Graduate School of Fundamental Science and Engineering
Waseda University

博　士　論　文　概　要
Doctoral Dissertation Synopsis

論　文　題　目
Dissertation Title

A Study of Inner Representations in Deep Neural Networks for Comprehending Network Behaviors Using Persistent Homology

ディープニューラルネットワークの挙動解析を目的としたパーシステントホモロジーを用いたネットワーク内部構造の研究

申　請　者
(Applicant Name)
Satoru WATANABE
渡辺　聡

Department of Computer Science and Communications Engineering, Research on Parallel and Distributed Architecture

April, 2022

Introduction: This thesis comprises a series of studies on inner representations in deep neural networks (DNNs) using persistent homology (PH). DNNs have demonstrated remarkable performance in various fields, including image analysis, speech recognition, and text classification. However, the inner representations in DNNs are indecipherable, which makes it difficult to tune DNN models, control their training processes, and interpret their outputs.

This thesis investigates the inner representation of DNNs using topological data analysis (TDA). TDA employs results from geometry and topology, which has provided new insights in various fields such as neuroscience, proteomics, and material science. PH is a prominent method in TDA owing to its three advantages: theoretical foundation, practical computability, and robustness with small perturbations. These advantages are beneficial to investigate DNNs; theoretical foundation and practical computability are fundamental to extracting knowledge from empirical observations, while robustness is indispensable for investigating DNNs involving parameter perturbations. Thus, PH is a prominent method to investigate DNNs for improving the comprehension of network behaviors.

Goals: The inner representation of DNNs constitutes network parameters calibrated with network training. The inner representation retains the knowledge learned through the training and determines the network behaviors. Therefore, the investigation of the inner representation is essential in improving the comprehension of network behaviors. Meanwhile, PH is designed to be applied to simplicial complexes in topological spaces. Due to the difference between network parameters in DNNs and simplicial complexes in topological spaces, PH cannot be straightforwardly applied to the investigation of DNNs. Consequently, this thesis aims to fulfill the following goals.

(G1): Developing a method to investigate the inner representation of DNNs using PH
(G2): Improving the comprehension of network behaviors using PH

DNNs work as knowledge-distilling pipelines. Alternatively, the degree of feature abstraction increases with the depth of DNN layers, where higher-level features are obtained by combining lower-level features. For example, images of cats are incrementally abstracted from pixels to diagonal lines and from diagonal lines to ear shapes. Then, DNNs can detect cats using a combination of high-level features, such as ear and body shapes. In this process, the neurons in the DNNs represent the features. Thus, the combination of the neurons, described by the network parameters, represents the implementation of knowledge retained in DNNs.

PH is a method for computing the topological features of a simplicial complex. One-dimensional PH counts the number of holes in a simplicial complex. The combination of the neurons develops holes in the simplicial complex, where the neurons and the connections between the neurons are considered as vertexes and edges in the simplicial complex, respectively. Further, the one-dimensional PH measures the stability of the holes, which varies depending on the number of vertexes and the distance among the vertexes. Accordingly, PH reveals the combinational effects of multiple neurons

in DNNs, which are difficult to capture without using PH. Thus, the comprehension of network behaviors can be improved through the investigation of the inner representation of DNNs using PH.

Contributions: To achieve the goals, this thesis proposes a construction method for clique complexes on DNNs and analyzes the changes in PH involved in different network parameters of DNNs. Additionally, this thesis proposes two investigation methods of DNNs using PH to improve the comprehension of the inner representation of DNNs.

C1. Proposal of a construction method for clique complexes on DNNs (Chapter 2): This thesis develops a construction method for clique complexes on DNNs through the introduction of two techniques: normalization and propagation. These techniques are inspired by the deep Taylor decomposition method, which has been designed to reveal the influential inputs to the outputs of DNNs. Furthermore, this thesis enhances the method for investigating the inner representation of DNNs and mathematically proves the correctness of the construction method.

Additionally, this thesis formalizes the PH calculation method of the clique complexes constructed on DNNs, considering three types of layers: dense, convolutional, and pooling layers. These layers are prevalent in many DNN applications, such as image analysis, speech recognition, and text classification. The construction of clique complexes and formalization in PH calculation provide a foundation for studying on the inner representation of DNNs using PH.

C2. Analysis on the changes in PH involved in different network parameters of DNNs (Chapter 3): This thesis analyzes the changes in PH involved in different network parameters of DNNs using PH diagrams. PH diagrams enable us to review the number and the stability of holes that appeared in the clique complexes obtained from trained DNNs. This analysis reveals that PH changes correspond to the problem's difficulty for which the DNNs are trained. Thus, this analysis indicates that PH reflects the inner representation of trained DNNs.

