Transfer Learning for Few-Shot Image Classification

Shang Xiang *

School of Electronics and Information, Northwest University, Xi'an, 710069, China

* Corresponding Author Email: 2023117546@stumail.nwu.edu.cn

Abstract. Image classification technology has numerous vital applications across various industries and research fields. However, in many practical cases, it is often challenging to obtain sufficiently high-quality labelled data to train reliable models. As a classic and widely used paradigm of few-shot learning, transfer learning effectively mitigates the serious problem of over-fitting caused by limited data samples. It significantly enhances model generalisation ability by transferring knowledge learned from large-scale source domains (e.g., ImageNet) to few-shot target domain tasks. This paper systematically and comprehensively investigates transfer learning-based few-shot image classification methods, categorising them into three main types: feature-representation-based transfer, parameter-based transfer, and relational-knowledge-based transfer. It has also been clearly found that transfer learning methods, such as Adversarial Discriminative Domain Adaptation (ADDA) and knowledge distillation, serve as practical and efficient tools for addressing few-shot image classification problems. However, several critical challenges, such as negative transfer, multimodal heterogeneous transfer, and domain shift, still urgently require careful consideration and effective resolution in the future.

Keywords: Transfer Learning; Few-Shot Learning; Image Classification; Domain Adaptation; Knowledge Distillation.

1. Introduction

Image classification is a fundamental technology in computer vision with extensive applications in areas such as medical diagnosis, industrial inspection, and autonomous driving. Deep learning, particularly Convolutional Neural Networks (CNNs), has become the mainstream approach for image classification tasks due to its powerful capability for learning hierarchical features. However, deep learning models typically rely heavily on large amounts of labelled data to achieve superior performance. In fields like rare disease diagnosis, industrial defect detection, and novel species classification, obtaining sufficient high-quality training data poses a formidable obstacle due to the high costs, extensive time, intensive labour, or sheer impracticality involved. This problem of data scarcity directly leads to issues, including over-fitting under limited sample conditions, poor generalisation ability, and difficulty in handling unseen samples effectively. Although techniques like data augmentation can alleviate the problem, they struggle to fundamentally overcome the bottlenecks of insufficient representational learning and limited generalisation capability imposed by small datasets.

In response to the challenges, Few-Shot Learning (FSL) has emerged, aiming to train robust models efficiently with limited labelled samples. As one of the core paradigms of FSL, Transfer Learning addresses these issues by transferring knowledge, such as general visual features, object structures, and inter-sample relationships, learned from large-scale source domain datasets (often through pre-trained models) to data-scarce target domain tasks. This approach reduces the demand for extensive target domain data and effectively mitigates the risk of over-fitting caused by limited samples. The concept of transfer learning has long been established, but it has seen unprecedented development in the era of deep learning. Its applications have evolved from early, relatively simple "pre-training + fine-tuning" models to more detailed and diversified strategies [1]. Researchers continue to explore areas such as source domain selection and the optimisation of transfer strategies, giving rise to methods like Adversarial Discriminative Domain Adaptation (ADDA), variants of knowledge distillation, and structured knowledge transfer. A notable trend is the adoption of integrated strategies that combine multiple transfer mechanisms to fully use source domain knowledge and adapt to complex target domain discrepancies [2].

This paper aims to provide a systematic review of transfer learning-based methods for few-shot image classification. It focuses on three core transfer mechanisms: the feature-representation-based, the relational-knowledge-based, and the parameter-based. The principles, innovations, and applicable scenarios of representative methods will be elaborated, along with a comparative analysis of their performance and limitations. Furthermore, the paper discusses key challenges in current transfer learning research, such as negative transfer and multimodal heterogeneous transfer, and suggests potential future research directions. It is hoped that this survey will help organise the landscape of transfer learning approaches for few-shot image classification and facilitate further development in this field.

2. Different types of transfer learning

A central challenge in transfer learning lies in how to effectively transfer knowledge from the source domain to the target domain. Sinno Jialin Pan and Qiang Yang previously proposed a categorisation of transfer learning into four types: instance transfer, feature representation transfer, parameter transfer, and relational knowledge transfer [1]. In this paper, people regard instance weighting as a preprocessing step that can be incorporated into every type of transfer. Furthermore, to align more closely with deep learning-based image classification, people focus on the latter three categories, feature-representation-based, relation-based, and parameter-based transfer, according to the type of knowledge being transferred during the process.

