Advancements in Image Classification: From Machine Learning to Deep Learning

Haoran Cheng

Zhejiang University - University of Illinois Urbana-Champaign Institute, Zhejiang University, Zhejiang, China

Abstract. Image classification, as an essential task within the realm of computer vision, has evolved from traditional machine learning methods to deep learning techniques. This paper systematically reviews the growth of image classification technology, beginning with the introduction of commonly used datasets such as CIFAR-10, ImageNet, and MNIST, and exploring their impact on algorithm development. Subsequently, the paper provides an in-depth analysis of image classification methods based on machine learning, including traditional algorithms such as Support Vector Machine (SVM), Random Forest, and Decision Tree. These methods achieve image classification through two stages: feature extraction and classification, but they encounter limitations when confronted with large-scale datasets and complicated tasks. Convolutional Neural Networks (CNNs) have gradually replaced traditional methods in image classification due to the rise of deep learning, resulting in improved accuracy and robustness. The paper also focuses on discussing classic deep learning models such as AlexNet, VGGNet, ResNet and ViT, analyzing their strengths and weaknesses. By comparing the performance of different methods, this paper aims to provide references for researchers in the realm of image classification, promoting further development in this area.

1 Introduction

Image classification is one of the most essential yet crucial tasks in the field of computer vision, with the primary goal of accurately assigning input images to predefined categories. As a cornerstone of computer vision, image classification is not only an independent research area but also lays the foundation for visual tasks that are more complex, like object detection, image segmentation, and scene understanding. The development of image classification techniques reflects the overall progress in the field of computer vision: from the early use of manually designed features and traditional machine learning algorithms for image classification, to the advent of the deep learning era, where end-to-end learning models, represented by Convolutional Neural Networks (CNNs), achieve automatic mapping from raw image pixels to class labels. This technological revolution has not only significantly improved classification accuracy but also greatly enhanced the model's generalization capability and adaptability in complex scenarios. Especially with the support of big data and

Corresponding author: haoranc.21@intl.zju.edu.cn

high-performance computing resources, modern image classification models are capable of handling larger and more complex datasets, demonstrating unprecedented performance levels, making image classification increasingly important in various application scenarios.

In addition, image classification has permeated various aspects of lives. In medical imaging diagnostics, image classification technology can help doctors identify lesion areas, enhancing the precision and effectiveness of diagnosis [1]. In the agricultural sector, image classification technology is used for detecting plant diseases and crop monitoring, enhancing agricultural production efficiency and yield [2].

This paper aims to systematically review the machine learning and deep learning methods in the field of image classification by examining existing technologies, comparing the benefits and drawbacks of different methods, and exploring their performance in practical applications to showcase the practical value and potential of these technologies, providing references for future related research and promoting further research and applications in this field.

This paper's narrative logic is divided into the following aspects: first, it introduces the commonly used datasets in image classification. These datasets not only serve as benchmarks for algorithm evaluation but also provide a unified testing environment for researchers, facilitating the testing of different methods. Next, it introduces commonly used image classification methods in machine learning, including Support Vector Machine (SVM), Random Forests, and Decision Trees. Then, it introduces image classification methods based on deep learning, covering various stages of deep learning, including AlexNet, VGGNet, ResNet, and ViT. Additionally, the paper compares and analyzes the strengths and weaknesses of these models from multiple dimensions, providing researchers with reference points for selecting appropriate models in practical applications.

2 Commonly used datasets for image classification

2.1 Overview of datasets

Algorithm development and evaluation are greatly influenced by using common datasets for image classification tasks. These datasets not only provide a unified testing environment, making it possible to compare different methods horizontally but also offer researchers abundant training resources, contributing to the creation and enhancement of novel algorithms. Selecting the appropriate dataset is vital for evaluating an algorithm's performance, as the image quality, number of categories, and data distribution in different datasets may significantly influence the algorithm's performance. Commonly used datasets include CIFAR-10, CIFAR-100, SVHN, ImageNet, MNIST, etc. This section focuses on the CIFAR-10, ImageNet, and MNIST datasets, that are frequently employed in research.

