Selective Information-Driven Learning for Producing Interpretable Internal Representations in Multi-Layered Neural Networks

Ryotaro Kamimura
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 Univerisity
4-1-1 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan
email: ryotarokami@gmail.com

***Abstract*—This paper aims to propose a new type of information-theoretic method to control information content stored and transmitted in neural networks. To make the meaning of information concretely interpretable, we introduce the selective information and a method to control it, called "selective information-driven learning". The new method is more suited for modeling neural learning than conventional information-theoretic measures, such as mutual information, because we can easily maximize and minimize the information, and we can interpret the meaning of information more concretely. The method was applied to the well-known wine data set. The experimental results show that the selective information could be maximized and minimized, and we could easily interpret the meaning of information in terms of the number of strong weights. In addition, the partial compression of multi-layered neural networks revealed that maximum information tended to be focused on output information, while minimum information tried to consider input information in addition to output information. Finally, collective weights, averaged over all compressed weights obtained in learning, were similar to the original correlation coefficients between inputs and targets, meaning that the selective information can disentangle complicated connection weights into simple, linear, and independent ones to be easily interpreted.**

Keywords-Selective information; mutual information; partial compression; collective interpretation; correlation coefficient

I. INTRODUCTION

Due to the requirement of right of explanation [1], many attempts have been made to explain how neural networks try to produce outputs. Because the inner mechanism of neural networks is far from being clarified, the black-box property has been taken into granted, in particular, in practical applications such as medical ones [2]. In those applications, the black-box has been considered to be not so critical as has been expected, because even the human body is a kind of black-box. The most important thing for application lies in the usability and prediction performance of adopted models. Though in neural networks, as well as machine learning, there have been many different types of methods to explain the network behaviors, they seem to suppose implicitly or overly this type of black box model. For example, in the field of convolutional neural networks, the majority of methods for interpretation, have been focused on the explanation of network behaviors with implicitly supposed black-boxed models. Since it is impossible to clarify the inner mechanism at the present stage, all we can do is to check how outputs can be changed in accordance with the inputs, namely, external explanation. The well-known methods such as the activation maximization, surrogate functions, local perturbations, layer-wise relevance propagation, and so on [3]–[7], seem to detect features, corresponding to the specific inputs. Though they try to make the maximally informative explanation [8], they seem to suppose implicitly the black boxed properties of inference mechanism. Those methods have been very effective in practical applications, in particular, in the cases of image data sets, because it is easy to understand intuitively the meaning of features detected by those methods. However, even though those methods with strong visual power can show how some important features can be detected in multiple layers in neural networks, it is far from understanding the inner mechanism of our intelligence [9].

In addition, when we try to deal with data sets whose meaning cannot be easily understood such as business data sets, more interpretable models to make the black-box whiter are needed [9], [10]. Even in the seemingly interpretable image data sets, the well-known Clever Hans phenomena [11] cannot be easily explained without understanding the overall context in the prediction. Naturally, there have been also some attempts to interpret the main and inner mechanism of human nervous systems, as well as human cognitive processes, dating back to the so-called "connectionism" approch to human cognition [12]–[14]. However, those attempts could not necessarily clarify the inner structure by which complicated cognitive processes can be explained [15].

Parallel to this connectionism approach to human cognition, there were important attempts to interpret the inner structure from the information-theoretic points of view. Linsker's maximum information preservation had an impact to the studies to understand and explain visual information processing, as well as human information processing in general [16]–[19]. Linsker's approach was a trigger to produce many different types of information-theoretic methods in neural networks [20]–[26]. Though the attempts may be promising in exploring the general information processing properties behind neural networks, the complexity of information measures, such as mutual information in computation have prevented those information-theoretic methods from being widely used in neural networks. In addition, we have another problem of difficulty in understanding the meaning of final internal repre-

 ISBN: 978-1-61208-847-1

sentations. Ironically, by introducing the information-theoretic methods, the inner structure has become uninterpretable due to the abstract properties of those information-theoretic measures. Thus, we can say that the abstract property of information measures, such as mutual information made the information-theoretic methods themselves black-boxed, though they tried to understand and explain the inner mechanism of human information-theoretic processing.

In this context, the present paper tries to make the concept of information as concrete as possible and as interpretable as possible in terms of components of neural networks. We suppose that the information content can be described in terms of selectivity of components of neural networks. When the selectivity of components increases, they tend to have more information content. The research on the selectivity have been well discussed in neuro-sciences [27]–[33]. In addition, in the interpretation methods in the field of convolutional neural networks, many interpretation methods have actually focused on the detection of selectivity of some parts of neural networks to the coming inputs [34]–[39].

We here represent information content stored in neural networks in terms of selectivity of components. When the selectivity increases, more information can be stored. Thus, the information dealt with in this paper is called "selective information", and a learning method by using this selective information should be called "selective information-driven learning". Because we can maximize completely this concrete selective information and at the same time minimize it, we can interpret states with maximum and minimum information by which we can infer the actual internal representations with intermediate information content.

