BEAN: Interpretable Representation Learning with Biologically-Enhanced Artificial Neuronal Assembly Regularization

Anonymous authors
Paper under double-blind review

Abstract

Deep neural networks (DNNs) are known for extracting good representations from a large amount of data. However, the representations learned in DNNs are typically hard to interpret, especially the ones learned in dense layers. One crucial issue is that neurons within each layer of DNNs are conditionally independent with each other, which makes the analysis of neurons at higher modularity difficult. In contrast, the dependency patterns of biological neurons in the human brain are largely different from those of DNNs. Neuronal assembly describes such neuron dependencies that could be found among a group of biological neurons as having strong internal synaptic interactions, potentially high semantic correlations that are deemed to facilitate the memorization process. In this paper, we show such a crucial gap between DNNs and biological neural networks (BNNs) can be bridged by the newly proposed Biologically-Enhanced Artificial Neuronal assembly (BEAN) regularization that could enforce dependencies among neurons in dense layers of DNNs without altering the conventional architecture. Both qualitative and quantitative analyses show that BEAN enables the formations of interpretable and biologically plausible neuronal assemblies in dense layers and consequently enhances the modularity and interpretability of the hidden representations learned. Moreover, BEAN further results in sparse and structured connectivity and parameter sharing among neurons, which substantially improves the efficiency and generalizability of the model.

1 Introduction

Deep neural networks (DNNs) are known for extracting good representations from a large amount of data (Bengio et al. (2013)). Despite the success and popularity of DNNs in a wide variety of fields, including computer vision (Krizhevsky et al. (2012); He et al. (2016)) and natural language processing (Collobert & Weston (2008); Young et al. (2018)), the representations learned in DNNs are typically hard to interpret, especially the ones in dense (fully connected) layers. Recently, the attempt to build a more intrinsically interpretable convolutional unit has received much attention (Zhang & Zhu (2018); Sabour et al. (2017)), yet little has been explored of the representation learned in the dense layer. In fact, dense layers are the fundamental and critical component of most of the state-of-the-art DNNs, which are typically used for the late stage of the network's computation, akin to the inference and decision-making process of the network (Krizhevsky et al. (2012); Simonyan & Zisserman (2014); He et al. (2016)). Thus the advancement of the interpretation and visualization of the dense layer representation is crucial if we are to fully understand the behavior of DNNs.

However, interpreting the representations learned in dense layers of DNNs is typically a very challenging task. One crucial issue is that neurons are conditionally independent with each other within each layer since dense layers typically assume an all-to-all feed-forward neuron activity (and an all-to-all feedback weight adjustment). In this comprehensively 'vertical' connectivity, every node is independent and abstracted 'out of the context' of the other nodes. This issue limits the analysis of the representation learned in DNNs to single-unit level, instead of a higher modularity level such as neuron population level. Moreover, recent study on single unit importance study seems to suggest that individually selective units may have little correlation with the overall network performance (Morcos et al. (2018); Zhou et al. (2018)).

On the other hand, understanding the neuron correlations in biological neural networks (BNNs) has long been a subject of intensive interest for neuroscience researchers. Circuitry blueprints in the real brain are 'filtered' by the physical requirements of axonal projections and the consequent need to minimize cable while maximizing connections. One could naively expect that the non-all-to-all limitations imposed in natural neural systems would be detrimental to their computational power.

Instead, it makes them superiorly efficient and allows cell assemblies to emerge. Neuronal assembly or cell assembly (Hebb (1949)) describes such neuron dependencies that could be found among a group of biological neurons as having strong internal synaptic interactions, potentially high semantic correlations that are deemed to facilitate the memorization process (Braitenberg (1978)).

In this paper, we show such a crucial gap between DNNs and BNNs can be bridged by modeling of the neuron correlations within each layer of DNNs. By leveraging biologically inspired learning rules in neuroscience and graph theories, we propose a novel Biologically-Enhanced Artificial Neuronal assembly (BEAN) regularization that can enforce dependencies among neurons in dense layers of DNNs without altering the conventional architecture. Specifically, the advantages are threefold:

- **Enhance interpretability and modularity at neuron population level.** Modeling neural correlations and dependencies will enable us to better interpret and visualize the learned representation in hidden layers at the neuron population level instead of the single neuron level. Both qualitative and quantitative analyses show that BEAN enables the formations of interpretable and biologically plausible neuronal assembly patterns in the hidden layers, which consequently enhances the modularity and interpretability of the hidden representations of DNNs models.
- **Achieves efficient neuron connectivity.** Efficient connectivity presents topologically in most BNNs due to the physical restrictions of dendrites and axons (Rivera-Alba et al. (2014)) and for energy efficiency. In DNNs, sparsity is also a promising property that supports energy-saving and network compression. Current studies on network compression (Han et al. (2015); Iandola et al. (2016)) achieve sparsity with conventional sparsity regularization that typically focuses on the scale of each individual weights, paying little attention to the global structure of the connectivity of the system. In the paper, we show that BEAN can help the model learn sparse and structured neuron connectivity as well as parameter sharing among neurons, which substantially improves the efficiency of the model at a higher modularity level.
- **Improves model generalizability.** Humans and animals can learn and generalize to new concepts with just a few trials of learning, while DNNs generally perform much poorly on such tasks. Current few-shot learning techniques in deep learning relay heavily on a large amount of additional knowledge to work well. For example, transfer learning based methods typically leverage a pre-trained model trained with a large amount of data (Xian et al. (2018); Socher et al. (2013)), and meta learning based methods require a large amount of additional side tasks (Finn et al. (2017); Snell et al. (2017)). Here we explore BEAN with a substantially more challenging *few-shot learning from scratch* task defined by Kimura et al. (2018), where no additional knowledge is provided beside the few training observations. Extensive experiments show that BEAN has a significant advantage of improving model generalizability over conventional techniques.

