EPiC Series in Computing

Volume 81, 2022, Pages 24–40

Proceedings of 11th International Congress
on Advanced Applied Informatics

Forced Selective Information Reduction for Interpreting Multi-Layered Neural Networks

Ryotaro Kamimura[1] and Ryozo Kitajima[2]

[1] Kumamoto Drone Technology and Development Foundation, Techno Research Park, Techno Lab 203,
1155-12 Tabaru Shimomashiki-Gun Kumamoto 861-2202
and IT Education Center, Tokai University, 4-1-1 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan
ryotarokami@gmail.com

[2] Tokyo Polytechnic University, 1583 Iiyama, Atsugi, Kanagawa 243-0297
ryozo.kitajima@gmail.com

Abstract

The present paper aims to reduce unnecessary information obtained through inputs, supposed to be inappropriately encoded, for producing easily interpretable networks with better generalization. The proposed method lies mainly in forced reduction of selective information even at the expense of a larger cost to eliminate unnecessary information coming from the inputs in the initial stage of learning. Then, in the later stage of learning, selective information is increased to produce a small number of really important connection weights for learning. The method was preliminarily applied to two business data sets: the bankruptcy and the mission statement data sets, in which the interpretation is considered as important as generalization performance. The results show that selective information could be decreased, though the cost to realize this reduction became larger. However, the accompanying selective information increase could be used to compensate for the expensive cost to produce simpler and interpretable internal representations with better generalization performance.

1 Introduction

The selectivity of neural components has been widely known [12, 5], and many experimental results to show the existence of the selectivity have been accumulated [33, 26, 7, 13, 4, 32]. In neural networks, the concept of selectivity has recently received much attention for interpreting convolutional neural networks (CNN). From our viewpoint, almost all interpretation methods in the field of CNN can be considered ones, based on the selectivity of components such as neurons and weights [10, 22, 19, 23, 25, 24] and sensitivity detection [14, 3, 27, 2, 16, 1, 29, 6, 28, 30, 20], to cite a few. This means that the majority of interpretation or visualization methods in the CNN have tried to determine which components are the most responsive to the inputs or outputs. In the CNN, inputs are transformed by the feature detection layers such as convolutional ones to extract the salient features, such as orientation features. Intuitively, this information on the features should be naturally necessary in learning, because they are obtained through learning itself. Thus, the selectivity is essential in improving their performance, in particular, generalization performance.

However, due attention has recently been paid to the problem of whether the selectivity of components in neural networks is necessary in learning [21, 17, 18, 31]. In spite of the well-recognized importance of selectivity [31], there are some reports saying that the selectivity should be undermined and eliminated as much as possible, especially for improving generalization performance [21, 17, 18]. This problem of the importance of selectivity can be considered in terms of the property of data sets. As mentioned above, in the convolutional neural networks, the selective features are obtained through learning processes in the CNN, and correspondingly and naturally, those selective features should play some important roles in learning, because they are created as the results of the learning itself. On the contrary, in other types of data sets, such as business data sets, dealt with in this paper, the input variables are created manually, and eventually the actual data should be extracted through those artificially and subjectively created input variables. Thus, we cannot necessarily say that these input variables can be created for them to be appropriately related to the targets and to represent well the characteristics of data sets. We have a high possibility that the input variables cannot fully and appropriately represent the data set itself. Even worse, the inputs may be very harmful to learning in neural networks due to the poor encoding of characteristics of data sets. As has been well known [11], the intuition or knowledge on our outer circumstances is strictly conditioned by our cultural and physical contextual knowledge to keep our existence as secure as possible. This means that it is impossible to see the data itself as it is, because the data harmful to the existence of human beings should be evaded as much as possible. Then, encoding data through artificial input variables can be of no use, and we can say that it is harmful to the learning of neural networks. One of the possible ways to deal with these problems of human bias toward or against the data sets is to decrease information content obtained through the inputs as much as possible, which is related to the conventional information minimization methods, dating back to Deco's pioneer works [8, 9]. This consideration has led us to conclude that information content obtained through the input variables should be reduced as much as possible, once we know that learning cannot be appropriately performed. Because we cannot know before learning what inputs should be appropriate to the target problem, we have a high possibility that harmful information can be introduced through artificially created inputs.

Then, the reduction of information in terms of weights is realized by distributing the strength of connection weights or neuron activations as evenly as possible or making the strength as small as possible. Concerning the connection weights, it may be difficult to reduce all connection weights as evenly as possible, because to make connection weights as uniform as possible is contrary to error minimization between outputs and targets. The error minimization can be realized by selectively increasing or decreasing the weight strength naturally. Traditionally, the regularization methods such as weight decay have been used to reduce the strength of connection weights as much as possible. Though those methods can push large weights toward smaller ones, actually a small number of selectively chosen weights tends to be stronger due to the necessity of error minimization. Then, we should state that selectivity reduction or information reduction cannot be easily attained in the conventional framework of neural learning. Considering those problems in selectivity, the present paper tries to show that the selective information should be reduced in the first place, and then the usual information augmentation should be applied for improved generalization and interpretation. Contrary to the conventional hypothesis of regularization, for realizing selective information reduction, we considerably increase the cost in terms of the strength of connection weights or at any costs. We should repeat here that, contrary to the conventional hypothesis of cost reduction, for example, in terms of adding regularization terms, the present paper stresses that we need to decrease information on inputs at the expense of a larger cost, because the information in data sets may not be encoded appropriately, and weight strength reduction is of no use in actually reducing selective information.

