Deep Learning Technology for Small Data

SATO Atsushi

Abstract

The recent emergence of deep learning has brought significant improvements in the accuracy of pattern recognition technologies including that of image recognition. Although training a large amount of data is required in order to achieve a high accuracy, the preparation of such large data amounts is often difficult in applications to real problems. Issues in how to improve accuracy with limited amounts of data are thereby created. This paper introduces two technologies developed for the effective deep learning of a small amount of training data. One is the layer-wise adaptive regularization" method that sets the regularization strength. This varies depending on the layer, according to the structure of the deep-layer network (a deep neural network). The other method is the "adversarial feature generation" that performs training in the middle layers by generating hard-to-recognize features. This paper demonstrates their validity through experiments on the public datasets for handwritten digit recognition (MNIST) and general object recognition (CIFAR-10).

Keywords

Deep learning, neural network, regularization, adversarial data generation

1. Introduction

The recent emergence of deep learning has brought significant improvement in the accuracy of pattern recognition technologies including image recognition. To achieve a high recognition accuracy by means of deep learning, it is required to prepare a large amount, thousands or tens of thousands of training samples that are composed of the input patterns and their correct answers. Those that are suitable for training (are hereinafter referred to as "training data"). However, when the application to real problems is considered, various factors make it hard to prepare such a large amount of training data. For example, abnormal data with low occurrence frequencies takes a long period for data collection, and it is only the specialist physicians who can give correct answers to medical data. From the viewpoint of early go-live, it is difficult to be capable of reserving sufficient time for collecting and building up the required large amount of training data. As a result, it has become a critical issue for expanding applications of deep learning technology that as high as possible accuracy even with a small amount of training data is achieved.

When the amount of training data is small, excessive fitting to the trained data makes the phenomenon called "overfitting" noticeable, by which the accuracy of non-trained data drops. In typical deep learning, the overfitting is reduced by the regularization that applies restrictions so as to decrease the sum of squares of the weighting parameters of the deep neural network. This has however resulted in the issue with which the mixed presence of the layers troubled by under-fitting due to excessive regularization and those troubled with overfitting due to weak regularization will limit the accuracy improvement. The technique called the data augmentation is often used for an artificial data increase by rotating images or changing their sizes. However, this technique is not always capable of generating data that can contribute to the accuracy improvement.

The present paper introduces two technologies that have been developed for effective deep learning that can solve the issues posed in the training of small data, called the layer-wise adaptive regularization and the adversarial feature generation. The validities of these technologies are also demonstrated through evaluation experiments using public datasets for handwritten digit

recognition (MNIST) and general object recognition (CIFAR-10).

2. Layer-wise Regularization

2.1 Deep learning and overfitting

This section gives a simple description of the mechanism of deep learning (**Fig. 1**). First, the training data composed both of data and the correct answers are prepared in advance. Then, the connections of neurons should be tuned so that the output from the network after data input matches the correct answers. This action is called the training, and the training will be provided to each training data until the error between the output and the correct answer becomes sufficiently small.

If the amount of training data is small, a phenomenon

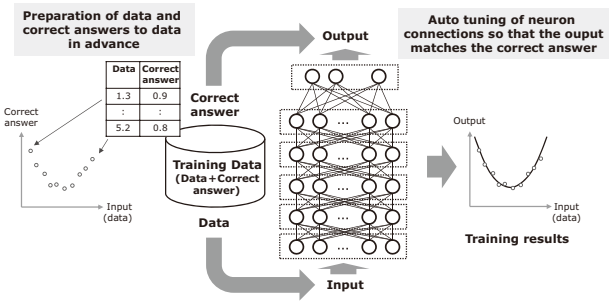

Fig. 1 Mechanism of deep learning input.

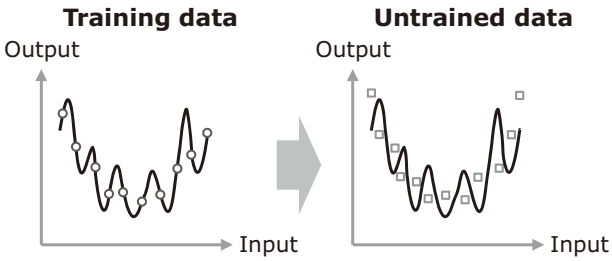

Fig. 2 Example of overtraining.