2 Biologically-Enhanced Artificial Neuronal Assembly Regularization

In this section, we first propose the overall objective of Biologically-Enhanced Artificial Neuronal Assembly (BEAN) regularization inspired by neuroscience and graph theories. And then we propose the necessities for implementing it on DNNs, which are *Layer-wise Neuron Correlation* to model the implicit correlations and dependencies between neurons within the same layer and *Layer-wise Neuron Co-activation Divergence* to characterize the co-activation rate between neurons.

2.1 The Layer-wise Neuron Co-activation Divergence

Due to the physical restrictions imposed by dendrites and axons (Rivera-Alba et al. (2014)) and for energy efficiency, biological neural systems are "parsimonious" and can only afford to form a limited number of connections between neurons. Based on the neuron connectivity patterns, their activation patterns are often formed based on the principle of "*Cells that fire together wire together*", which is known as **cell assembly theory**. It explains and relates to several characteristics and advantages of BNN architecture such as modularity (Peyrache et al. (2010)), efficiency, and generalizability, which are just the aspects where the current DNNs are usually struggling in (LeCun et al. (2015)). To enjoy the architectural merit in BNNs and overcome the existing drawbacks of DNNs, we propose the Biologically-Enhanced Artificial Neuronal assembly (BEAN) regularization that ensures neurons that "wire" together with a high outgoing weight correlation also "fire" together with minimal pairwise co-activation divergence. An example of the artificial neuronal assembly achieved by our method can be seen in Figure 1(d). The regularization is formulated as follows:

$$L_c^{(l)} = 1/(SN_l^2)\sum_s \sum_i \sum_j A_{i,j}^{(l)} \times d(H_{s,i}^{(l)}, H_{s,j}^{(l)}) \tag{1}$$

where the term $A_{i,j}^{(l)}$ is for characterizing the wiring (i.e., the higher the stronger) while the term $d(H_{s,i}^{(l)}, H_{s,j}^{(l)})$ is for modeling the divergence of firing patterns (i.e.,g the higher the more different). Thus, by multiplying these two functions, we are penalizing those neurons with strong connectivity while having high divergence in terms of their activation, which is biologically plausible and echoes the cell assembly theory. S is the total number of input samples while N_l is the total number of hidden neurons in Layer l.

Specifically, $A_{i,j}^{(l)}$ defines the connectivity relation among the neuron i and neuron j in DNN, which is instantiated by our newly proposed "Layer-wise Neuron Correlation" and will be elaborated in Sections 2.2 and 2.3. On the other hand, to model the "co-firing" correlation, $d(H_{s,i}^{(l)}, H_{s,j}^{(l)})$ is defined as "Layer-wise Neuron Co-activation Divergence" which denotes the difference in the activation patterns in lth layer between $H_{s,i}^{(l)}$ and $H_{s,j}^{(l)}$ of neuron i and neuron j, respectively. Here $H_{s,i}^{(l)}$ represents the activation of neuron i in layer l for a given input sample s. And the function $d(x, y)$ can be a common divergence metric such as absolute difference or square difference. In this study, we show the results for a square difference in the Experimental Study Section; the absolute difference results follow a similar trend.

Model Training: The general objective function of training a DNN model along with the proposed regularization on fully connected layer l can be written as: $L = L_{DNN} + \alpha L_c^{(l)}$, where L_{DNN} represents the general deep learning model training loss and the hyper-parameter α controls the relative strength of the regularization.

Equation 1 can be optimized with backpropagation (Rumelhart et al. (1988)) using the chain rule:

$$\frac{\partial L_c^{(l)}}{\partial W^{(l+1)}} = \frac{\partial A^{(l)}}{\partial W^{(l+1)}} D^{(l)}, \ \frac{\partial L_c^{(l)}}{\partial W^{(l)}} = A^{(l)} \frac{\partial D^{(l)}}{\partial H^{(l)}} \frac{\partial H^{(l)}}{\partial W^{(l)}}, \ ... \tag{2}$$

where $D^{(l)} \in \mathbb{R}^{S \times N_l \times N_l}$ of which each element is $D_{s,i,j}^{(l)} = d(H_{s,i}^{(l)}, H_{s,j}^{(l)})$.