This paper is organized as follows. In Section 2, we present how to compute the informational potentiality and the corresponding selective information for connection weights. Then, we present how to implement the decrease and increase of selective information. The method was applied to two data sets, namely, the bankruptcy and mission statements data sets. The experimental results show that the selective information was forced to be reduced by increasing the cost in the initial learning steps, and then it was forced to be increased. The connection weights were equally activated in terms of their absolute strength, and a small number of stronger weights appeared. The final compressed weights were close to the correlation coefficients between inputs and targets of the original data set. The results show that the present information reduction and augmentation can be used to produce very simple, linear, and independent internal representations, which can be easily interpreted.

2 Theory and Computational Methods

2.1 Potentiality and Information

The selectivity in neural networks can be described by using the selective information to measure to what extent a neuron is connected with other neurons in a hidden layer. We focus on connection weights of hidden layers because information content in hidden layers can be freely controlled, while information in the input and output layer is severely restricted by inputs and outputs. In addition, the selectivity of connection weights is dealt with because we try to consider the selectivity of neural components themselves, independently of any input patterns.

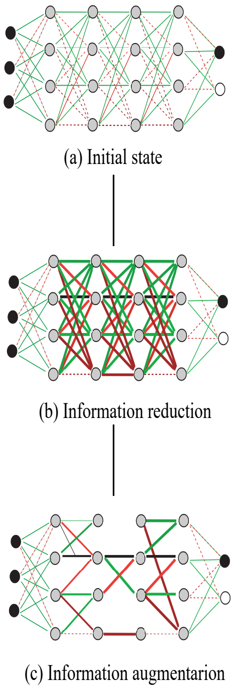


Figure 1: Selective information reduction (b) and selective information augmentation (c) with a six-layered neural network.

For the first approximation, we suppose that the importance of weights can be obtained by their absolute values, and for simplicity we focus on connection weights between the second and third layer (2,3), as shown in Figure 1.

$$u_{jk}^{(2,3)} =\mid w_{jk}^{(2,3)} \mid \tag{1}$$

Then, normalized absolute weights are defined by

$$g_{jk}^{(2,3)} = \frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} \tag{2}$$

We call this importance “potentiality,” because it has some power to increase or decrease the selectivity. By using this potentiality, selective potentiality can be computed by

$$G^{(2,3)} = \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \left[\frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} \right] \tag{3}$$

When this selective potentiality increases, the number of connection weights between neurons increases gradually. On the other hand, when this potentiality decreases, the number of weights decreases, and finally only one weight remains.

In addition, we define the complementary potentiality

$$\bar{g}_{jk}^{(2,3)} = 1 - \frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} \tag{4}$$

By using this complementary potentiality, selective information can be computed by

$$\bar{G}^{(2,3)} = \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \left[1 - \frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} \right] \tag{5}$$

When only one potentiality becomes one, while all the others are zero, the selective information becomes maximum. A neuron is connected with the other neuron explicitly by one connection weight. For simplicity, we suppose that at least one connection weight should be larger than zero.

Then, we should compute the cost associated with this selective information. In this paper, the cost is computed simply by the sum of all absolute weights.

$$C^{(2,3)} = \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} u_{jk}^{(2,3)} \tag{6}$$

Here, we control the ratio of selective information to its cost.

$$R^{(2,3)} = \frac{\bar{G}^{(2,3)}}{C^{(2,3)}} \tag{7}$$

2.2 Computational Methods

Connection weights are changed by multiplying them by the individual potentiality or complementary potentiality. In this paper, we suppose that the inputs cannot be appropriately encoded to contain important information for learning. We first decrease the selective information, and then it is increased in the end.

To decrease the selective information, we can use the complementary potentiality.

$$w_{jk}^{(2,3)}(n+1) = \bar{g}_{jk}^{(2,3)}(n)\, w_{jk}^{(2,3)}(n) \tag{8}$$

However, some instability has been observed by this direct method, and we introduce a modified one with three parameters.

$$\bar{h}_{jk}^{(2,3)} = \theta_1 \left[1 - \frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} + \theta_3 \right]^{\theta_2} \tag{9}$$

The parameter θ_3 is introduced to eliminate zero potentiality, θ_2 is for stabilizing the learning, and θ_1 is for increasing or decreasing the strength of average weights. Then, we have a modified one

$$w_{jk}^{(2,3)}(n+1) = \bar{h}_{jk}^{(2,3)}(n)\, w_{jk}^{(2,3)}(n) \tag{10}$$

This method can push stronger weights toward weaker ones, and finally, all weights tend to be evenly distributed. In our experiment, the parameter θ_1 is larger than one in the first place, and the weights tend to be larger and evenly distributed in the initial learning epochs.

In the remaining learning epochs, we use the usual and individual potentiality with two parameters.

$$d_{jk}^{(2,3)} = \theta_1 \left[\frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} \right]^{\theta_2} \tag{11}$$

Then, we use the same type of potentiality assimilation process.

$$w_{jk}^{(2,3)}(n+1) = d_{jk}^{(2,3)}(n)\, w_{jk}^{(2,3)}(n) \tag{12}$$

In those learning steps, because of the smaller parameter value of θ_1, all connection weights tend to be smaller, and in particular, smaller ones become much smaller, and finally a smaller number of relatively stronger weights remains.

