Theory and Practice of Science and Technology
2025, VOL. 6, NO. 5, 118-122
DOI: 10.47297/taposatWSP2633-456923.20250605

From full fine-tuning to parameter-efficient adaptation: a taxonomy of deep transfer learning strategies

Wenkai Zhang

Electronic Information Science and Technology, Nanjing Agricultural University, Nanjing, Jiangsu, 211800, China

ABSTRACT
This paper focuses on the field of deep transfer learning, aiming to study the dual dilemma of traditional full-parameter fine-tuning in terms computing power demand and overfitting risk. By establishing a three-level taxonomy of parameter-efficient transfer learning, this paper analyzes in detail the methods such as parameter addition strategy parameter reparameterization, and parameter subset selection. It also refines the general principles such as minimal disturbance of feature space, task-knowledge decoupling, and incremental knowledge injection Simultaneously, it points out the current challenges of lack of theoretical interpretability (black-box nature) and weak cross-architecture generalization ability, and proposes future directions such as-explanation-based interpretable adaption and neural architecture search (NAS)-driven automatic adaption, aiming to construct a more systematic and efficient theoretical framework for deep transfer learning.

KEYWORDS
Deep transfer learning; Parameter-efficient adaption; Taxonomy; Lightweight adaption

1 Introduction

Deep transfer learning has emerged as a powerful technique in the field of artificial intelligence, enabling the reuse of knowledge learned from one or more source tasks to boost performance on related target tasks. While full-parameter fine-tuning has shown effectiveness in certain cases, it presents notable drawbacks. Its high computational demand makes challenging to implement in many resource-constrained scenarios, and the risk of overfitting often reduces the model's generalization ability. [1-2]

To address these issues parameter-efficient adaptation methods have been proposed. These methods aim to adapt pre-trained models to new tasks with minimal adjustments to the model parameters, thereby reducing computational costs and mitigatingfitting. However, existing research lacks a comprehensive and systematic taxonomy, making it difficult to compare and choose appropriate methods across different applications.

This paper aims to fill this gap by a three-level taxonomy for parameter-efficient transfer learning. We will analyze a range of representative methods, distill their core principles, and explore the challenges and future directions this field.

2 Taxonomy Design

2.1 Overall Framework

Parameter-efficient transfer learning can be broadly categorized into three main classes: parameter-addition strategies, reparameterization, and parameter subset selection. To facilitate understanding, we can refer to a tree-like diagram that visually represents the taxonomy of parameter-efficient transfer learning, as in Figure.

Parameter addition strategy

This category involves adding new parameters to pre-trained models. Two representative methods are adapters and prompt learning.

Adapter

Proposed by Houlsby et al., an adapter is a small neural network module added to pre-trained models. The output of the adapter-enhanced layer calculated as where is the original layer's output and denotes the adapter function. This method is particularly suitable for multi-task continuous learning scenarios. By adding an adapter, the model learn task-specific information without changing the original pre-trained parameters, thereby reducing the risk of catastrophic forgetting. [3,4].

Prompt learning

Prompt is a technique that modifies the model input by adding task-related prompts. It enables pre-trained models to perform new tasks without the need for large parameter updates. The prompt be a text sequence or other data representation that guides the model to predict. This method is flexible and can be widely applied to various natural

* **Corresponding Author:** 476196943@qq.com

language processing tasks. [5,6]

Parameter Re-parameterization

The parameter re-parameterization methods transform the original parameters of the model in a more efficient. Low-rank decomposition and sparse mask are typical examples.

Low-rank decomposition LoRA

LoRA approximates weight updates through low-rank matrix approximation. In conventional parameter updates, it is usually the weight matrix W that is directly updated, which requires a large amount of computational resources and storage space when the model size is large. However, theRA approach introduces two low-rank matrices A and B, and approximates the update amount of the weight matrix through their product. Since A and B are low-rank matrices, have a relatively small number of parameters, thus greatly reducing the parameter scale that needs to be updated.