2.1. Feature-representation-based transfer

Feature-based transfer learning aims to identify a common feature representation between the source and target domains, transferring knowledge through feature transformation. A widely adopted approach in combination with convolutional neural networks (CNNs) is the use of a feature extractor [3]. In this paradigm, the convolutional layers of a pre-trained model are frozen to preserve their ability to extract generic image features, including low- and mid-level representations such as edges and textures. Only the newly added classification layers (typically fully connected layers) are trained on the target domain. This allows the model to use the feature representation capability learned from the source domain without modifying the core parameters of the source model.

Choi and his team proposed a Structured Set Matching Networks (SSMN), which uses CNNs with shared weights to extract local features from both source and target images [4]. A bidirectional LSTM is used to generate context-aware feature maps within a unified space. Image classification is achieved by combining local similarity (f_a)and global consistency (f_{gc}).

Liu and his team introduced a transfer learning method called Substructure Correlation Adaptation (SCOAD), based on K-means clustering [5]. This approach clusters data from both the source and target domains into K subdomains using the K-Means algorithm. The data distribution is then aligned by matching the second-order covariance matrices of the centres of these subdomains.

Tzeng and his team proposed a method called Adversarial Discriminative Domain Adaptation (ADDA), which demonstrates significant advantages in cross-modal image classification [6]. They adopted a decoupled adversarial training strategy, utilising independent encoders for the source and target domains. The target domain encoder M_t is optimised such that the features it generates cannot be distinguished by a domain discriminator D from those produced by the source domain encoder M_s within a shared feature space. The corresponding loss function is shown in Equation (1) [4].

$$\mathcal{L}_{adv_D} = -E_{x_s \sim X_s}[\log D(M_s(x_s))] - E_{x_s \sim X_s}[\log(1 - D(M_s(x_s)))] \tag{1}$$

2.2. Relation-based transfer

Relation-based transfer learning builds mappings between samples from the source and target domains. These mappings capture relationships such as similarity or correlation. Instead of transferring features or parameters directly, this approach transfers relational knowledge. Hinton and his team introduced a widely used technique known as knowledge distillation [7]. A large teacher model transfers its expertise to a small student model. The transferred knowledge includes representations of inter-class relationships. A temperature parameter T is introduced to soften the teacher's output. This converts challenging class labels into a soft probability distribution. The student model is then trained to mimic this soft distribution using the KL divergence. Equation (2) illustrates this temperature scaling [7].

$$q_i = \frac{exp(^{z_i}/_T)}{\sum_j exp(^{z_i}/_T)} \tag{2}$$

Fukuda and his team improved the standard knowledge distillation framework. They introduced a mechanism for transferring privileged information. This enables a narrow-band student model to utilise soft targets encoded with high-frequency information from a wide-band teacher [8].

Li and his team applied knowledge distillation to hyperspectral image classification [9]. The teacher utilised two graph neural networks (GNNs) and student GNNs to extract superpixel features from different perspectives. The teacher GNN transfers its knowledge to the student GNN through distillation. This helps the student model capture relationships among unlabelled super pixels.

2.3. Parameter-based transfer

Parameter-sharing-based transfer learning involves sharing specific model parameters between the source and target domains. This approach includes the reuse of convolutional layer weights or the assumption that parameters in both domains follow the same prior distribution.

Fine-tuning is a widely used method in image classification. It is often applied with convolutional neural networks [3]. This approach reuses weights from a pre-trained model as initialisation. Some or all parameters are then updated on the target domain data. For example, the last few convolutional layers may be unfrozen during training. Parameter adjustment allows the model to adapt to the target domain's distribution. This is useful for domains with unique features, such as medical image textures.

Hussain and his team directly reused the convolutional layer parameters of Inception-v3 [10]. They only replaced and retrained the last fully connected classification layer. This fine-tuning approach achieved efficient transfer on the CIFAR-10 and Caltech Faces datasets. Swati and his team proposed a block-wise fine-tuning strategy [11]. They used a VGG19 model pre-trained on ImageNet. The model was divided into six blocks (B_1–B_6) based on the positions of the pooling layers. Initially, only the top block B_6, which consists of fully connected layers, was fine-tuned. The fine-tuning scope was then gradually expanded toward shallower blocks. This strategy freezes shallow layers and fine-tunes deeper ones. It balances the adaptation between general features and medical-specific features. It also addresses the domain gap between medical images and natural images.