2.2 CIFAR-10 dataset

CIFAR-10 is a diffusely used small-scale image classification dataset, consisting of 60,000 32x32 color images in 10 different categories. Training requires 50,000 images and testing requires 10,000 images. The categories in the CIFAR-10 dataset include cars, dogs, frogs, ships and so on. These images are preprocessed to ensure a balanced number of samples in each category, providing a unified evaluation benchmark. In recent years, the CIFAR-10 dataset has been used in many studies. For example, Siripuri Divya et al. [3] proposed a new Convolutional Neural Network (CNN) architecture for image enhancement and classification, validated using the CIFAR-10 dataset. Their research demonstrated that image enhancement techniques could significantly improve the performance of classification models, ultimately

achieving higher classification accuracy. Giuste et al. [4] successfully improved the image classification accuracy on the CIFAR-10 dataset to 94.6% by combining various feature extraction methods such as HOG, pixel intensity, and transfer learning-optimized deep learning models. Additionally, the CIFAR-10 dataset is frequently employed to evaluate the effectiveness of various algorithms, thereby driving continuous progress in image classification technology.

2.3 ImageNet dataset

ImageNet is a large visual database containing over 14 million labeled images, covering more than 20,000 categories. The ImageNet Challenge (ILSVRC) attracts many research teams every year, promoting the rapid growth of image classification technology. Through the ImageNet Challenge, many groundbreaking deep learning models, such as ResNet, VGGNet, and AlexNet, have emerged, significantly enhancing the accuracy of image classification and laying the foundation for deep learning in computer vision. Lucas Beyer et al. [5] explored the potential overfitting issues in current image classification research based on the ImageNet dataset, particularly the model's dependency on the dataset's labels. By reassessing the labels of the ImageNet validation set, they developed a more robust labeling method called Real Labels and used these new labels to reevaluate the results of recently proposed image classification models.

2.4 MNIST dataset

The MNIST dataset is a tool for recognizing handwritten digits and there are 60,000 training images and 10,000 test images in the collection, with each image being a 28x28 pixel grayscale image. Due to its simple structure and clear classification objectives, the MNIST dataset is regarded as an introductory dataset for image classification tasks, making it an ideal choice for evaluating new algorithms. For example, Chen et al. [6] used the MNIST dataset in their research, proposing an enhanced FGSM-based adversarial training method. Their method not only significantly accelerated the training speed, approximately five times faster than traditional methods, but also improved the model's robustness against adversarial attacks. Kadam et al. [7] evaluated the performance of various CNN models using the MNIST and Fashion-MNIST datasets, demonstrating the superiority of these models in tasks such as handwritten digit and clothing image recognition. They proposed five different CNN architectures, testing the performance of these models by combining different convolutional layers, filter sizes, and fully connected layers. Wan et al. [8] studied the Drop Connect regularization method using the MNIST dataset and proved that this method can significantly improve the accuracy of neural networks in image classification tasks.

3 Machine learning-based image classification methods

3.1 Introduction to machine learning-based Image classification

Two main stages are typically used in traditional machine learning-based image classification methods: feature extraction and classification. In the feature extraction part, researchers apply various techniques and algorithms based on prior experience and expertise to extract representative features from raw images. These features typically include color, texture, shape, etc., and are then converted into numerical feature vectors for subsequent processing. After feature extraction is complete, researchers input these extracted feature vectors into a

classification model for classification. The classification model is trained and learned based on these features, enabling it to identify and classify various objects or categories in images.

