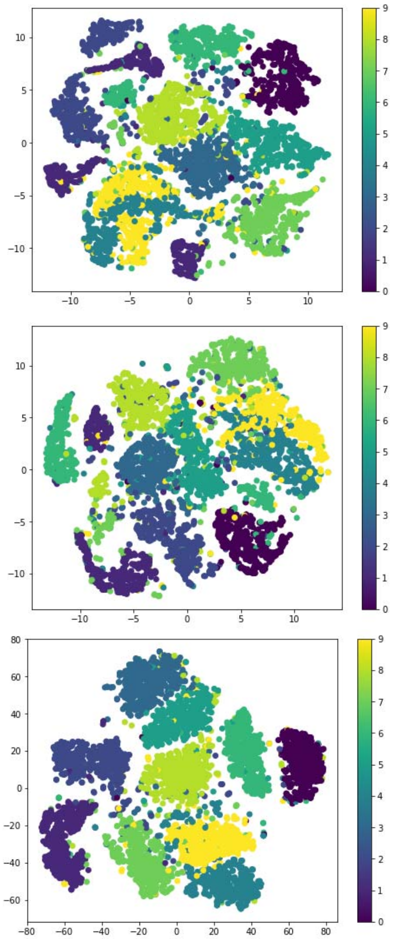

Fig. 2: Visual Analysis of MNIST Datasets;
From Top to Bottom: PCA, AE, VAE.

B. The Performance of Base Classifier

At present, there are mainly two types of model input for deep learning in image: original pixel and normalized pixel. In order to further explore the representation ability of variational autoencoder, we separately extract the latent variables of different dimensions for these two different kinds of data and then evaluate the influence of the dimensions of the latent variables on the representation ability of variational autoencoder. Moreover, we study the effects of variational autoencoder on these types of data.

1) Original Pixels: In this paper, we extract the features of MNIST from 20 to 500 dimensions respectively and compare the classification results of MNIST by the nearest neighbor classifier. With the performance of the classifier, we evaluate the representation ability of variational autoencoder.Since the intrinsic dimension of MNIST data is estimated to be 13 by the maximum likelihood method, we extract the 13-dimensional features of the MNIST data at the same time.

In this paper, we use 50000 train samples to train principal component analysis, autoencoder and variational autoencoder respectively, then we extract the features of train and test sets using the trained model. Through the performance of the nearest neighbor classifier with new features of 10000 test samples, we compare the effect of feature extraction with three methods. Among them, autoencoder and variational autoencoder are trained 10 times by Adam [22] algorithm respectively, and we compare the average effect of multiple experiments. In addition, in order to prove the accuracy of the experimental results, we analyze the classification results of the numbers ”1” and ”2” in the MNIST dataset at the same time, and evaluate the stability of our model. With the numbers ”1” and ”2”, the classification performance are more stable. The classification results of the nearest neighbor classifiers in category ”1”, ”2” and all categories are shown in Figure 4.2, in which autoencoder and variational autoencoder take multiple average effects.

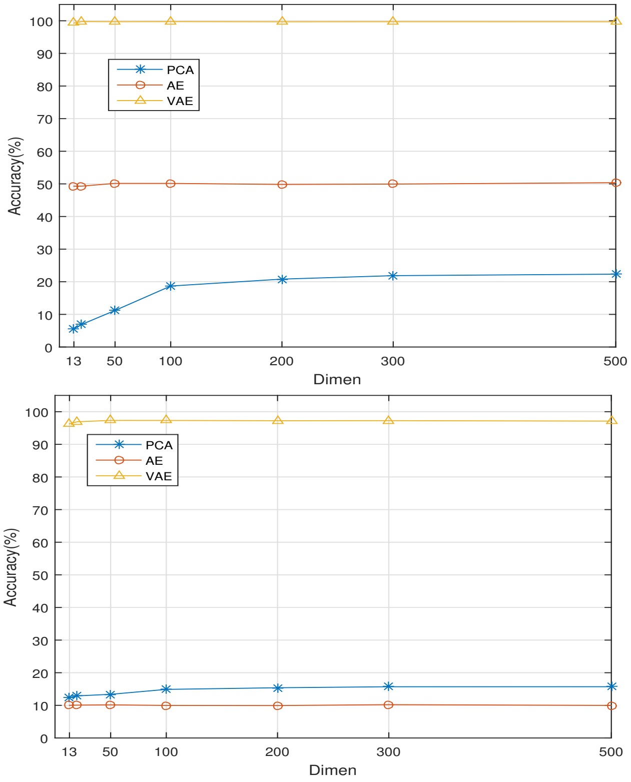

Fig. 3: The Result of The Nearest Neighbor Classifier on The Features Extracted by The Three Methods;
Top: The Result on Numbers 1 and 2;
Bottom: The Result on All Numbers.