The paper is organized as follows. In Section 2, we first define the selective information and how to increase and decrease this selective information in the name of selective-information-driven learning. Then, we briefly explain how to compress connection weights partially and step by step to examine the information flow in multiple layers of neural networks. In Section 3, we present the results on the well-known application to the classification of wines. We first explain that the selective information could be maximized and minimized, producing different connection weights. When the selective information increases, more individually separated weights appeared, while more collective behaviors of several neurons appeared when the selective information decreased. Finally, we show that generalization and selectivity-based interpretation may be contradictory to each other, but the contradiction can be solved by supposing two types of information for different levels.

II. Theory and Computational Methods

We here explain the concept of selective information and how to compute it for multi-layered neural networks. In addition, we present how to compress partially multi-layered neural networks to examine the outputs from hidden layers.

A. Selective Information-Driven Learning

Now, let us begin with the definition of selectivity and selective information. For simplicity's sake, we compute the selectivity between the second and third layer denoted by the notation (2,3) in Figure 1. For the first approximation, we suppose that the strength of connection weights can be obtained by their absolute values. When the strength of absolute values of weights are larger, neurons connected with these weights are more strongly connected. As shown in Figure 1(a), in an initial stage of learning, all connection weights are randomly connected in all layers. When, the selective information is maximized, only one connection weight is connected with the corresponding neuron. When the selective information is minimized, we can have different types of states with minimum information. For example, as shown in Figure 1(c), all connection weights have equal and strong absolute values. Stronger connection weights may cause troubles in improving generalization and interpretation, we decrease the strength of connection weights as much as possible as shown in Figure 1(d). In this paper, we try to decrease the strength of connection weights when we try to minimize the selective information.

Then, we can obtain the absolute values of original weights from the second to the third layer,

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

where the notation (2,3) denotes the transition from the second to the third layer. Then, we normalize these values by their maximum one, which can be computed by

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

where maximum operation is over all connection weights between two layers.

Now, by using this normalized strength, selective information for the second to the third layer (2,3), can be computed by

$$G^{(2,3)} = n_2 n_3 - \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} g_{jk}^{(2,3)} \tag{3}$$

where n_2 and n_3 denote the number of neurons in the second and the third layer. Then, we try to increase or decrease this selective information, and in particular, we actively control this information to produce appropriate internal representations. When only one connection weight has some strength, while all the others are zero, the selective information is maximized $(n_2 n_3 - 1)$. On the contrary, when all connection weights have the same strength, the selective information become zero. In the extreme case, when no connection weights exist between two layers, the selective information is minimized by definition, because all connection weights have the same strength of zero.

To maximize the selective information, we must control the normalize strength g_{jk}. However, when we try to decrease this selective information, we need to reduce the strength

for all neurons between the layers. For this, we introduce an complement case of the original normalized strength

$$\bar{g}_{jk}^{(2,3)} = 1 - g_{jk}^{(2,3)} \tag{4}$$

This means that, when the strength increases, this inverse one decreases. This inverse equation have an effect to reduce the strength of larger connection weights. When the strength becomes larger, the corresponding connection weights are forced to be smaller. Then, by combining those two, we have a unified one

$$h_{jk}^{(2,3)} = \alpha g_{jk}^{(2,3)} + \bar{\alpha}\bar{g}_{jk}^{(2,3)} \tag{5}$$

where the parameter α ranges between zero and one. Then, the connection weights at the $t+1$th learning step is simply computed by

$$w_{jk}^{(2,3)}(t+1) = h_{jk}^{(2,3)} w_{jk}^{(2,3)}(t) \tag{6}$$

When the parameter α increases, weights tend to be more affected by the force to increase the strength. On the contrary, when the α decreases, the weights are more strongly affected by the force to decrease the strength as shown in Figure 1(d). When the information is maximized, only one connection weight is used to connect one with the other, while in the minimum state, all neurons are equally connected with each other. Usually, the learning starts with randomly initialized weights, corresponding to an intermediate information state. From this intermediate state, we can increase and decrease information flexibly, and we interpret the meaning of information in terms of the number of strong weights. On the contrary, information, for example, even when mutual information can be defined for the connection weights, it is extremely difficult to understand the actual meaning in terms of components of neural networks. However, the selective information, though defined very simply, can represent the content of mutual information concretely in terms of the number of strong connection weights.

1) Partial Compression : To examine information flow in multiple hidden layers, we here use the partial compression in which a multi-layered neural network is compressed gradually up to the simplest one without hidden layers in Figure 2.

Let us show how to compress a multi-layered neural network gradually from the input to the output layer. For this, we suppose that the number of neurons in all the hidden layer is the same. This does not necessarily exclude the variable number of neurons. Actually, the number of neurons can be reduced by the effect of selective information maximization, where the number of strong connection weights can be increased or decreased.