3.2 Support vector machine

Support Vector Machine (SVM) is a typical machine learning model, renowned for its unique concept and wide range of applications. The basic concept of SVM is to transfer the original data to a high-dimensional feature space and then create an optimal hyperplane in that high-dimensional space, which can effectively separate different types of data points. Maximizing the minimum distance between data points and the hyperplane is what determines this optimal hyperplane, ensuring the maximum boundary for classification. To address the computational complexity that may arise during the mapping process from low-dimensional space to high-dimensional space, SVM introduces the concept of kernel functions, which allow direct computation in low-dimensional space without explicitly calculating high-dimensional features. Common kernel functions include linear, polynomial, and Gaussian (RBF) kernels, each suitable for different types of data distributions and classification tasks. In specific applications, SVM often performs well on small sample datasets. Due to its classification by constructing an optimal hyperplane, SVM exhibits strong generalization ability, adapting well to unseen data [9]. Additionally, the optimization objective of SVM is a convex optimization problem, eliminating the risk of local minima, making the training process more stable and reliable.

3.3 Decision tree

A Decision Tree is a tree-based machine learning model that recursively divides the dataset into multiple subsets, gradually generating a tree structure for classification or regression. A decision is made by each node based on a specific feature, the branches represent data splits based on different values of that feature, and the leaf nodes correspond to the final classification or regression results. To select the optimal split point for the decision tree, criteria like information gain or Gini index are utilized to maximize the purity of the subsets after each split. However, decision trees are prone to overfitting, especially when handling complex data, where they may generate overly deep tree structures that overfit the training data. To address this, pruning techniques are often used to remove unnecessary nodes, thereby improving the model's generalization ability. For example, Jijo et al. [10] used the decision tree algorithm to classify the Luzhou dataset, improving the model's classification accuracy to 80.84% by combining the Principal Component Analysis methods and Minimum Redundancy Maximum Relevance.

3.4 Random forest

Random Forest is an ensemble learning technique that is based on decision trees. It constructs multiple independent decision tree models and integrates their prediction results, such as through majority voting, to improve classification accuracy and stability. Random Forest introduces the mechanism of randomly selecting training samples (Bootstrap sampling) and randomly selecting feature subsets, enhancing the model's diversity and robustness, particularly when dealing with high-dimensional data. By integrating the consequences of multiple decision trees, Random Forest effectively decreases the overfitting problem commonly associated with single decision tree models, while also demonstrating strong noise resistance and generalization performance. Random Forest also provides an evaluation of feature importance, helping to understand the key features in the data and supporting further model optimization. In the application of Random Forest, Chandrasekaran et al. [11]

compared the performance of Random Forest and SVM in remote sensing image classification, finding that Random Forest performed exceptionally well in handling SAR data, achieving a classification accuracy of 91.60%.

3.5 Advantages and limitations of machine learning

The advantages of machine learning lie in its efficiency and flexibility, especially when dealing with small to medium-sized datasets. Machine learning algorithms can usually quickly build models and achieve classification. Additionally, machine learning algorithms are becoming a variety of data types and tasks, including not only image classification but also text classification, speech recognition, and more. However, machine learning also has certain limitations. Firstly, it is highly dependent on feature selection, with model performance largely influenced by the choice and extraction of features, which in turn relies on the researchers' personal experience, introducing some subjective factors into related experiments. Secondly, machine learning algorithms are prone to overfitting, particularly when handling large-scale data and complex tasks, where their performance is limited, making it difficult to extend to more complex scenarios.

4 Deep learning-based image classification methods

4.1 Introduction to deep learning-based image classification

Deep learning-based image classification is a fully automated process, unlike traditional machine learning methods. Deep learning models learn features from raw images automatically and perform classification through multilayer neural networks. Deep learning models are particularly suitable for the needs of the big data era, competent in handling large numbers of image data, and excelling in image classification tasks. The use of CNNs in deep learning is widespread and widely acknowledged due to their outstanding performance in image classification. The deep learning image classification process typically includes three steps: data preprocessing, model training, and classification. During the data preprocessing stage, image data is subjected to operations such as normalization, standardization, and data augmentation, ensuring the data meets the requirements for model training. In the training stage, the model extracts features through stacked convolutional layers, continuously refining parameters through iterative training to achieve effective feature approximation. In the classification stage, the model classifies features through fully connected layers, thereby achieving image classification.