2.3 Collective Compression

For interpretation, we use a new type of network compression method called "collective compression." We compress here a multi-layered neural network into the simplest one without hidden layers, as shown in Figure 2. In addition, this compression should be applied to any moment of the learning steps, and we average all those connection weights to obtain the final ones.

In the first compression in Figure 2(a), weights from the first layer to the second layer $w_{ij}^{(1,2)}$ and from the second layer to the third layer $w_{jk}^{(2,3)}$ are combined into

$$w_{ik}^{(1,3)} = \sum_{j=1}^{n_2} w_{ij}^{(1,2)} w_{jk}^{(2,3)} \tag{13}$$

where (1,3) represents a route from the first to the third layer. Suppose that $w_{ik}^{(1,3)}$ denotes a compressed weight from the first layer to the third layer. Then, this compressed weight is again compressed with a connection weight from the third to the fifth layer in Figure 2(b).

$$w_{il}^{(1,4)} = \sum_{k=1}^{n_3} w_{ik}^{(1,3)} w_{kl}^{(3,4)} \tag{14}$$

In the same way, we can obtain the compressed weight from the first layer to the fifth layer $w_{im}^{(1,5)}$. By combining this compressed weight with connection weights to the output layer in Figure 2(e), we have

$$w_{in}^{(1,6)} = \sum_{m=1}^{n_5} w_{im}^{(1,5)} w_{mn}^{(5,6)} \tag{15}$$

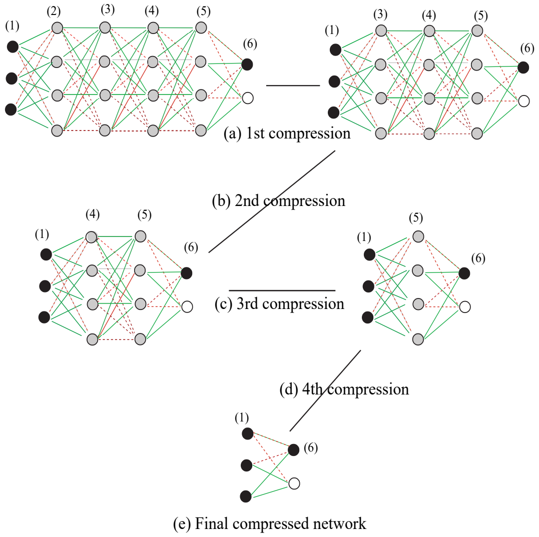


Figure 2: A process of network compression from the first (a) to the final compression (d) for a network architecture with six layers, including four hidden layers.

3 Results and Discussion

3.1 Bankruptcy Data Set

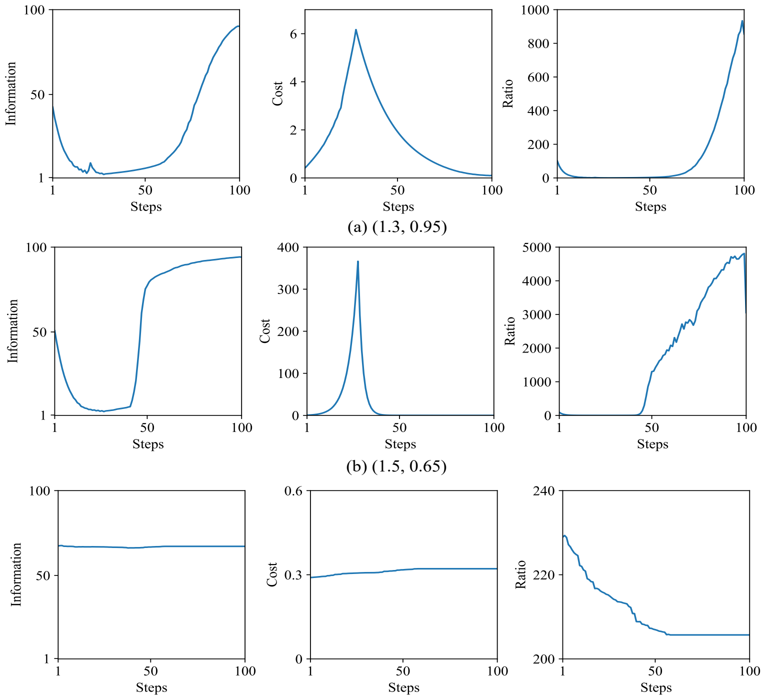
3.1.1 Experimental Outline

For examining the effectiveness of the present method, we conducted a preliminary experiment with the bankruptcy data set, where we tried to predict the bankruptcy of companies [15]. The experiment tried to examine which input variables could contribute to improved generalization and how the new method could disentangle connection weights for easy interpretation. Though the data set was small, the conventional methods such as logistic regression analysis and random forest could not show better generalization performance; the new method could improve generalization performance and in addition, connection weights were disentangled for easy interpretation. The number of companies was 100, and six input variables were used in the experiments. Seventy percent of the data set was for training, and the remaining data was for testing. The learning was divided into two parts. In the first part, which was composed of a one-third period of all the learning epochs, the selective information was reduced by increasing the cost in terms of strength of weights. In the second part, the selective information was increased as much as possible. For easy reproduction of the results, we used the scikit-learning package with all default parameter values except for the number of learning epochs and the learning method.

3.1.2 Selective Information and Cost Control

This section shows that selective information was first reduced by increasing the corresponding cost in terms of weight strength. Then, we showed that selective information could be increased by reducing the cost.