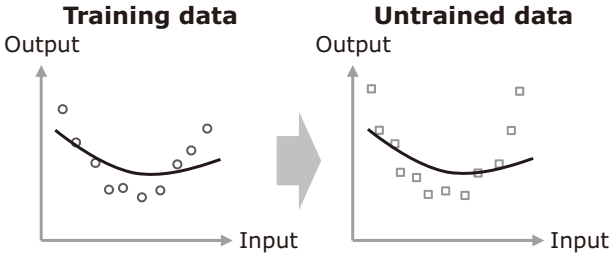

Fig. 3 Example of excessive regularization.

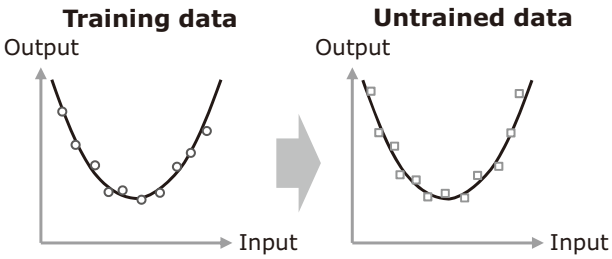

Fig. 4 Example of optimum regularization.

called the overfitting becomes noticeable, with which overfitting to even the noise contained in the data may deteriorate the accuracy for the untrained data (**Fig. 2**). The conventional technology for reducing it is the L2 regularization. This technique however brings about minimization of square sum of the trained parameters (connection weights with the deep learning technology) as well as the errors between correct answers and outputs. This can decrease overfitting by preventing the parameter values from increasing excessively. However, on the other hand, too strong regularization hinders the advancement of training and drops the adaptability to data (**Fig. 3**). Therefore, to achieve a high accuracy, it is critical to set the optimum strength for the regularization (**Fig. 4**).

2.2 Issues of the conventional L2 regularization

Deep learning is a method of training a neural network with a deep layered structure. It updates the connection weight of each layer by propagating the error between the output and the correct answer to the layer above it. This technique is referred to as back propagation, by which the updating is performed as shown below;

$$w_i \leftarrow w_i - \mu(\Delta w_i + \lambda w_i)$$

w_i being the connection weight of the i-th layer, Δw_i is the gradient with respect to w_i calculated by back propagation, μ is the training ratio determining the scale of updating, and λ is the regularization coefficient determining the L2 regularization strength. Δw_i acts as the accelerator for advancing the training so as to minimize the error between the output and the correct answer, and λw_i acts as the brake for reducing the accelerating action. Consequently, in order to obtain an optimum regularization effect, value λ should be set to balance Δw_i and λw_i appropriately.

When Δw_i is calculated with the back propagation of error, the scale of propagation varies depending on the

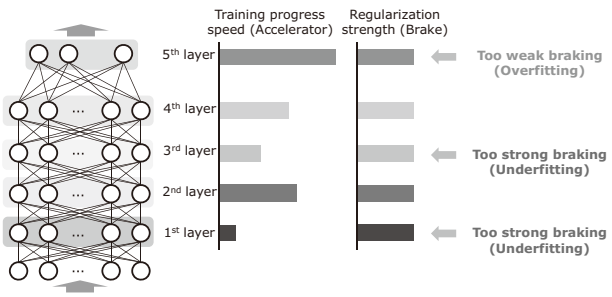

Fig. 5 Issues in conventional L2 regularization.

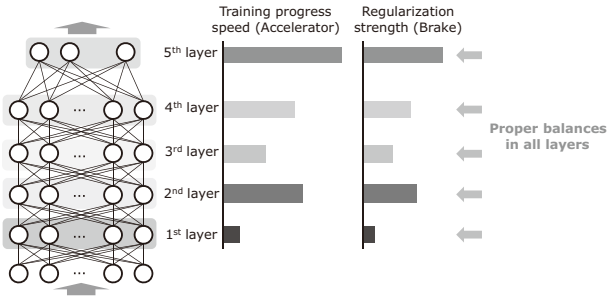

Fig. 6 Effect of adaptive regularization of each layer.

network structure in the upper layers than the *i*-th layer. On the other hand, since the value of λw_i is determined exclusively depending on the connection weights of the *i*-th layer, the balance between the two values varies between layers. Nevertheless, since the conventional deep learning uses the same λ in all of the layers, the balance is different between the layers. This means that a mixed presence of the layers with excessive regularization and those with insufficient regularization results. This issue becomes more serious as the layer depth increases (**Fig. 5**).