Remark 1. *BEAN regularization enjoys several merits. First, it enforces interpretable and biologically plausible neuronal assemblies without the need to introduce sophisticated handcrafted designs into the architecture, which is justified later in Section 4.1. In addition, modeling the neuron correlations and dependencies further results in efficient connectivity in dense layers, which substantially improved the generalizability of the model when insufficient data and knowledge is provided, which is demonstrated later in Section 4.2. Lastly, the Layer-wise Neuron Correlation can be efficiently computed with matrix operations, as per Equations 5 and 7, which enables modern GPUs to boost up the speed during model training.*

2.2 The First-Order Layer-wise Neuron Correlation

In the following two sections, we introduce how we formulate the layer-wise neuron correlation, namely $A_{i,j}^{(l)}$ between any pair of neurons i and j.

In the human brain, the correlation between two neurons depends on the wiring between them (Buzsáki (2010)), and hence is typically treated as a binary value in BNNs studies with "1" indicating the presence of a connection and "0" the absence, so the correlation among a group of neurons can be represented by the corresponding adjacency matrix. Although there is typically no direct connection between neurons within the same layer of DNNs, it is possible to model neurons correlations based on their connectivity patterns to the next layer. Just like in network science, where it is useful to also consider the relationships between nodes based on their common neighbor nodes besides their direct connections. One classic concept that is widely used to describe such pattern is called *triadic closure* Granovetter (1977). As shown in Figure 1(b), triadic closure can be interpreted here as a property among three nodes i, j, and k, such that if connections exist between $i - k$ and $j - k$, there is also a connection between $i - j$. This is closely related to the concepts of *clustering coefficient* (Watts & Strogatz (1998)) and *transitivity* (Holland & Leinhardt (1971)) in graph theory.

Inspired by this, we take it a step further to model the correlations between neurons within the same layer by their connections to the neurons in the next layer. In fact, this can be considered loosely as analogous to the degree of similarity of the axonal connection pattern of biological neurons in BNNs. To simulate the relative strength of such connections in DNNs, we introduce a function $f(\cdot)$ which

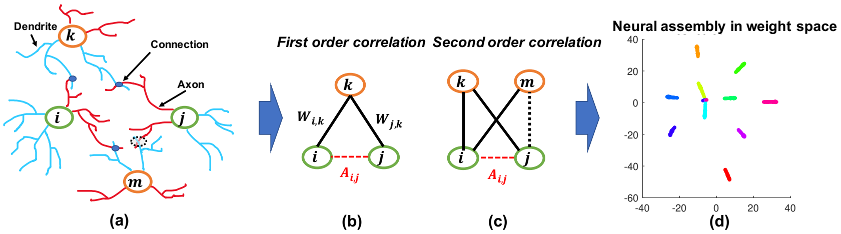

Figure 1: An illustration of how the proposed constraint drew inspiration from BNNs and bipartite graphs. **(a)** neuron correlations in BNNs correspond to connections between dendrites, which are represented by blue lines, and axons, which are represented by red lines. **(b)** and **(c)** analogy of figure (a) represented as connections between layers in DNNs; although nodes i and j cannot form direct links, they can be correlated by a given node k as a first-order correlation, or by two nodes k and m as a second-order correlation which is also equivalent to a 4-cycle in bipartite graphs. **(d)** an example of a learned neuronal assembly in neurons outgoing weight space, with the dimensionality reduced to 2D with T-SNE (Maaten & Hinton (2008)). Each point represents one neuron and the neurons are colored according to their highest activated class in the test data.

converts the actual weights into a relative connectivity strength. Suppose matrix $W^{(l+1)} \in \mathbb{R}^{N_l \times N_{l+1}}$ represents all the weights between neurons in layers l and $l+1$ in DNNs, where N_l and N_{l+1} represent the numbers of neurons respectively. The relative connectivity strength can be estimated by the following equation[1]:

$$f(W^{(l+1)}) = |tanh(\gamma W^{(l+1)})| \tag{3}$$

where $|\cdot|$ represents the element-wise absolute operator; $tanh(\cdot)$ represents the element-wise hyperbolic tangent function; and γ is a scalar that controls the curvature of the hyperbolic tangent function. The values of $f(W^{(l+1)}) \in \mathbb{R}^{N_l \times N_{l+1}}$ will all be positive and in the range of $[0, 1)$ with the value simulating the relative connectivity strength of the synapse between neurons. Base on this, we can now give the definition for the *layer-wise first-order neuron correlation* as:

Definition 1. Layer-wise first-order neuron correlation. For a given neuron i and neuron j in layer l, the layer-wise first-order neuron correlation is given by:

$$A_{i,j}^{(l)} = (1/N_{l+1}) \sum_{k=1}^{N_{l+1}} f(W_{i,k}^{(l+1)}) \times f(W_{j,k}^{(l+1)}) \tag{4}$$

The above formula can be expressed as the product of two matrices:

$$A^{(l)} = (1/N_{l+1}) f(W^{(l+1)}) \cdot f(W^{(l+1)})^T \tag{5}$$

where $\cdot$ represents the matrix multiplication operator.

