Controlling Individual and Collective Information for Generating Interpretable Models of 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 Univeirsity
4-1-1 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan
email: ryotarokami@gmail.com

***Abstract*—The present paper aims to control selective information to understand the main mechanism of information processing in multi-layered neural networks. We propose two types of selective information, namely, individual and collective selective information, or simply, individual and collective information. The individual information represents to what degree a neuron is connected specifically to another one, and it should be increased as much as possible. Then, we try to use this abundant information as impartially as possible, reducing the specificity of collective neurons and reducing collective information. By controlling the ratio of individual and collective information, we can realize a number of different types of states to be interpreted, leading to the interpretation of the inference mechanism. The method was applied to the bankruptcy data set. In the experiments, we successfully increased individual information and decreased collective information. By examining partially compressed weights, we could see how neural networks, by controlling the selective information, can process information content in multi-layered neural networks. This examination of information flow can lead us to understand the main inference mechanism of neural networks.**

***Keywords*—individual, collective, information, selectivity, partial compression, interpretation, generalization**

I. Introduction

The black-box property of neural networks, as well as machine learning have caused much confusion in their applications, as well as model evaluation [1]–[6]. The black-box presupposition for human intelligence has been one of the major concerns in the possible introduction of machine learning into our society [7]–[10]. From a certain viewpoint, the black box is necessary, because the studies on human intelligence have been so far very limited and premature. However, we think that this confusion seems to be due to the different types of objectives in model creation and application. When we try to model human intelligence, our nervous systems, and cognitive systems in terms of neural networks, it is absolutely necessary to understand and interpret the main inference mechanism inside. The objective of these types of studies is to reduce the black-box properties as much as possible. This approach was very active in the early stage of development of neural networks, as has been well known in terms of connectionism [11]–[14]. The neural networks were considered models to create new information or knowledge by which we could deepen the understanding our nervous and cognitive systems.

On the contrary, if we try to apply models to practical problems, the objective is not necessary to understand human intelligence or cognitive processes, but to apply well and appropriately the models to practical problems naturally. From this point of view, the recent development of neural networks, as well as machine learning seem to be focused on the application to the practical problems. For example, the Convolutional Neural Networks (CNN) have been greatly developed recently, but they have been based on the simplified models of visual nervous systems [15]–[18], developed in the early seventies. However, the simplicity of the models, inherited only partially from the properties of our nervous systems [18], has made it possible to improve their prediction performance in an unexpected way. In other words, they have tried to extend some parts of our visual nervous systems to many different types of practical applications, keeping the main mechanism of optical systems only partially known.

As mentioned above, those methods have aimed not to understand the main nervous systems but to improve recognition and prediction performance, and the black-box property has been not so serious as had been expected. Naturally, there have been many attempts to interpret the inference mechanism in the field of convolutional neural networks due to the urgent need to respond to the right of explanation [9]. However, the majority of methods have been focused on the individual interpretation of neural networks for specific input patterns. Many attempts have been made to examine what kind of features among given input patterns are extracted in components of neural networks [19]–[26]. Since we have not known the main inference mechanism inside neural networks, all we can do is to uncover partially and step by step the characteristics related to specific input patterns. It should be repeated that this prevailing approach is very natural, because the convolutional neural network itself does not aim to clarify the human visual nervous framework itself, but it tries to enhance the simple and easily accessible parts of nervous systems as much as possible to strengthen the model performance itself.

Thus, we can say that one of the main problems of interpretation is that the objects to be interpreted have been ambiguous. If our objective is to apply the model to practical applications, the present models of interpretation, prevailing in the convolutional neural networks, may be sufficient. But if we try to understand the main mechanism of human intelligence and cognitive processes in terms of neural networks, we should think of a different type of interpretation.

Actually, there were serious attempts in neural networks to understand the inference mechanism of living systems, namely, the information-theoretic methods. One of the most important approaches is the maximum information preservation by Linsker [27]–[30]. He tried to understand the visual nervous mechanism by supposing that neural networks try to increase mutual information as much as possible. Then, from this maximum information principle, he tried to explain the actual phenomena observed in living systems. In terms of model interpretation, maximum information preservation tried to interpret human nervous systems in terms of information storage and transmission in our neural systems.

Though this approach seemed to be a good starting point for interpretation, the definition and computation of mutual information inside cannot be necessarily and successfully used in the field of neural networks. Though many different types of methods have been developed so far [31]–[37], one of the main problems is the too-abstract property of mutual information, as well as other information measures applied to the neural networks. Thus, if we try to understand the inference mechanism in terms of information storage and transmission, we need to make this abstract property of mutual information more concrete and easily interpretable to be applicable to the interpretation of neural networks.

We should repeat again that one of the major problems with the Linsker-type maximum information preservation, when it is applied to neural networks, is how to measure information content in concrete components in neural networks. The present paper aims to realize the concept of mutual information in more concrete ways. In mutual information maximization in terms of neurons' information processing, two contradictory terms of entropy maximization and conditional entropy minimization exist. Using more concrete terms, individual neurons should respond very specifically to inputs, and at the same time, those neurons should respond very uniformly to any inputs on average or collectively. In living systems, neurons should be used as equally as possible on average, and at the same time, each neuron should respond as unequally as possible individually. The living systems must cope with many different types of new inputs and situations, and thus they need to use any resources inside as much possible, and at the same time, they need to know the properties of incoming inputs as much as possible. Thus, this information maximization principle states that the specific responses or larger information to targets, should be compensated for by the uniform use of components or smaller information from a collective point of view. In other words, it is necessary to have larger information, but this larger information should be distributed over as many components as possible.