Figure 3(a) shows selective information (left), cost (middle), and ratio (right) by the present method with the parameter θ_1 (1.3, 0.95) with the original correlation coefficients between inputs and targets. The selective information (left) decreased to the minimum point with 30 learning steps, and then it



(a) (1.3, 0.95)

(b) (1.5, 0.65)

(c) Conventional

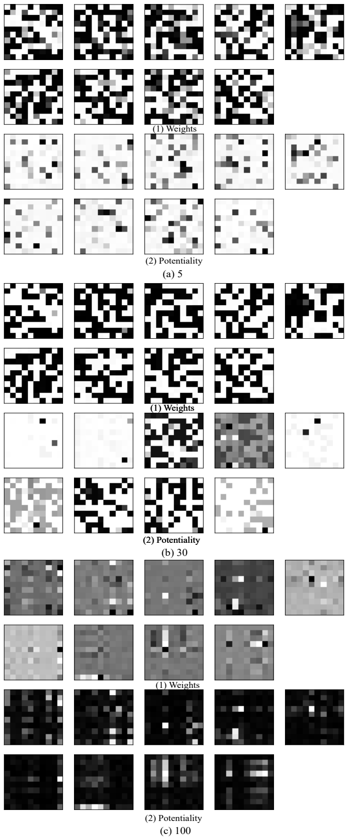
Figure 3: Selective information (left), cost (middle), and ratio of information to its cost (right) as a function of the number of learning steps, when the parameter θ_1 was (1,3, 0.95) for the first and the second stage (a), (1,5, 0.65) (b), and by the conventional method (c) for the bankruptcy data set.

increased rapidly in the later stage of the learning steps. The figure in the middle shows the cost, and it increased rapidly in the first period of learning and then decreased rapidly. The ratio of selective information to its cost remained small in the initial stage of learning, and it increased rapidly in the later stage of learning in the right-hand box in Figure 3(a).

This tendency was enhanced when the criterion of improved generalization was adopted in Figure 3(b), where we tried to choose the parameters to increase generalization performance as much as possible. The selective information (left) plunged to the lowest level and the cost (middle) increased considerably and became close to 400. The ratio also increased to almost the 5,000 point in the right-hand box in Figure 3(b). This suggests that we need to decrease selective information more strongly in the initial learning steps to improve generalization performance. Finally, Figure 3(c) shows the results by the conventional method. The values of selective information (left) and its cost (middle) remained unchanged throughout all the learning steps. Then, as can be seen in the right-hand box in Figure 3(c), the ratio decreased constantly, though slightly. These results confirmed that the cost augmentation or information reduction was necessary for neural networks to appropriately choose a smaller number of connection weights by the cost reduction or information augmentation.

3.1.3 Connection Weights

The connection weights were, in the initial stage of learning, evenly distributed in terms of their absolute strength by pushing all weights toward ones with larger absolute strength. In the later stage of learning, the number of stronger weights became smaller, which can be considered ones to be important for


(1) Weights

(2) Potentiality

(a) 5

(1) Weights

(2) Potentiality

(b) 30

(1) Weights

(2) Potentiality

(c) 100

Figure 4: Weights for all hidden layers by the selective information when the number of steps was 5 (a), 30 (b), and 100 (c) and when the parameter θ_1 was 1.3 and 0.95 for the bankruptcy data set.

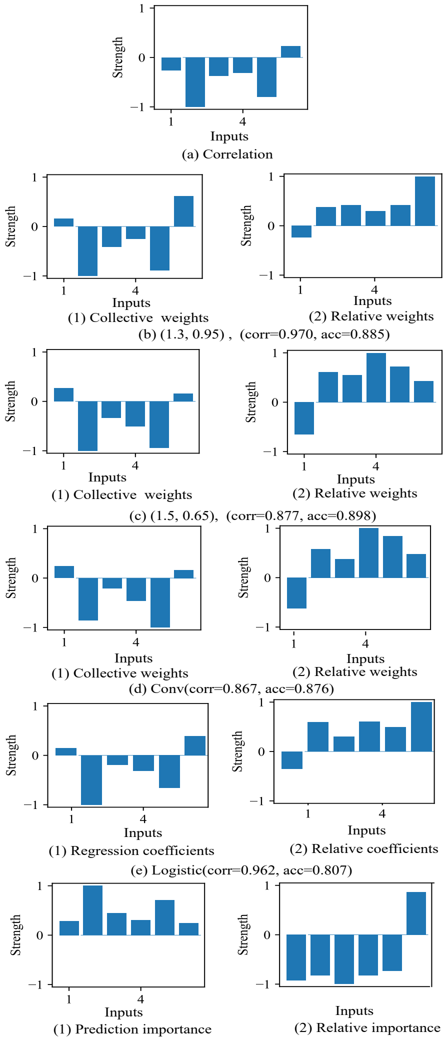
learning.

Figure 4 shows connection weights and the corresponding potentialities when the parameter θ_1 was 1.3 and 0.95 with the largest correlation coefficients. As can be seen in the figure, connection weights (1) and potentialities (2) became more uniformly activated up to the 30th learning step due to the larger parameter value. In particular, in the potentiality values in Figure 4(a2) and (b2), the number of larger weights in white increased considerably. Finally, in the later stage in Figure 4(c), the number of stronger weights became smaller by the effect of selective information augmentation with a smaller parameter value.

