International Journal of Smart Computing and Artificial Intelligence
International Institute of Applied Informatics
2025, Vol. 9, No. 1, IJSCAI886

Selective Potentiality and Moving Focus for Interpreting Multi-Layered Neural Network

Ryotaro Kamimura [*]

Abstract

The present paper aims to demonstrate the existence of simplification forces in neural networks. These simplification forces can be represented by the simplest network, called "prototype." To extract the prototype, we need to identify necessary and important information during learning. The structural potentiality has been proposed to reduce information, aiming to reduce unnecessary information, but one of its problems lies in excessive information reduction. To preserve important information, we need to maximize or at least weaken the excessive information reduction. To solve this problem, we introduce a new potentiality called "selective potentiality," which allows us to move a focus field where a group of connection weights can be flexibly reduced. This method aims to replace the troublesome contradictory operations of potentiality reduction and augmentation with more concrete and manageable ones. The method was applied to an artificial dataset, in which linear and non-linear relations were introduced. The results confirmed that selective potentiality could be increased to weaken structural potentiality reduction. The selective potentiality showed strong forces of simplification throughout the entire learning process. By seeking the simplest prototype, additional results were obtained, where networks tried to infer the outputs, enhancing both linear and non-linear inputs for better generalization.

Keywords: structural potentiality, selective potentiality, moving focus, prototype, interpretation

1 Introduction

1.1 Simplification and Prototype

The present paper aims to demonstrate that there are strong forces for the simplification of network configurations in neural networks. The simplest form is realized by the prototype, which has the simplest network configuration, and all surface networks are assumed to be generated from this prototype network through transformational rules. One of our primary objectives is to determine whether we can extract this prototype from actual multi-layered neural networks.

The simplification principle seems to be observed in a variety of phenomena, for example, in terms of implicit bias or regularization toward simpler solutions [1]–[7]. In neural

[*] Tokai University, Kanagawa, Japan

networks, the components and learning mechanisms are considerably simplified, realized by imitating only a tiny fraction of our cognitive functions. This inherent simplicity certainly contributes to their success in various applications, because it has been easy to develop the partially imitated functions to the extreme point. Considering this fact, the simplicity principle, here stated, seems to be more easily accepted. Technically, this simplification principle has the merit that the interpretation of neural networks becomes much easier. Due to complicated network configurations and input patterns, the interpretation becomes a troublesome problem in neural networks. If the simplification in terms of prototype is valid, the actual interpretation process becomes much easier because all we have to do is to understand the simplest form of networks and how to transform them into more complicated surface ones.

1.2 Structural and Selective Potentiality

To extract the presumed simplest prototype concealed within complex surface networks, we have developed a structural potentiality method to reduce the complexity of connection weights. This structural potentiality primarily aims to reduce weight complexity, with the expectation of eliminating unnecessary information. However, as the method may not adequately consider the quality of information in connection weights, important information could be lost during the potentiality and complexity reduction processes.

To address the issue of excessive information reduction caused by the structural potentiality reduction, we propose a new type of potentiality, termed "selective potentiality." Rather than simply reducing the structural potentiality, the selective potentiality actively controls a specific group of weights for controlling the potentiality. In essence, this method aims not to increase potentiality but to mitigate the impact of potentiality reduction. In terms of information content, we try to combine and unify information maximization and minimization by controlling the focused fields for potentiality reduction. Although reconciling potentiality reduction and augmentation may seem challenging, the selective potentiality addresses this issue by controlling a field, focusing on a group of connection weights to be reduced. This flexible control of focused fields enables the realization of potentiality maximization within the framework of potentiality minimization. In other words, we aim to replace the problem of potentiality reduction and augmentation with the problem of moving focused fields. The ambiguous nature of potentiality reduction and augmentation can be replaced by a more concrete operation involving the movement of focused fields. This selective potentiality is expected to solve the problem of excessive information reduction, where information can be reduced while keeping important information intact.

1.3 Outline with Main Contributions

Then, we here outline the paper with our main contributions as follows:

- This paper attempts to demonstrate the existence of simplification forces in neural networks by considering the simplest form of a prototype.

- Because the simplest prototype should retain necessary and important information, we need to carefully select important information from available information. The structural potentiality method has focused on information reduction to reach the prototype. However, this reduction does not necessarily guarantee the acquisition of necessary information.

- To mitigate the force of structural potentiality reduction, we introduce selective potentiality to resolve the contradiction between potentiality reduction and augmentation by moving focused fields. The moving focus aims to control the groups of weights to be reduced, corresponding to the potentiality control.
- The selective potentiality was applied to an artificial dataset with both linear and non-linear inputs. The results confirmed that the increase in the selective potentiality entailed a moderate increase in the structural potentiality during the structural potentiality reduction process.
- The method produced estimated prototypes that were significantly closer to the supposed prototype. In addition, improved generalization was observed by enhancing both non-linear and linear relationships.
- The final results confirmed that the selective potentiality could be used to increase the structural potentiality while simultaneously reducing it. The interpretation process was simplified by relying on the prototype interpretation, and improved generalization was achieved by enhancing both important linear and non-linear inputs.

2 Related Work

2.1 Prototype-based Interpretation

The prototype finding attempts to transform the diverse interpretation approaches of neural networks into the interpretation of the simplest network, hidden behind a great number of surface neural networks. One of the major challenges in neural networks lies in solving the difficulty of understanding their inference mechanism. Even if multi-layered neural networks demonstrate success in prediction, we may encounter unexpected problems in actual processing due to the incomprehensibility of the final inference. To address this problem, numerous interpretation methods have been proposed [8]–[13]. From our perspective, the majority of these methods seem to be oriented towards local interpretation [14], meaning that they attempt to understand the inference mechanism for a specific input, although some have sought to interpret neural networks globally [15], [16]. For these diverse surface networks, correspondingly a large number of local interpretations may be required. However, when we examine the components of neural networks, one of their most important characteristics is simplicity. All elemental components, such as units and connection weights, are relatively simple to interpret, even though complex components may be employed in recent developments of convolutional neural networks [17] and natural language processing models [18]. Neural networks have achieved significant success in applications by focusing on a small fraction of our cognitive functions. This simplicity has been a key factor in accelerating their remarkable development.

2.2 Selectivity and Selective Attention

Prototype finding is based on increasing the potentiality of neural network to select important information, which is realized by a new measure of selective potentiality. Our selective potentiality tries to focus on specific groups of connection weights to be weakened. The selectivity is one of the major principles to regulate the behaviors of neural networks, where a specific neuron tends to respond to specific inputs [19], [20]. This is because the acquisition of specific behaviors of neurons should be one of the major characteristics of neural networks. Then, there have been many studies to use the selectivity principle in neural networks, cognitive sciences and neurosciences [21]–[25]. In particular, the selective attention, focusing on specific parts of inputs, has been widely used in the natural language processing and vision [18], [26]–[28]. Those methods should be classified as the passive selectivity detection, where the selectivity is based on an idea that a neurons respond well to a specific input. A neuron should be chosen selectively by responding to an specific input. When we adopt this idea of selectivity, the selectivity of neurons should be represented by a great number of different types of selectivity, corresponding to different input patterns.

On the contrary, our selective potentiality is a mechanism inherent to neural networks, meaning that neural networks should have an ability to focus selectively on their components. The components are not selected by specific neurons but a network can choose which inputs should be responded by some specific components. The selectivity in this paper is derived not from the specific inputs but from the specific components in neural networks. This eventually means that we try to understand the limitations of neural networks beyond which it is impossible to continue the learning processes.