4.2 AlexNet

AlexNet was proposed by Krizhevsky et al. [12] and significantly outperformed other competing models in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC), bringing about a new age of deep neural networks in computer vision. In the model, there are eight layers of neural networks, with five convolutional layers and three fully connected layers. These convolutional layers achieve multi-level feature extraction from images by stacking layers, enabling the model to capture richer image information. AlexNet's innovation lies not only in its deep structure but also in its use of the Rectified Linear Unit (ReLU) activation function, replacing traditional Sigmoid or Tanh activation functions. The introduction of ReLU significantly accelerated the training process and improved the network's nonlinear expression capabilities, allowing the network to better fit complex features. Additionally, AlexNet effectively alleviated the overfitting problem in

neural networks through Dropout technology. By applying Dropout in the fully connected layers, random connections between neurons were discarded, reducing the model's overreliance on drilling data and improving its generalization ability on new data. AlexNet's proposal also laid the foundation for subsequent models such as VGGNet and ResNet.

4.3 VGGNet

VGGNet was proposed by Simonyan et al. [13] in 2014. Compared to AlexNet, VGGNet used multiple small convolutional kernels (3x3) instead of the large convolutional kernels commonly used before, keeping the receptive field intact while reducing the number of parameters. This network also increased the network depth, improving the model's representation capability. VGGNet is commonly used in two structures, VGG-16 and VGG-19, with VGG-16 consisting of three connected layers and thirteen convolutional layers, while VGG-19 consists of three fully connected layers and sixteen convolutional layers. The multiple versions of VGGNet and the consistency and modular design between versions make it more efficient in image classification, and due to its deeper network, the model can effectively capture complex visual information, improving classification accuracy. VGGNet demonstrated the potential of deep models in computer vision and provided important references for subsequent model designs.

4.4 ResNet

Under the leadership of AlexNet and VGGNet, the depth of models was continuously deepened. In a certain sense, with the deepening of the model, the features of the image should be more effectively extracted, and the classification accuracy should be higher. However, as researchers continued to stack convolutional layers, the model began to experience performance degradation issues after reaching a certain network depth. That is, the model's training error increases with the depth of the network, and problems such as vanishing gradients and exploding gradients begin to appear. To handle this problem, He et al. [14] designed the ResNet network in 2015, introducing the residual module, which allows information to be transmitted across layers through direct shortcut connections, solving the degradation problem of deep layers. In the specific residual module, input information can be directly added to the output, meaning that the original input information can be transmitted directly to subsequent layers through the shortcut path. This design mitigates the problem of information loss during convolutional stacking, ensuring that even if some convolutional layers have slow weight updates or fail to learn effective features, the original input information can still be transmitted directly to the output layer via the shortcut path. ResNet not only solved many issues related to deep network training but also significantly improved the accuracy of image classification. Since its introduction, ResNet has become a significant milestone in deep learning, profoundly influencing subsequent network structure designs.

4.5 Vision transformer

Vision Transformer (ViT) is an innovative image classification model proposed in recent years [15]. It leverages the powerful characteristics of the Transformer architecture, breaking away from the dominant role of traditional CNNs in image processing. Unlike traditional CNNs that rely on convolution operations, ViT employs a novel approach by dividing input images into fixed-size patches, then linearly unfolding and embedding these patches as vectors. The Transformer model processes these embedding vectors later. This approach allows ViT to effectively find long-range dependencies and information between image patches, capitalizing on the strengths of Transformer in handling sequential data. On large-

scale datasets, ViT has demonstrated excellent learning capabilities, particularly achieving high classification accuracy even without extensive pre-training. The introduction of ViT not only brought a new solution to image classification tasks but also inspired a deeper exploration of the potential applications of Transformers in visual tasks. The success of this model marks a diversification of deep learning models in image classification and opens new directions for future research [16].