The first partial compression in Figure 2(b) lies in connecting the input and output layer

$$w_{ir}^{(1,2,7)} = \sum_{q=1}^{n_6} w_{iq}^{(1,2)} w_{qr}^{(6,7)} \tag{7}$$

where the notation (1,2,7) represents compressed weights up to the second layer, and n_6 is the number of neurons in the sixth

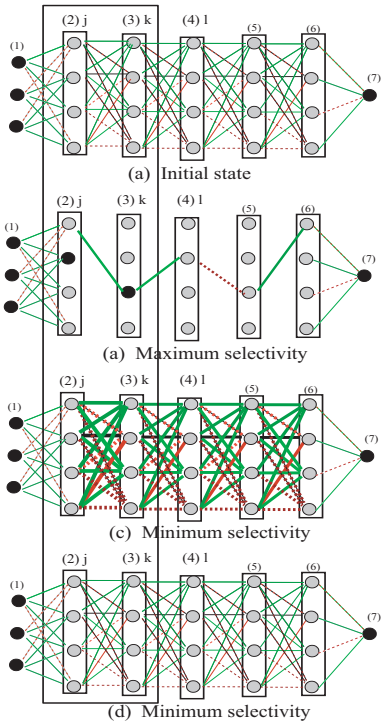

Fig. 1. Network architecture with seven layers, including five hidden layers (a) for active information maximization (b), information minimization No.1 (c) and selective informaiton minimization No.2 (d).

layer. The second partial compression in Figure 2(c) begins with connecting the first two connection weights

$$w_{ik}^{(1,2,3)} = \sum_{j=1}^{n_2} w_{ij}^{(1,2)} w_{jk}^{(2,3)} \tag{8}$$

where the notation (1,2,3) denotes compression up to the second layer. Then, by combing it with weights to the output layer, we have the second partial compression

$$w_{ir}^{(1,3,7)} = \sum_{q=1}^{n_6} w_{iq}^{(1,3)} w_{qr}^{(6,7)} \tag{9}$$

where the notation (1,3,7) shows the compression up to the third layer with the weights to the output layer. In the same way, we gradually combing the remaining connection weights, and finally, we can compress all connection weights

$$w_{ir}^{(1,6,7)} = \sum_{q=1}^{n_6} w_{iq}^{(1,5,6)} w_{qr}^{(6,7)} \tag{10}$$

where the notation (1,6,7) denote compression up to the sixth layer with the seventh layer or output layer, and (1,5,6) denotes compression up to the fifth layer.

III. Results and Discussion

We applied the method to the well-known wine data set to show how well the method could maximize and minimize selective information, and we could easily interpret the meaning of connection weights. In addition, the method could

 ISBN: 978-1-61208-847-1

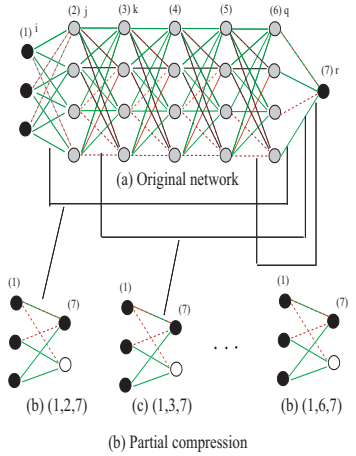

Fig. 2. Network architecture with seven layers, including five hidden layers (a) to be compressed gradually and partially into the simplest ones (b).

produce weights, which was close to the original correlations between inputs and targets. This means that the method could disentangle complicated relations into the simple ones.

A. Experimental Outline

The experiments aimed to predict red and white wine from the north of Portugal [40], based on twelve input variables. The number of patterns was 4898 white and 1599 red wines. As shown in Figure 3, the number of hidden layers was ten with ten neurons for each layer. We compressed multi-layered neural networks partially (b) and fully (c) for interpreting the information flow in multiple layers. The parameter α decreased from 1 to 0, which means that the selective information could be maximized and minimized. The number of learning steps was 100, and within each learning step, there were several sub-steps to assimilate the effect of selective information, ranging from 5 to 10. Though we can completely maximize and minimize the selective information, the selective information minimization was accompanied by the weight strength reduction, making the learning impossible when the information becomes closer to a minimum state. Thus, we made the effect of selective information minimization weaker for the stable learning. We should note that without considering the stability of learning, we could completely minimize the selective information. We used the scikit-learn neural network package with all default setting except the number of learning steps (epochs) and tangent-hyperbolic activation function, and naturally, connection weights were modified by the composite function to control the selective information. Those default values were used to make the reproduction of the present results as easy as possible.

B. Selective Information and Mutual Information

First, we try to show that the selective information can be controlled flexibly, reducing the strength of connection weights. In addition, the selective information can be easily understood in terms of the number of weights, while mutual information cannot give concrete meaning to the final results.

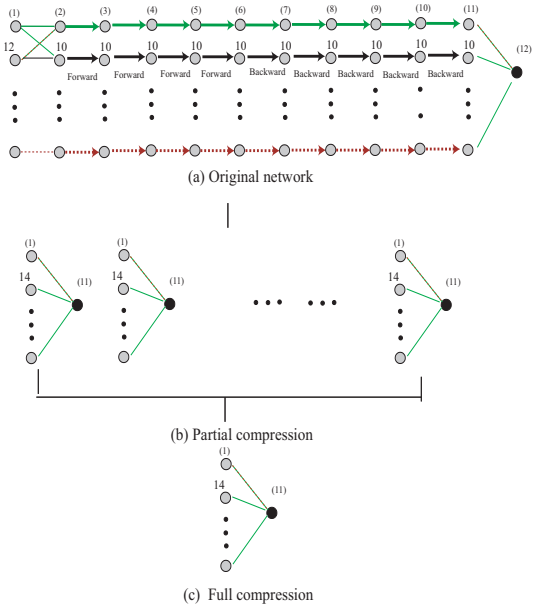

Fig. 3. Network architecture with 12 layers (10 hidden layers) in which each hidden layer has 10 neurons, a series of partial compression (b) and full compression (c) for the wine data set.

