論 文 の 内 容 の 要 旨

論文題目　Analysis of Deep Learning from the Viewpoint of Model Structures
（モデル構造的見地からの深層学習解析）

氏　名　大野　健太

Machine learning (ML) is a technology for having machines recognize, decide, and act intellectually by making full use of knowledge inherent in data. With the increase of data generated in the world, the importance of ML technology is rapidly growing as a way of knowledge discovery from data. ML theory has revealed that it is essential to exploit prior information about data for successful learning. It motivated us to explore methodologies for effectively reflecting biases of specific tasks (known as *inductive bias*) into ML models.

Deep Learning (DL) has become a significant movement in the ML community since the late 2000s. Most practically successful DL models are designed to reflect the inductive bias of problems to their architectural *structures*. However, we would argue that there is a considerable gap between empirical triumph and theoretical understanding of the role of structures in DL. On the one hand, researchers and engineers worldwide deeply analyze their models and tasks and work out novel DL models with unique structures capturing the tasks' inductive bias. However, their justification is made mostly only through the empirical evaluation and is not examined via rigorous theory. On the other hand, many studies have clarified the strength of DL theoretically using various mathematical and statistical tools. However, their focus was mostly on FNN, the most general architecture in DL. They have limitations in understanding the power of practically successful DL models with specific structures.

Filling the gap between practice and theory of DL is a critical issue we must address. In science and engineering, it often happens that theoretical justification lags behind practical usefulness. However, without theory, practitioners must rely on the tacit knowledge of domain experts. This problem is particularly critical for DL because DL is notoriously hard for tuning hyperparameters. Theory clarifies the essential enablers and limiting factors of the technology. It provides us trustful guidance on how to tame the technology and develop it further. The rigorous theory also makes the technology trustworthy because it enables us to identify its potential pitfalls; even domain experts are unaware. Therefore, we believe that the theoretical understanding of models' structures contributes to spreading the DL technology to the real-world.

The current situation mentioned above motivates us to pose the following question in this dissertation: How do structures in DL models affect their characteristics? Among others, we focus on the following three structures because of their practical importance and lack of theoretical understanding. Namely, we elucidate skip connections in Convolutional Neural Networks (CNNs), node aggregation operation of Graph Neural Networks (GNNs), and skip connections in multi-scale GNNs.

Residual Networks. Training of deep CNNs has several difficulties, such as vanishing and exploding gradient problems. These problems limited the depth of classical CNNs to at most twenty layers. Residual Network (ResNet) was designed to solve the performance degradation of deep CNNs. ResNet consisted of subnetworks representing functions of the form id + F and concatenated them in series. Here, id is the identity function and F is a dense (and typically shallow) subnetwork called a *residual block*, or *resblock*. ResNet employed a skip connection, which bypassed the network F to realize the identity function. This architectural trick enabled us to learn ResNets with more than 100 layers. Inspired by the success of ResNet, many CNN models with skip connections have been proposed.

Many studies have focused on understanding the success of ResNet-type CNNs theoretically and explained the practical effectiveness of skip connections. However, to the best of our knowledge, few of them have given explanations from the viewpoint of statistical learning theory. Our first problem is why skip connections promote the predictive performance of CNNs. In Chapter 3, we answer this question from the viewpoint of *sparsity*. We introduce a special type of sparse structures, a *block-sparse* structure. We hypothesize that the architecture of ResNet inherits the block-sparse structure and that it promotes theoretical superiority of ResNet-type CNNs.

Graph Neural Networks and Over-smoothing. Graphs are a universal data structure representing relationships between entities. Inspired by the success of DL, there has been a movement to apply DL models to structured data that is more complicated than grid-like data. Applying ML models to graphs has several difficulties, such as node order invariance and inhomogeneous neighborhood structures. Graph Neural Networks (GNNs) are a collective term of DL models for graph-structured data that solve these problems. GNNs have benefits in common with classical DL models such as FNNs and CNNs, such as end-to-end differentiability. GNNs have been empirically successful for graph data analysis in various fields such as biochemistry, computer vision, and knowledge graph analysis.

Message Passing Neural Network (MPNN) is a subtype of GNNs. Its architecture consists of interleaving two types of operations: node *aggregation* and node *update* operations. Given representations for each node on a graph, each node collects representations of its neighboring nodes in the aggregation operation. Then, each node revews its representation using the collected information in the update operation. MPNN-type GNNs are popular due to its empirical performance and implementation simplicity. Among others, Graph Convolution Networks (GCN) is the most popular MPNN-type GNNs and has been applied in broad fields.

Node aggregation operations in MPNN-type GNNs resemble convolution operations of CNNs. However, as opposed to CNNs, some MPNN-type GNNs reportedly cannot perform well practically when they go deep (for example, more than ten layers). Several studies attributed this performance degradation to the degeneration of node representations, a phenomenon they coined *over-smoothing*.

We can think of the over-smoothing phenomenon as a side effect of the architectural design of GNNs. When we apply the aggregation operation, it smoothens node representations and makes them close to each other. Finally, node representations go indistinguishable when we repeat it many times. That is, node representations are over-smoothened.