The layer-wise neuron correlation matrix $A^{(l)}$ is a symmetric square matrix that models all the pairwise neuron correlations in layer l. Each entry $A_{i,j}^{(l)}$ takes a value in the range $[0, 1)$ and models the correlation between neuron i and neuron j in terms of the similarity of their connectivity patterns. The higher the value, the stronger the correlation between the two.

In this setting, two neurons i and j from layer l will be linked and correlated by an intermediate node k from layer $l+1$ if and only if both edges $f(W_{i,k}^{(l+1)})$ and $f(W_{j,k}^{(l+1)})$ are non zero, and the relative strength can be estimated by $f(W_{i,k}^{(l+1)}) \times f(W_{j,k}^{(l+1)})$, which will be in the range $[0, 1)$. Since there are N_{l+1} neurons in layer $l+1$, where each neuron k can contribute to such connections, run over all neurons in layer $l+1$ we obtain Equation 4 and Equation 5.

2.3 The Second-Order Layer-wise Neuron Correlation

Although the first-order correlation is able to estimate the degree of dependency between each pair of neurons, it may not be sufficient to strictly reflect the degree of grouping or assembly of the neurons. Thus, here we further propose a second-order neuron correlation based on the first-order correlation defined in Equation 4 and 5, as:

Definition 2. Layer-wise second-order neuron correlation. For a given neuron i and neuron j in layer l, the layer-wise second-order neuron correlation is given by:

$$A_{i,j}^{(l)} = (1/N_{l+1}^2) \sum_{k,m} f(W_{i,k}^{(l+1)}) \times f(W_{j,k}^{(l+1)}) \times f(W_{i,m}^{(l+1)}) \times f(W_{j,m}^{(l+1)}) \tag{6}$$

The above formula can be expressed as the product of four matrices:

$$A^{(l)} = (1/N_{l+1}^2)(f(W^{(l+1)}) \cdot f(W^{(l+1)})^T) \odot (f(W^{(l+1)}) \cdot f(W^{(l+1)})^T) \tag{7}$$

[1] Similar to ReLU activation function, our formulation introduces a non-differentiable point at zero, we follow the conventional setting by using the sub-gradient for model optimization.

where $\odot$ represents the element-wise multiplication of matrices.

The second-order correlation is more strict in terms of correlating neurons, as it requires at least two common neighbor neurons from the layer above that appear to have strong connectivity in order to form correlations, as compared to the first-order correlation the requires just one common neighbor. Moreover, the second-order neuron correlation is closely related to concepts in both graph theory and the neuroscience learning rule. As shown in the following remarks:

Remark 2. ***Interpretation from graph theory.*** *Modeling the correlations and dependencies between neurons in one layer via the connectivity pattern with the neurons in the layer above in DNNs can be closely related to the analysis of the relationship between nodes from the same mode in two-mode networks or bipartite graph. In graph theory, the definition of global clustering coefficients for the two-mode network proposed by Robins & Alexander (2004) is defined as the ratio between the number of 4-cycles and the number of 3-paths, as illustrated in Figure 1(c)* [2]*. The solid lines represent a 3-path in two-mode networks. If the dashed line was present, the 3-path would be closed and part of a 4-cycle. Formally, this coefficient is:*

$$C_4 = (\textit{number of 4-cycles})/(\textit{number of 3-paths}) \tag{8}$$

As can be easily seen, 4-cycle in two-mode networks is directly related to the second-order neuron correlation in terms of modeling the grouping tendency of two nodes from the same mode.

Remark 3. ***Interpretation from neuroscience theory.*** *The BIG-ADO rule proposed by Mainetti & Ascoli (2015) and the discrete neuronal circuits studied by Pulvermüller & Knoblauch (2009) are two similar neuroscience learning mechanisms that closely related to our formulation of second-order neuron correlation. Specifically, the BIG-ADO learning rule quantifies the tendency of neurons to form potential synapses by defining a proximity function as:*

$$\pi(j,m) = \sum\nolimits_{i,k} (W_{j,k} \times W_{i,k} \times W_{i,m}) \tag{9}$$

Figure 1 (a) illustrates a scenario of the BIG-ADO learning rule in BNNs. The solid blue circle represents a connection that was formed between two neurons (i.e. a synapse), while the dashed circle between neuron j and m represents an Axo-Dendritic Overlap (ADO) (i.e. a potential synapse) between the two neurons. The BIG-ADO rule was designed to encourage a potential synapse between j and m in such a pattern to form a connection. Noticeably, both of the aforementioned papers mentioned such learning mechanism could be related to the formation of cell assemblies in the brain, which ultimately linked to our observation of neuronal assemblies patterns in DNNs hidden representation where BEAN regularization was imposed, as introduced later in Section 3.1.

