Generating Interpretable Prototype Networks by Comprehensive Compression for Multi-Layered Neural Networks

Ryotaro Kamimura

Tokai University, Japan

TOKAI UNIVERSITY

ryotarokami@gmail.com

Japan

IARIA

Ryotaro Kamimura

Tokai University, Japan

ryotarokami@gmail.com

Japan

Ryotaro Kamimura is an emeritus professor of Tokai University in Japan. He has been interested in the interpretation of multi-layered neural networks by new methods with information-theoretic methods

Outline

- **Problems of interpretation**
  - Conventional methods are exclusively focused on a fixed point of view to consider only one aspect of data sets
- **Multiple viewpoints**
  - Multiple viewpoints are needed to consider all possible representations and to extract main characteristics behind complicated ones
  - Neural networks have the potential for the multi-viewpoints
- **Prototype generation**
  - All representations by multiple viewpoints are unified into a prototype network
  - Interpretation is replaced for finding a prototype network too unify all possible representations
- **Application to customer data set**
  - New inputs were considered important by examining the prototype

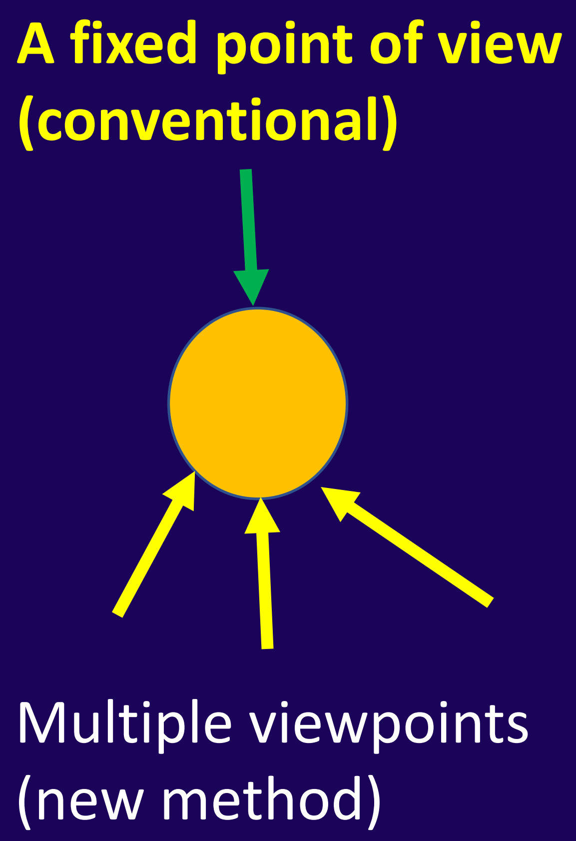
Problems of Interpretation

- **Problem of interpretation**
  - As the complexity of neural networks becomes larger, the interpretation of neural networks has become a very serious problem

- **Necessity of Interpretation**
  - **Research Objective**
    - The objective of neural networks is to understand how neural networks and the corresponding our brain work. Thus, it is necessary to understand the inference mechanism of neural networks
  - **Critical decision making**
    - For application areas with critical decision making such as medical and business applications, interpretation is much more important than improved generalization.
    - It is necessary to interpret and explain the inference mechanism of neural networks

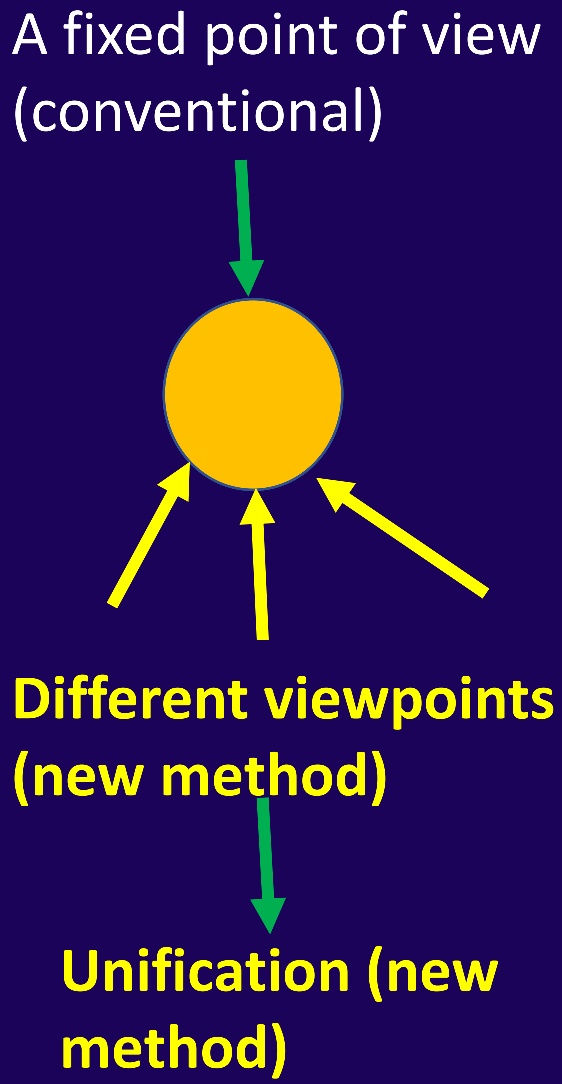
Conventional and Specific Interpretation

