Neurocomputing 248 (2017) 76–79

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Roles of pre-training in deep neural networks from information theoretical perspective

Yasutaka Furusho, Takatomi Kubo, Kazushi Ikeda*

*Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan*

ARTICLE INFO

*Article history:*
Received 28 June 2016
Revised 11 November 2016
Accepted 17 December 2016
Available online 8 March 2017

*Keywords:*
Deep neural networks
Pre-training
Information theory

ABSTRACT

Although deep learning shows high performance in pattern recognition and machine learning, the reasons remain unclarified. To tackle this problem, we calculated the information theoretical variables of the representations in the hidden layers and analyzed their relationship to the performance. We found that entropy and mutual information, both of which decrease in a different way as the layer deepens, are related to the generalization errors after fine-tuning. This suggests that the information theoretical variables might be a criterion for determining the number of layers in deep learning without fine-tuning that requires high computational loads.



## 1. Introduction

Deep learning, which is a multi-layered neural network, has been changing the history of pattern recognition and machine learning in accuracy performance [1] and can be applied to computer vision, automatic speech recognition and translation, and so on [2,3]. However, the reasons for its high performance remain relatively unclarified since layered models have singular points that are difficult to treat statistically [4,5]. In addition, the performance improved by combining several heuristics, such as drop-out [10] and pre-training [1,6,7].

Pre-training is an unsupervised learning algorithm that improves the classification performance. How pre-training affects classification performance through fine-tuning has been studied in several ways [7–9], suggesting that it works as a kind of regularization that resembles manifold learning.

Since regularization is determined by representation, we focused on it in the hidden layers in this paper. Representation is transformed layer by layer through the connections set by pre-training. In other words, a pre-trained network works as an encoder from the information theoretical viewpoint. Consider the simplest case where the network respectively has input, hidden and output layers with $I$, $J$ and $I$ nodes, and the following activation functions are linear:

$$h_j = \sum_{i=1}^{I} w_{ji} x_i, \qquad j = 1, \ldots, J, \tag{1}$$

$$y_i = \sum_{j=1}^{J} v_{ij} h_j, \qquad i = 1, \ldots, I, \tag{2}$$

as shown in Fig. 1. Then minimizing the squared errors, $\sum_i |x_i - y_i|^2$, leads to principle component analysis (PCA) and projects vector $x = \{x_i\}$ onto the space spanned by the $J$ principal components of the training samples. This autoencoder with a linear activation function and is equivalent to the maximum entropy method under mild conditions. In fact, the autoencoder maximizes the mutual information between a layer and the succeeding layer for each local connection [11].

In discussion from the information theoretical viewpoint, another key is the discriminability among classes. Since deep neural networks are mainly used as classifiers, the labels should be taken into account. The key is the variance among the classes and the variance within them in Fisher discriminant analysis (FDA) that assumes all the distributions are homocedastic Gaussian [12]. In the framework of the information theory, the variance within a class is the conditional entropy of the data that are given labels to be minimized and the variance among the classes is the mutual information to be maximized. Such information theoretical variables give the following lower bound of the classification error probability:

$$1 - \frac{I(h,t) + 1}{H(t)}, \tag{3}$$

which is known as Fano's inequality [13].

Although the results above suggest the relationship between information theory and representation, the representation of pre-training has not been studied quantitatively from the information theoretical viewpoint. In this paper, we investigated how

* Corresponding author.
*E-mail addresses:* kazushi@is.naist.jp, kazushi.ikeda@ieee.org (K. Ikeda).

http://dx.doi.org/10.1016/j.neucom.2016.12.083
0925-2312/

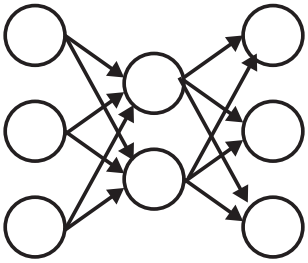


**Fig. 1.** Linear neural network.

pre-training affects the "encoding" of the input data in the hidden layers by evaluating the information theoretical variables of the representation and the labels. We trained a deep neural network with popular training methods using actual data (MNIST, NORB) and calculated the conditional entropy and the mutual information in each of the hidden layers. We found that the entropies of the representations of the data and their labels remain high in the early layers and decrease in the late layers and the mutual information between the representations and the labels as well as the prediction performances decrease in the earlier layers. This implies that information theoretical variables may be an alternative criterion for cross-validation to determine the number of layers in deep learning since they do not need fine-tuning that requires high computational loads.