2.3 Layer-wise adaptive regularization

To solve the issue described in section 2.2, the authors developed a technology called the layer-wise regularization. This technology determines the optimum regularization of the coefficient layer by layer, so that the ratio between the gradient and the regularization term is constant in each layer as shown below;

$$\frac{|\lambda_i w_i|}{|\Delta w_i|} = c \quad \text{that is} \quad \lambda_i = c\frac{|\Delta w_i|}{|w_i|}$$

λ_i being the regularization coefficient of an *i*-th layer, and c is the constant independent from the layer. However, because the gradient scale $|\Delta w_i|$ is unknown before training, it is not possible to obtain λ_i directly. Therefore, the ratio between the gradient scales of adjacent layers is estimated and the ratio between regularization coefficients is estimated based on the former ratio. This means that, once the regularization coefficient of the last layer is determined, the regularization coefficients of the other layers can be determined automatically and optimally according to the obtained ratio (**Fig. 6**). This method requires the regularization coefficient tuning of only the last layer. This means that the regularization coefficient of each layer can be determined adaptively with a similar amount of labor to the conventional L2 regularizaiton[1)].

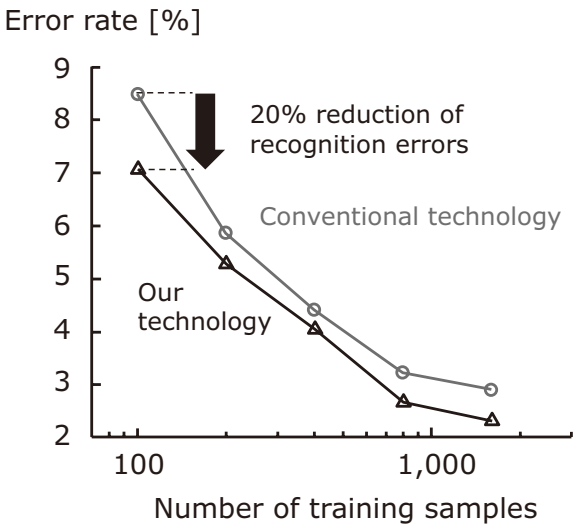

Fig. 7 Comparison between layer-wise regularization and conventional L2 regularization.

2.4 Experiment

The authors evaluated the validity of this technology by an experiment using the MNIST handwritten digits dataset for comparison of the recognition accuracy with the conventional L2 regularization (**Fig. 7**). The horizontal axis represents the number of training samples and the vertical axis the error rate with respect to the test data. The graph confirms the effectiveness of this technology that can reduce the error rate from that of the conventional technology by nearly 20%.

3. Adversarial Feature Generation

3.1 Traditional data augmentation

With the image recognition, the significance of objects in the image do not change even when the image is somewhat distorted. In order to deal with this issue,

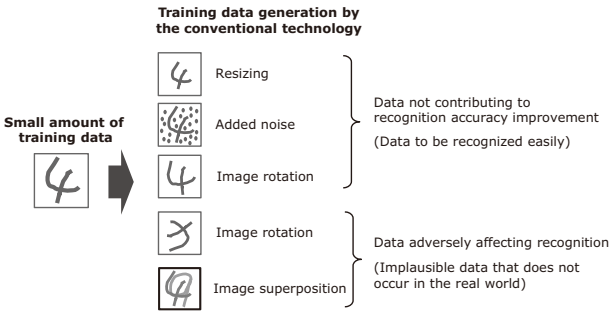

Fig. 8 Traditional data augmentation.

the deep learning often uses a technique called the data augmentation in order to increase data artificially by rotating or resizing the image.

It is effective for the recognition accuracy improvement to train a large amount of hard-to-recognize data. However, the improvement by the data augmentation has been limited because it does not always generate such types of data. In addition, it has been necessary to have specialists adjust the data generation method, so that the generated data does not exert adverse effects depending on the types of data such as image and audio (**Fig. 8**).