- **Specific interpretation**
  - Majority of methods are based on some specific conditions
  - With a set of initial weights, a network is trained to produce an internal representation to be interpreted
- **A fixed viewpoint for a data set**
  - Meaning that most methods aim to explain the responses of neural networks only for some specific input patterns
- **Multiple viewpoints**
  - A data set should be understood by seeing it from multiple and different viewpoints



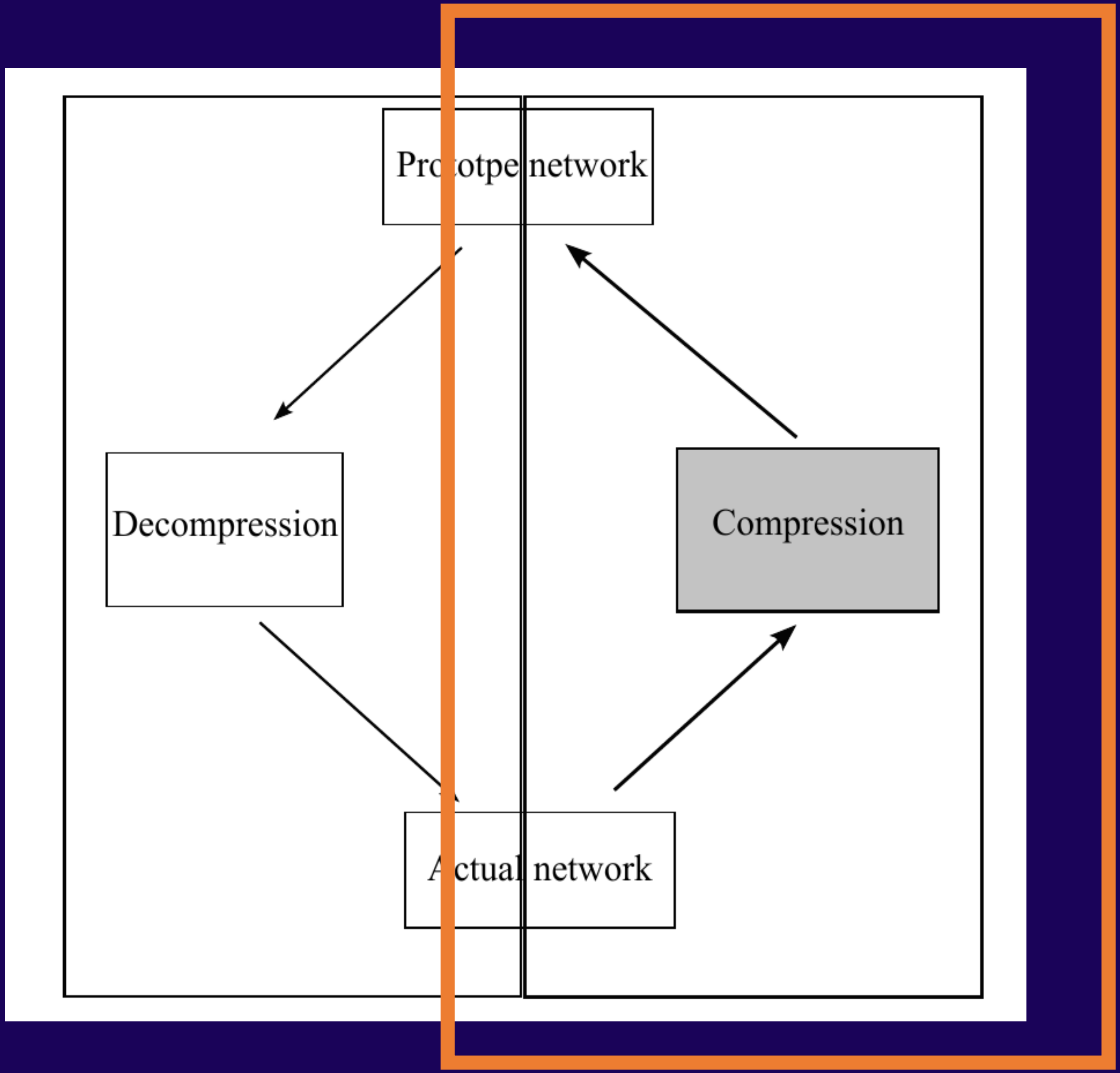
Multi-views by Neural Networks

- **Multiple viewpoints**
  - A data set should be seen by as many different view-points as possible to identify their true characteristics
- **Potentiality of neural networks**
  - The neural network can be used to see a data set from different view-points, because they can produce a number of different internal representations only with different conditions
- **Unification of all representations**
  - For understanding the productive neural network, we should unify representations by different initial conditions and different input patterns