Let us apply this knowledge of living systems from an information-theoretic view more concretely to neural networks. We suppose here that the information content is concretely measured in the selectivity of components such as neurons and connection weights. When we try to measure information content in neural networks, the selectivity of components should play an important role. Thus, information on inputs is stored and transmitted in terms of selectivity of components. When the selectivity of neurons toward inputs becomes higher, we should say that the neurons tend to have more information on inputs. Thus, we need to define the selectivity of components and how to control it.

The importance of selectivity in actual living systems, the interpretation methods, and in the field of improved generalization has been already pointed out. First, for actual living systems, it has been said that the selectivity should play an important role, discussed in the literature on neural sciences [38]–[44]. Second, as mentioned above, in the field of convolutional neural networks, to address the right to the explanation [9], a number of attempts have been made to interpret information processes in the networks. We think that the majority of methods have been based on the selectivity of components of neural networks [45]. Roughly speaking, the majority of interpretation methods have tried to determine which parts or components in multi-layered neural networks represent the distinctive or common features of input patterns. In other words, they have tried to determine which parts or components try to respond to inputs selectively. The interpretation of neural networks in this case corresponds to the determination of specific components for specific inputs. Third, though the paper does not deal with generalization, we should note that that generalization is directly related to the selectivity of components [45]–[50]. For example, when a neuron responds too specifically to some inputs, generalization performance cannot be improved, because it cannot respond appropriately to ambiguous input patterns. For improving generalization, we need to weaken the specific responses of neurons naturally.

This consideration leads us to say that the selectivity should be appropriately controlled in living systems for coping with uncertain conditions. Thus, we need to understand how this selectivity can be controlled in living systems and how they try to deal with uncertain conditions and unseen situations. The present paper aims to control the selectivity or selective information to explore how information should be stored and transmitted in neural networks.

One of the main hypotheses in this paper is to suppose that there is a variety of types of selectivity, and we should control those variants to cope with coming unseen inputs . In this paper, we suppose two types of selective information, namely, individual and collective selective information, or more simply, by eliminating the word "selective," individual and collective information. Individual information represents how much a component responds to an input specifically, namely, the information content an individual component has. On the contrary, collective information represent information pooled collectively by many neurons. Individual and collective information in neural networks are not necessarily in harmony with each other as is the case with human society. Thus, we need to make a compromise between individual and collective information in actual neural learning.

For simplicity's sake, we suppose here that individual information is naturally information on a specific and individual input, and it should be increased as much as possible. On the contrary, regarding collective information, it is supposed that we to collect as much information as possible on any inputs, and thus it is necessary for collective information to be as non-specific as possible to inputs. More technically, we suppose that individual information should be increased, while collective information should be decreased as much as possible. Then, we should try to examine what properties can be extracted when controlling two types of selective information.

As mentioned above, the maximum information principle by Linsker does not necessarily state that information should be simply maximized. Information in the individual level can be increased under the condition that, in the collective level, information should be decreased. More concretely, information increase for individual components must be compensated for by information decrease from a collective point of view. In complex living systems, there are many different types of contradictions, where a contradiction can be resolved in a level, but in another level, the contradiction can be stronger. Thus, for the maximum information principle, we need to develop a composite information function by which information can be increased and decreased at the same time.

The present paper does not simply increase the selectivity of components. We try to control the selectivity of components individually and collectively. Then, we try to examine how information is stored and transmitted by changing the selectivity individually and collectively. We try to understand the main information processing mechanism in multi-layered neural networks by which we can explain the generation of individual inference mechanisms for different input patterns.

The paper has been organized as follows. In Section 2, we present how to compute two types of selective information, namely, individual and collective information. Then, we try to present how to modify connection weights by changing the ratio of individual to collective information. This method to modify connection weights is called “selective information-driven learning.” Then, to understand the information flow in multi-layered neural networks, we show how to compress multi-layered neural networks step by step, namely, partial compression. In Section 3, we applied the method to the bankruptcy data set. In the experiments, we tried to show that the selective information could easily be controlled, and this control was directly related to which parts of information were obtained by multi-layered neural networks. By changing the ratio, we could produce compressed networks whose connection weights were quite similar to the original correlation coefficients between inputs and targets. In addition, we could extract and choose a few weights necessary only for prediction. The results show that the selective information control can generate more interpretable networks by which we can understand how information is processed in a multi-layered neural network.

II. Theory and Computational Methods

This section describes the selective information with individual and collective information. Then, we explain how to compress fully and partially networks into the simplest ones.

A. Selective Information Control

As discussed above, information contained in neural networks can be represented in terms of selectivity or selective information in their components. Thus, selective information is supposed to represent information content stored and transmitted in neural networks. Depending on given objectives, we need to control selective information appropriately, corresponding to the information control in neural networks. For example, too much selectivity of components is not good for responding to new inputs, but it is needed to specify the roles of components for the interpretation. In addition, for the efficient use of components, we need to treat those components as equally as possible. This means that we need to reduce the selectivity of components when we try to use the resources of components as much as possible. Then, the selectivity of components should be reduced, where each component should have equal importance collectively. Thus, we suppose that the selective information should be increased when components are treated individually. On the contrary, when they are treated collectively, the selectivity should be reduced.

