UDC 621.391

METHODS OF LOW DATA IMAGE CLASSIFICATION WITH NEURAL NETWORK

Y.M. CHEN

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus

Received April 04, 2025

Abstract. Image classification has evolved from relying on handcrafted feature engineering to a data-driven deep learning paradigm, with recent breakthroughs extending its capabilities to zero-shot and few-shot learning scenarios. This paper introduces emerging paradigms for low-data scenarios: (1) zero-shot learning methods, including multimodal models (e.g., CLIP) that align visual and textual embeddings for open vocabulary classification and generative frameworks for synthesizing features of unseen classes; and (2) few-shot learning strategies, such as meta-learning (e.g., MAML), metric-based networks (e.g., ProtoNet), and fast adaptation techniques (e.g., Tip-Adapter) that leverage pre-trained knowledge for fast adaptation. While these methods reduce the reliance on labeled data, challenges remain in domain adaptation, fine-grained classification, and computational efficiency.

Keywords: Image classification, Deep learning, Zero-shot learning, Few-shot learning, CLIP, Convolutional Neural Networks (CNNs), Vision Transformers (ViTs)

Introduction

Image classification, the task of assigning semantic labels to digital images based on their visual content, has become an essential task in modern computer vision. Its applications range from medical imaging and autonomous driving to social media content moderation, driving the continuous demand for accurate, scalable, and adaptable solutions. Traditional approaches are rooted in feature engineering techniques such as scale-invariant feature transform (SIFT) and histogram of oriented gradients (HOG), relying on manually designed descriptors combined with machine learning classifiers such as support vector machines (SVM). While these methods are effective for constrained scenes, they face significant limitations in handling real-world variations including occlusions, illumination changes, and viewpoint changes due to their reliance on domain expertise and rigid feature representations.

The advent of deep learning marked a paradigm shift. Convolutional neural networks (CNNs), exemplified by the breakthrough of AlexNet in the ImageNet competition, demonstrated an unprecedented ability to learn hierarchical features directly from raw pixels. Subsequent architectures such as VGGNet, ResNet, and DenseNet further improved accuracy through innovations such as networks and deeper connections. Recently, Vision Transformers (ViTs) challenged the dominance of CNNs by leveraging self-attention mechanisms to model global dependencies, achieving state-of-the-art performance on large-scale datasets. However, these data-hungry models require large amounts of labeled training data, which limits their remaining usefulness in scenarios where annotations are scarce, expensive, or dynamically changing common challenges in medical, ecological, or industrial applications.

To address these limitations, the field has increasingly focused on low-data schemes, particularly zero-shot learning (ZSL) and few-shot learning (FSL). Zero-shot methods aim to classify unseen categories by leveraging auxiliary information such as semantic attributes or textual descriptions, while few-shot techniques adapt models to new tasks using minimal labeled examples. Pioneer methods in ZSL, such as attribute embeddings and generative adversarial networks (GANs), have laid the foundation for bridging semantic and visual spaces. Meanwhile, meta-learning frameworks such as model-agnostic meta-learning (MAML) and prototypical networks have redefined few-shot adaptations

by learning transferable initialization parameters or similarity metrics. The advent of multimodal models such as Contrastive Language-Image Pretraining (CLIP) has further revolutionized the field by unifying vision and language through contrastive learning, enabling open vocabulary classification without the need for task-specific fine-tuning.

Methods

As image classification tasks gradually shift from relying on large amounts of data to low-data scenarios, a variety of innovative methods have been proposed to address the challenges of zero-shot and few-shot learning. These methods significantly improve the generalization ability of models under limited supervision through strategies such as multimodal alignment, meta-optimization, prototype matching, feature generation, and lightweight adaptation.

These methods not only expand the boundaries of traditional deep learning but also provide flexible solutions for dynamic needs in practical applications.

CLIP (Contrastive Language-Image Pretraining), developed by OpenAI, represents a paradigm shift in image classification by unifying vision and language through multimodal contrastive learning. The model jointly trains a visual encoder (e.g., Vision Transformer or ResNet) and a text encoder (e.g., Transformer) on a large dataset of image-text pairs, mapping both modalities into a shared embedding space. Figure 1 shows the approach of their method. During inference, CLIP achieves zero-shot classification by comparing image embeddings with text embeddings of user-defined class descriptions (e.g., "satellite photo of a rainforest"), eliminating the need for task-specific fine-tuning. This approach performs well in open vocabulary scenarios, allowing flexible adaptation to new categories via natural language cues. However, its performance on fine-grained tasks (e.g., distinguishing between bird subspecies) is still limited by the semantic granularity of the text cues, and biases in the training data from the network can propagate to downstream applications.