3.1.4 Interpreting Collectively Compressed Weights

The results show how the compressed weights and other importance values were close to the original correlations, and in what points the compressed weights were different from the original ones in Figure 5(a).

Figure 5(b)-(f) shows compressed weights and the other related importance measures (1) and relative compressed weights and related ones (2) by five methods. The collective weights or the other importance measures and correlation coefficients were quite similar to each other, except those by the random forest in Figure 5(f). The random forest could not principally produce negative values, but when the importance values for inputs No.2 and No.5 were changed to negative, the prediction importance was



(a) Correlation

(1) Collective weights (2) Relative weights

(b) (1.3, 0.95) , (corr=0.970, acc=0.885)

(1) Collective weights (2) Relative weights

(c) (1.5, 0.65), (corr=0.877, acc=0.898)

(1) Collective weights (2) Relative weights

(d) Conv(corr=0.867, acc=0.876)

(1) Regression coefficients (2) Relative coefficients

(e) Logistic(corr=0.962, acc=0.807)

(1) Prediction importance (2) Relative importance

(f) Random(corr=-0.919, acc=0.785)

Figure 5: The original correlation coefficients between inputs and targets (a), compressed weights (1), and relative weights (2) when the parameter θ_1 was (1.3, 0.95); with larger correlation coefficients (b), the parameter θ_1 was (1.5, 0.65) with better generalization (c); conventional method (d), regression coefficients by the logistic regression analysis (e); and prediction importance by the random forest method (f) for the bankruptcy data set.

very close to those by the other methods. Thus, all five methods produced final weights or importance values close to the original correlation coefficients. However, the present method produced a higher correlation, and in addition, the generalization accuracy was better than those by the other methods, which will be explained in the following section.

Then, we should examine in which points the weights and the other measures were different from the original correlation coefficients. Figure 5(2) shows the relative collective weights and related importance measures, which were obtained by dividing the compressed weights and the other measures by the original correlation coefficients, as shown in Figure 5(a). As can be seen in the figures (b), (e), and (f), by the present method, logistic regression and random forest, those relative weights were very close

Table 1: Summary of experimental results on correlation coefficients and generalization performance, averaged over 30 different initial conditions and different training data subsets for the bankruptcy data set.

Method	Accuracy	Corr
(1.3,0.95)	0.885	**0.970**
(1.5,0.65)	**0.898**	0.877
Conv	0.876	0.867
Logistic	0.807	0.962
Random	0.785	-0.919

to each other, where input No.6 had relatively high values. All those methods could produce positive relative values for input No.6, meaning that the strength of weights was in the same direction as the original correlation coefficients. However, those by the present method with better generalization (c) and conventional method (d) showed that input No.4 had larger relative importance.

This suggests that, to improve generalization performance, this input No.4 should play an important role, because with this input, the highest accuracy was obtained, which will be shown in Table 1. The results suggested that, by disentangling connection weights, we could see which inputs could contribute to the outputs.

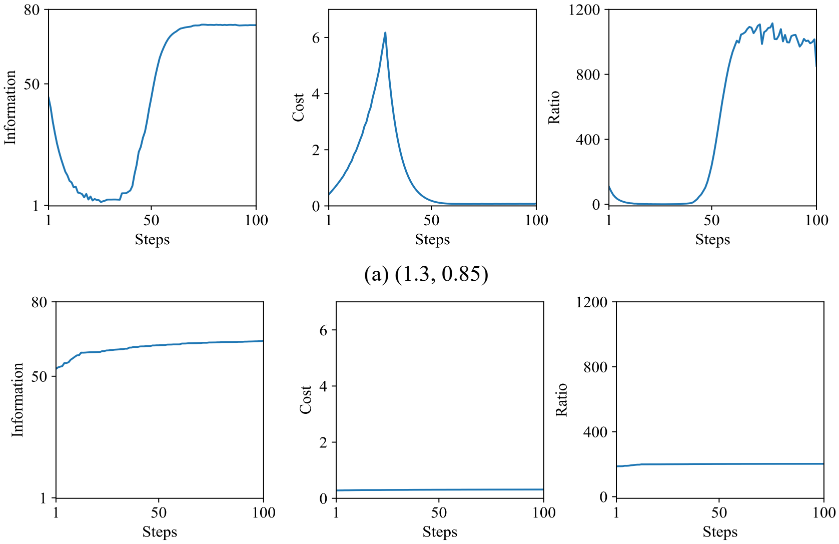
3.1.5 Summary of Generalization and Correlation

Table 1 shows the summary of results by five methods. The method aiming to increase the correlation when the parameter θ_1 was (1.3, 0.95), produced the best correlation of 0.970, which was higher than the 0.962 by the logistic regression analysis. In addition, this method could produce a much higher generalization value of 0.885, larger than the 0.807 by the logistic regression. This means that the present method could improve generalization performance, and at the same time, all connection weights were disentangled and each connection weight in the compressed weights was more separately and more independently connected with the output than by the logistic regression analysis. In addition, when the parameter θ_1 was (1.5, 0.65), controlled to produce the best generalization, the generalization accuracy was the highest at 0.898, but the correlation coefficient decreased to 0.877. Using the conventional method, the correlation and generalization accuracy were further decreased to 0.867 and to 0.876. The logistic regression analysis produced the second highest correlation value of 0.962 but the second lowest generalization accuracy of 0.807. Finally, we should note that the random forest produced the worst prediction accuracy of 0.785, and the correlation coefficient was the lowest, namely, -0.919. However, considering the difficulty in describing negative values by the importance of random forest, the importance seemed to represent well the original correlation coefficient if the importance values were the inverse, as mentioned above.