For realizing this situation, we suppose two type of selective information. In one type of selective information, called “individual information,” the information should be increased as much as possible. Each component should respond very specifically to inputs. On the contrary, in the other type of selective information, called “collective information,” the information should be decreased as much as possible. All components should have the same importance; that is, all components should respond equally to inputs. By mixing these two types of information, we can control the selectivity of components. Naturally, it is not so easy to maximize and at the same time minimize the selective information in the same level. However, the selective information increase and decrease are performed in different levels, namely, individual and collective ones, and it is possible to make a compromise between two types of different selectivity control.

B. Individual Information

We explain here how to compute individual information. The individual information measures how much a neuron is connected with the corresponding one specifically. For this, we use a network architecture shown in Figure 1, in which the number of layers is seven, including the first input and the last seventh output layer. For simplicity’s sake, we define them by connection weights between the second and third layer and neurons in the third layer in Figure 1. In our experiments, we tried to control the selective information only for hidden layers, because it is easy to do it in the hidden layers due to less input and output information.

First, the strength of connection weights can be computed by the absolute value

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

where $(2, 3)$ denotes a transition from the second to the third layer. This strength denotes the strength of connecting a neuron with the other specific one. Because the strength is changeable, depending on different neurons, we normalize it by its maximum value

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

By this normalized strength, we have individual information for the jth neuron

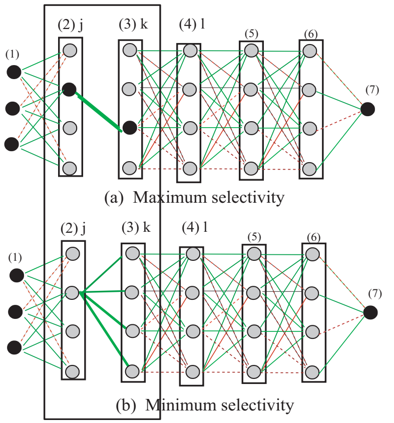
$$h_j^{(2,3)} = n_3 - \sum_{k=1}^{n_3} z_{jk}^{(2,3)} \tag{3}$$

where n_3 is the number of neurons in the third layer. This individual information increases when the number of stronger neurons decreases. Finally, when a neuron is connected only with one specific neuron, the selective information becomes maximum $(n_3 - 1)$ in Figure 1(a). On the contrary, when the neuron is connected equally with all neurons, the selective information becomes minimum (zero) in Figure 1(b). When all connection weights are zero, the selective information should be zero by definition, because all connection weights have the same value of zero.

For the overall property of this selective information, by averaging it, we have the final individual information

$$h^{(2,3)} = \frac{1}{n_2} \sum_{j=1}^{n_2} h_j^{(2,3)} \tag{4}$$

This individual information roughly corresponds to conditional entropy in mutual information. However, this definition of information in terms of selectivity is more easily interpreted. Finally, we should note that, when we try to control the individual information, we should change weights according to the normalized value of z_{jk}.



(a) Maximum selectivity

(b) Minimum selectivity

Fig. 1. Individual information with maximum (a) and minimum (b) selectivity.

C. Collective Information

We increase individual information and at the same time decrease the collective information. This is because we try to use all components as equally as possible for their efficient use, but we try to understand the meaning of each component as much as possible for interpretation. Simultaneous information maximization and minimization has difficulty for a compromise to be made between them. However, when those contradictory operations are performed in different levels, namely, individually and collectively , it is possible to do it.

Then, the collective information should denote information, independently of specific neurons, as well as connection weights. Thus, as shown in Figure 2(a) for the collective information, we sum the strength of connection weights

$$u_k^{(2,3)} = \sum_{j=1}^{n_2} u_{jk}^{(2,3)} \tag{5}$$

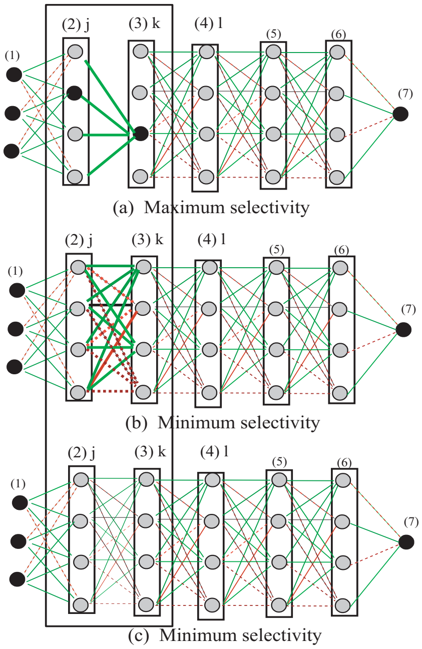
where n_2 denotes the number of neurons in the second layer. The normalized strength is computed by

$$v_k^{(2,3)} = \frac{u_k^{(2,3)}}{\max_{k'} u_{k'}^{(2,3)}} \tag{6}$$

Using this normalized strength, collective information is defined by

$$g^{(3)} = n_3 - \sum_{k=1}^{n_3} v_k^{(2,3)} \tag{7}$$

When this collective information, all connection weights to the corresponding neurons in the third layer becomes stronger, as shown in Figure 2(a). On the contrary, when the information decreases, all connection weights to the neurons become equal, as shown in Figure 2(b). As mentioned above, all connection weights happen to be very strong when the information is maximized, as shown in Figure 2(b). Thus, we need to reduce the strength of weights as much as possible for the minimum information states shown in Figure 2(c). Collective information minimization is a good candidate for information minimization, because this minimization aims to make all connection weights equal only collectively, meaning that some weights may be relatively stronger. This property can be good at compromising between individual and collective information.