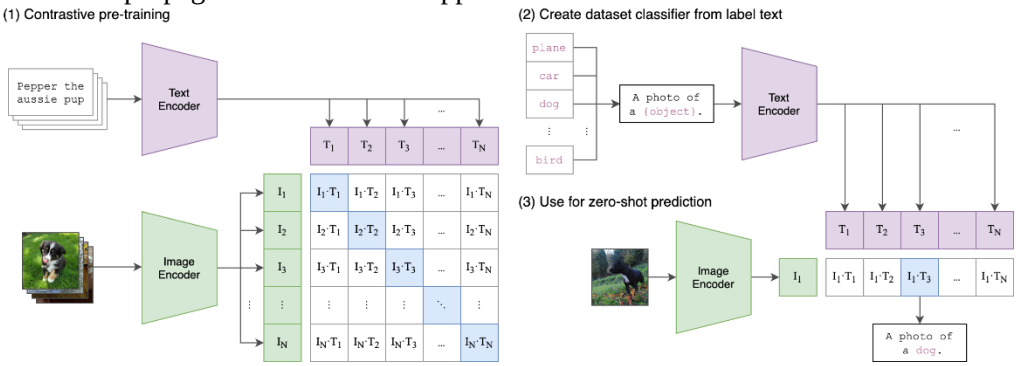

Figure 1. Summary of CLIP's approach [1]

MAML (Model-Agnostic Meta-Learning [2]) is a foundational meta-learning framework that solves the few-shot learning problem by “learning to learn”. Instead of training a model for a specific task, MAML optimizes a set of initial parameters that can quickly adapt to new tasks with minimal labeled examples. By training across scenarios of different tasks, the model internalizes a generalizable initialization, enabling efficient fine-tuning using gradient updates on a small support set. While MAML performs well in domains ranging from robotics to medical imaging, it is computationally expensive as it requires second-order derivative computations during meta-training. Furthermore, performance depends on the diversity and quality of the meta-training task distribution, which poses a challenge in scenarios with limited task variation.

ProtoNet (Prototypical Network) simplifies few-shot classification by leveraging principles of metric learning. For each class in the support set, the network computes a "prototype" as the average feature vector of its labeled examples. Figure 2 demonstrated the classification mechanism of the prototype network in few-shot and zero-shot scenarios. During inference, a query sample is classified based on its Euclidean or cosine distance to these prototypes in the embedding space. This approach bypasses complex architectural modifications and instead relies on a well-structured feature space learned through episodic training. The simplicity and efficiency of ProtoNet make it a popular baseline for few-shot benchmarks, but its performance degrades when the support samples are noisy or

insufficient to capture intra-class variation. Extensions to this framework, such as leveraging attention mechanisms, have been proposed to enhance robustness.

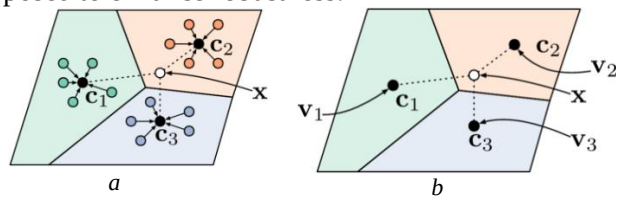

Figure 2. Prototypical networks in the few-shot and zero-shot scenarios [3]:
a – few-shot scenario; *b* – zero-shot scenario

f-CLSWGAN (Feature-Conditioned Least Squares Wasserstein GAN [4]) tackles zero-shot learning by synthesizing features for unseen classes, effectively transforming the problem domains into a supervised learning task. By conditioning a Wasserstein GAN on semantic attributes or text descriptions, the model generates synthetic visual features that mimic real data distributions. These generated features are then used to train a classifier for unseen categories, bridging the gap between semantic and visual. While this approach mitigates the "domain shift" problem common in zero-shot learning, its success heavily depends on the quality and diversity of generated features. Imperfections in feature synthesis, such as mode collapse or unrealistic variations, can propagate errors to downstream classification.

Tip-Adapter [5] bridges the gap between large-scale pre-trained models and small-sample adaptation by combining feature caching with lightweight neural networks. The approach stores pre-trained image features of supporting examples in a key-value cache and then trains a small adapter network to map query features to cached prototypes through non-parametric retrieval. This hybrid design fully leverages the rich representations of models such as CLIP or ResNet while avoiding extensive fine-tuning, achieving high accuracy with minimal computational overhead. The efficiency of Tip-Adapter makes it particularly suitable for edge devices or scenarios that require fast deployment, although its performance stagnates when the pre-trained features lack discriminative power for the target task.

These methods expand the boundaries of traditional deep learning and provide flexible solutions for dynamic real-world applications.

Conclusion

Image classification has advanced from manual feature design to methods that work with little data. New approaches, like aligning images with text or learning from a few examples, show promise for real-world use. Challenges remain, such as reducing biases and improving efficiency. Future work should focus on making models more adaptable, fair, and efficient. By addressing these issues, image classification can better serve diverse applications, from healthcare to environmental monitoring, even with limited resources.

References

1. Radford, A., et al. // International conference on machine learning. 2021. P. 8748–8763.
2. Finn, Chelsea, et al. // International conference on machine learning. 2017. P. 1126–1135.
3. Snell, J., Swersky, K., Zemel, R. // Advances in neural information processing systems. 2017. P. 4077–4087
4. Xian, Y., Lorenz, T., Schiele, B., & Akata, Z. // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. P.5542–5551.
5. Zhang, R., et al. // European conference on computer vision. 2022. P. 493–510.