4.6 Advantages and limitations of deep learning

Compared to machine learning, deep learning is a fully automated method, with feature extraction and classification processes integrated into a single flow managed by a fixed model. This fully automated approach is more in line with the requirements of the big data era, facilitating the training of models on large datasets. Moreover, feature selection in deep learning is also automated, with the model autonomously selecting the most effective features through continuous training, thereby reducing the influence of personal subjective factors and ensuring more objective feature selection. Continuous learning can lead to optimal solutions for deep learning models trained on large-scale data, resulting in higher classification accuracy. However, because of the higher complexity of deep learning models and the larger datasets required for training, they demand more calculating resources and a longer training time. Additionally, models are prone to overfitting when the data volume is insufficient.

5 Conclusion

This paper reviews the evolution of image classification technology from traditional machine learning to deep learning. Through the analysis of classical algorithms, it reveals the advantages and limitations of machine learning and deep learning methods in the image classification. Machine learning methods that are traditional, like SVM and Decision Trees, perform well on small and medium-sized datasets but are limited when facing large-scale and complex image data. With the rise of deep learning, models such as CNNs and VITs have gradually become mainstream in image classification field. Deep learning models have meaningfully enhanced the accuracy and generalization ability of image classification by automatically learning features, especially with the support of big data. However, deep learning also faces challenges, such as high computational resource requirements and potential overfitting issues.

References

1. L. Cai, J. Gao, D. Zhao, A review of the application of deep learning in medical image classification and segmentation. Ann. Transl. Med. **8**, 713 (2020).
2. S.-J. Liu, et al. Research on the classification algorithm of crop pathology images based on convolutional neural network. Hubei Agric. Sci. **60**, 131 (2021).
3. S. Divya, B. Adepu, P. Kamakshi, Image Enhancement and Classification of CIFAR-10 Using Convolutional Neural Networks, in Proceedings of the 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), (2022).
4. F.-O. Giuste, J.-C. Vizcarra, CIFAR-10 image classification using feature ensembles, in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), (2020).

5. L. Beyer, O.-J. Hénaff, A. Kolesnikov, Are We Done with ImageNet? arXiv preprint arXiv:2006.07159 (2020).
6. E.-C. Chen, C.-R. Lee, Towards Fast and Robust Adversarial Training for Image Classification, in Proceedings of the Asian Conference on Computer Vision, (2020).
7. S.-S. Kadam, A.-C. Adamuthe, A.-B. Patil, CNN Model for Image Classification on MNIST and Fashion-MNIST Dataset. Int. J. Comput. Sci. Netw. Secur. **21**, 19-24 (2021).
8. L. Wan, M. Zeiler, S. Zhang, et al. Regularization of neural networks using dropconnect, in Proceedings of the 30th International Conference on Machine Learning (ICML), (2013).
9. M. Arya, S.-S. Bedi, Survey on SVM and their application in image classification. Int. J. Inf. Technol. **13**, 1867-1877 (2021).
10. B.-T. Jijo, A.-M. Abdulazeez, Classification Based on Decision Tree Algorithm for Machine Learning, J. Appl. Sci. Technol. Trends. **1**, 56-61 (2021).
11. S. Chandrasekaran, Raman, Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review, in Proceedings of the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, (2020).
12. M.-Z. Alom, T.-M Taha, C. Yakopcic, et al. The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv preprint arXiv:1803.01164 (2018).
13. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
14. K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition, in Proceedings of the IEEE conference on computer vision and pattern recognition, (2016).
15. A. Dosovitskiy, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929 (2020).
16. A. Fred, Deep learning using Vision Transformers. Appl. Sci. **13**, 987-1005 (2022).