Figure 4 shows the selective information (left), mutual information (middle) and averaged absolute weight strength (right), when the parameter α decreased from 1(a) to 0(e). As shown in the leftmost figure of Figure 4(a), the selective information increased rapidly and close to its maximum value (10 by 10 by definition) in the end. Then, when the parameter decreased from 0.7(b) to 0(e), the selective information decreased gradually. When the parameter was zero in the leftmost figure of Figure 4(k), the selective information became slightly larger than that obtained when the parameter was 0.2 in Figure 4(j). As mentioned, to stabilize the learning, we reduced the effect of selective information when the parameter decreased and in particular, close to zero. Without this constraint on the selective information minimization, it could be reduced to almost an minimum point of zero. Thus, when the parameter decreased, the selective information was forced to be smaller as can be expected. In the same way, the figures in the middle shows mutual information, where mutual information increased immediately up to almost its maximum value. Then, when the parameter decreased, mutual information decreased gradually, and final close to zero. Then, the rightmost figures show the average strength of absolute weights, which were forced to be smaller by decreasing the parameter α.

As above mentioned, we can easily interpret the values of selective information. When the selective information is higher, the number of stronger weights becomes smaller. On the contrary, when the selective information is smaller, the number of stronger weights connecting with specific neurons becomes smaller, and all the weights become equal in their strength. At the beginning of learning, the random initial states are usually given, and the selective information in this case should be close to the middle of 50, meaning that the

 ISBN: 978-1-61208-847-1

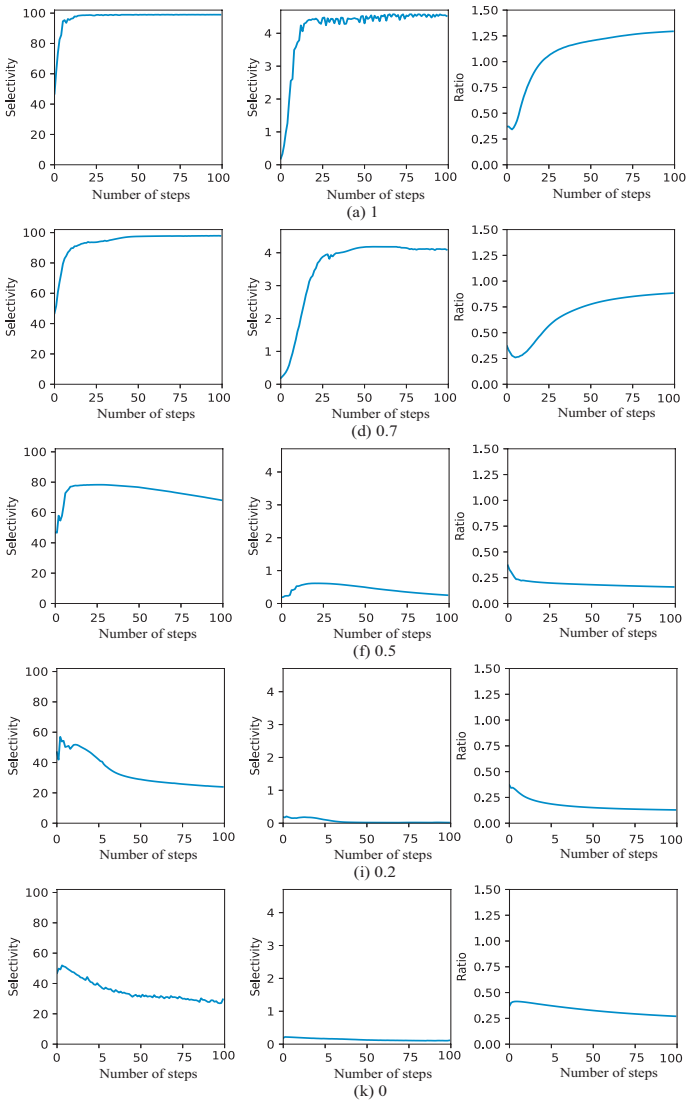

Fig. 4. Selective information (left), mutual information (middle) and averaged weights (right) when the parameter α decreased from 1(a) to 0(k) for the wine data set.

fifty percent of all connection weights should have stronger weights, though the strength of weights are different from each other. Then, when the selective information increases, the number of stronger weights becomes smaller and smaller and in the end when the maximum information state is reached, where only one connection weight becomes strong, while all the others are zero. On the contrary, when the selective information becomes smaller, the number of strong weights becomes smaller, and all weights become equally small. In the extreme case, all connection weights becomes zero, which is also a minimum selective information state by definition.

Then, we should examine how connection weights changed when the parameter also changed. Figure 5(a) shows connection weights when the parameter was one, namely, when only selective information maximization was applied. As can be expected, only a small number of connection weights among many became stronger, while all the others became very small. When the parameter decreased to 0.7 in Figure 5(b), a neuron connected with many neurons appear, which could be seen over weights close to the input and output layer. When the parameter was 0.5 in Figure 5(c), we could have an explicit pattern that all neurons in the precedent layers tended to be connected with many neurons in the subsequent layers, and vice versa. When the parameter was decreased to 0.2 in Figure 5(d), this tendency was further enhanced, and all neurons were explicitly connected with all the other neurons. Finally, when the parameter was set to zero, and only selective information minimization was applied, the tendency became slightly weaker. This can be explained by the fact that we made the effect of selective information weaker for the stability of learning. These results show that when the selective information increases, individual connection weights tend to behave independently of other weights. Then, when the selective information becomes smaller, connection weights behave collectively, and a neuron tend to be connected with many other neurons.