(a) Maximum selectivity

(b) Minimum selectivity

(c) Minimum selectivity

Fig. 2. Collective information with maximum selective information (a) and minimum selective information (b) and (c).

D. Selective Information-Driven Learning

We must control two types of selective information by actually changing the weights. However, to control the selective information, all we have to do is to change weights by the normalized strength of connection weights z and v. For this purpose, we introduce a composite measure, combining the normalized strength of weights

$$d_{jk}^{(2,3)} = \alpha z_{jk}^{(2,3)} + \bar{\alpha}\bar{v}_k^{(3)} \tag{8}$$

where parameter α ranges between zero and one, $\bar{\alpha} = 1 - \alpha$, and $\bar{v} = 1 - v$. When the parameter α increases, the effect of individual information increases. On the contrary, when the parameter α decreases and $\bar{\alpha}$ increases, the effect of collective information increases. When the effect of $\bar{\alpha}$ increases, the strongest weights are forced to be smaller, keeping smaller ones relatively the same. Eventually, all connection weights become equal and smaller.

Then, for the connection weights from the second to the third layer in the tth learning step, the weights are changed simply by

$$w_{jk}^{(2,3)}(t+1) = d_{jk}^{(2,3)}(t)\ w_{jk}^{(2,3)}(t) \tag{9}$$

When the composite measure d is applied, weights should be updated with the normal error minimization method to assimilate the effect of the composite measure.

E. Partial Compression

To examine more carefully the information flow by the selective information control, we try to see how connection weights are changed when going through multiple hidden layers. For this purpose, we introduce partial compression, where an original multi-layered neural network is gradually compressed into the simplest one without hidden layers for interpretation. For simplicity's sake, the number of neurons in any hidden layers was the same. This assumption of an equal number of neurons in hidden layers does not necessarily mean that we do not consider a different number of neurons in hidden layers. The reduction in the number of neurons can be actually realized by suppressing the number of connection weights by the present method.

Figure 3 shows how to compress a multi-layered neural network step by step. In the first partial compression, as shown in Figure 3(a), we immediately compress the input and output layers, skipping all hidden layers,

$$w_{iq}^{(1,7)} = \sum_{q=1}^{n_6} w_{iq}^{(1,2)} w_{qr}^{(6,7)} \tag{10}$$

This shows only information contained in the input and output layers.

Then, we suppose that two connection weights from the first to the second layer, represented by (1,2), and between the second and the third layer, represented by (2,3), are combined into

$$w_{ik}^{(1,2,3)} = \sum_{j=1}^{n_2} w_{ij}^{(1,2)} w_{jk}^{(2,3)} \tag{11}$$

where (1,2,3) represents compression that is performed up to the second layer. In the second compression in Figure 3(b), we combine this compressed weight with the final output layer

$$w_{ir}^{(1,3,7)} = \sum_{q=1}^{n_6} w_{iq}^{(1,2,3)} w_{qr}^{(6,7)} \tag{12}$$

where (1,3,7) denotes that compression is performed up to the third layer. We compress the remaining connection weights in the same way in Figure 3(c) and (d), and finally, we can compress all layers into the simplest ones in Figure 3(e)

$$w_{ir}^{(1,6,7)} = \sum_{q=1}^{n_6} w_{iq}^{(1,6,6)} w_{qr}^{(6,7)} \tag{13}$$

where $(1, 6, 7)$ shows that compression is performed up to the sixth layer, namely, full compression. Those compressed weights aim to represent the main characteristics of overall connection weights. The compressed weights can be computed by multiplying connection weights of all routes from an input to the corresponding output.