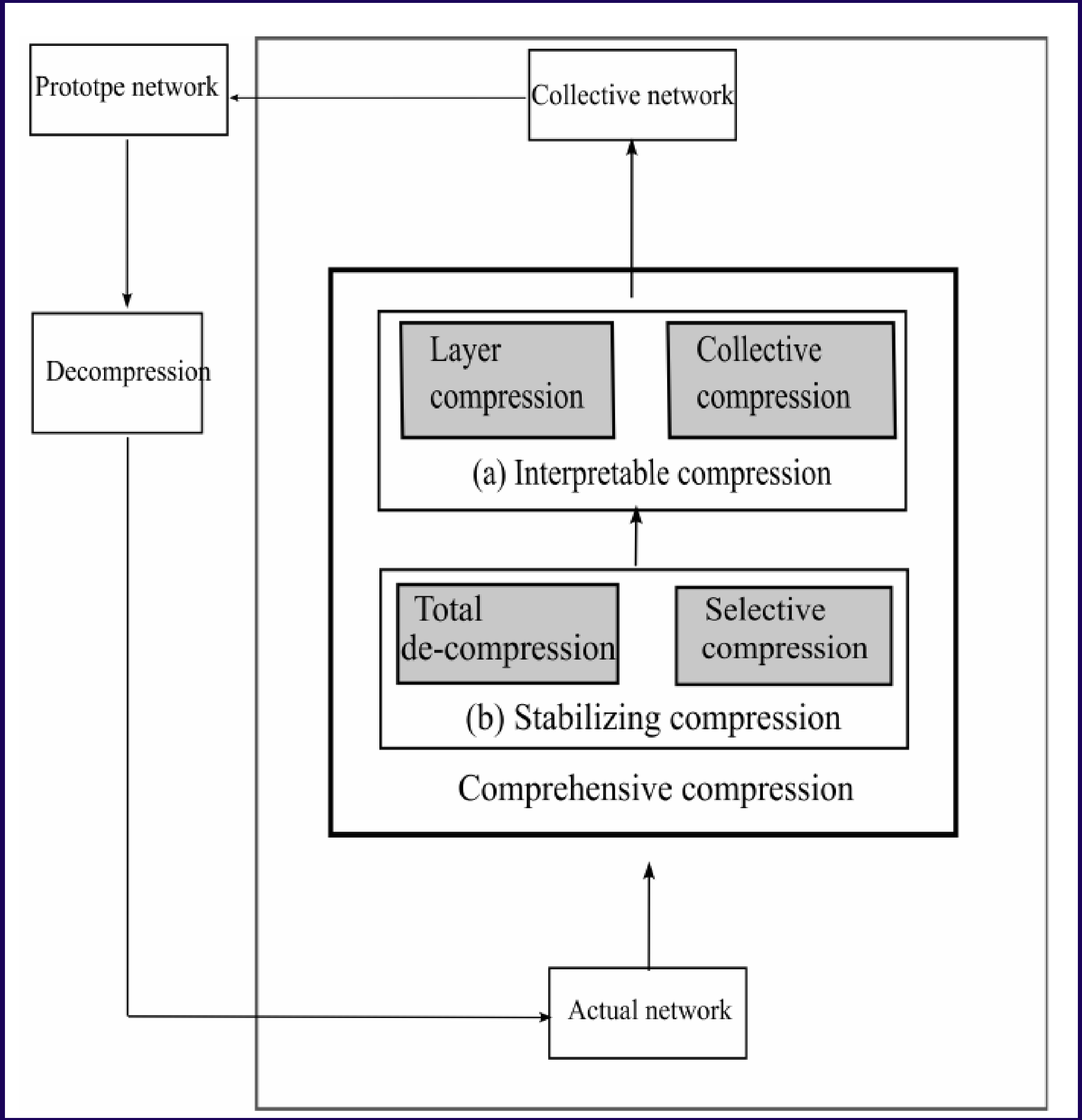
Prototype Generation by Unification

- Prototype generation
  - Many different types of networks are produced to examine their productivity and unified for reaching one prototype
  - Interpretation to be replaced for generating prototype network
  - The prototype network is obtained by compression and decompression
- Compressing
  - A prototype is obtained by compressing a multi-layered neural network into the simplest one
- Decompression
  - This prototype network should be decompressed to have the original multi-layered neural network
- Compression is only treated in this presentation