C. Partial Compression

Then, we tried to examine how information from the inputs and outputs were transmitted in multiple layers. The results show that the selective information maximization tried to capture output information mainly, while information minimization tried to take into account input information in addition to output information.

Figure 6 shows partially compressed weights when the parameter decreased from 1 (a) to 0 (e). When the parameter was one in Figure 6(a), and the selective information is maximized, partial compressed weights in the intermediate layers, were very weak in their strength, and only in the final compression state, namely, in the full compression, the weights became stronger. This means that information on the outputs should be necessary to form the compressed weights. When the parameter decreased from 0.7 (b) to 0.2 (d), gradually, in the initial stages of partial compression, located on the left hand side, connection weights became relatively stronger. This means that when the selective information becomes smaller, information on inputs was taken into account. Finally, when the parameter became zero, namely, when only the selectivity minimization was applied, the states of partial compression became similar to those by the selective information maximization only used in Figure 6(a).

For more clearly presenting this tendency, we computed the standard deviation of absolute connection weights of partially compressed weights in Figure 7. For example, Figure 7(a) shows the standard deviation of compressed weights when the parameter was one, and only selective information maximization was used. As can be seen in the figure, the standard deviation of weights from the first compression to the ninth compression was very small. When only all weights were compressed fully, the standard deviation became larger, meaning that the final connection weights played important roles to form the appropriate compressed weights. Then, when the parameter decreased from 1 (a) to 0.1 (j), gradually,

 ISBN: 978-1-61208-847-1

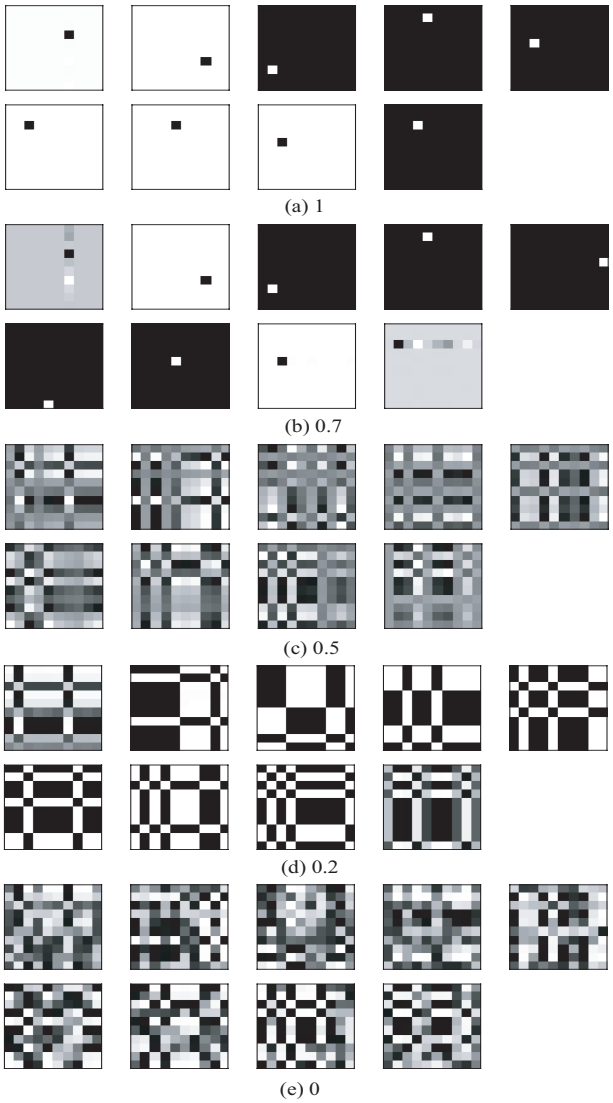

Fig. 5. Weights of hidden layers when the parameter α decreased from 1 (a) to 0 (e) for the wine data set.