## 2. Material and methods

### 2.1. MNIST database

We used the MNIST database for our experiments [14]. Each image was decimated to one fourth of its original (196 pixels) by merging four pixels to one to reduce the complexity for brevity. We divided 70,000 images in ten categories to three sets for training (50,000 images), validation for early-stopping (10,000 images) and test for evaluation (10,000 images).

### 2.2. NORB database

The NORB database was also used for our experiments [15]. Each image was decimated to 144 pixels to reduce the complexity for brevity. 48,600 images in five categories were divided into three sets for training (20,000 images), validation for early-stopping (4300 images) and test for evaluation (24,300 images).

### 2.3. Deep neural network

Our deep neural network has one input layer (196 nodes for MNIST and 144 nodes for NORB), six hidden layers (150, 120, 90, 60, 30, 10 nodes for MNIST and 120, 100, 80, 60, 40, 20 nodes for NORB, respectively) and one output layer (ten nodes for MNIST and five nodes for NORB), corresponding to the number of pixels and categories for each dataset (Fig. 2). The activation function of each node in the hidden layers was the sigmoid, while the category was determined by winner-take-all in the output layer.

### 2.4. Learning algorithms

Pre-training is the essence of deep learning. Among the restricted Boltzmann machine [1], the autoencoder [6] and their variants [16], we chose the following two representative methods below:

1. Deep Belief Network (DBN) [1]
2. Stacked Autoencoder (SAE). [6]

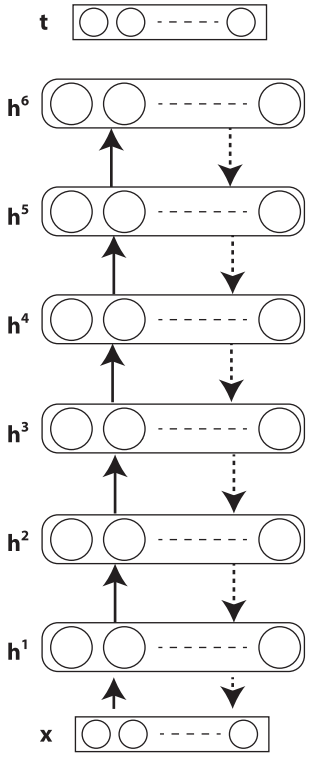



**Fig. 2.** Structure of our deep neural network.

After pre-training with either of the above, the network was fine-tuned using the stochastic gradient method with early-stopping using validation data.

### 2.5. Evaluation

After pre-training and fine-tuning, we calculated the information theoretical variables and the prediction errors of each algorithm at initialization to identify the relationship between the information theoretical variables and the prediction performance.

Here, the information theoretical variables here were the conditional entropy and the mutual information:

$$H(h^i|t) = \mathrm{E}\big[-\log_2 P(h^i|t)\big], \tag{4}$$

$$I(h^i, t) = \mathrm{E}\left[\log_2 \frac{P(h^i, t)}{P(h^i)p(t)}\right], \tag{5}$$

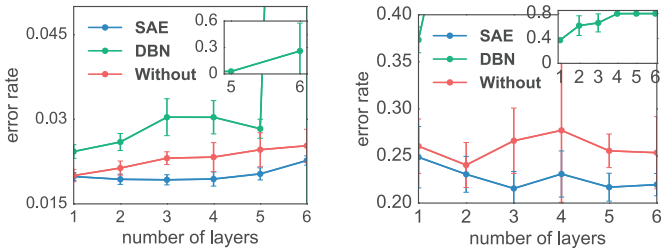
which were calculated for the deep neural network with each algorithm, where $h^i = \{h^i_j\}$ was the activation of the $j$th node in the $i$th layer and binarized to 0 or 1 using the threshold 0.5.

We also evaluated that the prediction error of the deep neural network as a classifier. Here, the prediction error of each layer was calculated using a linear classifier that had a layer as input and trained it using fine-tuning.