2.3 Mutual Information

In addition, our selective potentiality is closely related to attempts to unify information maximization and minimization. This is because the potentiality has been developed to approximate the well-known entropy function for easier computation and understanding. In learning theory, the entropy, on which our proposed potentiality is based, has played one of the most important roles from the beginning [29]. Actually, learning has been considered a process of entropy reduction, or in our terms, potentiality reduction. However, a process of learning has turned out to be a more complex process mixing entropy minimization and maximization. Then, in neural networks, mutual information has gained more popularity in describing learning processes [33]–[36]. In the formulation of mutual information, there exist two types of entropy, and to maximize mutual information, entropy should be maximized, and at the same time, conditional entropy should be minimized. Because of those contradictory operations inside mutual information and the computational burden, mutual information seems to be not appropriately used in describing learning since the pioneering work by Linsker [30]–[32].

Furthermore, a more complex formulation of mutual information has been well accepted in neural networks in the name of the information bottleneck [37]–[43]. In those methods, mutual information should be minimized to have compact representations, but mutual information should be maximized to have enough information for predicting the outputs. As mentioned above, one of the major problems of those methods is that it is quite challenging to compromise entropy maximization and conditional entropy minimization and even more challenging to compromise mutual information maximization and minimization. Though the computational burden of those methods exists, some important results were reported, related to this paper. For example, stochastic gradient descent should have two learning phases, namely, a fitting phase to increase target information and a compression phase to decrease input information [44]. Our results in this paper show that there exist two learning phases: prototype and non-prototype phases. Those two phases seem to be similar to those described by the information bottleneck principle, though they differ from those of the information bottleneck in terms of the acquisition of the simplest form in the first place.

As discussed above, mutual information has produced many important results in learning, but it has the serious problem of computational burden and serious contradictory operations inside mutual information. Our method of selective potentiality seems to solve this problem. First, the computational burden of entropy is replaced by a simpler potentiality function, behaving similarly to the entropy. More importantly, the meaning of processing information becomes clearer. Actually, we try to minimize the structural potentiality, corresponding to the entropy, and at the same time to maximize it. Because the compromise between two contradictory operations is quite challenging, we try to replace this optimizing operation with the operation of moving focus. The structural potentiality is forced to decrease as a property of potentiality, but by controlling the size and strength of weights to be selectively focused on, the potentiality reduction is weakened.

2.4 Least Effort Principle

Finally, we should mention the relations between simplification forces and the least effort principle in quantitative linguistics [45]. This principle states that human linguistic behavior should be governed by the least effort, realized by the relations between word frequency and its rank. This linguistic phenomenon has been applied to many other fields [46]–[52], and the ubiquitous presence of this principle has been gradually confirmed.

This paper tries to show the existence of simplification forces in multi-layered neural networks. For simplification, a variety of facts and experimental results have been accumulated, related to these simplification forces from different viewpoints such as architecture, optimizing methods, initialization, and so on [1]–[7], [53]–[61], to cite a few. Mainly, they have tried to explain, from technical viewpoints, the fact that generalization cannot be degraded even if the network architecture becomes more complicated and overparameterized, contrary to our intuition. Then, there are a number of different hypotheses for explaining this implicit simplification.

Contrary to those hypotheses from technical viewpoints, we think that the simplification is due to the collective principle of our cognitive behaviors. This least effort principle shows that simplicity is inherited by the collective properties of our behaviors, which neural networks try to imitate. The reason why neural networks have produced seemingly successful results in many applications is that they try to simplify relations among many components inside as much as possible, and only this extreme simplification is the cause of good results of neural networks over many applications. Paradoxically, for dealing with complicated data sets, any components should be as simple as possible.

3 Theory and Computational Methods

3.1 Concept of Selective Potentiality

As mentioned above in the section on the related work, there is almost always a serious contradiction in information control in neural networks. First, this contradiction can be solved or, more exactly, weakened by focusing on simpler aspects. For example, we have almost always tried to resolve one factor among many contradictory ones to be optimized. From the pioneering studies on the mutual information-theoretic methods by Linsker [30], we have tried to deal only with one factor in mutual information, where the entropy is almost always maximized, without considering the corresponding conditional entropy due to its simplicity. Thus, mutual information maximization has been replaced by simple entropy maximization. Second, multiple contradictory terms are adjusted by introducing a

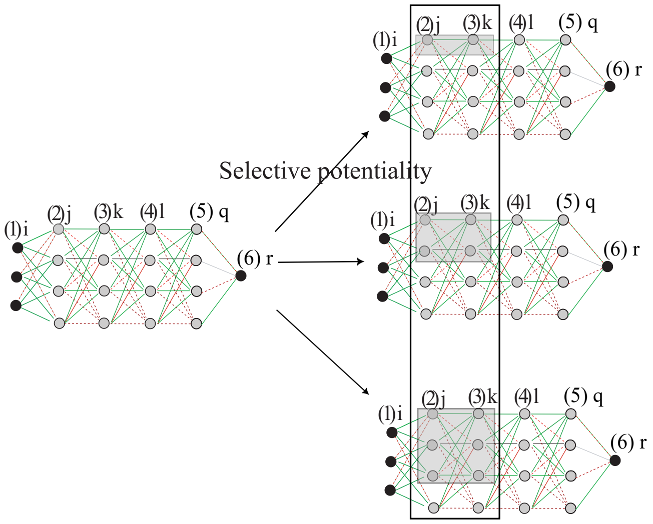


Figure 1: An example of network configurations with different moving focuses by the selective potentiality control.

controlling parameter to compromise contradictions between them. As mentioned above, the information bottleneck has principally adopted this approach [37].

Contrary to those methods, we propose a new method that involves replacing the contradiction with more concrete and manageable operations. In our method, the potentiality increase and decrease, corresponding to information minimization and maximization, are replaced by the operation of how many weights should be focused on. This operation can actually control a new type of potentiality, called "selective potentiality." The selective potentiality is introduced to compromise between potentiality reduction and augmentation. Figure 1 shows a concept of selective potentiality, where we try to examine which connection weights should be focused on. The selective potentiality aims to determine which connection weights are focused on by controlling the selective parameter. For example, in the upper figure, only the connection weights on the upper side are focused on, then the figure in the middle shows that two groups of connection weights on the upper side are focused on. Finally, in the lower figure, three groups of connection weights are focused on. The selective potentiality tries to change the focus point and its width flexibly. This moving focus field can be directly related to the structural and selective potentiality change. Naturally, as the number of weights to be focused on decreases, the potentiality tends to increase eventually by definition. By this method, potentiality reduction and augmentation can be controlled more concretely by controlling the number of weights to be focused on.

3.2 Difference in Structural and Selective Potentiality

The structural potentiality is introduced to describe the complexity of connection weights, and this potentiality is used to define the selective potentiality introduced here. We will explain the main differences between structural and selective potentiality roughly, with more computational procedures provided in the next section. Figure 2 shows the structural (a) and selective (b) potentiality as a function of the structural parameter β_{str}. As can be seen in Figure 2(a), the structural potentiality decreases gradually as the structural parameter increases. On the other hand, in Figure 2(b), the selective potentiality is plotted. One of the main characteristics is that the selective potentiality decreases in the beginning and then

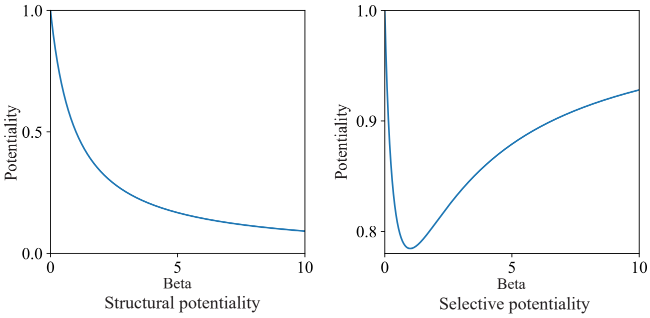


Figure 2: Structural potentiality (a) and selective potentiality (b) as a function of the structural parameter β_{str}.

