Higher-order mutual information reveals synergistic sub-networks for multi-neuron importance

Anonymous Author(s)
Affiliation
Address
email

Abstract

Quantifying which neurons are important with respect to the classification decision of a trained neural network is essential for understanding their inner workings. Previous work primarily attributed importance to individual neurons. In this work, we study which groups of neurons contain synergistic or redundant information using a multivariate mutual information method called the O-information. We observe the first layer is dominated by redundancy suggesting general shared features (i.e. detecting edges) while the last layer is dominated by synergy indicating local class-specific features (i.e. concepts). Finally, we show the O-information can be used for multi-neuron importance by re-training a synergistic sub-network results in a minimal change in performance. These results suggest our method can be used for pruning and unsupervised representation learning.

1 Introduction

Methods for understanding deep neural networks (DNNs) often focus on analysing properties of individual neurons to explain their good generalization performance. One approach is to quantify importance of neurons for the classification task based on their selectivity (i.e. neurons firing only for a specific class)[13]. Various studies have shown that DNNs might not rely on these selective neurons for generalization [9][10][12]. This suggest researchers should focus on properties that exists across multiple neurons in high-dimensional distributed representations. In this work, we provide the first quantifiable metric based on multivariate information theory for analysing DNNs and multi-neuron importance.

Recent work uses mutual information(MI) for individual neuron importance[9]. However, the MI can only identify the non-linear dependencies between two variables. When considering higher-order interactions of more then three variables the information is decomposed into the synergy and redundancy. Synergy arises from statistical interactions that can be found collectively in the whole (i.e. emergence) but not in the parts when considered in isolation, while redundancy refers to the information shared across variables [18]. In this paper, we argue that the generalization abilities of DNNs is due to the networks ability to make a decision that emerges from the higher-order interactions of multiple synergistic neurons (i.e. the whole) while none of the neurons would trigger in isolation (i.e. the part). In this work, we use a multivariate information theory measure called the O-information that scales to higher-order interactions[14] to study the properties of multiple neurons in artificial neural networks.

Current methods verify importance by removing important neurons from the network and evaluating the drop in classification performance. This approach results in a new model that was trained on data from a different training distribution. As a result, the test data is different from what

Submitted to Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

has been seen from training. Without re-training, it is difficult to determine if the drop in accuracy is due to the distributional shift or if those neurons are actually important. To avoid this issue, we focus on the lottery ticket hypothesis (LTH)[2]. The LTH claims each network contains a sub-network such that went trained in isolation, it matches or exceeds the performance of the original network. If the synergistic neurons are important for classification then re-training these neurons should result in a minimal drop in accuracy. Our contributions are as following :

1. We present, to the best of our knowledge, the first quantifiable metric for multi-neuron feature importance
2. We show the features in the early layers of a DNN are redundant while the last layers contain synergistic features suggesting the network specializes as the depth increase across the layers
3. We provide an alternative perspective to verify importance by showing re-training a sub-network based on the synergistic neurons results in a minimal drop in performance

Impact Knowing which groups of neurons are important can :

- Provide interpretability of the internal decisions in a network to end users[7]
- Explain generalization of DNNs[5]
- Enable compression of networks using pruning[8] and the lottery ticket hypothesis[2]

2 Related Work

Mutual Information Mutual information is commonly used in deep learning as a loss function for training, quantifying importance of individual neurons[9], provide generalization bounds[1] and study the learning dynamics of DNNs using the information bottleneck[19]. The limitation of all these approaches is that they rely on pairwise interactions which ignores the information present in groups of neurons.

Partial Information Decomposition Few works have focused on multivariate information theory for analysing neural networks but they rely on the partial information decomposition which is difficult to scale when using more than three variables[22][17].

3 Methodology

In this work, we use the O-information to identify which combination of n multiple neurons $x_l^1, .., x_l^n$ in each hidden layer l are important or redundant for classification of the networks decision y. Importantly, we consider x_l^n to be a vector before applying the activation function to avoid non-singular values resulting from a ReLU activation function. Both x_l^n and y are evaluated based on the test set. We also show that the synergistic neurons are causing the network to generalize.

3.1 Summary of models and datasets

We use a fully connected neural network architecture consisting of three fully connected hidden layers : fc_1, fc_2, fc_3 and the output layer fc_{out} with 10 neurons. The hidden layers consist of 20 neurons while the output layer consists of 10 neurons. Each hidden layer uses ReLU as its activation function, except for the last layer that uses a softmax function. We train 20 networks on MNIST for 50 epochs each reaching 96 percent accuracy on the training and test set. Training was performed on the MNIST dataset using stochastic gradient descent with a learning rate of 0.01 and a batch size of 32. The test set consists of 10000 images while the training set consists of 60000 images. We did not include any regularization techniques. For each experiment, we use 20 networks trained with a different random seed.