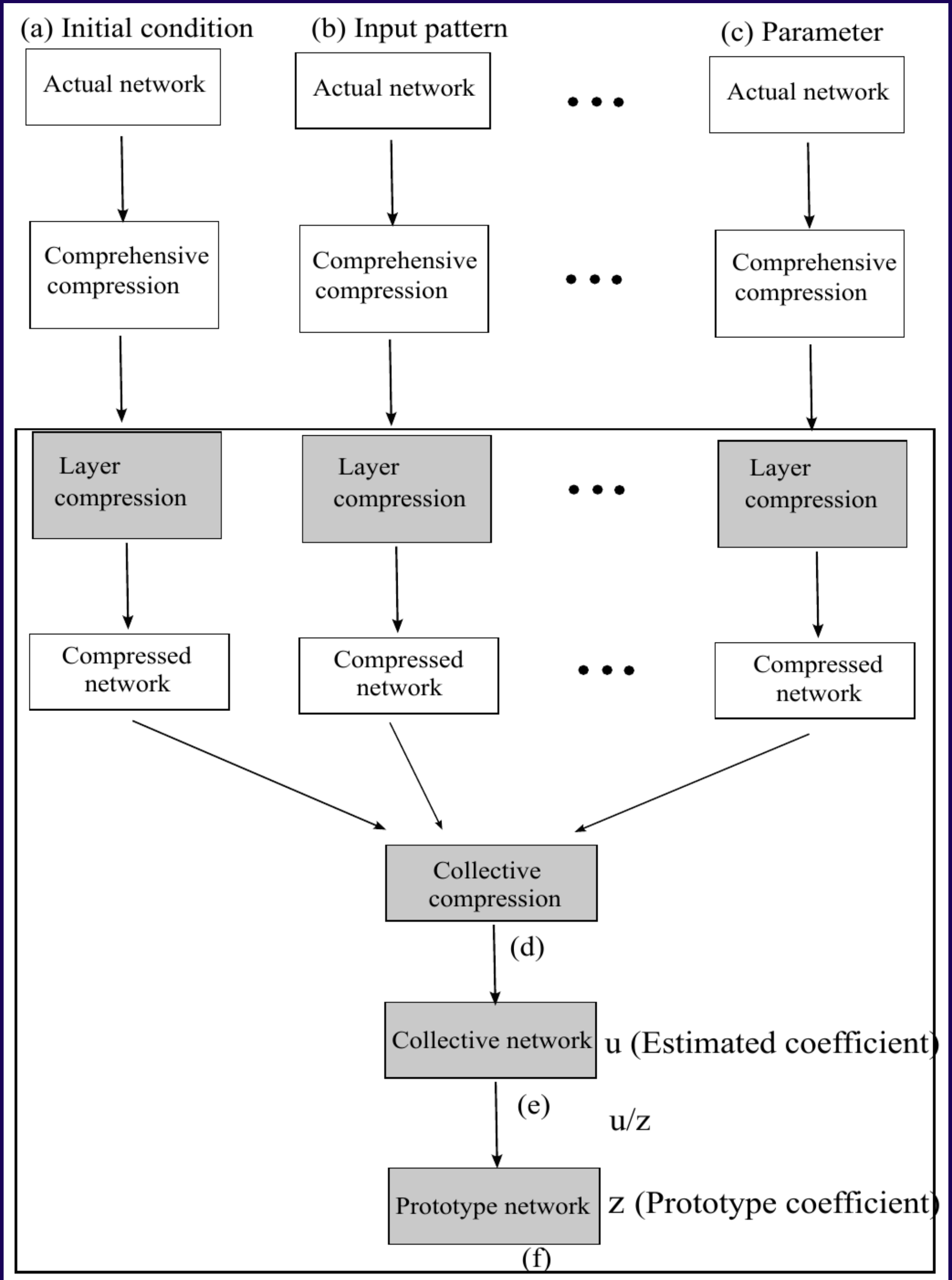
Extended Prototype Generation

- Extended prototype generation
  - In actual experiments, the simple prototype generation is extended to more elaborated models for dealing with the practical problems, in which the interpretable and stabilizing compression are introduced, called “comprehensive compression.”



