07 Interpreting deep learning from network structure

- Detecting communities in trained layered neural network -

Abstract

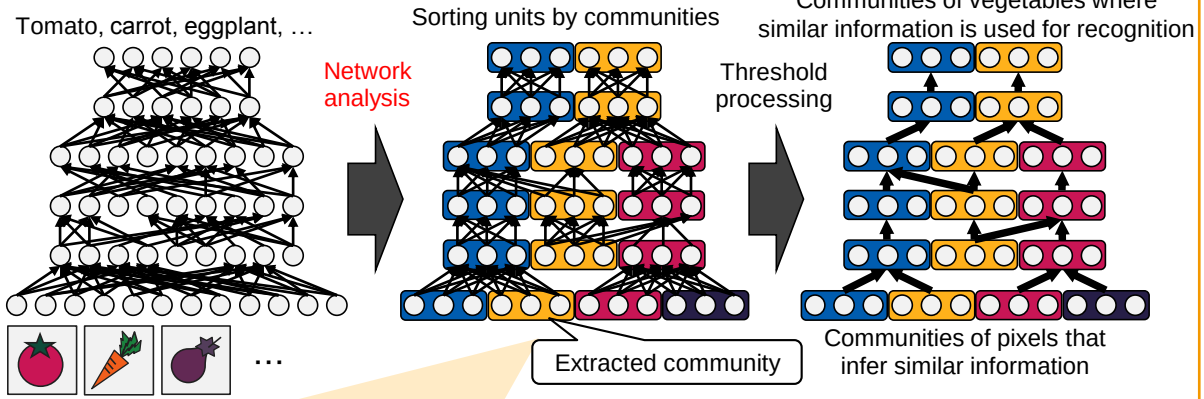
The effectiveness of layered neural networks is widely acknowledged for a wide range of tasks, including image recognition and natural language processing. However, the internal inference mechanism of layered neural networks is black-boxed, and we cannot understand the way in which they represent complex input-output relationships hidden in practical high-dimensional data sets. In this study, we propose a method for extracting the simplified global structure of a trained neural network by employing a network analysis method. Our proposed method first decomposes the units in each layer into communities according to their connection patterns with the adjacent layers, and then unites the multiple connections between a pair of communities into a single bundled connection. We aim for a future in which neural networks can be used over a wider application area than at present, such as automatic driving and medical care, where an explanation of the inference result is strongly required.

1. Training neural network with given data set

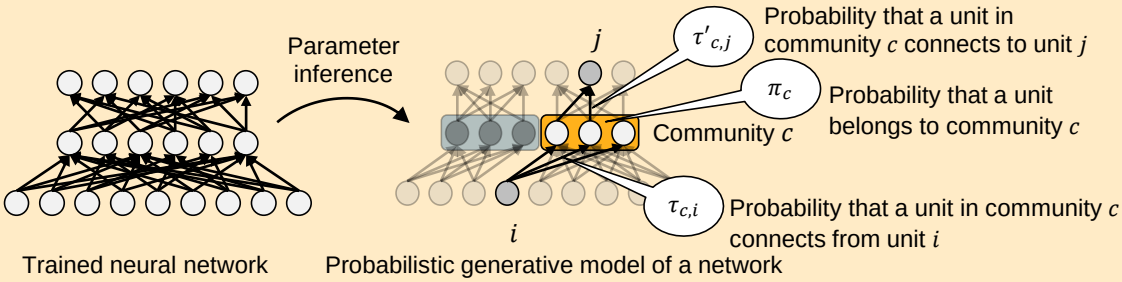
2. Detecting communities of units with similar connection patterns from/to adjacent layers

3. Uniting multiple connections between communities

E.g. Image recognition of vegetables



■A network community detection method is applied to a trained layered neural network
■The simplified global structure is extracted, which helps us better “understand” the network



References

[1] C. Watanabe, K. Hiramatsu, K. Kashino, “Modular representation of layered neural networks,” *Neural Networks*, Vol. 97, pp. 62-73, 2018.
[2] C. Watanabe, K. Hiramatsu, K. Kashino, “Community detection in layered neural networks based on signs of connection weights,” In *Forum on Information Technology*, 2017.
[3] C. Watanabe, K. Hiramatsu, K. Kashino, “Recursive extraction of modular structure from layered neural networks using variational Bayes method,” in *Proc. Discovery Science*, Lecture Notes in Computer Science, Vol. 10558, pp. 207-222, 2017.

Contact

Chihiro Watanabe Recognition Research Group, Media Information Laboratory