Taking a weight matrix W with a dimension as an example, if W updated directly, then parameters need to be updated. While using the LoRA method, assuming that A is of dimension and B is of dimension , where , then the number parameters to be updated is only , which is much less than the number of parameters to be updated when the weight matrix is updated directly. In this way, when fine-tuning language models, the calculation efficiency can be significantly improved and the computational cost can be reduced under the premise of ensuring the performance of the model. [7,8]

Sparse

The core idea of sparse mask is to introduce a binary mask matrix to screen and reconstruct the parameters of the model. Suppose the original weight matrix is W, and the mask matrix introduced is M, and the elements in M can only take two values: 0 and 1. The weight matrix after sparse mask processing can be expressed as , denotes the element-wise multiplication. In this way, when the element in the mask matrix is 0, the corresponding original weight parameter is "masked" in the calculation process and does not participate in the calculation and update; only when the element in the mask matrix is 1, the original weight parameter will take effect.

In practical applications, sparse can be learned and optimized in a variety of ways. A common method is to incorporate the learning of the mask matrix into the training process of the model and update it with the other of the model through the backpropagation algorithm. In the early stage of training, the mask matrix can be relatively dense, and as the training proceeds, the proportion of 0 the mask matrix is gradually increased, making the model parameters sparser. This approach can significantly reduce the number of parameters and computational power of the model without sacrificing too much model performance.

Parameter Subset Selection

Parameter subset selection methods aim to select a subset of the model's parameters to update instead of updating all parameters. Layer selection neuron selection are common implementations.

Layer Selection: In the layer selection approach, specific layers in the model are selected for parameter updates based on the task requirements. For example, some extreme low-resource scenarios, only the last few layers of the model are updated because these layers are closer to the task output and can quickly adapt to new tasks, while avoiding modifications to the general features learned in earlier layers. [9,10]

Neuron Selection: Neuron selection, on the other hand, involves selecting and updating only neurons most relevant to the new task. This approach can further reduce the computational load and the number of parameter updates, especially in applications where computational resources are extremely sensitive. [11,12]

2.2 Methodological Details

To gain an in-depth understanding of the technical essence and application boundaries of the parameter-efficient transfer learning strategy, this will systematically analyze the three core classification methods. By deconstructing their mathematical expressions, applicable scenarios, advantages and limitations, and case studies, a comprehensive methodological knowledge map will be. At the same time, combined with the structured comparison in Table 1, a horizontal comparison of the key characteristics of each method will be conducted from the dimensions of parameter update, computational complexity, and task adaptability, with the aim of providing a clear technical selection reference framework for researchers.

2.3 Extraction of Common Principles

Through the study of various efficient transfer learning methods for different parameters, we distilled the following common principles:

Principle 1: Minimization of Feature Space Perturbation

To ensure that the model does not cause excessive damage to learned feature representation when adapting to new tasks, it is necessary to constrain the model's parameter updates to satisfy the Lipschitz condition. The Lipschitz condition is essentially a constraint on the of change of a function, and in the context of deep transfer learning, it requires that the

model's output change is bounded with respect to the input change before and after parameter.

Category	Representation Method	Mathematical Expression	Applicable Scenarios
Parameter Addition	Adapter	$(h' = h + f_{adapter}(h))$	Scenarios of multi - task continuous learning where pre - trained knowledge needs to be retained
Parameter Addition	Prompt Learning	$(x' = [p; x])$	Natural language processing, few - shot learning tasks
Parameter Reparameterization	Low - Rank Adaptation	$(\Delta W = A \cdot B^T)$	Fine - tuning of large language models in scenarios sensitive to computational resources
Parameter Reparameterization	Sparse Masking	$(W' = W \odot M)$	Model compression, inference acceleration scenarios
Parameter Subset Selection	Layer Selection	Selectively update parameters of specific layers (such as the last k layers)	Scenarios in resource - constrained environments where quick adaptation to new tasks is required
Parameter Subset Selection	Neuron Selection	Update \(\theta_{selected}\) based on neuron importance screening	Ultra - lightweight model deployment in scenarios with strict limitations on the number of parameters

Mathematically, a function satisfies the Lipschitz condition if there exists a constant K such that for any and , the inequality holds, where K is called the Lipschitz constant In deep neural networks, when we update the model's parameters, the mapping relationship between the model's input and output changes. If not constrained, excessive parameter updates may cause changes in the mapping relationship of the model in the feature space, destroying the effective feature representation originally learned in the source task.