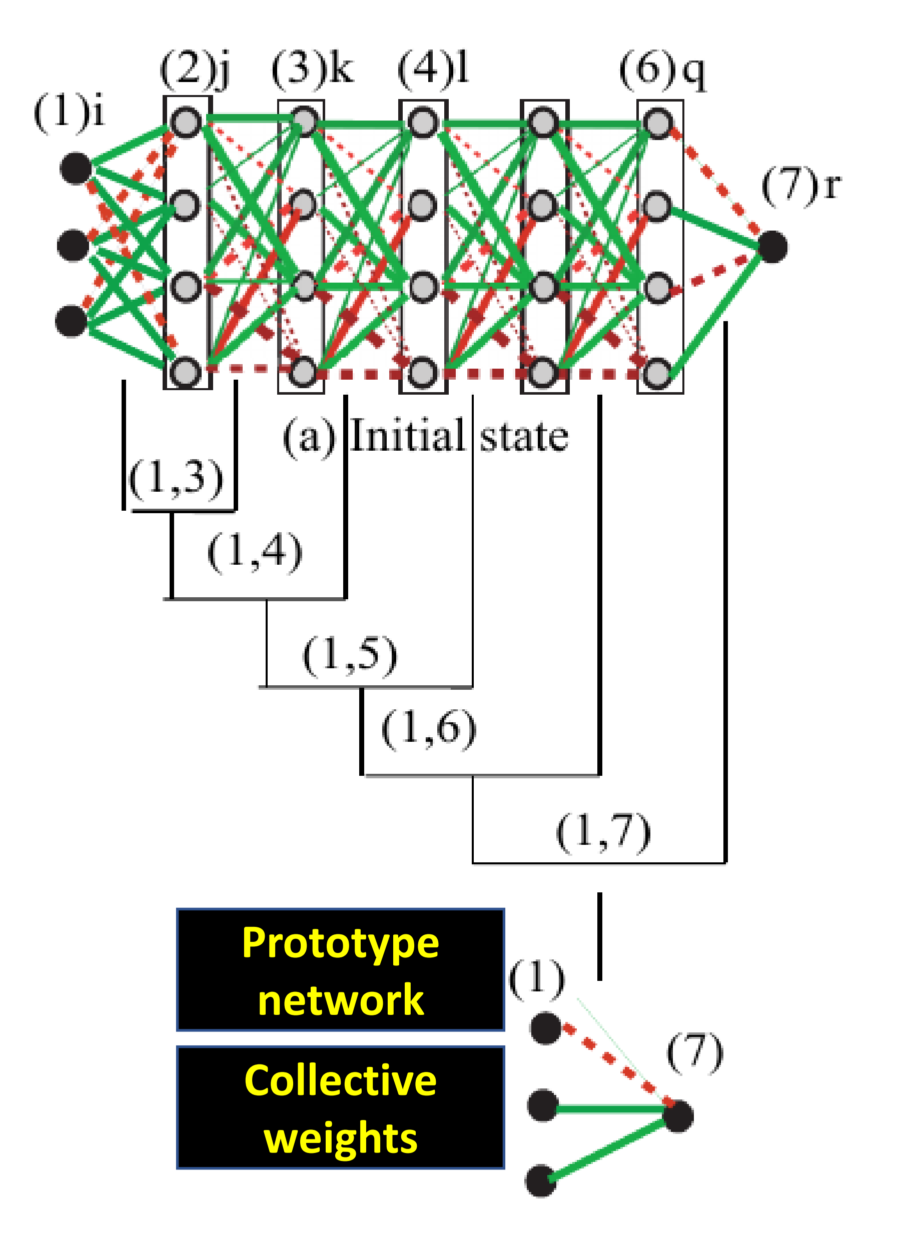
Concept of Prototype Generation

- Multiple networks generation
  - By different initial conditions, input patterns, network configurations, multiple networks are generated
- Layer compressing
  - Multiple networks are compressed into the simplest ones
- Collective compression
  - All compressed networks are unified as collective networks, composed of a prototype network