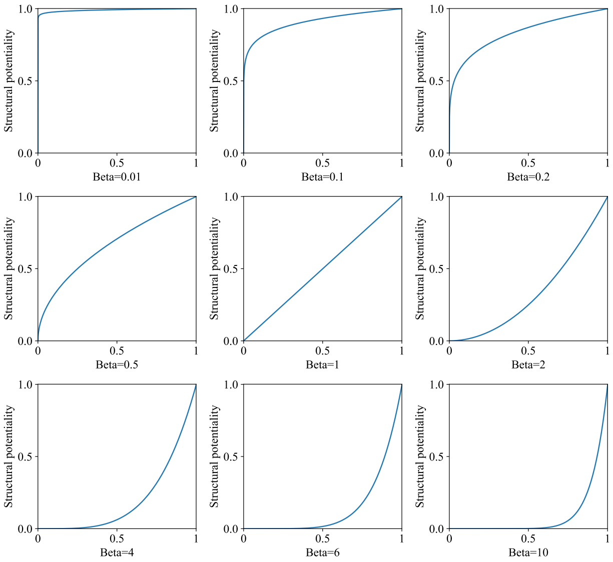


Figure 3: Individual structural potentiality as a function of simple individual potentialities, when the structural parameter β_{str} increased from 0.01 to 10.

begins to increase. This means that when we can use relatively larger structural parameter values, the selective potentiality can be increased, and correspondingly the structural potentiality does not necessarily increase but at least its reduction force can be weakened.

Let us explain more concretely how the potentiality can be changed. Figure 3 shows the individual structural potentiality when the structural parameter increases from 0.01 to 10 from top left to bottom right. When the parameter is 0.01, all the individual structural potentialities are close to their maximum value. As the parameter increases, the number of smaller individual structural potentialities becomes larger and larger. This means that the structural potentiality tries to reduce the number of stronger connection weights as much as possible. Finally, only one connection weight with the strongest potentiality remains, while all the other weights become close to zero. Compared with conventional weight decay, this reduction is stronger, eliminating as many weights as possible. However, this structural potentiality reduction is sometimes too strong, causing the problem of eliminating important connection weights. To solve this problem, we introduce the selective potentiality.

On the contrary, Figure 4 shows the individual selective potentiality when the structural parameter increases from 0.01 to 10. When the parameter increases from 0.01 to 0.2, the

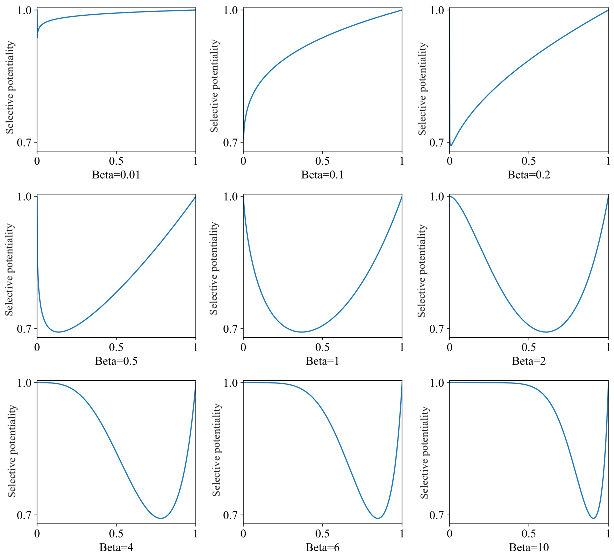


Figure 4: Individual selective potentiality as a function of simple individual potentialities, when the structural parameter β_{str} increased from 0.01 to 10.

selective potentiality is close to the structural potentiality, though minimum values are different. Then, when the parameter increases from 0.5 to 10, the focus field for the reduction in terms of lower potentiality values gradually moves from left to right. We call this phenomenon “moving focus.” As explained below, both the structural and selective potentiality are the sum of individual potentialities. As the structural parameter increases, the width of the focus field becomes smaller and s maller. This means that the number of stronger individual potentialities increases, leading to an increase in the selective potentiality in terms of the sum of all individual potentialities. In actual learning cases, this has the effect of pushing some stronger weights toward smaller ones, which can be related to the reduction of structural potentiality. Roughly speaking, the selective potentiality can be increased by increasing the structural parameter. At the same time, even in those cases, the structural potentiality tends to reduce its potentiality.

3.3 Structural Potentiality

Before going into details on the selective potentiality, we need to explain the structural potentiality, on which the selective potentiality is based. The structural potentiality is introduced to reduce the complexity of connection weights. We now consider a hidden layer from the nth layer to the $n+1$th layer, termed $(n,n+1)$. The individual structural potentiality is defined by using the absolute values of connection weights:

$$u_{jk}^{(n,n+1)} = \mid w_{jk}^{(n,n+1)} \mid, \tag{1}$$

where, for simplicity, all connection weights are supposed to be larger than zero. Then, the relative individual structural potentiality can be obtained by dividing the individual potentiality by its maximum value:

$$g_{jk}^{(n,n+1)} = \frac{u_{jk}^{(n,n+1)}}{\max_{j'k'} u_{j'k'}^{(n,n+1)}}, \tag{2}$$

where the maximum operation is over all connection weights in the layer. The parametric form of the individual structural potentiality is obtained by raising the individual potentiality to the power of the structural parameter β_{str}. We introduce a new function "str" to simplify the following notation and to clarify which variables are focused on by hiding the structural parameter, because the structural parameter is always present in the formulation of structural potentiality. Then, the structural potentiality is defined by:

$$\mathrm{str}\left(g_{jk}^{(n,n+1)}\right) = \mathrm{power}\left(g_{jk}^{(n,n+1)}, \beta_{str}\right), \tag{3}$$

where the function power(x,a) means raising x to the power of a. By summing all relative individual structural potentialities, we have the final form of structural potentiality:

$$G^{(n,n+1)} = \sum_{jk} \mathrm{str}\left(g_{jk}^{(n,n+1)}\right). \tag{4}$$

As explained above, the structural potentiality decreases gradually as the structural parameter increases. The structural potentiality is introduced principally to reduce its strength for reducing the complexity of connection weights. In addition, we aim to replace the well-known entropy with this structural potentiality for simpler computation and interpretation. When all the individual potentialities become the same, the potentiality becomes maximum. On the other hand, when one individual potentiality becomes larger than any other ones, the structural potentiality becomes smaller. Compared with entropy, the potentiality is simpler without logarithmic functions inside, and the meaning of potentiality can be more easily understood. For example, when entropy is maximized, all the probabilities should be the same, irrespectively of the strength of connection weights. When the structural potentiality is maximized, the strength of all the corresponding weights should be maximized.

This potentiality can be used to change the strength of connection weights simply by multiplying the weights by the corresponding individual structural potentialities. Then, the weight change aims mainly to reduce the structural potentiality:

$$w_{jk}^{(n,n+1)}(t+1) = \mathrm{str}\left(g_{jk}^{(n,n+1)}\right) w_{jk}^{(n,n+1)}(t), \tag{5}$$

where the weight at the $(t+1)$th learning step can be obtained by multiplying the weight by the corresponding individual structural potentiality. By the use of simplified notation for the structural potentiality str$(g_{jk}^{(n,n+1)})$, this method tries to replace the connection weights with the corresponding potentialities. We should note that the multiplication by the weight is only for making learning easier, and it is possible to eliminate it in the actual learning. However, the parameter setting becomes challenging, and we adopt this weight multiplication for easy implementation at the present stage.