3.2 Mission Statement Data Set

3.2.1 Experimental Outline

In this experiment, we tried to analyze the relations between companies' mission statements and their profitability. We collected 300 companies' mission statements, listed in the first section of the Tokyo stock exchange, by using the natural language processing systems. After adding the profitability scores, we chose only 100 companies divided evenly into higher and lower ones, based on their average profitability. The selective information can be controlled by controlling mainly the cost associated with the



(a) (1.3, 0.85)

(b) Conventional

Figure 6: Selective information (left), cost (middle), and ratio of information to its cost (right) as a function of the number of steps, when the parameter θ_1 was 1.3(0.85) (a) and by the conventional method (b) for the mission statements data set.

selective information. In the initial learning epochs, the cost augmentation and selective information reduction were applied, and in the remaining learning steps, we only used the cost reduction without the selective information augmentation, because we could get better results by this method. For the initial 30 learning epochs, we tried to decrease the selective information with a higher parameter value larger than one, and for the remaining learning epochs, the parameter decreased to the normal value of 0.85 to increase the selective information.

3.2.2 Selective Information Control

The results show that the present method could decrease the selective information by increasing the corresponding cost in terms of the strength of connection weights. Then, in the later stage of learning, the selective information could be increased, and the cost could be decreased.

Figure 6(a) shows the selective information (left), cost (middle), and ratio of information to its cost (right) when the parameter θ_1 was set to 1.3 (initial learning epochs), followed by 0.85 (remaining epochs). As can be seen in the figure, the selective information decreased gradually up to 30 learning steps, and then the information increased rapidly. This is because the parameter θ_1 was set to 1.3 for the initial learning epochs, and then the parameter was reduced to 0.85 for the remaining learning steps. The cost (middle) increased for the first 30 learning steps, and then it was forced to decrease rapidly. Finally, the ratio of information to its cost was low for the first 30 learning steps, and the ratio increased rapidly for the later stage of learning. On the other hand, Figure 6(b) shows the results by the conventional method without selective information and its cost. The selective information increased very slightly, and the cost remained low compared with those by the present method. Finally, the ratio also remained low for all learning steps. These results show that the present method could decrease the selective information by increasing the cost for the initial learning steps, and then the selective information was forced to be larger for the remaining learning steps.

3.2.3 Connection Weights

The absolute strength of connection weights (cost) became larger and evenly distributed in the initial stage. Then, the number of strong weights became smaller, and higher selective connection weights could be obtained.

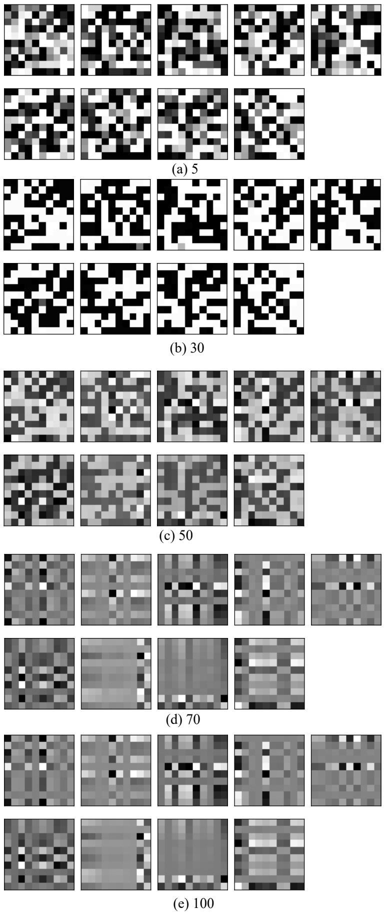
Figure 7 shows connection weights when the number of learning steps increased from 5 (a) to 100 (e). When the learning steps increased from 5 (a) to 30 (b), and the selective information was decreased by increasing the cost and decreasing the selective information, connection weights were equally positive and negative. This is because the cost minimization was effective in increasing the absolute strength of connection weights. The selective information was computed by using the absolute strength of connection weights. This shows that all connection weights had equal individual potentialities, leading to minimum selective information. When the learning steps increased from 50 (c) to 100 (e), where the parameter θ_1 was reduced to 0.85, all the connection weights were pushed toward smaller values, and in the end only several weights remained relatively large. As shown in Figure 7(e), several groups of stronger weights could be seen in the end. Experimental results show that the present method could reduce the number of stronger connection weights in the end, and in addition, those stronger weights formed several groups.

3.2.4 Interpreting Collectively Compressed Weights

The present method could produce compressed weights closest to the original correlation coefficients between inputs and targets. In addition, we could see that input No.1 should have much importance, but also that input No.2 should play some other role in generalization.