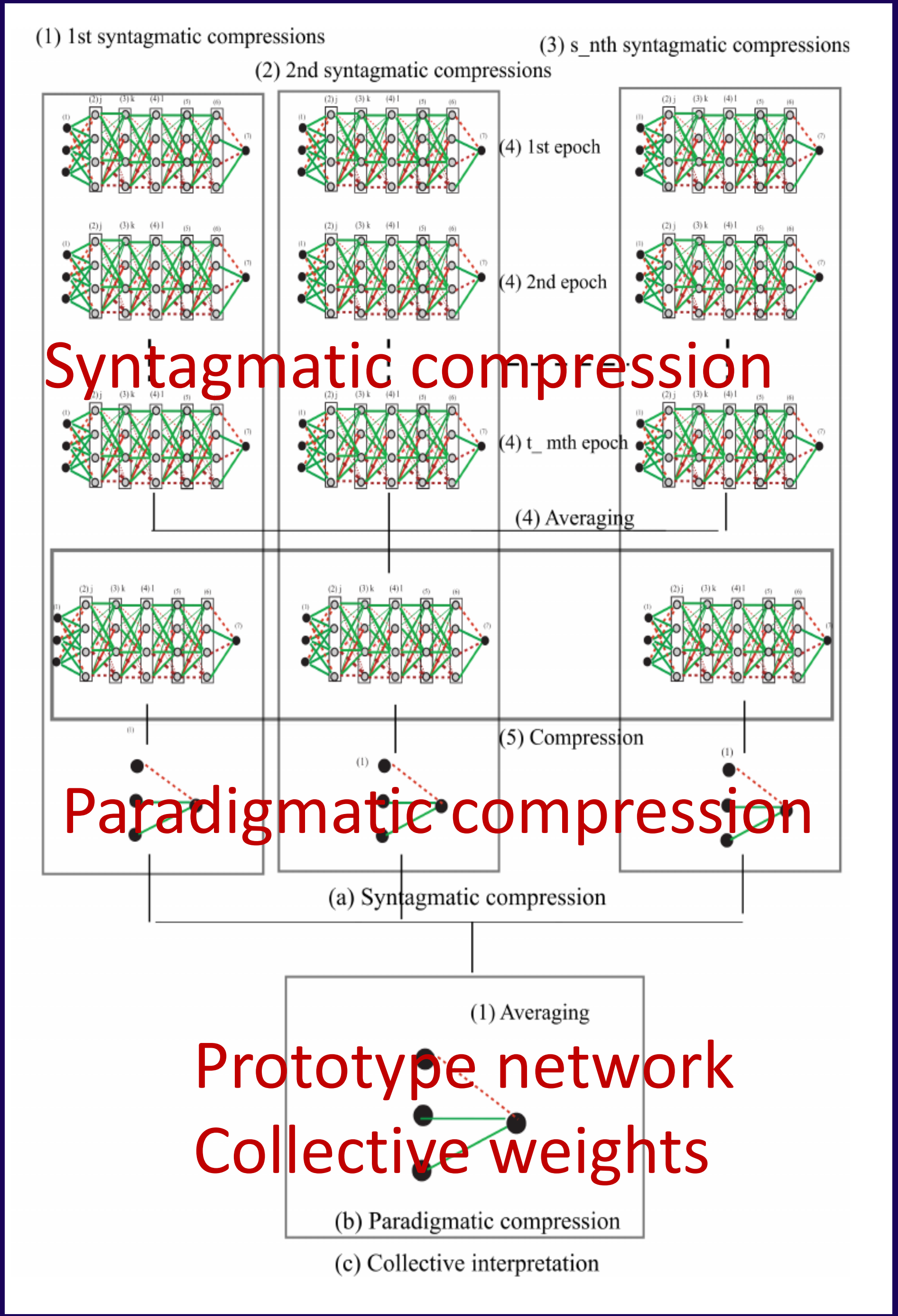
Example of Network Compression

- **Collective weights for a prototype**:
  - All weights in all layers are gradually multiplied and summed to produce collective weights of prototype network
- **No hidden layers**:
  - Any multi-layered neural networks can be reduced to networks without hidden layers
  - To be interpreted as the conventional regression analysis



Example of Actual Prototype Generation

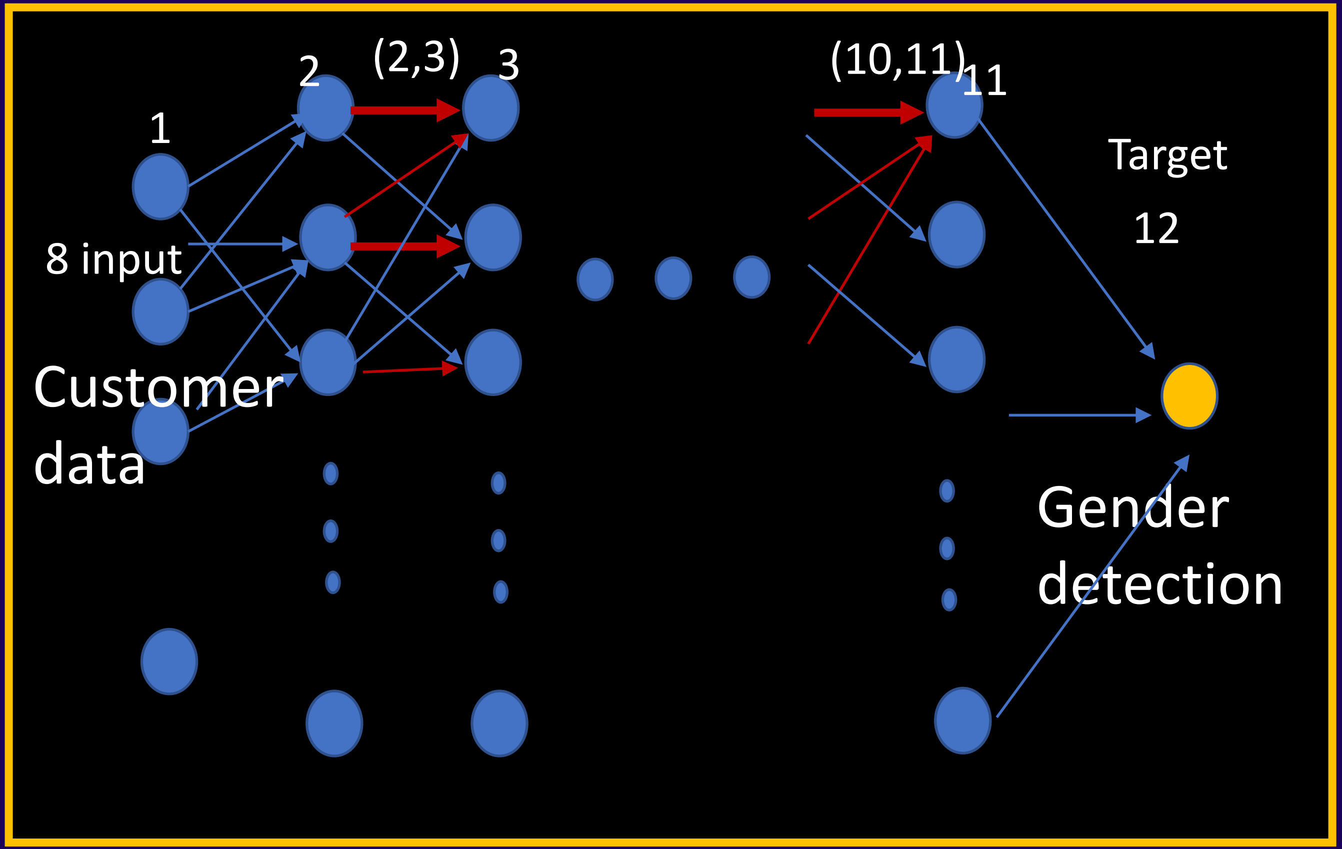
- **Syntagmatic compression**
  - All weights for each learning step are compressed
- **Paradigmatic compression**
  - Averaging all syntagmatically compressed weights to produce the final collective weights
- **As many representations as possible**
  - for comprehensive interpretation