3.4 Selective Potentiality and Selective Focus Moving

The selective potentiality aims to see and determine which connection weights should be selectively focused on. As shown in Figure 2(a), the structural potentiality decreases when the structural parameter β_{str} increases. On the contrary, the selective potentiality decreases initially and then gradually increases as the structural parameter increases, as shown in Figure 2(b). Thus, it is possible to increase the potentiality by choosing larger structural parameter values. We should note that we do not necessarily increase the structural potentiality directly, but since the selective potentiality is based on the structural potentiality, changes in the selective potentiality naturally affect the structural potentiality.

Now, let us define the selective potentiality for the $(n, n+1)$ layer. The individual selective potentiality is obtained by raising the individual structural potentiality to the power of itself:

$$\mathrm{sel}\left(g_{jk}^{(n,n+1)}\right) = \mathrm{power}\left(\mathrm{str}\left(g_{jk}^{(n,n+1)}\right), \mathrm{str}\left(g_{jk}^{(n,n+1)}\right)\right). \tag{6}$$

By summing all the individual selective potentialities, we have the final form of selective potentiality:

$$H^{(n,n+1)} = \sum_{jk} \mathrm{sel}\left(g_{jk}^{(n,n+1)}\right) \tag{7}$$

Now, weights at the $t+1$th learning step can be obtained by multiplying the weights at the tth step by the corresponding potentiality:

$$w_{jk}^{(n,n+1)}(t+1) = \mathrm{sel}\left(g_{jk}^{(n,n+1)}\right) w_{jk}^{(n,n+1)}(t). \tag{8}$$

As shown in Figure 4, the selective potentiality's focus moves gradually from left to right as the parameter increases. When the structural parameter is small, the potentiality focus is similar to the structural potentiality in Figure 3, where, as the parameter becomes smaller, all individual potentialities become the same and flat, close to maximum potentiality. Then, as the structural parameter increases gradually, the focus point moves to the right, meaning the focus is on larger individual potentialities except the maximum ones. The potentiality is not effective for the maximum individual one, but the strength of the surrounding individual ones is effectively reduced. This may be closely connected with soft-type competitive learning, and this property is one of the main differences between potentiality and other regularization methods. We should reiterate that the selective potentiality method tries to reduce connection weights except the maximum one, which makes learning more stable because, without this property, the weights become smaller and smaller. The selective potentiality is introduced to determine which connection weights should be focused on with the very flexible control of reducing the strength of connection weights.

4 Results and Discussion

4.1 Experiment Outline

The data set was artificially created by imitating the input variables of the actual bankruptcy data set [62], focusing on the clearer understanding of the inference mechanism of neural networks, as shown in Figure 5. We prepared seven input variables, and one of the major characteristics is that the first three input variables were created with three different types of random values, and used without any modifications. However, the remaining four inputs were modified through the non-linear transformations of the original inputs. For example, input No. 4 was created based on the sigmoid function, input No. 5 was created based on the squared function, input No. 6 was based on the square root function, and input No. 7 was based on the uniform distribution with different ranges. We tried to examine how the selective potentiality control detects those complicated relations, distinguishing between linear and non-linear ones.

The number of inputs was seven, the number of hidden layers was ten with ten neurons for each hidden layer, and the number of input patterns was 1,000. We used the Pytorch program package with the default parameter setting for almost all cases to easily reproduce the final results presented here. The activation function was the ReLU function because we

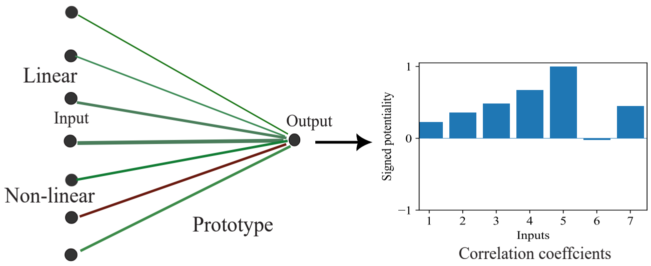


Figure 5: Supposed prototype (left) with the corresponding weights by the normalized correlation coefficients of training data set (right).

could produce clearer results with this activation function. The following results were all the averages obtained by five different trials, randomly chosen, for each parameter value. As the prototype is supposed to be the simplest one, the connection weights of the prototype were computed by taking the correlation coefficients between inputs and targets of the training data set, as shown in Figure 5. Then, we tried to estimate the supposed prototype by compressing multi-layered neural networks with ten hidden layers. For the procedures of compression, see Appendix A.

The main findings can be summarized as follows:

- The results confirmed that while the structural potentiality decreased, the selective potentiality increased. In addition, when the structural potentiality decrease was weakened, generalization accuracy increased gradually.

- In all cases, the ratio potentiality, measuring similarity to the supposed prototype, took higher values in the beginning of learning. In the later stage of learning, the ratio potentiality decreased, but by the selective potentiality control, the ratio increased again. This means that the strong forces of simplification could be clarified for the entire learning step by controlling the selective potentiality.

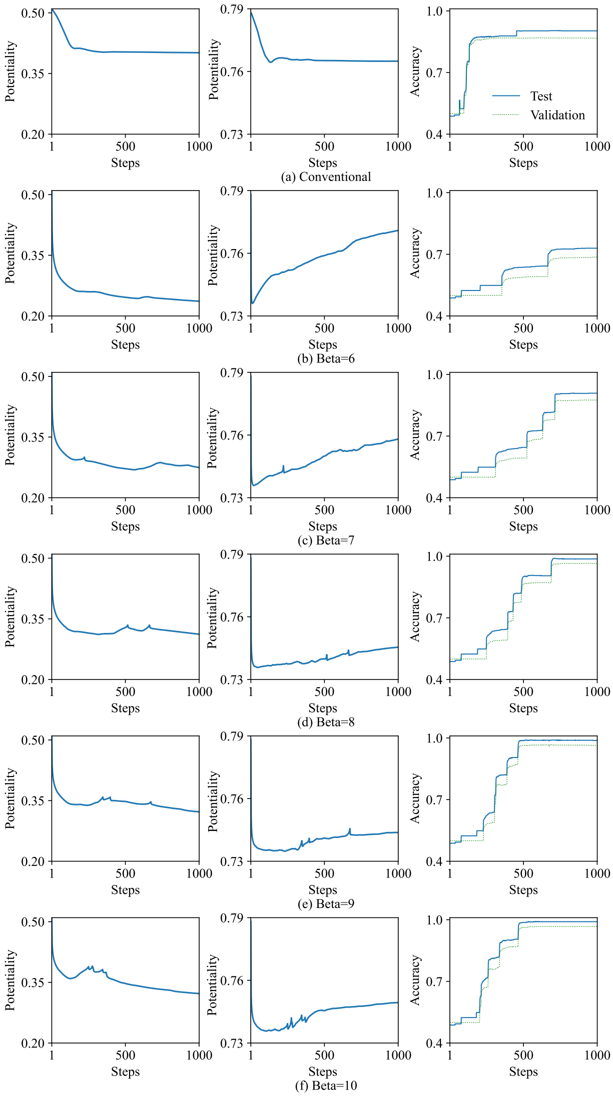
- By examining individual ratio potentialities, the network tried to extract a non-linear input in the first place, and then linear inputs were focused on. Finally, when the structural parameter was ten with the best generalization, the network tried to enhance both linear and non-linear inputs. This means that to improve generalization, the properties of both linear and non-linear inputs must be enhanced.

- These results confirmed that the detection of the simplest prototype was related not only to easier interpretation due to the simple and linear relations but also to the discovery of important relations beyond the prototype for better generalization.

4.2 Potentiality and Generalization

The results showed that the structural potentiality decreased and, at the same time, the selective potentiality increased. Generalization performance was improved when the structural parameter increased. Note that we present the results when the structural parameter was larger than six because we had better results only when the structural parameter was larger than six.

Figure 6 shows the structural potentiality (left), selective potentiality (middle), and generalization and validation accuracy (right) as a function of the number of steps (epochs).