## 3. Results

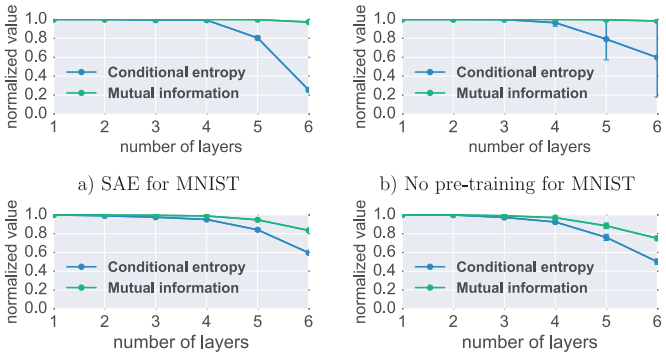
### 3.1. Prediction error of classification

We compared the prediction errors of the classifiers made from the hidden layers that were pre-trained with SAE, DBN, or neither (Fig. 3). SAE improved the performance, but DBN did not, suggesting that both of the classification problems in this study were unsatisfactory for DBN. Hence, we omitted DBN from the following analysis.




a) MNIST b) NORB

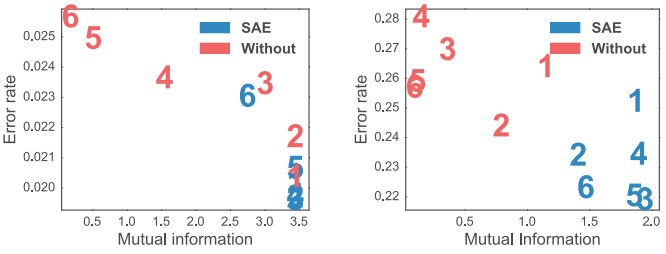
**Fig. 3.** Prediction errors of the classifiers.




a) SAE for MNIST b) No pre-training for MNIST

c) SAE for NORB d) No pre-training for NORB

**Fig. 4.** Information theoretical variables of the classifiers.




a) MNIST b) NORB

**Fig. 5.** Prediction error vs. mutual information before fine-tuning.

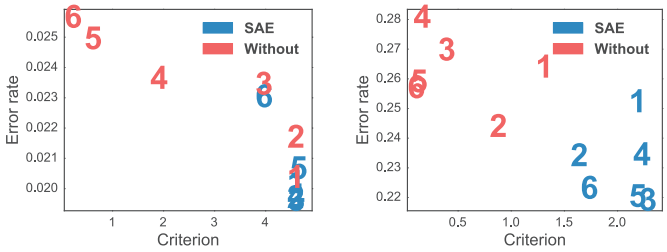
For the MNIST dataset, the classifier that has the third hidden layer as an input had the minimum prediction error when it was pre-trained with SAE, but it was comparable with the others, except that at the sixth hidden layer. The same tendency was found in experiments with the NORB database. The mutual information between the representations and the labels decreased as the layer deepened, as with the conditional entropy, too (Fig. 4). However, although pre-training only slightly affected the mutual information, it did affect the conditional class entropy.

## 4. Discussion

### 4.1. Entropy, mutual information and performance

Mutual information $I(h^i, t)$ expresses the amount of information of labels $t$ included in representation $h^i$. Hence, $I(h^i, t)$ and the prediction error are negatively correlated ($R = -0.48$ for MNIST, $R = -0.58$ for NORB), where the greater $I(h^i, t)$ is, the smaller the prediction error is (Fig. 5).

However, this ignores the discriminability of the labels considered in the FDA. Since conditional entropy $H(h^i|t)$ is a kind of variance within the classes and the mutual information $I(h^i, t)$ corresponds to the variance among them, we plotted the performance




a) MNIST b) NORB

**Fig. 6.** Prediction error vs. normalized mutual information before fine-tuning.

vs. ratio $I(h^i, t)/(H(h^i|t)/H(h^i))$, where $H(h^i|t)$ was normalized by total entropy $H(h^i)$ (Fig. 6). We found that the normalized mutual information is negatively correlated to the prediction error.

Note that the information theoretical measures were calculated before fine-tuning. Since fine-tuning requires high computational loads, these measures predicted the classifier's performance with less computational complexity.

### 4.2. Differences in pre-training algorithms

DBN shows a much larger prediction error than SAE or even without pre-training (Fig. 3), probably because DBN is based on energy minimization, while SAE explicitly considers encoding and the mutual information. In addition, the datasets treated here are rather simple and can be presented even in deeper layers (Fig. 4). These results imply that the learning algorithm should consider keeping the information of the data in such a case.