Application to Customer Data Set

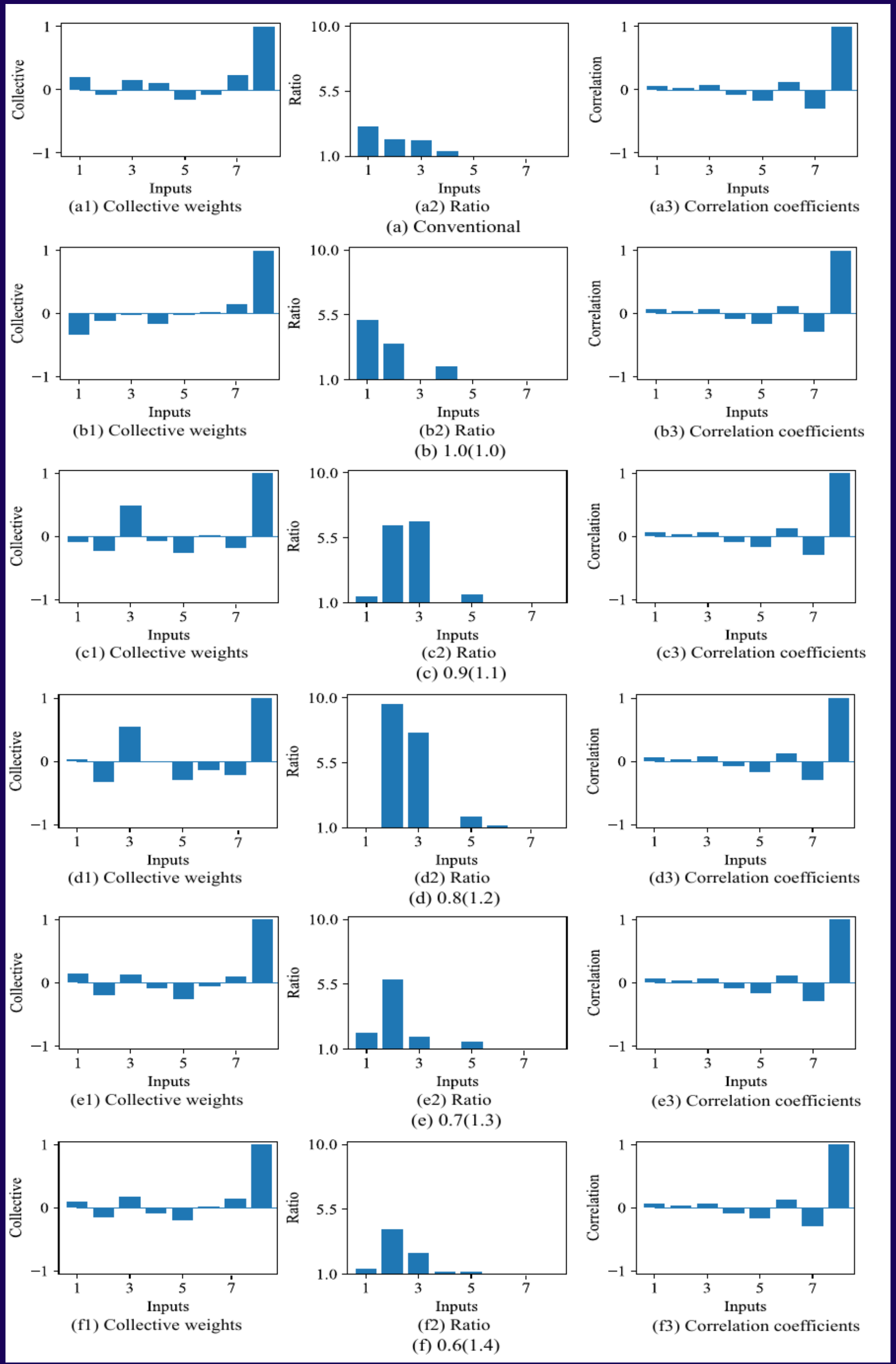
- **Data description**
  - The experiments were performed, using a data set with a sport gymnasium's customers. We tried to discriminate between genders, and to infer the characteristics of female customers.
- **Network and experiment description**
  - 10 hidden layers with 10 neurons
  - 20 runs with different initial conditions and different subsets of data set