(a) Conventional

(b) Beta=6

(c) Beta=7

(d) Beta=8

(e) Beta=9

(f) Beta=10

Figure 6: Structural potentiality (left), selective potentiality (middle), and generalization and validation accuracy (right) as a function of the number of steps, when the conventional method without the selective potentiality control was used (a), and the structural parameter β_{str} increased from 6 (b) to 10 (f).

When the conventional method without selective potentiality control was used in Figure 6(a), the structural potentiality decreased, but it remained at a relatively high level throughout the entire learning process. The selective potentiality in the middle decreased and remained unchanged in the later stage of learning. Generalization accuracy (right) increased immediately but moderately and remained unchanged in the later stage of learning. The structural potentiality decreased gradually, and the selective potentiality increased gradually when the number of learning steps increased and when the structural parameter increased from 6 (b) to 10 (f). In addition, we observed that the decrease in structural potentiality and the increase in selective potentiality were gradually weakened as the structural parameter increased. The effect of the selective potentiality decreased as the structural parameter increased because the size of the group of connection weights to be focused on became smaller when the structural parameter increased. Generalization accuracy increased step-by-step, and when the structural parameter increased, generalization tended to increase. When the structural parameter was 10 in Figure 6(f), the highest generalization accuracy was obtained.

The results confirmed that although the effect of selective potentiality decreased, we observed that the structural potentiality decreased and, at the same time, the selective potentiality increased. Eventually, generalization performance improved as the structural parameter increased gradually.

4.3 Ratio Potentiality

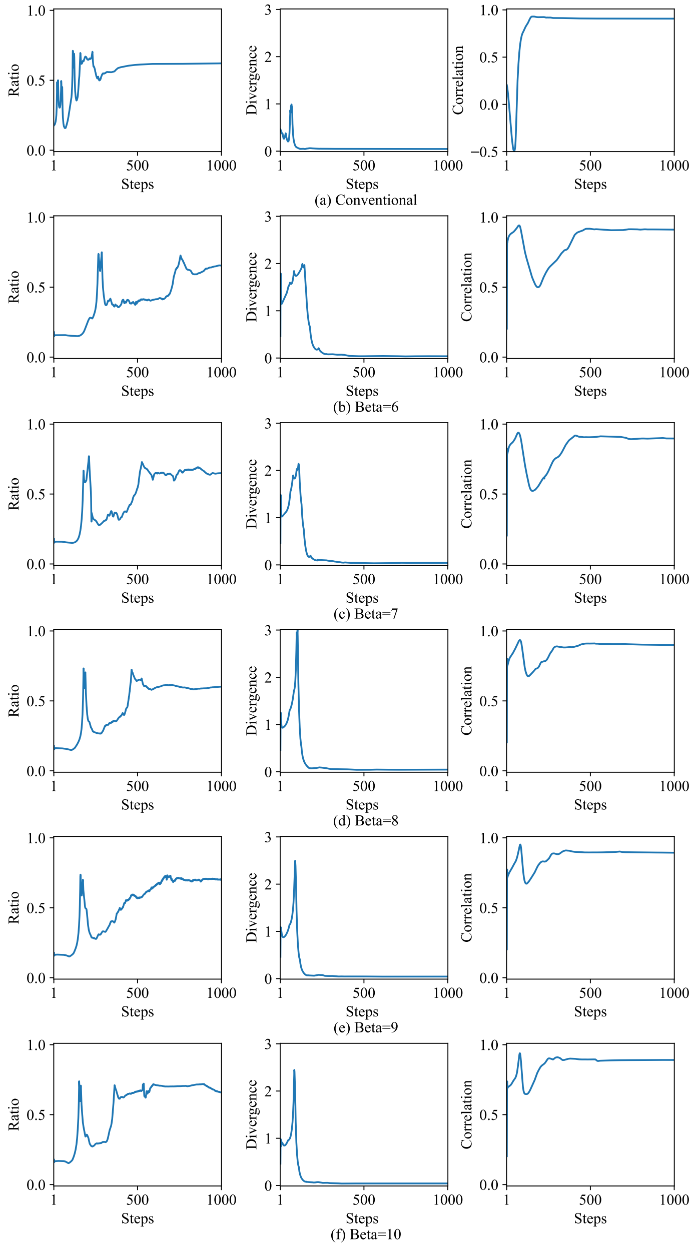
The ratio potentiality (see Appendix C), similarity to the supposed prototype, became higher in the beginning for all cases, meaning that the prototype was detected in the beginning of learning by all methods. By using the selective potentiality, after the first peak in the ratio potentiality, the higher values of ratio potentiality continued to be observed in the later stage of learning. This means that the forces of simplification continued to exist throughout the entire stage of learning, which could be found by the selective potentiality control.

Figure 7 shows the ratio potentiality (left), KL-divergence (middle), and correlation coefficients between estimated and supposed prototype (right) as a function of the number of learning steps. When the conventional method was used in Figure 7(a), the ratio potentiality increased in the first place and then remained unchanged in the later stage of learning. The divergence in the middle increased at the beginning of learning and became much smaller in the later stage of learning. The correlation coefficients (right) increased immediately to higher values and remained unchanged. The conventional method could detect the existence of the prototype in the beginning of learning. Figure 7(b) to (f) shows the results when the structural parameter increased from 6 (b) to 10 (f). The ratio potentiality increased rapidly in the beginning of learning and then decreased considerably. Finally, it increased again and remained unchanged, keeping relatively higher values, in the later stage of learning. This means that by the selective potentiality control, the ratio potentiality showed a tendency for a large increase in the beginning, a large decrease in the middle of learning, and finally a large increase at the end of learning. The correlation coefficients (right) showed the same tendency of increase and decrease, though the variation in correlation coefficients was smaller compared with the ratio potentiality. However, the divergence in the middle could not necessarily detect the tendency, because the divergence produced much higher values in the beginning due to the unbounded property of divergence.

The results confirmed that the selective potentiality control successfully detected the prototype in the beginning and, even in the later stage of learning, the ratio potentiality became higher, showing the detection of the prototype even in the later stage of learning. This means that there exists a strong force for simplification in learning. The ratio potentiality could show this tendency more clearly than the divergence and correlation coefficients.

4.4 Individual Potentialities and Rotation

The results confirmed that the signed individual structural potentialities or normalized connection weights became close to the supposed prototype. Then, by increasing the structural parameter β_{str} or through the rotation of ratio potentialities, the number of strong potentialities became smaller and smaller. This simplification of final ratio potentialities helped us understand which inputs the networks tried to use for producing the outputs.