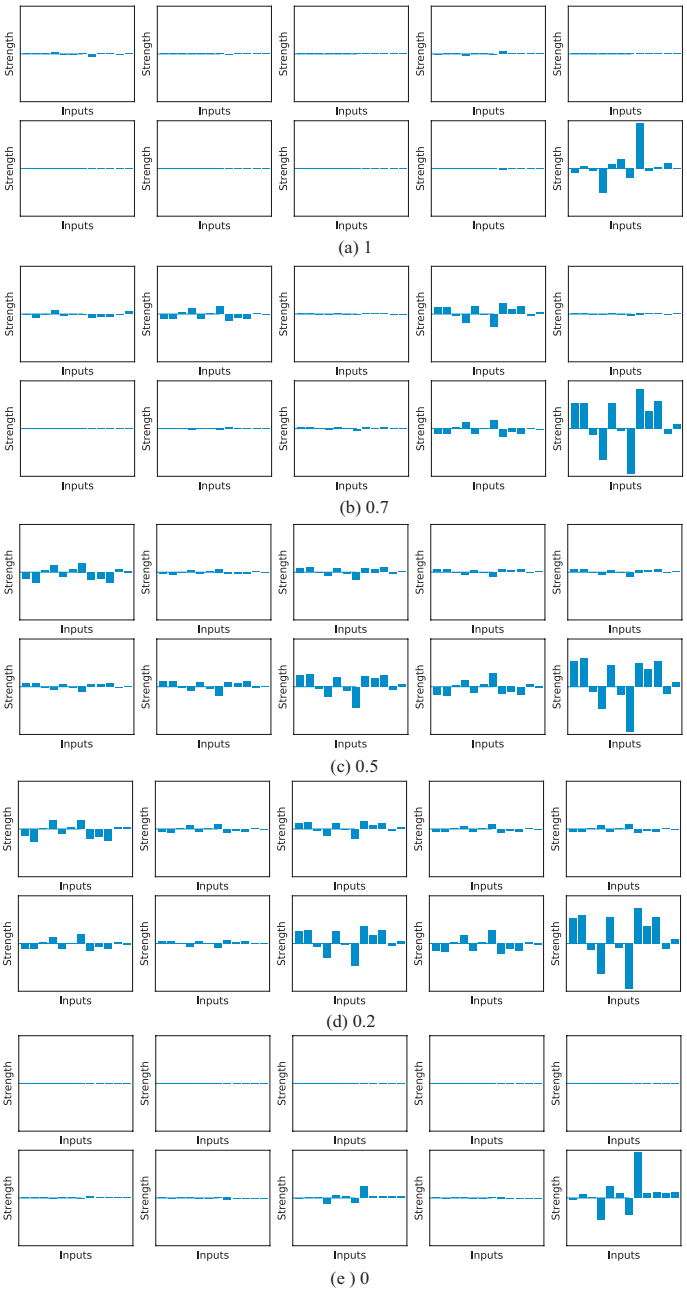

Fig. 6. Partially compressed weights when the parameter α changed from 1(a) to 0(e).

the standard deviation of absolute weights of initial stages of partial compression became stronger, though those compressed weights were still weaker. This means that when the selective information is decreased, connection weights in the intermediate layers tended to have some information, probably, on inputs. Finally, when the parameter was zero in Figure 7(k), the partially compressed weights became similar to those obtained when the parameter was one. This cannot be easily interpreted, but we infer that input information acquired in the intermediate layers, can be obtained only when selective information maximization and minimization effect are combined with each other. Or, as mentioned, we made the effect of selective information weaker, which may be the main cause of this state with the zero parameter.

D. Compressed Weight Comparison

We here compare the compressed weights with those by the other conventional methods. The results show that the final compressed weights were quite similar to the original correlation coefficients between inputs and targets. This means that the selective information has an effect to disentangle connection weights to have simpler, linear and independent relations between inpts and outputs.

Figure 8(a) shows the correlation coefficients between inputs and targets of the original data set. When the parameter was one in Figure 8(b), the correlation between the original correlations in Figure 8(a) and compressed weights was 0.673. When the parameter decreased to 0.7, 0.5, 0.2, the correlation increased to 0.953, 0.946 and 0.958, which were almost perfect correlations. Those correlations were higher than 0.937 of

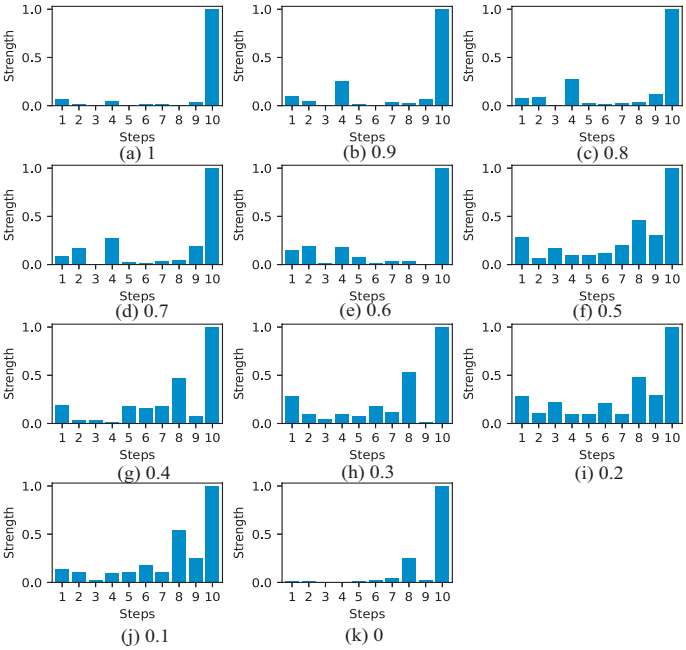

Fig. 7. Standard deviation of partially compressed weights when the parameter α decreased from 1 (a) to 0 (k) by 0.1 for the wine data set.