For *linear* GNNs, i.e., GNNs whose activation functions are the identity functions, several studies have proven that this intuition is correct. That is, linear GNNs provably suffer from the over-smoothing problem. In classical DL models such as FNNs and CNNs, non-linearity introduced by activation functions can contribute to the theoretical and empirical performance. Given that, one may hope that non-linearity can be beneficial to GNNs. However, to the best of our knowledge, there has been little theoretical explanation for the over-smoothing of non-linear GNNs. Our second problem is whether non-linearity can mitigate the over-smoothing of GNNs. In Chapter 4, we answer this question negatively in a specific situation. We interpret the forward propagation of a GNN as a discrete-time dynamical system and characterize the over-smoothing as the convergence to an invariance space of the dynamics that is "information-less" for node prediction tasks. We prove that the convergence rate to the invariant space for ReLU GCNs is the same as that for their linear counterparts, implying that ReLU GCNs are as vulnerable to over-smoothing as linear GCNs.

GNNs with Skip Connections. Over-smoothened GNNs are not appropriate for node prediction tasks because the loss of their expressive power can cause under-fitting. As we explained previously, skip connections enable us to stack many layers in the case of CNNs. Therefore, it is natural to think that we may solve the over-smoothing problem by introducing skip connections to GNNs. Several studies have shown that skip connections are practically useful to some extent. However, to the best of our knowledge, it has not been known for the theoretical explanations. Our third problem is what is the role of skip connections in GNNs. In Chapter 5, we focus on *multi-scale* GNNs to answer this question. The idea of multi-scale GNNs is to exploit the inductive bias of tasks that subgraphs of various scales are beneficial for predicting nodes' properties. The most standard approach for realizing it is to connect intermediate outputs of a model to its final output directly using skip connections.

Our analysis key is to interpret a multi-scale GNN as an ensemble of single-scale GNNs. This interpretation enables us to apply the boosting theory. We derive optimization and generalization bounds for multi-scale GNNs in node prediction tasks. These bounds give us insights when multi-scale structures induced by skip connections effectively counter the over-smoothing problem.

Organization of Dissertation. Chapter 1 is the introduction, which overviews the dissertation. In Chapter 2, we introduce the backgrounds of the following chapters. First, we explain the architecture of ResNet to explain its design intent. Then, we provide representative variants of ResNet-type CNNs. Regarding the basics of GNNs, we first overview ML tasks on graphs, particularly the classification of

graph types and task types, to clarify problem settings on which we focus. Among others, we mainly focus on node prediction tasks on a graph and formulate it as semi-supervised learning problems (more specifically, transductive learning problems). After that, we showcase representative models of GNNs. Finally, we briefly explain statistical learning theories of supervised learning problems for ResNet-type CNNs and transductive learning problems for GNNs.

The goal of Chapter 3 is to clarify the role of skip connections in ResNet-type CNNs. Toward this goal, we analyze the approximation and estimation errors in non-parametric regression problems using ResNet-type CNNs. The analysis key is FNNs with special sparse structures, which we coin block-sparse FNNs, that have statistically favorable properties. We associate the expressive power of ResNet-type CNNs with that of block-sparse FNNs. By doing so, we derive the approximation and estimation bounds of ResNet-type CNNs, which inherit the favorable properties of block-sparse FNNs. We show that when the true function is Hölder, ResNet-type CNNs can achieve the minimax optimality (ignoring log factors) without unrealistic sparse constraints. This optimality has not been known for FNNs. We hypothesize from the result of this chapter that the skip connections of ResNet-type CNNs implicitly induce special sparse structures, which contribute to their performance.

In Chapter 4, we analyze the expressive power of ReLU GCNs in node classification tasks to explain the over-soothing problem when we introduce non-linearity to GNNs. Our theory associates the representation power with underlying graph spectra. The key idea is to interpret the ReLU function as a projection on the cone. This fact works well with the underlying graph Laplacian. We prove that the convergence rate of ReLU GCNs to the invariant space that is "information-less" for node prediction tasks is the same as that of the corresponding linear GCNs. From this result, we suggest that in the case of ReLU GCNs, non-linearity does not help mitigate the over-smoothing problem.

In Chapter 5, we study the optimization and generalization performance of multi-scale GNNs to investigate how skip connections affect the performance of GNNs. The key to the analysis is that by interpreting a multi-scale GNN as an ensemble of single-scale GNNs, we can train it using boosting algorithms. We assume a weak-learning-type condition (w.l.c.), a standard assumption in the boosting theory. We derive an AdaBoost-like optimization and generalization performance bounds for multi-scale GNNs trained using boosting algorithms under this assumption. When the model satisfies w.l.c. with "good" parameters, the performance of multi-scale GNNs is monotonically increasing as the depth (the number of node aggregations) increases. It contrasts with the over-smoothing problem of single-scale GNNs we observe in Chapter 4. Our results advance the understanding of multi-scale structures induced by skip connections to counter the over-smoothing problems.

In the final chapter, we summarize each chapter and discuss future directions. We point out that analyzing continuous models, that is, the continuity limit of discrete models, is a promising approach for highlighting structures' characteristics. We conclude this dissertation by emphasizing the importance of unified theory for various architectures to understand the role of structures in DL models.