(a) Conventional

(b) Beta=6

(c) Beta=7

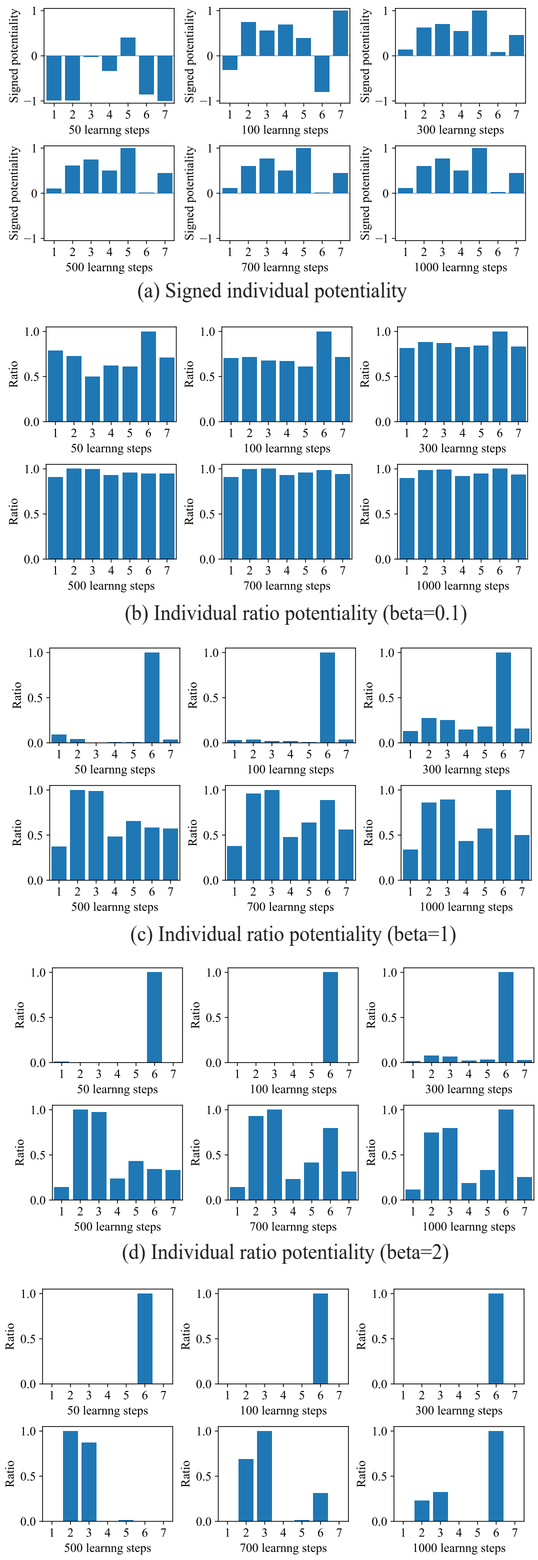
(d) Beta=8

(e) Beta=9

(f) Beta=10

Figure 7: Ratio potentiality (left), divergence (middle), and correlation coefficient between estimated and supposed prototypes (right) as a function of the number of steps (epochs), when the conventional method was used (a), and the structural parameter β_{str} increased from 6 (b) to 10 (f).

Figure 8(a) shows signed individual structural potentialities or normalized connection weights, when the structural parameter was set to 10 with the best generalization. As can be seen in the figure, the potentialities were close to the supposed prototype except for input No.4 in Figure 5. To more easily clarify the characteristics of those potentialities, we tried to change them by controlling the structural parameter for the ratio potentiality. When the structural parameter for the ratio potentiality increased from 0.1 (b) to 10 (e), the number of strong potentialities became smaller. This process of enhancement can be called



(a) Signed individual potentiality

(b) Individual ratio potentiality (beta=0.1)

(c) Individual ratio potentiality (beta=1)

(d) Individual ratio potentiality (beta=2)

(e) Individual ratio potentiality (beta=10)

Figure 8: Signed individual structural potentialities (a) and individual ratio potentialities, when the structural parameter β_{str} for the ratio potentiality increased from 0.1 (b) to 10 (2).

"potentiality rotation," and it can be used to examine which inputs are more important than the others. When the structural parameter was 0.1 (b), all individual structural potentialities were approximately the same. When the parameter increased to 1 (c), input No. 6 remained relatively stronger in the initial stage of learning, and in the later stage of learning, the linear inputs No.2 and 3 tended to be relatively higher. When the structural parameter increased to 2(d), in the later stage of learning, the linear inputs No.2 and 3 tended to be more accentuated. Finally, when the parameter was 10 (e), initially, the non-linear input No.6 was detected, and gradually the linear inputs No.2 and 3 were accentuated. In the final stage, the linear inputs No.2 and 3 were weakened, while the non-linear input No.6 remained strong. This means that to achieve the best generalization, the linear and non-linear inputs should be jointly enhanced, and they should cooperate with each other.

The results showed that the selective potentiality could produce the estimated network close to the supposed prototype. In addition, by increasing the structural parameter, the number of strong ratio potentialities became small for easier interpretation.

4.5 Rotation of Ratio Potentialities

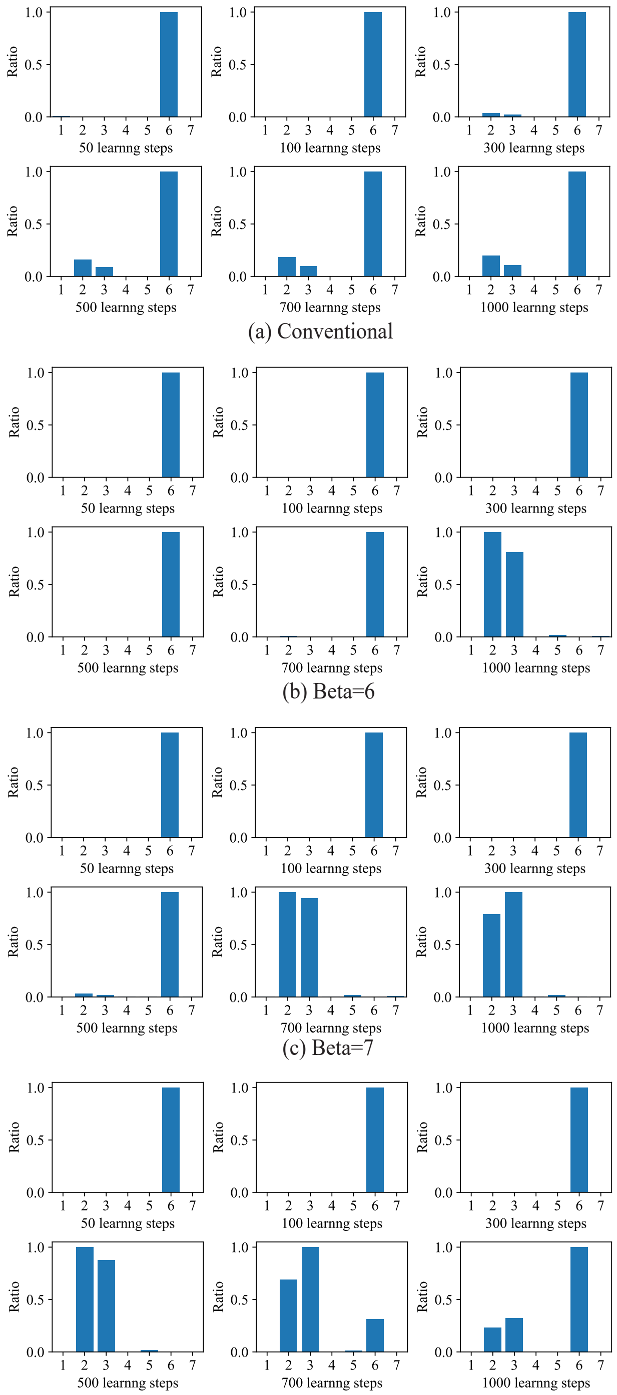
We show that by rotating the ratio potentialities, we can understand how networks tried to use specific inputs for better generalization.

Figure 9 shows the individual ratio potentialities by the conventional method (a), and when the structural parameter increased from 6 (b) to 10 (d). In all cases except the conventional method, the structural parameter for the structural potentiality was kept at 10 with the best generalization, as explained in the previous section. As shown in Figure 9(a), by the conventional method, the non-linear input No. 6 was strongly detected, and two linear inputs No. 2 and 3 were very weakly detected. This means that the conventional method focused on only one non-linear input No. 6 from the beginning. When the structural parameter increased from 6 (b) to 10 (d), in addition to input No.6, inputs No. 2 and 3 were detected. When the parameter was 6 (b), the non-linear input No. 6 was considered important, and only in the final stage of learning the linear inputs No. 2 and 3 were enhanced. When the parameter increased to 7 (c), the linear inputs No. 2 and 3 appeared earlier, with 700 learning steps. Finally, when the parameter increased to 10 (d), the linear inputs No.2 and 3 were combined with the non-linear input No. 6 from the middle of learning to produce the best generalization.