Principle 2: Decoupling Task Knowledge (Orthogonal Projection Theory)

Orthogonal projection theory provides an effective method to separate these two types of knowledge. Geometrically, we can view the's parameter space as a vector space, where the general knowledge learned by the pre-trained model corresponds to a subspace of this space, and the specific knowledge of the target task be regarded as another subspace. By using orthogonal projection operations, we can restrict the model's parameter updates on the target task to the subspace related to the target task without affecting parameters in the subspace of general knowledge of the pre-trained model.

Specifically, assuming the parameter vector of the model is , we can achieve decoupling by calculating the of the parameter vector in the subspace of general knowledge and the subspace of the target task. Let the projection matrix of the general knowledge subspace be , and the projection matrix of the task subspace be , then the parameter vector can be decomposed into , where represents the parameters of the general knowledge part, and represents the parameters of the specific knowledge part of the task. During the model training process, we can update and separately, so that the model can learn the specific knowledge of the target task without destroying the original general knowledge. For example in some multi-modal transfer learning tasks, orthogonal projection theory can be used to decouple and learn task knowledge from image modality and text modality separately, improving the performance of the model tasks of different modalities. [15,16]

Principle 3: Incremental Knowledge Injection (Dynamic Gating Mechanism)

The dynamic gating mechanism is employed to achieve knowledge injection. The gating mechanism allows new knowledge to be selectively injected into the model based on the needs of the task, avoiding the overfitting issue that can arise from large- parameter updates all at once. The core idea of the dynamic gating mechanism is to introduce a learnable gating unit to control the flow of information. The gating unit typically composed of a neural network module, and its input can be the intermediate layer output of the model, task-related feature vectors, or other task-related information. The output the gating unit is a weight value between 0 and 1, used to measure whether the new knowledge information should be injected into the model. For example, in an RNN the gating unit can determine whether to fuse the information of the current input with the information in the previous hidden state at each time step.

3 Current Challenges

3.1 Lack of Theoretical Interpretability (Black-box Nature)

Most current-efficient transfer learning methods suffer from the issue of insufficient theoretical interpretability, exhibiting black-box characteristics. Although these methods achieve decent performance in practical applications, it is challenging to explain from a theoretical perspective why they work effectively and the underlying reasons for their performance variations across different scenarios. For instance, it is difficult to accurately analyze how the parameter

update process affects overall behavior and performance of the model for some parameter reparameterization methods based on complex neural network structures [19,20].

3.2 Weak Cross-itecture Generalization Ability

Different neural network architectures have different feature representation capabilities and learning mechanisms. Current parameter-efficient transfer learning methods have weak cross-architecture generalization capabilities., as a classic neural network architecture, extracts local features of data such as images through convolutional layers and pooling layers and reduces the number of model parameters using the weight sharing mechanism. In, ViT is based on the Transformer architecture, dividing images into multiple patches and mapping them into a sequence of vectors, and then capturing global features using the self-attention mechanism. two architectures have significant differences in feature extraction and representation. Some parameter subset selection methods that perform well on CNN architectures may not fully play their advantages when directly applied to ViT, or even lead to a decrease in performance. This is because different architectures have different ways of utilizing parameters and different feature extraction patterns. For example, in CNN, layer selection methods usually based on the hierarchical structure of convolutional layers and the spatial information of feature maps. However, in ViT, due to the existence of its self-attention mechanism, the way interact and propagate is different from CNN, and the original layer selection criteria may no longer apply. Therefore, existing parameter-efficient transfer learning methods struggle to adapt effectively across different architectures, their application in more extensive scenarios.