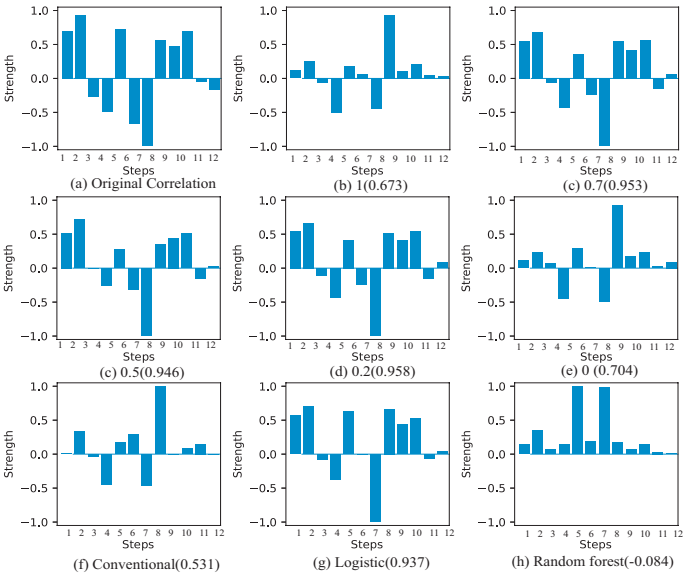

Fig. 8. Correlation coefficients (a), compressed weights for α=1 down to 0 (b)-(e), compressed weights by the conventional method (f), and the regression coefficients by the logistic regression analysis (g) and prediction importance by the random forest method (h) for the wine data set.

conventional logistic regression analysis in Figure 8(g). By the conventional method without selective information, the correlation was only 0.531 in Figure 8(f), and in addition, the random forest produced the worst correlation of -0.084 in Figure 8(h).

Finally, we examined relations between correlations and generalization. The results show that the interpretation and generalization were inversely correlated. Thus, we need to make an attempt to unify improve interpretation and generalization. Figure 9(a) shows the correlation coefficients between

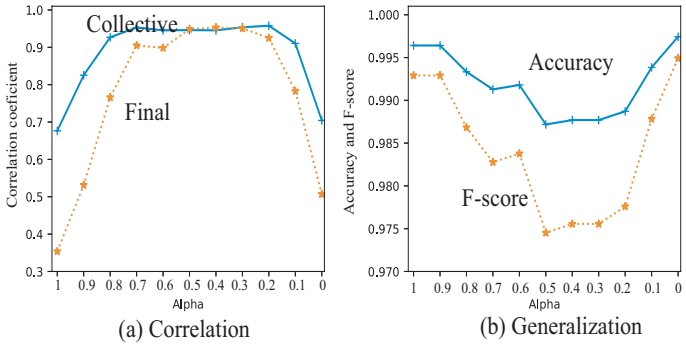

Fig. 9. Correlation coefficients (a) and generalization accuracy (b) for the wine dat set.

the correlations of original data set and compressed (final) and collective weights (collective). The compressed weights were ones, obtained when the learning steps was the final one, namely, 100 step in this case. On the other hand, the collective weights were obtained by averaging all compressed weights for all intermediate learning steps. As shown in the figure, the collective weights produced always higher correlations than the compressed weights. This means that the simple average of all compressed weights in learning can increase the correlation between the original correlation and compressed weights. Figure 9(b) shows generalization accuracy and F-measure, where the accuracy was always larger than the F-measure. Comparing Figure 9(a) and (b), we can conclude that the correlations were inversely related to generalization performance, though decrease in generalization was considerably small. This can be easily interpreted by using the selectivity of neurons and connection weights. When the selectivity of components of neural networks increases, and they tend to respond to the inputs very specifically, they cannot deal with less specific inputs naturally. Thus, we make a compromise between selectivity and generalization or interpretation and generalization. At the first glance, it seems to be impossible to make this kind of compromise between them, but this type of contradiction can be easily be solved by supposing selective information maximization and minimization operating in two different contexts or levels. We should explore this possibility of contradiction resolution as a future study of this paper.

IV. Conclusion

The present paper aimed to propose a new type of information-theoretic method called “selective information-driven learning”. The selective information is introduced to measure the selectivity of components in neural networks to replace conventional mutual information, because it can easily be interpreted in terms of the number of strong weights, while conventional mutual information can not be easily interpreted in terms of components of neural networks. The new method was applied to the well known wine data set. The experimental results showed that the selective information could be maximized and at the same time minimized within the same framework. The interpretation can be possible in terms of number of strong connection weights, which is easier to understand,

 ISBN: 978-1-61208-847-1

compared with the conventional mutual information. In addition, by partially compressing multi-layered neural networks, we could find that the selective information maximization is focused on output information, while the selective information minimization tried to detect input information as well.

The results confirmed that the explicit interpretation of internal representations is possible by the present method. However, it was observed that better interpretation is not necessarily followed by improved generalization. This is because the selectivity for improve interpretation may be harmful to improved interpretation, needing an ability to responding well non-specific and ambiguous cases. Thus, we need to make further studies on unifying improved interpretation and generalization.

REFERENCES

[1] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a right to explanation," *arXiv preprint arXiv:1606.08813*, 2016.

[2] M. Sendak, M. C. Elish, M. Gao, J. Futoma, W. Ratliff, M. Nichols, A. Bedoya, S. Balu, and C. O'Brien, "" the human body is a black box" supporting clinical decision-making with deep learning," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 99–109, 2020.

[3] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, Springer, 2014.

[4] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437, 2017.