3 Experimental Study

In this section, we first studied and analyzed the interpretability and biological plausibility of the learning outcomes of BEAN regularization on multiple classical image recognition tasks in Section 3.1. We then further studied the effect of BEAN regularization on improving the generalizability of the model on several few-shot learning from scratch task simulations in Section 3.2. We studied both BEAN variations, i.e. BEAN-1 and BEAN-2, based on the two proposed layer-wise neuron correlation defined by Equation 5, and Equation 7 respectively. The value for γ defined in Equation 3 was set to 1. All the experiments were conducted on a 64-bit machine with Intel(R) Xeon(R) W-2155 CPU 3.30GHz processor and 32GB memory and an NVIDIA TITAN Xp GPU.

3.1 Biologically plausibility and interpretability of BEAN.

To analyze and interpret the learning outcomes of BEAN regularization, we conducted experiments on two classical image recognition tasks on MNIST (LeCun et al. (1998)) and CIFAR10 (Krizhevsky & Hinton (2009)) datasets. Specifically, an MLP with one hidden layer of 500 neurons with ReLU activation function and a LeNet-5 (LeCun et al. (1998)) were used for the MINIST dataset; and a ResNet18 (He et al. (2016)) was used for the CIFAR10 dataset. The Adam optimizer (Kingma & Ba (2014)) was used with a learning rate of 0.0005 and a batch size of 100 for model training until train loss convergence was achieved; BEAN was applied to all the dense layers of each model.

[2] We are aware that there has been some debate on the definition between 4-cycle and 6-cycle (Opsahl (2013)). Here we accept the conventional definition and the associated discussion is outside the scope of this paper.

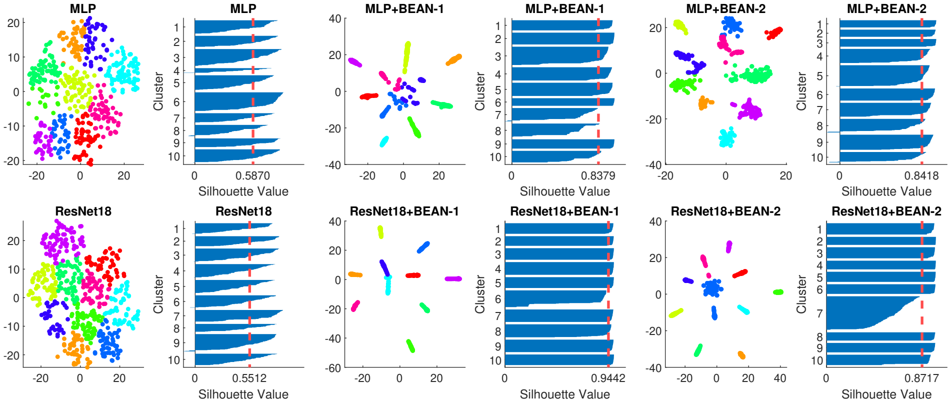

Figure 2: Neuronal assembly patterns found in neurons' weight space of the dense layer of different models on both MNIST (top) and CIFAR-10 (bottom) datasets, along with clustering validation via Silhouette score on 10 clusters K-means clustering. The dimensionality of neurons' weight space was reduced to 2D with T-SNE for visualization.

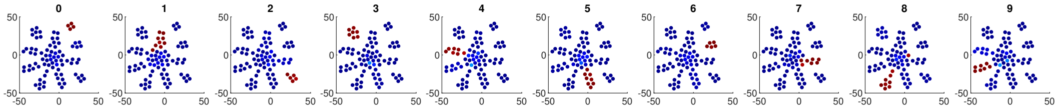

Figure 3: Neuron co-activation patterns found in the representation of the last dense layer of LeNet-5+BEAN-2 model. The dimensionality of neurons' weight space was reduced to 2D with T-SNE for visualization. Each point represents one neuron within the last dense layer of the model and is colored based on its activation scale. The 10 subplots show the average activation heat-maps when each digit's samples were fed into the model. The warmer color indicates a higher neuron activation.

3.1.1 Biological plausibility of the learned neuronal assemblies

By analyzing the neurons' connectivity patterns based on their out-going weights, we found that there are indeed neuronal assemblies among neurons in dense layers where BEAN regularization was enforced. Specifically, for both datasets, we found that the neuronal assemblies at the last dense layer could be best described by 10 clusters with K-means clustering (MacQueen et al. (1967)) validated by Silhouette co-efficient (Rousseeuw (1987)), as shown in Figure 2. Both BEAN-1 and BEAN-2 could enforce neuronal assemblies for various models on various datasets, yielding Silhouette indices around 0.9, which indicates strong clustering patterns among neurons in dense layers where BEAN regularization was applied. On the other hand, training conventional DNN models with the same architectures could only yield nearly 0.5 Silhouette indices, which indicate no clear clustering patterns in conventional dense layers of deep neuronal networks.