4 Future Directions

4.1 Interpretable Adaptation Based on Causal Inference

To the lack of theoretical interpretability, future work can explore the use of causal inference methods to achieve interpretable and parameter-efficient adaptation. By establishing causal models and analyzing the causal relationship parameter updates and model outputs, we can gain a clearer understanding of the model's decision-making mechanism during the transfer learning process. For instance, using causal diagrams to analyze causal impact of different parameter addition strategies on the model's prediction results can provide theoretical support for method improvement and optimization. For example, in a deep transfer learning task based on image, we can construct a causal diagram that includes nodes such as the image's feature parameters, the model's weight parameters, and the task's class labels. By using inference algorithms, such as methods based on structural causal models, we can compute the causal effects of different parameter updates on the model's classification accuracy. In this way, we can know which parameter adjustments truly help the model learn the target task and which may be unnecessary or even have negative effects. Based on these analysis results, we can further optimize the parameter- transfer learning methods to make them more interpretable.

4.2 Automated Adaptation Driven by Neural Architecture Search

To enhance cross-architecture generalization, neural architecture (NAS) technology can play a significant role. Through NAS, it is possible to automatically search for parameter-efficient transfer learning strategies suitable for different architectures. For example, combined with learning or evolutionary algorithms, it is possible to search for the optimal parameter reparameterization or subset selection scheme in different neural network architecture spaces, enabling the model to achieve efficient transfer learning various architectures. Additionally, the idea of evolutionary algorithms can be combined by encoding different parameter-efficient transfer learning methods and treating them as individuals for evolutionary operations. In each generation of, new individuals (i. e., new combinations of parameter-efficient transfer learning methods) are generated through operations such as crossover and mutation, and then screened based on their performance different architectures, with better-performing individuals being retained for the next generation. In this way, it is possible to automatically search for strategies that can achieve efficient transfer learning across multiple architectures thereby improving the cross-architecture generalization ability of parameter-efficient transfer learning methods.

5 Conclusion

This paper systematically classifies and studies the strategies of going from full fine-tuning to parameter-efficient adaptation in deep transfer. By establishing a three-level taxonomy, methods such as parameter addition strategy,

parameter reparameterization, and parameter subset selection are elaborated in detail, and the common principles as minimal perturbation of feature space, task-knowledge decoupling, and incremental knowledge injection are distilled. Meanwhile, the existing challenges such as lack of theoretical interpretability and weak crossarchitecture generalization ability are analyzed, and future directions such as causal-explanation-based interpretable adaption and neural architecture search driven automatic adaption are pointed out.

This provides a more systematic theoretical framework for the field of deep transfer learning, which helps researchers to understand and choose appropriate parameter-efficient transfer learning strategies more clearly, and pushes the field to development in both theory and application. Future research can be based on the taxonomy and directions proposed in this paper to explore more efficient and interpretable deep transfer learning methods to meet the demand for artificial intelligence applications.

References

[1] Wang Yitong. Research on Cross-domain Sentiment Classification Based on Pretraining and Adversarial Learning [D]. Beijing University of Posts and Telecommunications, 2023.

[2] Wu Peng. Research on Fine-grained Bird Classification Model Based on Transfer Learning [D]. Beijing Forestry University, 2021.

[3] Wang Keli, Yuan Hongchun. Image Recognition Method of Aquatic Animals Based on Transfer Learning [J]. Journal of Computer Applications, 2018, 38(5): 6. DOI: 10.11772/j.issn.1001-9081.2017102487.

[4] Zang Haixiang, Guo Jingwei, Huang Manyun, et al. State Estimation of Power Systems with Time-varying Topology Based on Deep Transfer Learning [J]. Automation of Electric Power Systems, 2021, 45(24): 8.

[5] Wang Pengde, Zhang Jingzhou. Research on Text Sentiment Analysis Based on MMD Dual-module Parameter Transfer [J]. Information Technology and Informatization, 2022(4): 65-68.