In comparison with Figure 3, we can see that principal component analysis has relatively low accuracy in the classification results of numbers ”1”, ”2” and all numbers, and the classification effect is closely related to the dimension of features. The classification effect of autoencoder is better in the number ”1” and ”2”, but as for all numbers, there is the worst result among the three methods, indicating that the features extracted by autoencoder have the small difference in different categories. Autoencoder does not extract the main features of various type data. The variational autoencoder has the best classification results for all methods. With the increase of the dimension, the accuracy of the classification improves slightly, but it keeps the relatively stable effect all the time. This shows that the effectiveness and stability of variational autoencoder representation ability from the side.

TABLE I: Classification Results on SVHN (%)

	PCA		AE		VAE	
Dimen	**Average**	**Best**	**Average**	**Best**	**Average**	**Best**
500	13.29	13.31	40.35	40.78	47.81	48.32
600	13.31	13.34	41.43	41.70	52.08	52.95
800	13.30	13.32	41.48	41.73	54.72	55.52
1000	13.28	13.30	43.22	43.56	57.01	57.83

TABLE II: Classification Results on CIFAR10 (%)

	PCA		AE		VAE	
Dimen	**Average**	**Best**	**Average**	**Best**	**Average**	**Best**
500	13.29	13.31	40.35	40.78	47.81	48.32
600	13.31	13.34	41.43	41.70	52.08	52.95
800	13.30	13.32	41.48	41.73	54.72	55.52
1000	13.28	13.30	43.22	43.56	57.01	57.83

2) Normalized pixels: SVHN and CIFAR10 are more complex than the MNIST dataset and have 3072 dimensions, which is relatively high. In this paper, we extract the feature of two datasets from 500 to 1000 dimensions respectively, and compare their classification results by the nearest neighbor classifier, then evaluate the representation ability of variational autoencoder to normalized pixel data. In this paper, we first divide the original pixels of all the samples in both datasets by 255 and map the original pixels into the range of 0 to 1, which is to normalize the original pixels, then we use all the train sets (SVHN: 73257; CIFAR10: 50000) to train PCA, AE, VAE. After the training is completed, we extract the features of train and test sets and compare the effect of the extracted features on the nearest neighbor classifier. Among them, autoencoder and variational autoencoder both train 5 times, and we compare the average and the best results respectively among them. The classification results of the nearest neighbor classifiers on the two datasets are shown in TABLE I and TABLE II.

Comparing the classification results in TABLE I and TABLE II, we can find that the nearest neighbor classifier has the best result for the variational autoencoder whether it is SVHN or CIFAR10, the autoencoder is second and the principal component analysis has the worst effect. What's more, with the increase of the dimension, the classification results of the nearest neighbor classifier are maintained at a relatively stable state. These shows that compared with autoencoder and principal component analysis, variational autoencoder can extract the features of the normalized pixel data effectively and steadily. This further confirms the validity and stability of the representation ability of variational autoencoder.

V. Conclusion

In this paper, we investigate the representation ability of variational autoencoder that is a kind of generative model and compare the influence of the dimension of latent variables on its the representation ability. At the same time, we extract the features from the original pixels and normalized pixels of the image. Through the performance of new features on the nearest neighbor classifiers, we evaluate the representation ability of variational autoencoder compared with autoencoder and principal component analysis. The results of experiments on multiple datasets show that the representation ability of variational autoencoder is more efficient than the principal component analysis and autoencoder, and it also confirms that variational autoencoder has stable representation ability.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (No.2017YFC1307705).

References

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[2] Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.

[3] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015, June). Going deeper with convolutions. Cvpr.

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[6] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2014, September). Spatial pyramid pooling in deep convolutional networks for visual recognition. In european conference on computer vision (pp. 346-361). Springer, Cham.

[8] Girshick, R. (2015, December). Fast R-CNN. In Computer Vision (ICCV), 2015 IEEE International Conference on (pp. 1440-1448). IEEE.

[9] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[10] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[11] Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017, October). Revisiting unreasonable effectiveness of data in deep learning era. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 843-852). IEEE.

[12] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.611

[13] Rezende, D. J., Mohamed, S., & Wierstra, D. (2014, January). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In International Conference on Machine Learning (pp. 1278-1286).

[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

[15] Zhu, Z., Wang, X., Bai, S., Yao, C., & Bai, X. (2016). Deep learning representation using autoencoder for 3D shape retrieval. Neurocomputing, 204, 41-50.

[16] Sun, J., Shao, Y., & Yan, Q. (2016). Induction motor fault diagnosis based on deep neural network of sparse auto-encoder. Journal of Mechanical Engineering, 52(9), 65-71.

[17] Hjelm, R. D., Plis, S. M., & Calhoun, V. C. (2016). Variational Autoencoders for Feature Detection of Magnetic Resonance Imaging Data. arXiv preprint arXiv:1603.06624.

[18] Xu, W., Sun, H., Deng, C., & Tan, Y. (2017, February). Variational Autoencoder for Semi-Supervised Text Classification. In AAAI (pp. 3358-3364).

[19] Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., & Carin, L. (2016). Variational autoencoder for deep learning of images, labels and captions. In Advances in neural information processing systems (pp. 2352-2360).

[20] Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.

[21] Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12(Jul), 2121-2159.

[22] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[23] Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), 2579-2605.