Besides, we found co-activation behavior of neurons within each neuronal assembly, which is both interpretable and biologically plausible. Figure 3 shows the visualization of neuron co-activation patterns found in the last dense layer of LeNet-5+BEAN-2 model on MNIST dataset. For the samples of each specific class, only those neurons in the neuron assembly corresponding to this class have high activation. This indicates a strong association between each unique assembly and each unique class concept, yielding good interpretability of the neuron populations in the dense layers. From the neuroscience perspective, such pattern is also biologically plausible as neuroscientists also found similar co-firing patterns (Peyrache et al. (2010)) as well as a strong association between neuronal assembly and concepts (Tononi & Sporns (2003)) in biological neural networks.

Lastly, we also found a strong correlation between neuronal assembly and class selectivity indices. Selectivity index was originally proposed and used in systems neuroscience (De Valois et al. (1982); Freedman & Assad (2006)). Recently, unit class selectivity is also studied by machine learning researchers (Morcos et al. (2018); Zhou et al. (2018)) as a metric for interpreting the behaviors of single units in deep neural networks. Mathematically, it is calculated as: $selectivity = (\mu_{max} - \mu_{-max})/(\mu_{max} + \mu_{-max})$, where μ_{max} represents the highest class-conditional mean activity and μ_{-max} represents the mean activity across all other classes. To better understand how high-level concepts are associated with the learned neuron assemblies, we further visualized the neurons by labeling each neuron with the class in which it achieved its highest class-conditional mean activity μ_{max} in the test data. Figure 4 shows the results for the last dense layer of the models trained with both datasets. We found that the neuronal assembly could be well described based on neurons selectivity. The strong association between neuronal assemblies and neurons' selectivity index further demonstrated the biological plausibility of the learning outcomes of BEAN regularization.

3.1.2 Quantitative analysis of interpretability

To quantitatively evaluate and compare interpretability, ablation study is a commonly used technique inspired by experimental Neuropsychology when studying brains, where parts of the brain were

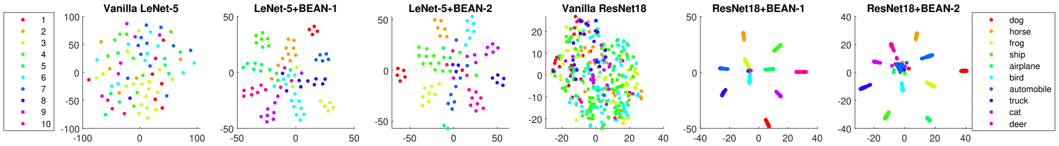

Figure 4: The strong association between neuronal assemblies and neurons' class selectivity index with BEAN regularization. Each point represents one neuron and the color represents the class where the neuron achieved its highest class-conditional mean activity in the test data.

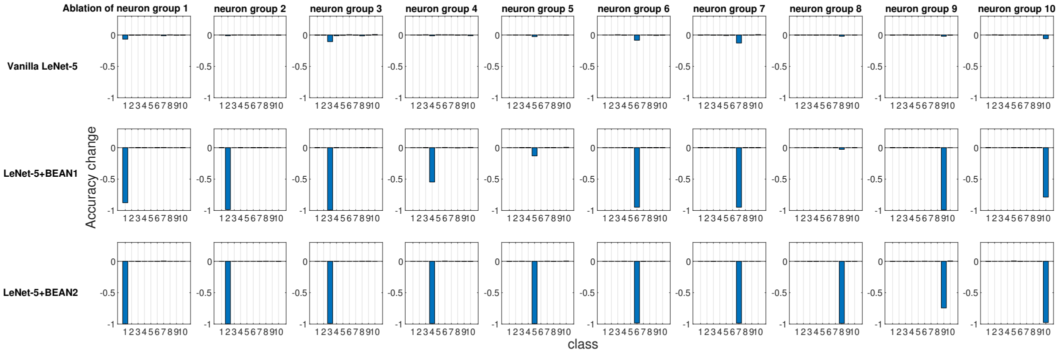

Figure 5: The ablation study at the neuron population level of the last dense layer of LeNet-5 models. Each time, one distinct group of neurons were ablated based on their most selective class and the model performance changes for each individual class were recorded.

removed to study the consequential effects. Likewise, ablation study has also been well-adapted for interpreting deep neural networks, such as understanding which layers or units are critical for model performance (Girshick et al. (2014); Morcos et al. (2018); Zhou et al. (2018)).

In the light of the neuronal assembly, we performed the ablation study at the neuron population level, where each time one distinct group of neurons were ablated and the model performance changes for each class were recorded. Like shown in Figure 4, we identified neuron groups via class selectivity and performed neuron population ablation accordingly. Figure 5 shows the results of all 10 ablation runs for each class in MNIST dataset. Just like Morcos et al. (2018) has discovered, for conventional deep neural nets, there is indeed no clear association between neuron's selectivity and importance to the overall model performance, even when neuron population ablation was conducted. However, when BEAN regularization was injected during training, such association became much clear and significant, especially for BEAN-2. This is because BEAN-2 could enforce neurons to form more strict neuron correlation than BEAN-1 with the second-order correlation, which enables the groups of neurons to represent more compact and disentangled concepts, such as handwritten digits.