Zhang and his team introduced a transfer term into the objective function of the traditional Least Squares Support Vector Machine (LS-SVM), which is shown in Equation (3) [12]. The optimised LS-SVM function is given in Equation (4) [12]. They also incorporated a transfer weight, denoted as γ_j, which indicates the relevance between the source and target domains. The transfer term encourages the hyperplane parameter ω of the target model to approximate that of the source model, ωj. The influence of different source models is controlled through the transfer weights γ_j.

$$\frac{1-\lambda}{2}\left\|\omega - \sum_{j=1}^{M} \gamma_j \omega_j^s\right\|^2 \tag{3}$$

$$\min_{\omega,b,\xi} Q(\omega, b, \xi) = \frac{\lambda}{2}\|\omega\|^2 + \frac{1-\lambda}{2}\left\|\omega - \sum_{j=1}^{M} \gamma_j \omega_j^s\right\|^2 + C\sum_{i=1}^{N} \delta_i \xi_i \tag{4}$$

3. Datasets and comparison of experimental results

3.1. Datasets and experimental results

Choi and his team trained their SSMN model using the DiPART (Domain-Invariant Part Labelling) and PPM (PASCAL-Part-Mini) datasets [4]. These datasets focus on part-level recognition and annotation in images. They represent typical few-shot scenarios for evaluating structured feature transfer capability. The SSMN model achieved an accuracy of 58.1% on the DiPART dataset. It reached 46.6% accuracy on the PPM dataset. In comparison, the traditional Matching Network (MN) with Hungarian algorithm achieved only 45.6% and 42.7% accuracy on the same datasets. The SSMN model demonstrated a significant improvement in performance.

During the training of the Adversarial Discriminative Domain Adaptation (ADDA) method, different datasets are used for the source and target domains [6]. The target domain uses an unlabeled dataset to evaluate its adaptation capability. The method was assessed primarily on digit datasets, such as MNIST, USPS, and SVHN. These datasets assess their unsupervised adaptation ability across different digit recognition domains. The source-target pairs included MNIST (handwritten digits) to USPS (postal mail digits); USPS to MNIST; and SVHN (street view house number digits) to MNIST. Significant domain differences exist between the SVHN and MNIST datasets. The adaptation accuracy reached 89.4% from MNIST to USPS, and 90.1% from USPS to MNIST. For the more challenging transfer from SVHN to MNIST, the accuracy was 76.0%. ADDA performed excellently when the domains were relatively similar, such as between MNIST and USPS. However, its performance declined significantly when transferring from SVHN to MNIST due to the large domain gap between the two datasets. Although ADDA utilises adversarial training to learn domain-invariant features, fundamental modal differences remain challenging to overcome fully.

Swati and his team utilised a publicly available CE-MRI (Contrast-Enhanced Magnetic Resonance Imaging) dataset [11]. This dataset comprises 3064 2D slices (512 × 512 pixels) obtained from 233 patients. The specific fine-tuning results are summarised in Table 1.

Table 1. Block-wise Fine-tuning Performance

Fine-tuning strategy	Accuracy (%)	Vs B_6	Key Findings
B_6	86.81	Baseline	Lowest performance
B_5-B_6	91.92	+5.11	Fine-tuning significantly improves
B_1-B_6	94.82	+8.01	Full domain adaptation

3.2. Brief comparison of the experimental results

Table 2. Dataset and Performance Comparison [4,6,7,8,11,12]

	Feature-representation-based transfer		Relation-based transfer	Parameter-based transfer	
Methods	SSMN	ADDA	Knowledge distillation	Block-wise fine-tuning	Improve LS-SVM
Core innovation points	Shared CNN for local feature extraction + Bidirectional LSTM for context modelling + Local similarity (f_a)/Global consistency (f_{gc})	Decoupled adversarial training: Independent encoders + Domain discriminator for feature distinction	Temperature parameter T for output softening + KL divergence for distribution matching	VGG19 is divided into six blocks (B_1–B_6) + Fine-tuning from the top to the bottom layers + Block-wise independent learning rates	Transfer term + Transfer weight γ_j
Dataset and Performance	DiPART/PPM DiPART: 58.1% Acc PPM: 46.6% Acc (↑10% vs MN+Hungarian)	MNIST→USPS: 89.4% SVHN→MNIST: 76.0%	MNIST/CIFAR: Teacher to student model compression with <3% accuracy drop	CE-MRI Brain Tumour: B_1–B_6: 94.82% Acc (↑8.01% vs only B_6)	Caltech-256: 6 Food Categories: 95% Acc (↑3% vs Single-KT) 10 Mixed Categories: Best Acc (↑5% vs Single-KT)

Based on the introduction of various transfer learning methods, people compare the datasets used and the final performance of several representative methods in Table 2. Examples include SSMN and

ADDA. It is observed that each method achieves performance improvement through its novel architecture or algorithm.