3.2 Adversarial feature generation

To solve the issue described in section 3.1, the authors developed a technology called the "adversarial feature generation". This technology generates hard-to-recognize training data artificially by intentionally varying the features obtained in the middle layers of deep neural networks, assuming that the output from a randomly selected middle layer is *h* and the network output for it is *g(h)*;

$$\boldsymbol{r}_{max} = \underset{\boldsymbol{r}}{\operatorname{argmax}} \; KL(g(\boldsymbol{h}), g(\boldsymbol{h}+\boldsymbol{r}))$$

r_{max} is the perturbation added to feature *h*. In the formula above, *KL* is the Kullback-Leibler divergence that decreases as the similarity between two values increases. Obtaining r to maximize it means that output *g(h + r)* for feature *h + r* should differ greatly from the output before perturbation addition *g(h)*. Since the data that produces most different output is regarded as hard-to-recognize data, the obtained $h + r_{max}$ is called the adversarial feature. However, since applying no restriction to the scale of *r* results in producing a perturbation having an adverse effect, r_{max} is obtained by applying the following restriction.

$$\text{subject to } \|\boldsymbol{r}\| \leq \epsilon \|\boldsymbol{h}\|$$

Here, ϵ>0 is the parameter defining the scale of *r* and should be set in advance. Since an adversarial feature in accordance with input data is generated every time data is input, the recognition accuracy can be improved by performing training so that the adversarial features can be recognized correctly[2].

Traditionally, the training data has been increased artificially by deforming the input data before the data is utilized in deep learning. On the other hand, the adversarial feature generation performs training while generating hard-to-recognize and difficult data inside the deep neural network. Because of automated training data generation based on values in the network instead

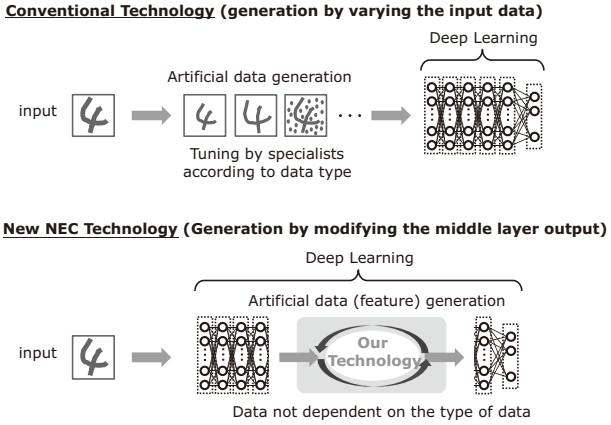

Fig. 9 Difference between traditional data augmentation and adversarial feature generation.

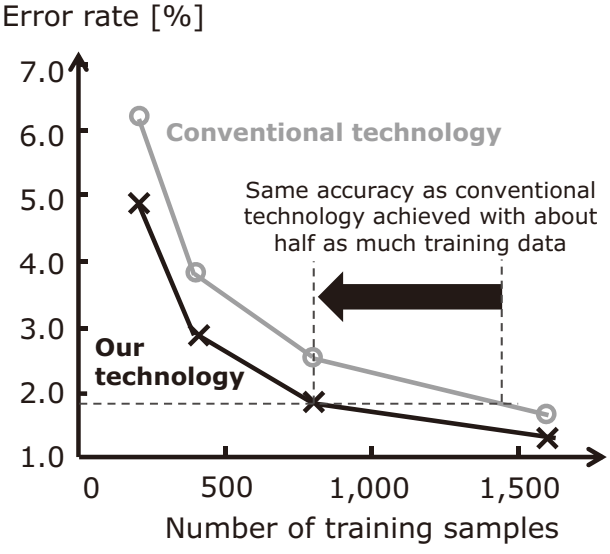

Fig. 10 Effect of adversarial feature generation (MNIST).