Figure 8(a) shows the correlation coefficients between inputs and targets of the original data set, where the first input had the largest absolute value. As shown in the left-hand box in Figure 8(b), the present method produced compressed weights close to the correlation coefficients of the original data set in Figure 8(a), where input No.1 had the largest importance in terms of weight strength. The right-hand box shows the normalized weights relative to the original correlation coefficients, where input No.2 had the largest value. Input No.2 had more importance relative to input No.2 of the original correlation coefficient. Figure 8(c) shows an optimal case with the highest generalization accuracy for a specific number of learning steps, where the generalization accuracy attained the value of 0.8 but the correlation coefficient was very low. The compressed weights were different from the original correlation coefficients in Figure 8(a), but input No.2 had the largest absolute strength, and we could also see in the right-hand box that input No.2 had the largest normalized relative weight value. This means that, of the inputs, input No.2 may play some roles in improving generalization performance. The conventional method, shown in Figure 8(d), and the logistic regression analysis, in Figure 8(e), produced compressed weights and regression coefficients that were very close to the correlation coefficients in Figure 8(a). In addition, input No.2 had the largest absolute strength in terms of normalized relative weights in the right-hand box. Finally, the random forest in Figure 8(f) shows prediction importance that was different from the other weights and coefficients. However, we could still see in the right-hand box that input No.2 had the largest absolute strength in terms of relative importance .

The results show that input No.1 played the most important role in learning, in particular, in producing simplified internal representations. However, the relative weights show that input No.2 had the largest value relative to the original correlation coefficient, meaning that input No.2 had some role in improving generalization performance, which could not be extracted by the original correlation coefficients.


(a) 5

(b) 30

(c) 50

(d) 70

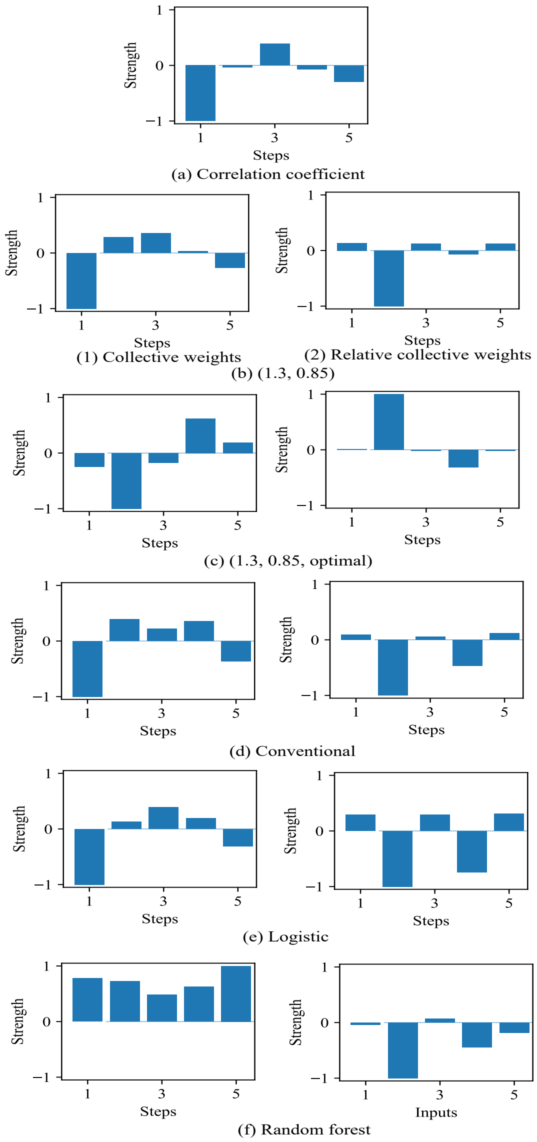
(e) 100

Figure 7: Weights for all hidden layers by the selective information when the number of steps was 5 (a), 30 (b), and 100 (c) and when the parameter θ_1 was 1.3 (up to 30 steps) and 0.85 (otherwise) for the mission statement data set.

3.2.5 Summary of Generalization and Correlation

The present method produced a higher correlation coefficient, close to that by the logistic regression analysis. In addition, the method could produce the highest generalization, which was considerably higher than that by the logistic regression analysis.

Table 2 shows the summary of final results by five methods. The present method produced the best average generalization accuracy of 0.680, and the correlation coefficient was 0.967, close to one, when the parameter θ_1 was 1.3 and 0.85. In particular, when we considered a case with the best generalization performance of 0.8 and with only thirty learning steps, the correlation coefficient decreased to -0.005. The conventional method produced the accuracy of 0.640, close to 0.680 by the present method, but the correlation coefficient decreased from 0.967 to 0.877. Logistic regression analysis produced the very high correlation coefficient of 0.975, higher than that of 0.967 by the present method. However, the accuracy decreased from 0.680 to 0.474. Finally, the random forest produced the accuracy of 0.584 and the negative correlation coefficient of -0.576. Thus, the experimental results confirmed that the new method could produce simple internal representations, keeping good generalization.



(a) Correlation coefficient

(1) Collective weights

(2) Relative collective weights

(b) (1.3, 0.85)

(c) (1.3, 0.85, optimal)

(d) Conventional

(e) Logistic

(f) Random forest

Figure 8: The original correlation coefficients between inputs and targets (a), when the parameter was 1.3 and 0.85 (b), and 1.5 and 0.95 with 30 learning steps (best generalization) (c), conventional (d), regression coefficients (e) by the logistic regression analysis, and prediction importance (f) by the random forest for the mission statement data set.

4 Conclusion

The present paper aimed to propose a new type of information-theoretic method to make easier the interpretation of the inference mechanism and to improve generalization. In the method, information or selective information should be minimized in the initial stage of learning at the expense of larger cost, and then selective information should be increased as much as possible. The method tried to eliminate information that is not necessary for learning, because artificially created input variables do

Table 2: Summary of experimental results on correlation coefficients and generalization performance, averaged over 30 different initial conditions and different training data subsets for the mission statement data set.