3.2 Towards few-shot learning from scratch with BEAN regularization

In an attempt to test the influence of BEAN regularization on model's generalizability, we conducted the *few-shot learning from scratch* task which refers to performing few-shot learning task with only the knowledge within the few training examples, without the help of any additional side tasks and pre-trained models Kimura et al. (2018). We conducted several simulations of *few-shot learning from scratch* task on MNIST (LeCun et al. (1998)) and fashion-MNIST (Xiao et al. (2017)) datasets. So far, such learning task has rarely been explored due to the difficulty of the problem setup as compared to other conventional few-shot learning tasks where additional data or knowledge could be accessed. Currently, only Kimura et al. (2018) did a preliminary exploration with the proposed imitation networks model. Beside imitation networks, we also compared BEAN with other conventional regularization techniques that were commonly used in deep learning literature. Specifically, we compared dropout (Srivastava et al. (2014)), weight decay (Krogh & Hertz (1992)), and ℓ_1-norm. The regularization terms were applied to all the dense layers and hyperparameters were chosen based on model performance on a validation set sampled from the original training base. The whole original 10K testing set was used for final performance evaluation.

Table 1 shows model performance on several *few-shot learning from scratch* experiments on MNIST and fashion-MNSIT datasets. Performance is averaged over 20 experiments of randomly sampled training data from the original training base. The best and second-best results for each few-shot learning setting are highlighted in boldface and italic font respectively. As can be seen, the proposed BEAN regularization advanced the state-of-the-art by a significant margin on all 4 *few-shot learning from scratch* tasks tested on both datasets. Moreover, BEAN advanced the performance more significantly when training samples were more limited. For instance, BEAN outperformed all comparison methods by 24% - 42% and 13% - 29% on 1-shot learning tasks on MNIST and Fashion MNIST datasets respectively, which demonstrates the promising effect of BEAN regularization on

Table 1: Few-shot learning from scratch experiments on MNIST (left) and Fashion MNIST (right) datasets. Performance is averaged over 20 simulations of randomly sampled training data from the original training base. The best and second-best results for each few-shot learning setting are highlighted in boldface and italic font respectively.

Dataset	MNIST				Fashion MNIST			
Model	1-shot	5-shot	10-shot	20-shot	1-shot	5-shot	10-shot	20-shot
Vanilla LeNet-5	38.63	70.21	78.97	86.68	39.32	59.02	64.50	70.23
LeNet-5 + Dropout	40.13	72.45	82.04	89.22	40.78	60.04	65.40	71.83
LeNet-5 + Weight decay	39.51	71.76	82.87	90.15	41.31	61.98	67.25	71.88
LeNet-5 + ℓ_1-norm	40.96	74.35	81.17	90.68	41.26	62.18	67.30	70.85
Imitation networks	44.10	70.40	80.00	86.70	44.80	62.10	68.00	72.50
LeNet-5 + BEAN-1	**54.79**	**83.42**	*87.51*	*92.79*	**50.57**	**66.95**	*69.21*	*74.25*
LeNet-5 + BEAN-2	*53.75*	*80.76*	**88.08**	**92.97**	*49.94*	*65.98*	**70.21**	**75.06**

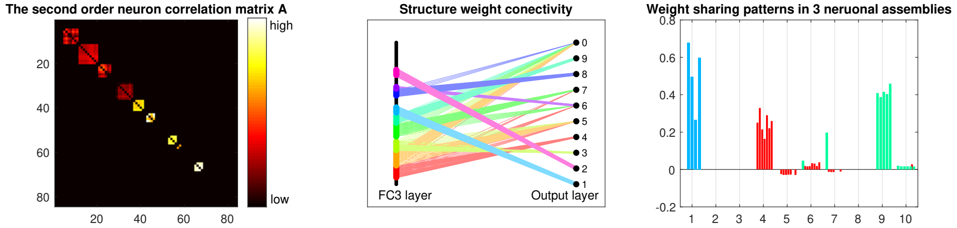

Figure 6: Analysis and visualization of the last dense layer of LeNet-5+BEAN-2 model on MNIST 10-shot learning from scratch task, BEAN regularization helped dense layer form efficient parameter usage via sparse and structured connectivity learning and weak parameter sharing among neurons. **(a)** The heat-map of the learned second-order neuron correlation matrix, neuron indices are re-ordered for best visualization of neuronal assembly patterns, BEAN is able to form plausible assembly structures even with such extreme limited sample size. **(b)** Visualization of the sparse and structured connectivity learned in dense layer, neurons are grouped and colored by neuronal assembly. **(c)** Visualization of the scales of neruons' out-going weights, the weights of neuron are colored to be consistent with neuron group in (b).

improving the generalizability of the over-parameterized deep nets. Another interesting observation is that BEAN-1 in general performed the best with extremely limited training samples, such as 1-shot and 5-shot learning tasks; while BEAN-2 regularization in general performed the best with a bit more training samples, such as 10-shot and 20-shot learning tasks. The reason behind this observation might be related to the fact that higher-order correlation is more strict in terms of correlating neurons, as it requires more common neighbor neurons that appear to have strong connections with both neurons. Thus with a bit more observations available, BEAN-2 could do a better job at estimating more realistic neuronal assembly and thus further improve the model performance.