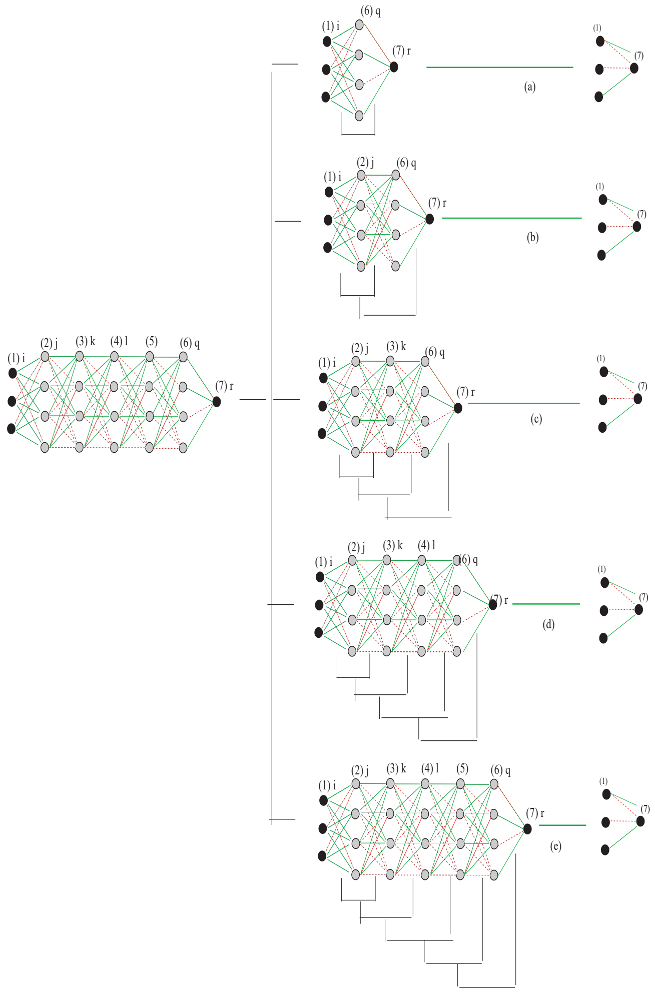


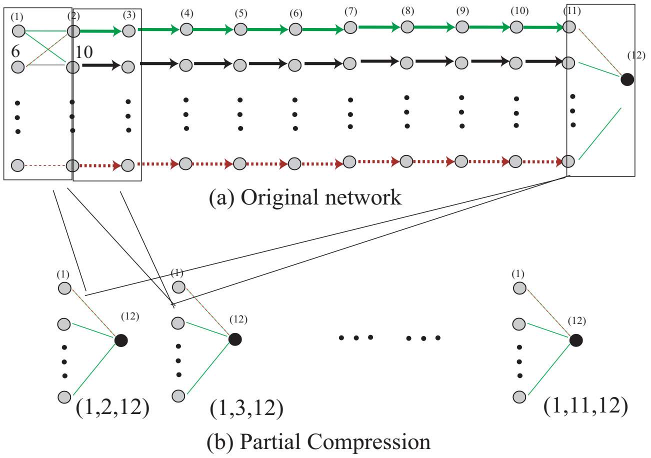
Fig. 3. Network architecture with seven layers, including five hidden layers (a) to be compressed step by step into the simplest ones (b)-(e).

III. Results and Discussion

This section present the experimental results applied to the bankruptcy data set, in which we tried to show that individual and collective information could be controlled to capture input or output information. In addition, behind complicated connection weights, we could find very simple, independent, and individual relations between inputs and outputs.

A. Results of Bankruptcy Data Set

1) Experimental Setting: The experiment aimed to predict bankruptcy by six qualitative input variables: industrial risk, management risk, financial flexibility, credibility, and operating risk [51]. As shown in Figure 4, the number of input, hidden, and output neurons was 6, ten, and one, and the number of hidden layers was ten. We used the partial compression in which compression was performed step by step by multiplying connection weights in higher hidden layers as shown on the lower side of the figure in Figure 4. The number of input patterns was 250, where 70 percent and the remaining 30



(a) Original network

(b) Partial Compression

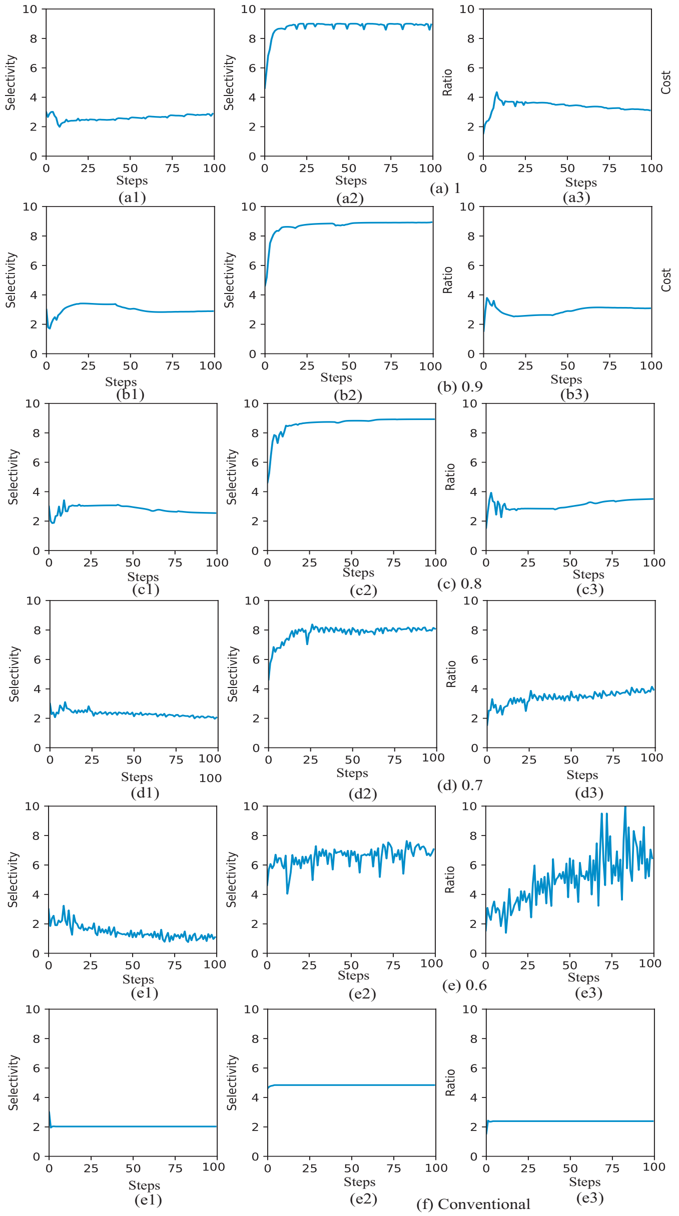
Fig. 4. Network architecture (a) and a process of partial compression (b) for the bankruptcy data set. The notation (1,2,12) denotes that compression is applied up to the second layer.

percent of the data set was for training and testing. The generalization performance was almost perfect by the present method, as well as the other conventional methods such as logistic regression and random forest. However, connection weights and other importance measures were different by different methods. Thus, this is a good benchmark data set for comparing the final representations by different methods. We used the scikit-learn neural network package with almost all parameters set to default values, except for the number of epochs and tangent-hyperbolic activation function, for making the reproduction of the present results as easy as possible. In addition, the effect of selective information is forced to be applied many times (up to 5 times), in direct proportion to the number of learning epochs. This property tended to make connection weights extremely large or small due to the repeated assimilation procedures. Thus, we reduced the effect of selective information as follows:

$$d_{jk}^{(2,3)} = \theta \left[\alpha z_{jk}^{(2,3)} + \bar{\alpha} \bar{v}_k^{(3)} \right]^{\beta} \qquad (14)$$

Newly added parameters θ and β were used only for reducing the excessive effects of selective information by repeating the assimilation processes (up to 5 times). The actual values of the parameter θ and β were 0.99 and 0.9, respectively.