3.2 O-information

Given a collection of random variables $\mathbf{Z} = \{X_1^1, .., X_k^l\}$, we compute the O-information as[14]:

$$\Omega(\mathbf{Z}) = (n-2)H(\mathbf{Z}) + \sum_{j=1}^{n}[H(Z_j) - H(\mathbf{Z}\backslash Z_j)] \tag{1}$$

where H is the entropy, and $\mathbf{Z} \backslash Z_j$ is the complement of Z_j with respect to $\mathbf{Z}$. If $\Omega_l(\mathbf{Z}) > 0$ the system is redundancy-dominated, while if $\Omega_l(\mathbf{Z}) < 0$ it is synergy-dominated. For a detailed description of the O-information, we refer to [14].

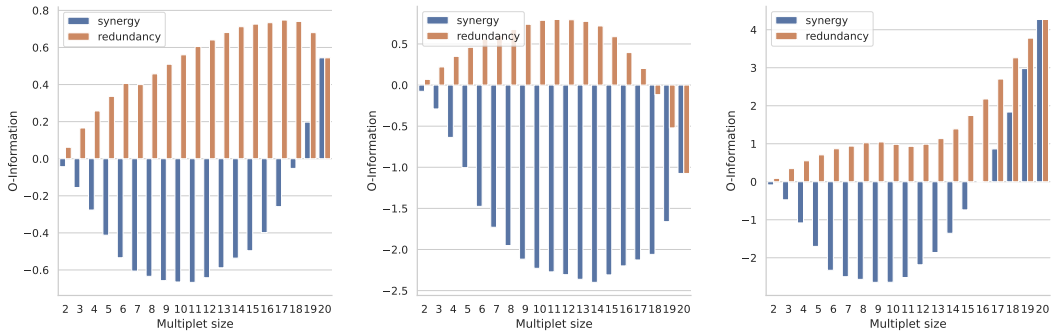
3.3 Layerwise multi-neuron importance using the O-information

In order to use the O-information for feature importance, we need a way to construct random variables based on the features x_l^n and target label y to estimate the joint entropy in equation 1. Most approaches using information-theoretic quantities rely on binning to transform the continuous samples into a discrete value. This way the discrete entropy can be estimated in closed-form. However, binning relies on parameter tuning and is therefore vulnerable to bias. It also relies on large amount of samples to gain an accurate estimate. To avoid these issues, we use the Gaussian Copula approach described in [4]. In summary, this method transforms the marginals to standard Gaussian variables, to which we can apply the computationally efficient closed-form solution of the parametric joint entropy.

4 Experiments

4.1 Layerwise synergy and redundancy using the O-information

In this experiment, we quantify the synergy and redundancy in the layers by measuring the O-information between the features and the predictive class labels. We use optimization to find the optimal combination for each multiplet of size k by performing an exhaustive search for $k = 6$, and larger k values are handled by adding one variable at a time to the best multiplet of $k - 1$ variables for respectively the synergy and redundancy.



(a) First hidden layer (b) Second hidden layer (c) Third hidden layer

Figure 1: Layerwise bar plots of the synergy and redundancy in each layer for multiplets of increasing size

As can be seen in figure 1, the initial layers are dominated by redundant information which seems to suggest the features in the first layer extract local features (i.e. edges) that are commonly shared among the classes. On the other hand, the last layer contains mainly synergistic information which can be explained by the network specializing for the classification decision. Our finding seem to contradict with the results in [9] claiming the last layers contain redundant information. We believe this result empirically validates that pairwise MI cannot efficiently measure the synergy and redundancy of a multivariate systems.

4.2 Evaluating importance through re-training based on the lottery ticket hypothesis

In this experiment, we empirically validate our hypothesis that the synergistic neurons are important for generalize by re-training the network with initial weights based on the these neurons. The remaining neurons are freezed during training. To achieve a fair comparison, we compare our method against three baselines. The first baseline randomly removes the neurons from the network. The second and third is based on the Gaussian Copula formulation of the mutual information (MI) between a single neuron and a discrete target[4]. To take into account the discrete target, the former uses a

mixture distribution, while the later is based on Anova. As in [4], we perform cumulative ablation by ranking the individual neurons according to the highest MI.