To better understand why BEAN regularization could help the seemingly over-parameterized model to generalize well on small sample set, we further analyzed the learned hidden representation of the dense layers where BEAN regularization was injected. We found that BEAN helped the model to gain better generalization power in two aspects: 1) by automatic sparse and structured connectivity learning and 2) by weak parameter sharing among neurons within each neuronal assembly. Both aspects enhanced the efficiency of dense layers' parameter usage in a biologically plausible way, which consequently prevented the model from over-fitting badly with small training sample size.

Figure 6 shows the learned parameters of the last dense layer of LeNet-5+BEAN2 on the MNIST 10-shot learning task. As shown in Figure 6 (b), instead of using all possible weights in the dense layer, BEAN enforced model to efficiently leverage the parameters, yielding a plausible sparse and structured connectivity pattern. This is because the learned neuron correlation helped disentangle the co-connections between neurons from different assemblies, as shown in Figure 6 (a). Besides, BEAN enhanced parameters sharing among neurons within each assembly, as demonstrated in Figure 6 (c). For instance, neurons in red-colored assembly all had high positive weights towards class 4, meaning that this group of neurons was helping the model to identify Digit 4. Similarly, neurons in the green-colored assembly were trying to distinguish between Digit 9 and 7. Such automatic weak parameter sharing not only helped prevent model from over-fitting but also enabled an overall interpretation of the behavior of the system as a whole from a higher modularity level.

4 Conclusion

In this work, we have taken a big step forward towards bridging the gaps between biological and modern artificial deep neural networks with the proposed Biologically-Enhanced Artificial Neuronal assembly (BEAN) regularization which enforces biologically plausible neuronal assembly patterns, enhances the efficency and modularity of the hidden representations, and improves generalizablity of DNNs. Our work opens up a new window for visualization and interpretation of the dense layer representation at a higher modularity level.

REFERENCES

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Valentino Braitenberg. Cell assemblies in the cerebral cortex. In *Theoretical approaches to complex systems*, pp. 171–188. Springer, 1978.

György Buzsáki. Neural syntax: cell assemblies, synapsembles, and readers. *Neuron*, 68(3):362–385, 2010.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160–167. ACM, 2008.

Russell L De Valois, E William Yund, and Norva Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vision research*, 22(5):531–544, 1982.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135. JMLR. org, 2017.

David J Freedman and John A Assad. Experience-dependent representation of visual categories in parietal cortex. *Nature*, 443(7107):85, 2006.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

Mark S Granovetter. The strength of weak ties. In *Social networks*, pp. 347–367. Elsevier, 1977.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pp. 1135–1143, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Donald O. Hebb. The organization of behavior. a neuropsychological theory. 1949.

Paul W Holland and Samuel Leinhardt. Transitivity in structural models of small groups. *Comparative group studies*, 2(2):107–124, 1971.

Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

Akisato Kimura, Zoubin Ghahramani, Koh Takeuchi, Tomoharu Iwata, and Naonori Ueda. Few-shot learning of neural networks from scratch by pseudo example optimization. *arXiv preprint arXiv:1802.03039*, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pp. 950–957, 1992.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 281–297. Oakland, CA, USA, 1967.

Matteo Mainetti and Giorgio A Ascoli. A neural mechanism for background information-gated learning based on axonal-dendritic overlaps. *PLoS computational biology*, 11(3):e1004155, 2015.

Ari S Morcos, David GT Barrett, Neil C Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*, 2018.

Tore Opsahl. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 35(2):159–167, 2013.

Adrien Peyrache, Karim Benchenane, Mehdi Khamassi, Sidney I Wiener, and Francesco P Battaglia. Principal component analysis of ensemble recordings reveals cell assemblies at high temporal resolution. *Journal of computational neuroscience*, 29(1-2):309–325, 2010.

Friedemann Pulvermüller and Andreas Knoblauch. Discrete combinatorial circuits emerging in neural networks: A mechanism for rules of grammar in the human brain? *Neural networks*, 22(2): 161–172, 2009.

Marta Rivera-Alba, Hanchuan Peng, Gonzalo G de Polavieja, and Dmitri B Chklovskii. Wiring economy can account for cell body placement across species and brain areas. *Current Biology*, 24 (3):R109–R110, 2014.

Garry Robins and Malcolm Alexander. Small worlds among interlocking directors: Network structure and distance in bipartite graphs. *Computational & Mathematical Organization Theory*, 10(1): 69–94, 2004.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pp. 3856–3866, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pp. 935–943, 2013.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Giulio Tononi and Olaf Sporns. Measuring information integration. *BMC neuroscience*, 4(1):31, 2003.

Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393 (6684):440, 1998.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Revisiting the importance of individual units in cnns via ablation. *arXiv preprint arXiv:1806.02891*, 2018.