The rotation of ratio potentialities made it clear that the enhancement of linear and non-linear inputs was important for better generalization. In addition, increasing the structural parameter means that the selective potentiality control tried to extract linear inputs in the middle of learning, accompanied by the enhancement of non-linear inputs. This also indicates that the strong force for simplification was present in the middle of learning.

5 Conclusion

The present paper aimed to unify potentiality reduction and augmentation in one framework. This unification is urgently needed to extract important information while excluding unnecessary information to obtain the simplest prototype, hidden in the deepest level of neural learning. The unification is realized by proposing a new type of potentiality called "selective potentiality" to flexibly control a group of connection weights whose strength should be reduced. We applied the selective potentiality to an artificial dataset, imitating



(a) Conventional

(b) Beta=6

(c) Beta=7

(d) Beta=10

Figure 9: Individual ratio potentialities, when the conventional method was used (a), and the structural parameter β_{str} was increased from 6 (b) and 10 (d).

the actual bankruptcy dataset, incorporating linear and non-linear inputs. The results confirmed that the selective potentiality could be increased, while the structural potentiality decreased. By increasing the structural parameter, the force to detect the prototype increased gradually. Then, by rotating the potentialities, we could see that both linear and non-linear inputs were focused on for predicting the outputs. These results mean that the simplest prototype detection was related to improved interpretation due to the simplicity of the prototype to be interpreted. The rotation of potentialities showed the importance of the combination of linear and non-linear inputs for improved generalization.

For future directions, we should mention three points. First, we used a structural parameter larger than six because we could obtain better results for those higher values. However, we need to examine the final results when the structural parameter is relatively small, for example, less than one. Second, as discussed in the section on the related work, the selective potentiality is close to selectivity and attention, which has received much attention due to the success of natural language processing. We need to examine how the selective potentiality can be used to control the attention in neural networks. Finally, we should use larger and practical datasets for evaluating the final results of our method. We used the artificially created dataset for clearly showing the performance of our method. However, it is necessary to check whether our method can be applied to real, practical and complicated datasets. Though much remains to be solved for practical implementation, the selective potentiality method can certainly contribute to an improved interpretation of neural networks.

Acknowledgments

This paper has been written based on the paper [63], presented in the IIAI conference.

References

[1] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro, “Implicit bias of gradient descent on linear convolutional networks,” *Advances in neural information processing systems*, vol. 31, 2018.

[2] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, “The implicit bias of gradient descent on separable data,” *Journal of Machine Learning Research*, vol. 19, no. 70, pp. 1–57, 2018.

[3] S. Arora, N. Cohen, W. Hu, and Y. Luo, “Implicit regularization in deep matrix factorization,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[4] G. Blanc, N. Gupta, G. Valiant, and P. Valiant, “Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process,” in *Conference on learning theory*, PMLR, 2020, pp. 483–513.

[5] A. Ali, E. Dobriban, and R. Tibshirani, “The implicit regularization of stochastic gradient flow for least squares,” in *International conference on machine learning*, PMLR, 2020, pp. 233–244.

[6] N. Razin and N. Cohen, “Implicit regularization in deep learning may not be explainable by norms,” *Advances in neural information processing systems*, vol. 33, pp. 21 174–21 187, 2020.

[7] S. L. Smith, B. Dherin, D. G. Barrett, and S. De, “On the origin of implicit regularization in stochastic gradient descent,” *arXiv preprint arXiv:2101.12176*, 2021.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you?: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, 2016, pp. 1135–1144.

[9] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint arXiv:1702.08608*, 2017.

[10] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, “A survey on neural network interpretability,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 726–742, 2021.

[11] C. Molnar, G. Casalicchio, and B. Bischl, “Interpretable machine learning–a brief history, state-of-the-art and challenges,” in *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14–18, 2020, Proceedings*, Springer, 2021, pp. 417–431.

[12] F.-L. Fan, J. Xiong, M. Li, and G. Wang, “On interpretability of artificial neural networks: A survey,” *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2021.

[13] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, and D. Dou, “Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond,” *Knowledge and Information Systems*, pp. 1–38, 2022.

[14] Y. Liang, S. Li, C. Yan, M. Li, and C. Jiang, “Explaining the black-box model: A survey of local methods for deep neural networks,” *Neurocomputing*, vol. 419, pp. 168–182, 2021.

[15] C. Yang, A. Rangarajan, and S. Ranka, “Global model interpretation via recursive partitioning,” in *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, IEEE, 2018, pp. 1563–1570.

[16] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, “From local explanations to global understanding with explainable ai for trees,” *Nature machine intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[17] Y. LeCun and Y. Bengio, “Convolutional networks for images, speech, and time series,” *The handbook of brain theory and neural networks*, vol. 3361, 1995.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[19] 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.

[20] Z. Wang, T. Zeng, Y. Ren, Y. Lin, H. Xu, X. Zhao, Y. Liu, and D. Ielmini, “Toward a generalized bienenstock-cooper-munro rule for spatiotemporal learning via triplet-stdp in memristive devices,” *Nature communications*, vol. 11, no. 1, pp. 1–10, 2020.

[21] C. Cadieu, M. Kouh, A. Pasupathy, C. E. Connor, M. Riesenhuber, and T. Poggio, “A model of v4 shape selectivity and invariance,” *Journal of neurophysiology*, vol. 98, no. 3, pp. 1733–1750, 2007.

[22] O. Barak, M. Rigotti, and S. Fusi, “The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off,” *Journal of Neuroscience*, vol. 33, no. 9, pp. 3844–3856, 2013.

[23] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi, “The importance of mixed selectivity in complex cognitive tasks,” *Nature*, vol. 497, no. 7451, pp. 585–590, 2013.

[24] M. V. Peelen and P. Downing, “Category selectivity in human visual cortex,” 2020.

[25] T. Vu and A. Eldawy, “Deepsampling: Selectivity estimation with predicted error and response time,” *arXiv preprint arXiv:2008.06831*, 2020.

[26] J. Zhang, X. Shi, J. Xie, H. Ma, I. King, and D.-Y. Yeung, “Gaan: Gated attention networks for learning on large and spatiotemporal graphs,” *arXiv preprint arXiv:1803.07294*, 2018.

[27] D. Bahdanau, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[28] J. D. M.-W. C. Kenton and L. K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of naacL-HLT*, Minneapolis, Minnesota, vol. 1, 2019, p. 2.

[29] S. Watanabe, *Knowing and guessing: A quantitative study of inference and information*. New York: John Wiley and Sons Inc, 1969.

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

[31] ——, “Perceptual neural organization: Some approaches based on network models and information theory,” *Annual review of Neuroscience*, vol. 13, no. 1, pp. 257–281, 1990.

[32] ——, “Improved local learning rule for information maximization and related applications,” *Neural networks*, vol. 18, no. 3, pp. 261–265, 2005.

[33] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, “Mutual information neural estimation,” in *International conference on machine learning*, PMLR, 2018, pp. 531–540.

[34] R. Fritschek, R. F. Schaefer, and G. Wunder, “Deep learning for channel coding via neural mutual information estimation,” in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, IEEE, 2019, pp. 1–5.

[35] J. Song and S. Ermon, “Understanding the limitations of variational mutual information estimators,” *arXiv preprint arXiv:1910.06222*, 2019.

[36] S. Molavipour, G. Bassi, and M. Skoglund, “Conditional mutual information neural estimator,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 5025–5029.

[37] N. Tishby and N. Zaslavsky, “Deep learning and the information bottleneck principle,” in *2015 ieee information theory workshop (itw)*, IEEE, 2015, pp. 1–5.

[38] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, “Deep variational information bottleneck,” *arXiv preprint arXiv:1612.00410*, 2016.