2) Selective Information Control: In the first place, we try to show that the present method can decrease collective information and at the same time increase individual information by changing the parameter α. Figure 5 shows collective and individual information and the ratio of individual to collective information for the last hidden layer, namely, from the tenth to the eleventh layer for the bankruptcy data set. We chose the most typical hidden layer, where selective information could be controlled the most explicitly by the present method. As shown in Figure 5(a1), collective information increased very slightly, and individual information increased immediately in the first place and remained almost the same in the later stages of the learning steps. Thus, the ratio of individual to collective information stayed almost the same throughout the entire learning steps. This can be explained by the fact that, when the parameter was one, only individual information was forced to be increased. When the parameter decreased from 0.9 (b) to 0.7 (d), individual information was forced to increase more slowly. Collective information gradually decreased, and the ratio of collective to individual information increased gradually. In particular, when the parameter was 0.6 in Figure 5(e3), the ratios increased considerably when the learning steps increased. However, we could see some fluctuations, showing difficulty in controlling two types of information. Note that, when the parameter was increased further, we could not obtain stable results. Finally, without information control, no changes in selective information and its ratios could be seen in Figure 5(f). These results show that the present method can control two types of selective information, where collective information can be decreased, and at the same time, individual information can be increased, though some difficulty could be seen when the parameter was forced to be smaller and we must compromise between individual and collective information.



(a) 1

(b) 0.9

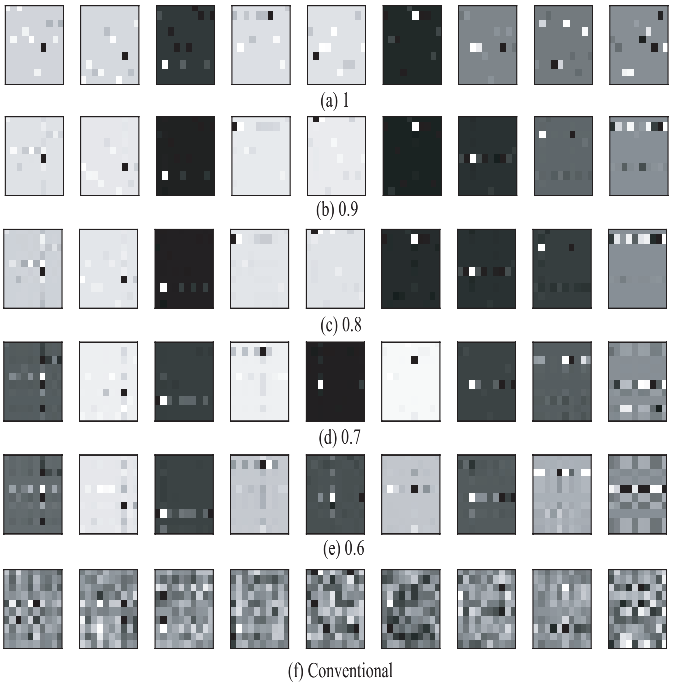
(c) 0.8

(d) 0.7

(e) 0.6

(f) Conventional

Fig. 5. Collective information (1), individual information (2), and the ratio of individual to collective information (3) when the parameter decreased from 1 (a) to 0.6 (e), and the conventional method (f) for the bankruptcy data set.



(a) 1

(b) 0.9

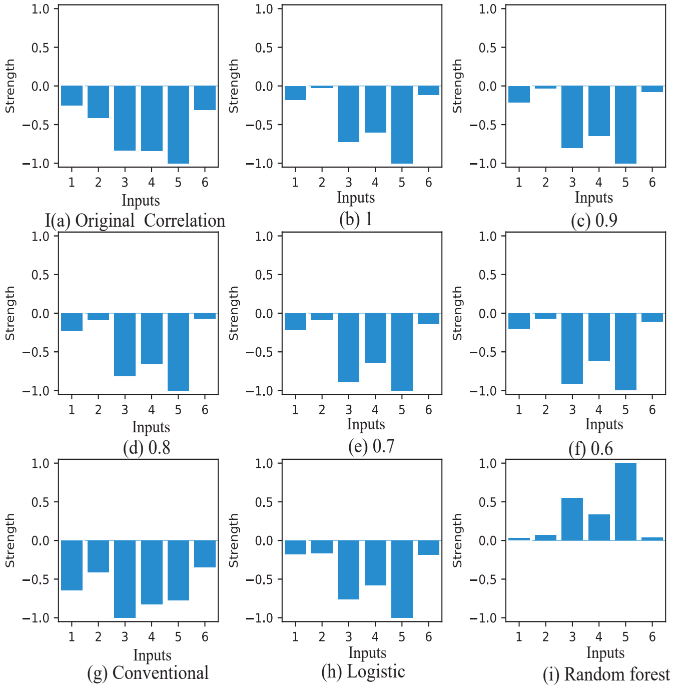
(c) 0.8

(d) 0.7

(e) 0.6

(f) Conventional

Fig. 6. Weights when the parameter decreased from 1 (a) to 0.6 (e) and by using the conventional method (f) for the bankruptcy data set.