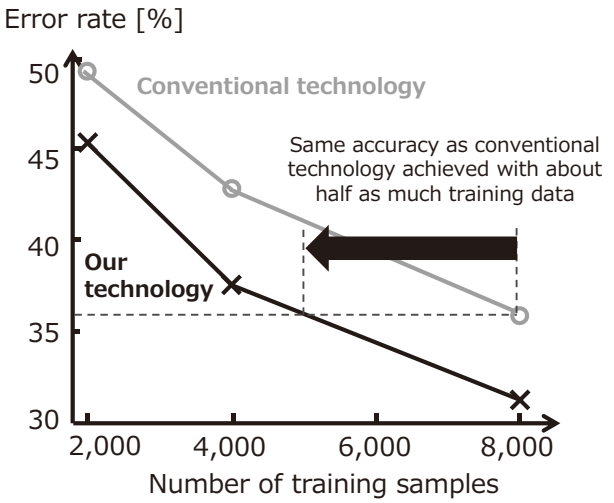

Fig. 11 Effect of adversarial feature generation (CIFAR-10).

of using the input data, this technology makes it possible to apply it universally and efficiently to various data, without the need for tuning by specialists (**Fig. 9**).

3.5 Experiment

The authors evaluated the validity of this technology by an experiment using the MNIST handwritten digits dataset and the CIFAR-10 object recognition dataset for comparison of the recognition accuracy with the traditional data augmentation (**Figs. 10 and 11**). The horizontal axis represents the number of training samples and the vertical axis the error rate with the test data. The graphs show that this technology achieves a lower error rate than the conventional technique and provides a similar accuracy to the conventional technique with about a half of the amount of training data.

4. Conclusion

This paper introduces the two technologies developed by the authors in order to enable efficient deep learning with a small amount of training data. The layer-wise regularization solves the previous issue of the mixed presence of layers with too strong and too weak regularizations, by allowing the regularization coefficients, which varies layer by layer according to the network structure that is to be set optimally. The adversarial feature generation solved the traditional problem of the difficulty of producing data that can contribute to the accuracy improvement, by performing training while automatically generating hard-to-recognize data based on values in the network. It also avoids the need for the tuning of data generation by a specialist. The authors will accelerate the practical implementation of these technologies as they enable the application of deep learning to real problems for which it is hard to prepare a large amount of training data.

References

1) Masato Ishii and Atsushi Sato: Layer-wise weight decay for deep neural networks, Pacific-Rim Symposium on Image and Video Technology, Springer, pp. 276–289, 2017
2) Masato Ishii and Atsushi Sato: Training Deep Neural Networks with Adversarially Augmented Features for Small-scale Training Datasets, International Joint Conference on Neural Networks, 2019

Author's Profile

SATO Atsushi
Research Fellow
Data Science Research Laboratories

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese | English

Vol.14 No.1 AI and Social Value Creation

Remarks for Special Issue on AI and Social Value Creation
Data — Powering Digitalization and AI

Vol.14 No.1
January 2020

Special Issue TOP

Papers for Special Issue

NEC's Efforts Toward Social Applications of AI
NEC's Commitment to Its New "NEC Group AI and Human Rights Principles" Policy
Human Resource Development in the Age of AI

AI-Enhanced Services/Solutions to Accelerate Digital Transformation
NEC Advanced Analytics Platform (AAPF) Promoting "AI Co-Creation"
Use of Individual Identification Based on the Fingerprint of Things Recognition Technology
Visual Inspection Solutions Based on the Application of Deep Learning to Image Processing Controllers
Remote Vehicle Surveillance Solution Based on Communication Prediction/Control Technology
NEC's Emotion Analysis Solution Supports Work Style Reform and Health Management
Facial Recognition Solution for Offices — Improved Security, Increased Convenience
Outline of an Auto Response Solution (AI Chatbot) for Assisting Business Automation and Labor Saving
AI for Work Shift Support — Accelerating the Transition to Human-Centered Business Value Creation
NEC Cloud Service for Energy Resource Aggregation Leveraging AI Technology
Patient Condition Change Signs Detection Technology for Early Hospital Discharge Support
Effective Data-Based Approaches to Disease Prevention/Healthcare Domains
Co-creation of AI-Based Consumer Insight Marketing Services
"Anokorowa CHOCOLATE" Lets People Savor Delicious Chocolates that Reflect the Mood of Special Moments in History

Cutting-Edge AI Technologies to Create the Future Together With Us
Heterogeneous Object Recognition to Identify Retail Products
Optical Fiber Sensing Technology Visualizing the Real World via Network Infrastructures
Intention Learning Technology Imitates the Expert Decision-Making Process
Graph-based Relational Learning
Retrieval-based Time-Series Data Analysis Technology
New Logical Thinking AI Can Help Optimize Social Infrastructure Management
Deep Learning Technology for Small Data
A Computing Platform Supporting AI