[6] Zheng Zongsheng, Hu Chenyu, Huang Dongmei, et al. Research on Typhoon Grade Classification Based on Transfer Learning and Meteorological Satellite Cloud Images [J]. Remote Sensing Technology and Application, 2020, 35(1): 9. DOI: CNKI:SUN:YGJS.0.2020-01-020.

[7] Zhou Qing, Zhou Jieqiong, Huang Yuan. Intelligent Project-based Learning: Immerse Every Child in "Real Learning"—Thoughts on the Project-based Teaching Case of "Making a Simple Intelligent Body Temperature Monitoring and Early Warning System" [J]. Inside and Outside the Classroom (Junior High School Edition), 2024(44): 1-3.

[8] Wang Huimei, He Jiawei, Liu Jian, et al. Small Sample DDoS Attack Detection Method Based on Deep Transfer Learning. CN202010943146.5 [Retrieval Date: 2025-05-09].

[9] Huang Yijing, Lü Junyan, Li Meng, et al. Auxiliary Diagnosis Algorithm for Diabetic Retinopathy Based on Transfer Learning [J]. Chinese Journal of Experimental Ophthalmology, 2019, 37(8): 5. DOI: 10.3760/cma.j.issn.2095-0160.2019.08.003.

[10] Wang Pin, Li Linyu, Li Yongming, et al. Deep Transfer Semi-supervised Domain Adaptation Classification Method for Histopathological Images: CN202110096888.3 [P]. CN112733859A [Retrieval Date: 2025-05-09].

11] Li Yuchen. Research on Motor Imagery EEG Classification Based on Convolutional Neural Network [D]. Xidian University, 2020.

[12] Dai Zuhua, Mou Qiaoling, Li Hongyi, et al. Deep Transfer Learning Method for Text Sentiment Classification: CN202010550138.4 [P]. CN111680160A [Retrieval Date: 2025-05-09].

13] Meng Zeqi. Developing Students' Cognitive Structure Based on Discovery Learning—Taking "Plane Rectangular Coordinate System" as an Example [J]. Innovation and Practice of Teaching Methods, 2020, 3(1): 72. DOI: 10.26549/jxffcxysj.v3i1.3150.

[14] Li Tong. Research on Teaching Strategies of Information Technology Classroom to Promote Deep Learning [J]. Advances in Education, 2025, 15(3): 7. DOI: 10.12677/ae.2025.153388.

[15] Dong Peihao, Jia Jibin, Zhou Fuhui, et al. Broadband Spectrum Sensing of Low-altitude UAV Swarms Based on Differential Privacy Federated Learning [J]. Journal of Electronics and Information Technology, 2025, 47(5): 1-11. DOI: 10.11999/JEIT241042.

[16] Wu Tianqi, Zhao Ruirui, Zuo Anxin, et al. Velocity Picking Based on Multi-feature Fusion Network [C]//Proceedings of the 3rd China Petroleum Geophysical Exploration Academic Annual Conference (IV). 2025.

[17] Yang Hailong, Yuan Yiping, Fan Panpan, et al. Foreign Object Recognition of Iron Removers for Mine Belt Conveyors Based on Transfer Learning and EfficientNet [J]. Coal Science and Technology, 2025, 53: 1-11. DOI: 10.12438/cst.2024-0772.

[18] Wang Jiatian, Li Fanchang. Research on a Linear Transfer Meta-learning Algorithm [J]. Journal of Computer Engineering & Applications, 2025, 61(5). DOI: 10.3778/j.issn.1002-8331.2310-0321.

[19] Feng Yunwen, Pan Weihuang, Lu Cheng, et al. Hydraulic Condition Monitoring and Fault Diagnosis of Civil Aircraft Based on Fault Logic [J]. Systems Engineering and Electronics, 2025, 47(3): 842-854. DOI: 10.12305/j.issn.1001-506X.2025.03.16.

[20] Wang Jing. Research on Text Representation and Classification Based on Deep Learning [D]. Beijing University of Posts and Telecommunications, 2019.