10 hidden with 10 neurons



Compressed and Collective Weights by Different Conditions

- Compressed weights with better generalization (left)
- Correlation coefficients between inputs and targets (right)
- Ratio of compressed weights of the corresponding correlation coefficients
- Though the input No.8 had the largest strength, the input No.2 and No.3 may have some strength in terms of the ratio

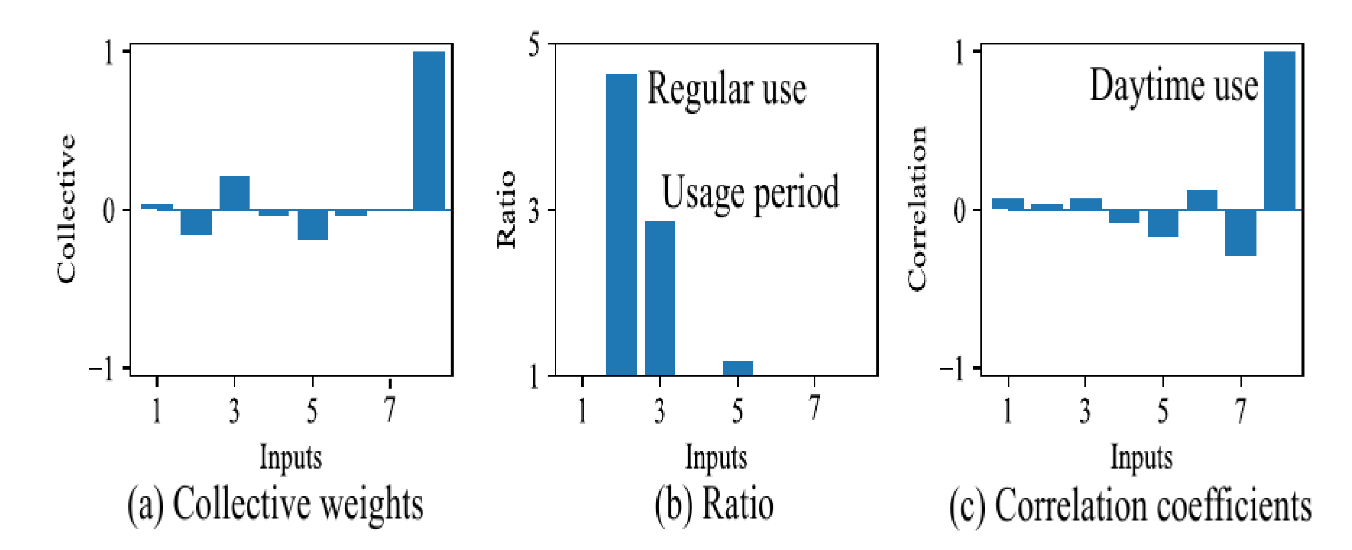


Collective Weights for the Prototype network

- **All compressed weights were averaged to have the final collective weights for the prototype**

- **Network with the largest correlation coefficient**
  - Negative relative weight from the input No.8
  - Daytime use was the largest
  - This sport center was naturally used in the daytime

- **Input No.2 (regular use) and No.3 (long usage period)**
  - Femail clients tended to use it regularly and for a long period



Summary of Experimental Results

- **Highest generalization and correlation coefficients**
  - Compression could produce prototype with the best generalization and the largest correlation coefficient.
  - This means that the prototype network represented the linear relations between inputs and outputs mainly
  - However, some inputs, not considered important, could be emphasized by our method

Methods	Param $\theta(2-\theta)$	Accuracy	Correlation
Compression	1.0(1.0)	0.684	0.807
	0.9(1.1)	**0.692**	**0.885**
	0.8(1.2)	0.691	0.833
	0.7(1.3)	0.690	0.878
	0.6(1.4)	0.690	**0.885**
Conventional		0.683	0.833
Logistic		0.678	0.855
Random		0.613	0.256

Summary

- **Multi-views for a data set**
  - A data should be seen form many different viewpoints
  - Neural networks have the potential to see the data set from many different viewpoints
  - However, the multi-views by neural networks have been considered one of the main shortcomings of neural network
  - Because there have been no methods to unify those multi-viewpoints in neural networks
- **Prototype generation**
  - Neural networks are used to produce many different types of representations
  - Then, compression is used to produce the simplest networks without hidden layers.
  - All those networks were unified into a prototype network
  - Interpretation is replaced for finding a prototype network to unify all possible representations
- **Application to customer data set**
  - Unified prototype networks could show that new inputs were considered important by examining the prototype

Conclusion and Future Directions

- **Decompression**
  - How to decompress a prototype network to the corresponding multi-layered neural network

- **Real interpretation**
  - Application to the large-scale interpretation problems