I(a) Original Correlation

(b) 1

(c) 0.9

(d) 0.8

(e) 0.7

(f) 0.6

(g) Conventional

(h) Logistic

(i) Random forest

Fig. 7. Correlation coefficients between inputs and targets in the original data set (a) and compressed weights when the parameter α decreased from 1 (b) to 0.6 (f) and by the conventional method without selective information control (g), and the regression coefficients by the logistic regression analysis (h) and prediction importance (i) by the random forest method for the bankruptcy data set.

Figure 6 shows connection weights for all hidden layers when the parameter decreased from 1(a) to 0.6(e). When the parameter was one in Figure 6(a), many strong connection weights were scattered. When the parameter decreased from 0.9 in Figure 6(b) to 0.7 in Figure 6(d), a neuron in the preceding layer tended to be connected with all neurons in the subsequent layer. In particular, the final connection weights from the tenth to the eleventh layer, located in the rightmost column, shows the most typical state. This tendency tended to prevail for all layers when the parameter decreased to 0.6 in Figure 6(e). This means that, when collective information is forced to decrease, a specific neuron tends to be connected with all neurons in the subsequent layer. Collective information minimization can be realized by connecting a neuron with as many different neurons as possible. Finally, when we did not use selective information, little regularity could be seen over connection weights in Figure 6(f). However, we could see the same tendency, that neurons in the precedent layers were connected with ones in the subsequent layers. The selective information method seems to enhance this tendency.

3) Compressed Weights: We show here that the compressed weights from the original multi-layered neural networks were very close to the original correlation coefficients between inputs and targets of the data set, meaning that the method could disentangle connection weights to get individual and independent relations between inputs and outputs.

Figure 7(a) shows the correlation coefficients between inputs and targets of the original data set. Figures 7 (b) to (f) show compressed weights when the parameter decreased from 1 to 0.6. In addition, Figure 7(h) shows the regression coefficients of the logistic regression analysis. Those compressed weights and regression coefficients were quite similar to each other, and they were close to the original correlation coefficients between inputs and outputs in Figure 7(a). On the contrary, the conventional method without the selective information produced compressed weights different from the original correlations and regression coefficients in Figure 7(g). This means that the selective information control can be used to disentangle connection weights and to produce individual and independent relations between inputs and outputs. Finally, Figure 7(i) shows the prediction importance by the random forest method, where the importance was all positive, because the method could not deal with the negative values.

4) Partially Compressed Weights: Then, we tried to partially compress connection weights to examine how neural networks tried to extract information when the hidden layer increased. The principal finding was that individual information tended to deal with information from outputs, while collective information tended to focus on information from inputs.

Figure 8(a) shows compressed weights for all layers when the parameter was one and only individual information was forced to be increased. Weights were partially and step by step compressed from the top- left to the bottom-right box. The top-left compressed weights were ones obtained by combining only input and output layers, and the bottom-right weights were obtained by compressed all connection weights of all layers, namely, full compression. As can be seen in the figures, partially compressed weights were small when we compressed weights up to the sixth layer, when the compressed weights became slightly stronger. Finally, we could obtain the final compressed weights by the full compression (bottom-right). This means that neural networks cannot extract necessary information up to the final layer, which suggests that the individual information is dependent on the acquisition of output information.

Figure 8(b) shows partially compressed weights when the parameter was set to zero and only collective information was used. Note that, when the parameter was set to zero, the neural networks could not finish the learning, producing large errors for the outputs. Thus, the results were used to show the effects of collective information minimization as clearly as possible. As can be seen in the figure, only when connection weights in the input and output layers were compressed (top-left), we could obtain strong connection weights, while in all the other cases, the compressed weights were very small. As was mentioned, the learning failed, because information from inputs could not be transmitted to the final layer.

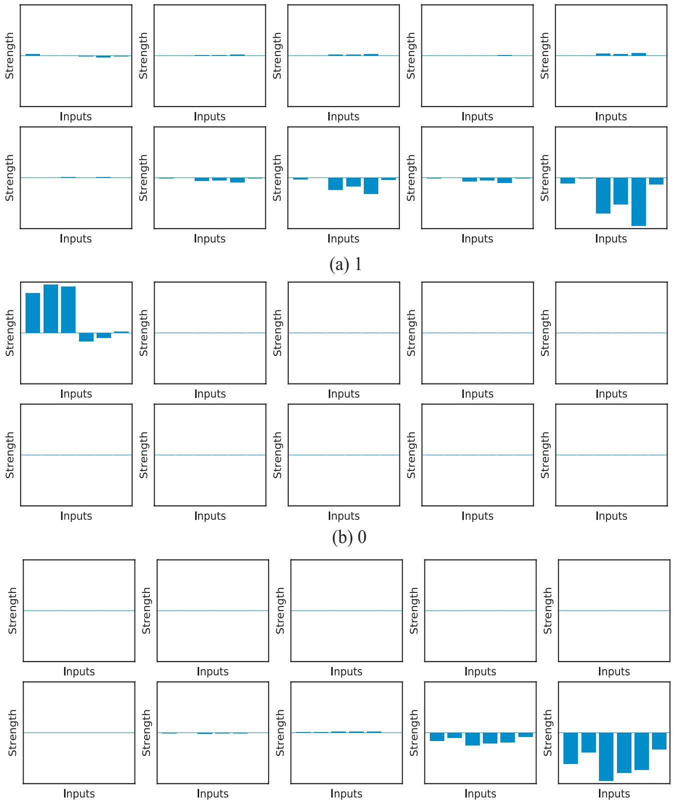
Figure 8(c) shows partially compressed weights by the conventional method without selective information. As can be seen in the figure, only when all connection weights in all layers were compressed (bottom right) were strong compressed weights obtained. This seems to us that the conventional method focused on information from outputs. As shown in Figure 8(a), individual information maximization also showed the same tendency of focusing toward output information.