[5] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv preprint arXiv:1702.04595*, 2017.

[6] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," *arXiv preprint arXiv:1711.06104*, 2017.

[7] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," in *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, Springer, 2019.

[8] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[9] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[10] M. N. Angenent, A. P. Barata, and F. W. Takes, "Large-scale machine learning for business sector prediction," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pp. 1143–1146, 2020.

[11] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.

[12] D. Rumelhart and J. M. et al., *Parallel Distributed Processing*, vol. 1. MA: MIT Press, 1986.

[13] D. E. Rumelhart and D. Zipser, "Feature discovery by competitive learning," *Cognitive Science*, vol. 9, pp. 75–112, 1985.

[14] D. E. Rumelhart, G. E. Hinton, and R. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing* (D. E. Rumelhart and G. E. H. et al., eds.), vol. 1, pp. 318–362, Cambridge: MIT Press, 1986.

[15] D. E. Rumelhart and J. L. McClelland, "On learning the past tenses of English verbs," in *Parallel Distributed Processing* (D. E. Rumelhart, G. E. Hinton, and R. J. Williams, eds.), vol. 2, pp. 216–271, Cambrige: MIT Press, 1986.

[16] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[17] R. Linsker, "How to generate ordered maps by maximizing the mutual information between input and output signals," *Neural computation*, vol. 1, no. 3, pp. 402–411, 1989.

[18] R. Linsker, "Local synaptic learning rules suffice to maximize mutual information in a linear network," *Neural Computation*, vol. 4, no. 5, pp. 691–702, 1992.

[19] R. Linsker, "Improved local learning rule for information maximization and related applications," *Neural networks*, vol. 18, no. 3, pp. 261–265, 2005.

[20] S. Becker, "Mutual information maximization: models of cortical self-organization," *Network: Computation in Neural Systems*, vol. 7, pp. 7–31, 1996.

[21] K. Torkkola, "Nonlinear feature transform using maximum mutual information," in *Proceedings of International Joint Conference on Neural Networks*, pp. 2756–2761, 2001.

[22] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of Machine Learning Research*, vol. 3, pp. 1415–1438, 2003.

[23] J. M. Leiva-Murillo and A. Artés-Rodríguez, "Maximization of mutual information for supervised linear feature extraction," *Neural Networks, IEEE Transactions on*, vol. 18, no. 5, pp. 1433–1441, 2007.

[24] M. M. Van Hulle, "The formation of topographic maps that maximize the average mutual information of the output responses to noiseless input signals," *Neural Computation*, vol. 9, no. 3, pp. 595–606, 1997.

[25] J. C. Principe, D. Xu, and J. Fisher, "Information theoretic learning," *Unsupervised adaptive filtering*, vol. 1, pp. 265–319, 2000.

[26] J. C. Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

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

[28] A. Schoups, R. Vogels, N. Qian, and G. Orban, "Practising orientation identification improves orientation coding in v1 neurons," *Nature*, vol. 412, no. 6846, pp. 549–553, 2001.

[29] L. E. White, D. M. Coppola, and D. Fitzpatrick, "The contribution of sensory experience to the maturation of orientation selectivity in ferret visual cortex," *Nature*, vol. 411, no. 6841, pp. 1049–1052, 2001.

[30] H. Ko, S. B. Hofer, B. Pichler, K. A. Buchanan, P. J. Sjöström, and T. D. Mrsic-Flogel, "Functional specificity of local synaptic connections in neocortical networks," *Nature*, vol. 473, no. 7345, pp. 87–91, 2011.

[31] J. F. Jehee, S. Ling, J. D. Swisher, R. S. van Bergen, and F. Tong, "Perceptual learning selectively refines orientation representations in early visual cortex," *Journal of Neuroscience*, vol. 32, no. 47, pp. 16747–16753, 2012.

[32] M. V. Peelen and P. Downing, "Category selectivity in human visual cortex," 2020.

[33] B. J. Bongers, A. P. IJzerman, and G. J. Van Westen, "Proteochemometrics–recent developments in bioactivity and selectivity modeling," *Drug Discovery Today: Technologies*, 2020.

[34] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, and M. Botvinick, "On the importance of single directions for generalization," *arXiv preprint arXiv:1803.06959*, 2018.

[35] I. Rafegas, M. Vanrell, L. A. Alexandre, and G. Arias, "Understanding trained cnns by indexing neuron selectivity," *Pattern Recognition Letters*, vol. 136, pp. 318–325, 2020.

[36] J. Ukita, "Causal importance of low-level feature selectivity for generalization in image recognition," *Neural Networks*, vol. 125, pp. 185–193, 2020.

[37] M. L. Leavitt and A. Morcos, "Selectivity considered harmful: evaluating the causal impact of class selectivity in dnns," *arXiv preprint arXiv:2003.01262*, 2020.

[38] W. J. Johnston, S. E. Palmer, and D. J. Freedman, "Nonlinear mixed selectivity supports reliable neural computation," *PLoS computational biology*, vol. 16, no. 2, p. e1007544, 2020.

[39] M. L. Leavitt and A. S. Morcos, "On the relationship between class selectivity, dimensionality, and robustness," *arXiv preprint arXiv:2007.04440*, 2020.

[40] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.

 ISBN: 978-1-61208-847-1