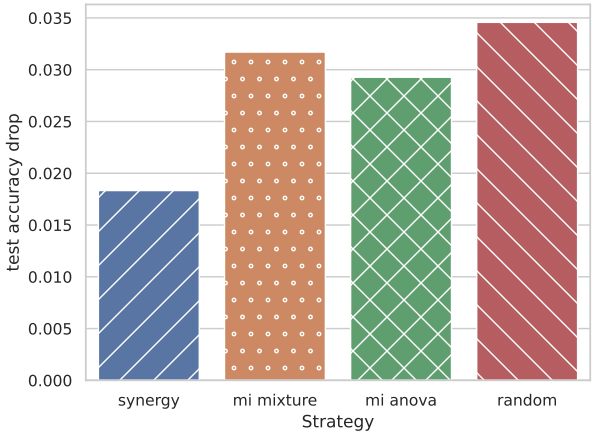


Figure 2: Evaluating importance by training the re-initialized weights of a subnetwork based on the best synergistic multiplets in each layer

As can be seen in figure , re-training the synergistic neurons results nearly zero accuracy drop. This seems to suggest the network is able to identify the important neurons in a network. On the other hand, both MI methods have similar performance then the random strategy. If we can scale our method to provide accurate estimation of the synergy, we believe it can be used to identify optimal sub-networks in a single iteration. This is in contrast to current pruning methods for the LTH which rely on multiple iterations to remove a fixed number of neurons.

5 Limitations

One of the main limitations of this work is the relatively small network and dataset. To increase the diversity of the experiments, we plan on extending this work to larger fully connected networks of 100 neurons on fashionMNIST[21] and VGG16[15] on CIFAR10[6]. This raises two issues. First, we need a way to scale the greedy search to larger multiplets without sacrificing the accuracy of the estimation. Second, we have to use convolutions which require reasoning about feature maps containing many neurons rather then just one. A naive solution for the latter might be to average the feature maps into a scalar value but this ignores the rich information contained in the combinations of neurons. To address both issues, we plan on using evolutionary search strategies based on Jax to enable efficient parallel optimization of the multiplets[16]. Another solution is to decompose the network into a modular representation using clustering as in [3].

6 Conclusion

In this work, to the best of our knowledge, we proposed the first method for quantifying multi-neuron feature importance[11]. We verified importance by re-training a sub-network based on the best synergistic neurons. In future work, we aim to explore the contribution of the features with respect to the individual classes. Another avenue for future research is to consider the training dynamics of DNNs using the information bottleneck principle[19]. Current methods consider the entire layer as input which ignores the patterns of the intermediate combinations of neurons. From a practical perspective, we believe the synergistic neurons can be used as a metric for pruning and as an optimization objective for unsupervised representation learning [20].

References

[1] A. Asadi, E. Abbe, and S. Verdú. Chaining mutual information and tightening generalization bounds. 31.

[2] J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks.

[3] S. Hod, S. Casper, D. Filan, C. Wild, A. Critch, and S. Russell. Detecting modularity in deep neural networks.

[4] R. A. Ince, B. L. Giordano, C. Kayser, G. A. Rousselet, J. Gross, and P. G. Schyns. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. 38(3):1541–1573. doi: 10.1002/hbm.23471. Publisher: Wiley Online Library.

[5] K. Kawaguchi, L. P. Kaelbling, and Y. Bengio. Generalization in deep learning.

[6] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Publisher: Citeseer.

[7] 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. URL `http://arxiv.org/abs/2103.10689`.

[8] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang. Pruning and quantization for deep neural network acceleration: A survey. 461:370–403. Publisher: Elsevier.

[9] K. Liu, R. A. Amjad, and B. Geiger. Understanding individual neuron importance using information theory.

[10] Matthew L. Leavitt, M. Leavitt, Matthew L. Leavitt, and A. S. Morcos. Selectivity considered harmful: evaluating the causal impact of class selectivity in DNNs. . MAG ID: 3009079798.

[11] Matthew L. Leavitt, M. Leavitt, Matthew L. Leavitt, and A. S. Morcos. Towards falsifiable interpretability research. . MAG ID: 3094509718.

[12] R. Meyes, M. Lu, C. W. de Puiseau, and T. Meisen. Ablation studies in artificial neural networks.

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

[14] F. E. Rosas, P. A. Mediano, M. Gastpar, and H. J. Jensen. Quantifying high-order interdependencies via multivariate extensions of the mutual information. 100(3):032305. doi: 10.1103/PhysRevE.100.032305. Publisher: APS.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition.

[16] Y. Tang, Y. Tian, and D. Ha. EvoJAX: Hardware-accelerated neuroevolution.

[17] T. Tax, P. A. Mediano, and M. Shanahan. The partial information decomposition of generative neural network models. 19(9):474. doi: 10.3390/e19090474. Publisher: Multidisciplinary Digital Publishing Institute.

[18] N. Timme, W. Alford, B. Flecker, and J. M. Beggs. Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. 36(2):119–140. doi: 10.1007/s10827-013-0458-4. Publisher: Springer.

[19] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE. doi: 10.1109/ITW.2015.7133169.

[20] M. Vera, P. Piantanida, and L. R. Vega. The role of the information bottleneck in representation learning. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1580–1584. IEEE.

[21] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

[22] S. Yu, K. Wickstrøm, R. Jenssen, and J. C. Príncipe. Understanding convolutional neural networks with information theory: An initial exploration. 32(1):435–442. Publisher: IEEE.