## 5. Conclusion

We analyzed the representations in the hidden layers of deep neural networks from the information theoretical viewpoint. We calculated the mutual information and the conditional entropy in each of the hidden layers when deep neural networks were pre-trained with SAE or DBN, or not trained and found that the entropies of the representations of the data and their labels remain high in the early layers and decrease in the late layers, while the mutual information between the representations and the labels decreases in the earlier layers. The mutual information and the normalized mutual information are related to the prediction error, suggesting that information theoretical variables may work as a criterion for model selection with less computation than accuracy-evaluation based criteria.

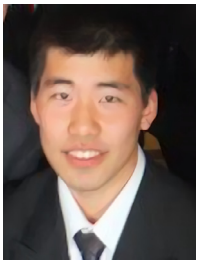
## Acknowledgments

This work was supported in part by JSPS KAKENHI 25280083, 25118019, 15H01620.

## References

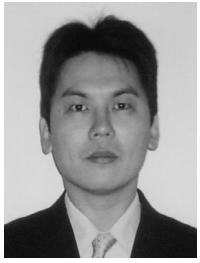
[1] G.E. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, Neural Comp. 12 (2006) 531–545.
[2] F. Seide, G. Li, D. Yu, Conversational speech transcription using context-dependent deep neural networks, in: Proceedings of the International Speech Communication Association INTERSPEECH (2011) 437–440.
[3] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proceedings of the European Conference on Computer Vision (2014) 818–833.
[4] S. Watanabe, Algebraic analysis for nonidentifiable learning machines, Neural Comp. 13 (2001) 899–933.
[5] K. Fukumizu, S. Akaho, S. Amari, Critical lines in symmetry of mixture models and its application to component splitting, Neural Inf. Process. Syst. 15 (2003) 857–864.
[6] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, Greedy layer-wise training of deep networks, Neural Inf. Process. Syst. 19 (2007) 153–160.

[7] D. Erhan, Y. Bengio, A. Courville, P.A. Manzagol, P. Vincent, S. Bengio, Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. 11 (2010) 625–660.
[8] H. Larochelle, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, S. Bengio, Exploring strategies for training deep neural networks, J. Mach. Learn. Res. 10 (2009) 1–40.
[9] Y. Bengio, Deep learning of representations for unsupervised and transfer learning, J. Mach. Learn. Res. 27 (2012) 17–36.
[10] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neuralnetworks by preventing co-adaptation of feature detectors, 2012, arxiv:1207.0580.
[11] P. Vincent, H. Larochella, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked demonising autoencoders:learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.
[12] L. Devroye, L. Györfi, G. Lugosi, A probabilistic theory of pattern recognition, in: Applications of Mathematics, Springer, 1996.
[13] T. Cover, J. Thomas, Elements of Information Theory, Wiley-Interscience, 2006.
[14] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2324.
[15] Y. LeCun, F. Huang, L. Bottou, Learning methods for generic object recognition with invariance to pose and lighting, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004 (2004) 97–104.
[16] S. Rifai, P. Vincent, X. Muller, X. Glorot, Y. Bengio, Contracting auto-encoders: explicit invariance during feature extraction, in: Proceedings of the International Conference on Machine Learning (2011) 833–840.



**Yasutaka Furusho** received his B.E. from National Institute of Technology, Kumamoto College, Japan, in 2015. He is currently working toward his M.E. at Nara Institute of Science and Technology in Nara, Japan. His research interests include analysis of neural network based on information theory, and its application to model selection of neural network.

**Takatomi Kubo** received his B.M. from Osaka University, Japan, in 2002 and his D.E. degree from Graduate School of Information Science, Nara Institute of Science and Technology, in Nara, Japan, in 2012. He had five years of medical experience in neurology. He completed a Human Resource Development Program for Medical Device Development Coordinators organized by Kobe University and Kyoto University, in Japan, in 2010. He is currently an Associate Professor on Project at Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include modeling and prediction of driving behaviors, neural engineering, and developing a communication device for dysarthric patients.



**Kazushi Ikeda** received his B.E., M.E., and Ph.D in Mathematical Engineering and Information Physics from the University of Tokyo in 1989, 1991, and 1994. He was a research associate with the Department of Electrical and Computer Engineering of Kanazawa University from 1994 to 1998. He was a research associate of Chinese University of Hong Kong for three months in 1995. He was with Graduate School of Informatics, Kyoto University, as an associate professor from 1998 to 2008. Since 2008, he has been a full professor of Nara Institute of Science and Technology. He was the editor-in-chief of the Journal of the Japanese Neural Network Society, and is currently an action editor of Neural Networks, and an associate editor of IEEE Transactions on Neural Networks and Learning Systems.