4. Conclusion

Each of the three methods — feature-representation-based transfer, relation-based transfer, and parameter-sharing transfer — has its own strengths and limitations. For example, feature-representation-based transfer is sensitive to label shift; relation-based transfer relies heavily on empirical design of metric functions; and parameter-sharing transfer often entails high training costs. Therefore, integrating these three paradigms and exploring more refined and efficient fusion mechanisms will be a key direction for future research. For instance, Wang and his team proposed the Multi-view Adaptive Subspace Alignment (MASA) method, which constructs feature subspaces and employs adaptive alignment to enable few-shot learning across multiple views.

Effectively suppressing negative transfer (NT) remains a long-standing and challenging problem in transfer learning. Current mainstream solutions—such as safe transfer, domain similarity estimation, distant transfer, and pseudo-labelling approaches—still have limited applicability across various scenarios. Therefore, developing theoretically grounded safe transfer algorithms for popular paradigms, such as unsupervised domain adaptation, few-shot transfer, and adversarial transfer, will be a crucial direction for future research.

In multimodal heterogeneous domain transfer, such as RGB images to depth maps or CT to MRI, the inherent structural differences between source and target modalities cause traditional feature alignment methods to fail. A promising direction for future research lies in the structured design of disentangled representations. This approach should explicitly separate modality-invariant semantics, such as object shape, structure, and semantic relationships, while applying contrastive constraints to enforce semantic consistency across modal feature spaces. Such a framework would help ensure effective transfer of core knowledge.

This paper provides a summary and organisation of transfer learning methods applied to current few-shot image classification tasks. By effectively leveraging knowledge acquired from large-scale source domain pre-training, transfer learning significantly reduces the demand for extensive target domain training data, offering a powerful solution for few-shot image classification. Based on the core mechanisms and operational subjects of transfer learning, the methods are categorised into three types in the paper: feature-representation-based, relation-based, and parameter-sharing approaches. An overview and performance analysis of specific methods under each category are provided. Finally, considering the existing limitations of transfer learning methods, the paper discusses current challenges and future research directions in few-shot image classification using transfer learning.

References

[1] Pan S J, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345–1359.

[2] Wang X, Liu J, Li L M. MASA: Multi-view adaptive subspace alignment for enhanced few-shot learning. Knowledge-Based Systems, 2025, 114148.

[3] Kim H E, Cosa-Linan A, Santhanam N, et al. Transfer learning for medical image classification: A literature review. BMC Medical Imaging, 2022, 22: 69.

[4] Choi J, Krishnamurthy J, Kembhavi A, et al. Structured set matching networks for one-shot part labelling. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3627–3636.

[5] Li J, Guo Y, Xu J, et al. Fine-grained image classification based on deep transfer learning with multi-scale attention. Expert Systems with Applications, 2023, 213: 118850.

[6] Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2962–2971.

[7] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv: 1503.02531, 2015.

[8] Fukuda T, Suzuki M, Kurata G, et al. Efficient knowledge distillation from an ensemble of teachers. Interspeech, 2017: 3697–3701.

[9] Li X, Ma H, Guo S, et al. Graph knowledge distillation-guided few-shot learning for hyperspectral image classification. Journal of Supercomputing, 2025, 81(11): 1–46.

[10] Hussain M, Bird J J, Faria D R. A study on CNN transfer learning for image classification. Advances in Intelligent Systems and Computing, 2019, 840: 224–233.

[11] Khan S Z N, Zhao Q, Kabir M, et al. Brain tumour classification for MR images using transfer learning and fine-tuning. Computerized Medical Imaging and Graphics, 2019, 75: 34–41.

[12] He J, Zhang X, Liu Q. An improved LS-SVM algorithm based on model parameter transfer and its application. Pattern Recognition Letters, 2020, 140: 51–58.