These results show that the conventional method, as well as individual information can produce the final compressed weights only when all information goes through all hidden layers. On the other hand, the collective information has a property to detect some information in the early stages of multi-layers. From these experimental results, we can infer that we can obtain different types of internal representations by focusing on input or output information.

IV. Conclusion and Future Work

The present paper aimed to propose a new type of information-theoretic method to control the selectivity of components of neural networks. To interpret the process of information storing and transmission of neural networks, we need to control the selectivity of components, or selective information. Only by controlling the selective information can we interpret the information processing of neural networks and thus interpret the main inference mechanism of neural networks. We prepared two types of selective information: individual and collective information. By the ratio of the two types of selective information, we tried to explore the main information processing mechanism. The method was applied to the bankruptcy business data sets. The experimental results showed that individual and collective information could be controlled by the present method. With this control, we could see how neural networks try to capture input information or output information differently. In addition, the results showed that, behind seemingly complicated representations in multi-layered neural networks, very simple, individual, and independent relations could be observed. It can be expected that the complicated representations in the surface level can be transformed from those simple basic representations. We should say that it is possible to transform complicated neural networks into the simple ones, whose basic structure can be easily interpreted, and the structure can be easily transformed to produce a variety of surface and complicated representations.

Finally, this paper was concerned with the interpretation, but it is better to unify the problem of interpretation with improved generalization performance. Thus, we need to propose a method to interpret the inference mechanism, followed by improved generalization.



(a) 1

(b) 0

(c) Conventional

Fig. 8. Partially compressed weights when the parameter was 1 (a) and 0 (b), and the conventional method without selective information (c) for the bankruptcy data set.

References

[1] A. Hart and J. Wyatt, “Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks,” *Medical informatics*, vol. 15, no. 3, pp. 229–236, 1990.

[2] M. Spining, J. Darsey, B. Sumpter, and D. Nold, “Opening up the black box of artificial neural networks,” *Journal of chemical education*, vol. 71, no. 5, p. 406, 1994.

[3] J. M. Benítez, J. L. Castro, and I. Requena, “Are artificial neural networks black boxes?,” *IEEE Transactions on Neural Networks*, vol. 8, no. 5, pp. 1156–1164, 1997.

[4] J. D. Olden and D. A. Jackson, “Illuminating the black box: a randomization approach for understanding variable contributions in artificial neural networks,” *Ecological modelling*, vol. 154, no. 1-2, pp. 135–150, 2002.

[5] F. Qiu and J. Jensen, “Opening the black box of neural networks for remote sensing image classification,” *International Journal of Remote Sensing*, vol. 25, no. 9, pp. 1749–1768, 2004.

[6] G. Bologna, “Is it worth generating rules from neural network ensembles?,” *Journal of Applied Logic*, vol. 2, no. 3, pp. 325–348, 2004.

[7] 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.

[8] M. Sendak et al., “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.

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

[10] K. R. Varshney and H. Alemzadeh, "On the safety of machine learning: Cyber-physical systems, decision sciences, and data products," *Big data*, vol. 5, no. 3, pp. 246–255, 2017.

[11] 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.

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

[13] 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.

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

[15] K. Fukushima, "Cognitron: a self-organizing multi-layered neural network," *Biological Cybernetics*, vol. 20, pp. 121–136, 1975.

[16] K. Fukushima, "Neocognitron: a hierarchical neural network capable of visual pattern recognition," *Biological Cybernetics*, vol. 20, pp. 121–136, 1975.

[17] K. Fukushima, "Neocognitron: a neural network model for a mechanism of visual pattern recognition," *IEEE Transactions on Systems, Man, and Cyvernetics*, vol. 13, pp. 826–834, 1983.

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

[19] A. Nguyen, J. Yosinski, and J. Clune, "Understanding neural networks via feature visualization: A survey," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 55–76, Springer, 2019.

[20] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[21] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, 2009.

[22] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[23] S. Bach, et al., "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[24] F. Arbabzadah, G. Montavon, K.-R. Müller, and W. Samek, "Identifying individual facial expressions by deconstructing a neural network," in *German Conference on Pattern Recognition*, pp. 344–354, Springer, 2016.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[26] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based grad-cam: Towards accurate visualization and explanation of cnns," *arXiv preprint arXiv:2008.02312*, 2020.

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

[28] 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.

[29] 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.

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

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

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

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

[34] 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.

[35] 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.

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

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

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

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

[46] A. S. Morcos, D. G. Barrett, M. Botvinick, and N. C. Rabinowitz, "On the importance of single directions for generalization," 2018.

[47] 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.

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

[49] 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.

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

[51] M.-J. Kim and I. Han, "The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms," *Expert Systems with Applications*, vol. 25, no. 4, pp. 637–646, 2003.