To confirm the robustness of the DNNs' investigation using PH, this thesis conducts each experiment 10 times with 30 different settings. This process is carried out using random initial values of network weights, resulting in a total of 300 experiments. These experiments reveal that the results obtained from the investigation using PH are robust with the network's settings and initial weights.

C3. Proposal of an overfitting measurement of DNNs using PH (Chapter 4): Overfitting reduces the generalizability of DNNs. Overfitting is generally detected by comparing the accuracies and losses of the training and validation data, where the validation data are formed from a portion of the training data. However, detection methods are ineffective for pretrained networks distributed without the training data. Thus, this thesis proposes a method to detect the overfitting of DNNs using the trained network weights inspired by the dropout technique. The dropout technique has been employed

to prevent DNNs from overfitting, where the neurons in the DNNs are invalidated randomly during their training. It has been hypothesized that this technique prevents DNNs from overfitting by restraining co-adaptations among neurons. This hypothesis implies that the overfitting of DNNs results from co-adaptations among neurons and can be detected by investigating the inner representation of DNNs. The proposed PH-based overfitting measure (PHOM) method constructs clique complexes on DNNs using the trained network weights. Furthermore, one-dimensional PH investigates co-adaptations among neurons. In addition, PHOM is enhanced to normalized PHOM (NPHOM) to mitigate the fluctuation in PHOM caused by the difference in network structures.

The proposed methods are applied to convolutional neural networks trained for the classification problems on the CIFAR-10, street view house number, Tiny ImageNet, and CIFAR-100 datasets. The experimental results demonstrate that PHOM and NPHOM can indicate the degree of overfitting of DNNs. Therefore, these methods enable us to filter overfitted DNNs without requiring the training data.

C4. Proposal of a network pruning method using PH (Chapter 5): The consumption of enormous computation resources prevents DNNs from operating on small computers such as edge sensors and handheld devices. Network pruning (NP), which removes parameters from trained DNNs, is a prominent method of reducing the resource consumption of DNNs. This thesis proposes a PH-based NP method (PHNP). PH investigates the inner representation of knowledge in DNNs, and PHNP utilizes the investigation in NP to improve the efficiency of pruning. PHNP prunes DNNs in ascending order of magnitudes of the combinational effects among neurons using PH to prevent the deterioration of accuracy. PHNP is compared to a global magnitude pruning method (GMP), which is a common baseline for evaluating pruning methods. The evaluation results reveal that PHNP can prune 95% of the edges from DNNs with 12% higher accuracy than the DNN pruned by the GMP method in our evaluation settings including dataset, network structure, and training process.

The outline of this thesis:

Chapter 1 describes the backgrounds and the goals of this thesis and provides a brief overview of the layout of the thesis.

Chapter 2 introduces a construction method for clique complexes on DNNs.

Chapter 3 analyzes the changes in PH involved in different network parameters of DNNs.

Chapter 4 proposes an overfitting measurement method of DNNs using PH.

Chapter 5 develops a network pruning method of DNNs using PH.

Chapter 6 concludes this thesis and provides viewpoints on future works.

List of research achievements for application of Doctor of Engineering, Waseda University

Full Name： 渡辺　聡　　seal or signature

Date Submitted(yyyy/mm/dd): 2022/6/18

種類別 (By Type)	題名、 発表・発行掲載誌名、 発表・発行年月、 連名者（申請者含む） (theme, journal name, date & year of publication, name of authors inc. yourself)
Int'l Journal (Academic Papers)	○**Satoru Watanabe** and Hayato Yamana, Overfitting Measurement of Convolutional Neural Networks Using Trained Network Weights, International Journal of Data Science and Analytics, pp. 1-18, Springer, 2022 (accepted), doi: 10.1007/s41060-022-00332-1. ○**Satoru Watanabe** and Hayato Yamana, Topological Measurement of Deep Neural Networks Using Persistent Homology. Annals of Mathematics and Artificial Intelligence 90(1), pp. 75-92, Springer, January 2022, doi: 10.1007/s10472-021-09761-3.
Int'lConf. (Lectures)	○**Satoru Watanabe** and Hayato Yamana, Overfitting Measurement of Deep Neural Networks Using No Data, 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), pp. 1-10, October 2021, doi: 10.1109/DSAA53316.2021.9564119. ○**Satoru Watanabe** and Hayato Yamana, Deep Neural Network Pruning Using Persistent Homology, 2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), pp. 153-156, December 2020, doi: 10.1109/AIKE48582.2020.00030. ○**Satoru Watanabe** and Hayato Yamana, Topological Measurement of Deep Neural Networks Using Persistent Homology, 2020 International Symposium on Artificial Intelligence and Mathematics (ISAIM), pp. 1-8, January 2020.