[39] M. Chalk, O. Marre, and G. Tkacik, “Relevant sparse codes with variational information bottleneck,” *Advances in Neural Information Processing Systems*, vol. 29, pp. 1957–1965, 2016.

[40] R. A. Amjad and B. C. Geiger, “Learning representations for neural network-based classification using the information bottleneck principle,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 9, pp. 2225–2239, 2019.

[41] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, “On the information bottleneck theory of deep learning,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, p. 124 020, 2019.

[42] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert, “Nonlinear information bottleneck,” *Entropy*, vol. 21, no. 12, p. 1181, 2019.

[43] Z. Goldfeld and Y. Polyanskiy, “The information bottleneck problem and its applications in machine learning,” *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 19–38, 2020.

[44] R. Shwartz-Ziv and N. Tishby, “Opening the black box of deep neural networks via information,” *arXiv preprint arXiv:1703.00810*, 2017.

[45] G. K. Zipf, *Human behavior and the principle of least effort: An introduction to human ecology*. Connecticut: Martino Fine Books, 2012.

[46] P. Vogt, “Minimum cost and the emergence of the zipf-mandelbrot law,” 2004.

[47] A. Hackmann and T. Klarl, “The evolution of zipf’s law for us cities,” *Papers in Regional Science*, vol. 99, no. 3, pp. 841–852, 2020.

[48] G. De Marzo, A. Gabrielli, A. Zaccaria, and L. Pietronero, “Dynamical approach to zipf’s law,” *Physical Review Research*, vol. 3, no. 1, p. 013 084, 2021.

[49] A. El Kaabouchi, F. X. Machu, J. Cocks, R. Wang, Y. Zhu, and Q. A. Wang, “Study of a measure of efficiency as a tool for applying the principle of least effort to the derivation of

the zipf and the pareto laws," *Advances in Complex Systems*, vol. 24, no. 07n08, p. 2 150 013, 2021.

[50] Q. A. Wang, "Principle of least effort vs. maximum efficiency: Deriving zipf-pareto's laws," *Chaos, Solitons & Fractals*, vol. 153, p. 111 489, 2021.

[51] G. M. Linders and M. M. Louwerse, "Zipf's law revisited: Spoken dialog, linguistic units, parameters, and the principle of least effort," *Psychonomic Bulletin & Review*, vol. 30, no. 1, pp. 77–101, 2023.

[52] A. Agouzal, T. Lafouge, and M. Bertin, "Relationship between the principle of least effort and the average cost of information in a zipfian context," *Journal of Informetrics*, vol. 18, no. 1, p. 101 478, 2024.

[53] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," *arXiv preprint arXiv:1803.03635*, 2018.

[54] J. Frankle, G. K. Dziugaite, D. M. Roy, and M. Carbin, "Stabilizing the lottery ticket hypothesis," *arXiv preprint arXiv:1903.01611*, 2019.

[55] X. Ma, G. Yuan, X. Shen, T. Chen, X. Chen, X. Chen, N. Liu, M. Qin, S. Liu, Z. Wang, *et al.*, "Sanity checks for lottery tickets: Does your winning ticket really win the jackpot?" *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[56] E. Malach, G. Yehudai, S. Shalev-Schwartz, and O. Shamir, "Proving the lottery ticket hypothesis: Pruning is all you need," in *International Conference on Machine Learning*, PMLR, 2020, pp. 6682–6691.

[57] J. Frankle, G. K. Dziugaite, D. Roy, and M. Carbin, "Linear mode connectivity and the lottery ticket hypothesis," in *International Conference on Machine Learning*, PMLR, 2020, pp. 3259–3269.

[58] X. Chen, Y. Cheng, S. Wang, Z. Gan, J. Liu, and Z. Wang, "The elastic lottery ticket hypothesis," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[59] U. Evci, Y. Ioannou, C. Keskin, and Y. Dauphin, "Gradient flow in sparse neural networks and how lottery tickets win," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 6577–6586.

[60] Y. Bai, H. Wang, Z. Tao, K. Li, and Y. Fu, "Dual lottery ticket hypothesis," *arXiv preprint arXiv:2203.04248*, 2022.

[61] A. da Cunha, E. Natale, and L. Viennot, "Proving the strong lottery ticket hypothesis for convolutional neural networks," in *International Conference on Learning Representations*, 2022.

[62] *Qualitative Bankruptcy*, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C52889, 2014.

[63] R. Kamimura, "Forced and natural creative-prototype learning for interpreting multi-layered neural networks," in *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, IEEE, 2024, pp. 343–350.

Appendix

A Compressed Network

To confirm the existence of the supposed prototype, we need to compare the prototype network with some estimated prototypes, extracted from the actual multi-layered networks. For comparing a supposed prototype network with the corresponding surface network, we need to compress the surface and multi-layered neural network into the simplest one. We compress the surface network step-by-step, using a six-layered neural network in Figure 1. For example, we compress the weights from the first layer to the third layer in the first place. The compressed weights, labeled (1,3), which connect the first to the second (1,2) layer and the second to the third layer (2,3), are computed by:

$$w_{ik}^{(1,3)} = \sum_j w_{ij}^{(1,2)} w_{jk}^{(2,3)}. \tag{9}$$

We can repeat this compression for all layers. Assuming that weights labeled (1,5) represent the compressed weights from the first to the fifth layer, the final compressed weights from the first to the sixth layer are given by:

$$w_i^{(1,6)} = \sum_m w_{im}^{(1,5)} w_m^{(5,6)}. \tag{10}$$

This compressed network should then be compared with the corresponding prototype network using the ratio potentiality.

B Structural Potentiality for Compressed Network

Here we explain how to compute the structural potentiality for the compressed network. The individual potentiality is defined as the absolute value of the corresponding compressed weights:

$$u_i^{(1,6)} = \left| w_i^{(1,6)} \right|. \tag{11}$$

Using this individual potentiality, we obtain the relative individual potentiality, computed by:

$$g_i^{(1,6)} = \frac{u_i^{(1,6)}}{\max_{i'} u_{i'}^{(1,6)}}. \tag{12}$$

The final structural potentiality is obtained by summing all individual potentialities:

$$G^{(1,6)} = \sum_i \mathrm{str}\left(g_i^{(1,6)}\right). \tag{13}$$

C Ratio Potentiality

In addition, we need to define another potentiality, called “ratio potentiality”, because we need to compare the compressed network with the corresponding prototype to examine whether learning can detect the supposed prototype. Then, the ratio of compressed to supposed prototype potentiality is:

$$v_i^{(1,6)} = \frac{g_i^{(1,6)}}{z_i^{(1,2)}}, \tag{14}$$

where z represents the individual structural potentiality of a supposed prototype. As shown in Figure 5, the supposed prototype is computed by taking the normalized correlation coefficients between inputs and targets of the training data set. For this v, we can obtain the relative potentiality r by dividing it by the corresponding maximum potentiality. For this new r, we can compute the ratio potentiality in the form of structural potentiality as follows:

$$R^{(1,6)} = \sum_i \mathrm{str}\left(r_i^{(1,6)}\right). \tag{15}$$

Naturally, the ratio potentiality is bounded in the framework of structural potentiality, and we compute the normalized ratio potentiality:

$$R_{nrm}^{(1,6)} = \frac{R^{(1,6)}}{R_{max}^{(1,6)}}. \tag{16}$$

This ratio potentiality becomes maximum when all individual ratio potentialities are the same, and the maximum value is equivalent to the number of weights. Comparison can be done by using the conventional KL divergence between the supposed and compressed network, but it is unbounded, preventing us from evaluating similarity or difference between the supposed and compressed networks appropriately. As shown in Figure 7, the divergence tended to be much larger, hiding detailed characteristics.