Method	Accuracy	Corr
(1.3,0.95,average)	0.680	0.967
(1.3,0.95,optimal)	**0.800**	-0.005
Conv	0.640	0.877
Logistic	0.474	**0.975**
Random	0.584	-0.576

not necessarily capture necessary information for learning.

The method was applied preliminarily to the bankruptcy and the mission statements analysis. In the experiments, the cost first increased and the selective information was decreased, and then the cost was reduced, accompanied by selective information increase. The connection weights in the first place were larger in terms of strength of weights, and all weights were forced to be equally distributed in terms of their absolute strength. These larger costs could eventually be used to distribute the absolute weights evenly, reducing the selective information due to the inputs. Then, the selective information maximization was applied to find a small number of important connection weights. The final results show that we could produce final connection weights close to the correlation coefficients between inputs and targets of the original data set. In addition, generalization performance could not be degraded.

The problem is that we cannot determine to what extent the cost should be increased in the first period of learning and to what extent the cost should be decreased in the later stage of learning. Thus, we should develop a process to control the cost more exactly. Though some problems should be solved for more practical applications, the present method shows a possibility and the necessity of higher cost, which we have so far tried to eliminate as much as possible in the conventional methods.

References

[1] Farhad Arbabzadah, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Identifying individual facial expressions by deconstructing a neural network. In *German Conference on Pattern Recognition*, pages 344–354. Springer, 2016.

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[3] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÃžller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

[4] Omri Barak, Mattia Rigotti, and Stefano Fusi. The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off. *Journal of Neuroscience*, 33(9):3844–3856, 2013.

[5] E. L. Bienenstock, L. N. Cooper, and P. W. Munro. Theory for the development of neuron selectivity. *Journal of Neuroscience*, 2:32–48, 1982.

[6] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pages 63–71. Springer, 2016.

[7] Charles Cadieu, Minjoon Kouh, Anitha Pasupathy, Charles E Connor, Maximilian Riesenhuber, and Tomaso Poggio. A model of v4 shape selectivity and invariance. *Journal of neurophysiology*, 98(3):1733–1750, 2007.

[8] G. Deco, W. Finnof, and H. G. Zimmermann. Unsupervised mutual information criterion for elimination of overtraining in supervised multiplayer networks. *Neural Computation*, 7:86–107, 1995.

[9] G. Deco and L. Parra. Non-feature extraction by redundancy reduction in an unsupervised stochastic neural networks. *Neural Networks*, 10(4):683–691, 1997.

[10] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341, 2009.

[11] Edward T Hall. Beyond culture. garden city, ny: Anchor, 1976.

[12] D. H. Hubel and T. N. Wisel. Receptive fields, binocular interaction and functional architecture in cat's visual cortex. *Journal of Physiology*, 160:106–154, 1962.

[13] Janneke FM Jehee, Sam Ling, Jascha D Swisher, Ruben S van Bergen, and Frank Tong. Perceptual learning selectively refines orientation representations in early visual cortex. *Journal of Neuroscience*, 32(47):16747–16753, 2012.

[14] Javed Khan, Jun S Wei, Markus Ringner, Lao H Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R Antonescu, Carsten Peterson, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673–679, 2001.

[15] T. Komazawa. *Quantitative theory by PC (in Japanese)*. Asakura Shoten, Tokyo, 1992.

[16] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Muller, and Wojciech Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2912–2920, 2016.

[17] Matthew L Leavitt and Ari Morcos. Selectivity considered harmful: evaluating the causal impact of class selectivity in dnns. *arXiv preprint arXiv:2003.01262*, 2020.

[18] Matthew L Leavitt and Ari S Morcos. On the relationship between class selectivity, dimensionality, and robustness. *arXiv preprint arXiv:2007.04440*, 2020.

[19] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[20] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209. Springer, 2019.

[21] Ari S Morcos, David GT Barrett, Matthew Botvinick, and Neil C Rabinowitz. On the importance of single directions for generalization. 2018.

[22] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4467–4477, 2017.

[23] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in neural information processing systems*, pages 3387–3395, 2016.

[24] Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding neural networks via feature visualization: A survey. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 55–76. Springer, 2019.

[25] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[26] Aniek Schoups, Rufin Vogels, Ning Qian, and Guy Orban. Practising orientation identification improves orientation coding in v1 neurons. *Nature*, 412(6846):549–553, 2001.

[27] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[28] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[29] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Interpretable deep neural networks for single-trial eeg classification. *Journal of neuroscience methods*, 274:141–145, 2016.

[30] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

[31] Jumpei Ukita. Causal importance of low-level feature selectivity for generalization in image recognition. *Neural Networks*, 125:185–193, 2020.

[32] Zhongqiang Wang, Tao Zeng, Yanyun Ren, Ya Lin, Haiyang Xu, Xiaoning Zhao, Yichun Liu, and Daniele Ielmini. Toward a generalized bienenstock-cooper-munro rule for spatiotemporal learning via triplet-stdp in memristive devices. *Nature communications*, 11(1):1–10, 2020.

[33] Leonard E White, David M Coppola, and David Fitzpatrick. The contribution of sensory experience to the maturation of orientation selectivity in ferret visual cortex. *Nature*, 411(6841):1049–1052, 2001.