Biomedical Testing & Sensing
2024, VOL. 2, NO. 2, 34-37
DOI: 10.47297/wspbtsWSP2755-952107.20240202

Research on the Optimization of Image Classification Algorithm Based on deep Learning

Dou Wang

School of Management, Liaoning University of International Business and Economics, Dalian 116052, Liaoning, China

ABSTRACT
Image classification is one of the core tasks in the field of computer vision. The rapid development of deep learning provides new possibilities for improving the performance of image classification algorithm. However, the existing deep learning models still face the problems of high calculating cost and insufficient generalization ability when dealing with complex image classification tasks. This paper takes the deep learning image classification algorithm as the research object and combines mainstream models such as convolutional neural network (CNN) and Transformer to propose an optimization method based on data enhancement, model compression and attention mechanism. Through theoretical analysis and technical improvement, this paper provides a new idea for the optimization of image classification algorithm and lays a foundation for its practical application.

KEYWORDS
deep learning; image classification; convolutional neural network

In recent years, with the rapid development of deep learning, image classification has been widely used in medical imaging, autonomous driving, security monitoring and other fields. However, the high-precision classification requirements in complex scenarios put forward higher requirements for algorithms, but the existing deep learning models still have shortcomings in computational efficiency and generalization ability. Regarding these problems, this paper improves the performance and practicability of image classification algorithm by optimizing the algorithm structure, introducing new techniques and optimizing training strategies. The research in this paper not only has theoretical significance, but also provides technical support for practical application scenarios of image classification algorithms.

1 The Theoretical Basis and Technology of Image Classification Algorithm

1.1 Convolutional Neural Networks (CNN)

Convolutional neural network (CNN) constructs the classical paradigm of hierarchical feature extraction through the core mechanism of local connection, weight sharing and spatial down-sampling. Its basic unit is composed of convolutional layer, pooling layer and activation function. The convolutional layer captures local spatial features through sliding windows, and weight sharing greatly reduces the number of parameters. The pooling layer achieves feature dimension reduction and translation invariance enhancement through maximum pooling or average pooling. Nonlinear activation functions (such as ReLU) introduce model expression ability. The typical architecture evolution is demonstrated by deep expansion and structural innovation. AlexNet first verifies the effectiveness of the deep web and uses Dropout and data enhancement to mitigate overfitting. In the VGG network, stacking 3×3 small convolution kernels replaces large convolutions, which increases the depth of the network while reducing the number of parameters. ResNet introduces residual connection to break through the bottleneck of gradation disappearance and make the network depth break through thousands of layers. Recent advances focus on the balance of efficiency and performance. MobileNet uses depthwise separable convolution to reduce computation to 1/8 to 1/9 of standard convolution. Dynamic Convolution (CondConv) realizes the adaptive aggregation of multi-scale features through the dynamic fusion of multi-expert weights[1]. Neural architecture search (NAS) uses reinforcement learning or evolutionary algorithms to automate the design of efficient network structures (such as EfficientNet) to achieve the optimal trade-off between accuracy and computational resources. These innovations have driven CNN to continuously break performance records on benchmark datasets such as ImageNet, and laid the foundation for edge device deployments.

1.2 Application of Transformer in Image Classification

Transformer uses self-attention to break through the local perception limitations of CNN and realize explicit modeling of the over dependence relation. Vision Transformer (ViT) divides images into 16×16 block sequences, retains spatial information through location coding, and uses multi-head self-attention to establish cross-regional associations to

outperform CNN after pre-training of the FT-300M large-scale data set. To solve the computational complexity, Swin Transformer innovatively introduces the hierarchical architecture and sliding window mechanism. It builds a hierarchical feature pyramid by merging blocks layer by layer, reduces the complexity of computing attention in local windows from O (n^2) to O(n), and retains the overall interaction capability through cross-window connection [2]. The hybrid architecture becomes a new trend. ConvNeXt achieves the Top-1 accuracy of 87% on ImageNet-1K by replacing the standard ResNet module with Transformer-like large kernel depthwise convolution and LayerNorm. MobileViT combines lightweight CNN with local global alternating attention to achieve a balance of precision and speed on the mobile terminal. Current research focuses include geometric perception improvement of location coding, sparse attention mechanism design and cross-modal pre-training strategies, which continue to expand Transformer's application boundaries in complex scenarios such as fine-grained classification and small-sample learning [3].

1.3 Data Enhancement Technology and Model Optimization Method

Data enhancement improves model generalization ability by expanding sample diversity, and its technological evolution shows a shift from manual design to automatic learning. Basic enhancement methods include geometric transformation (random cropping and rotation), color jittering (brightness/contrast adjustment), and texture distortion (elastic transformation). The advanced enhancement strategy focuses on semantic preservation and diversity maximization. Mixup generates virtual samples through linear interpolations to enhance the smoothness of decision boundaries. CutMix replaces local image blocks with other sample areas to improve the robustness of the model to occlusion. AutoAugment uses reinforcement learning to automatically search for optimal enhancement strategy combinations on data sets. Model optimization methods include training strategies and regularization techniques. Stochastic gradient descent with momentum (SGDM) and adaptive optimizers (such as AdamW) balance convergence speed and stability. Cosine annealing learning rate scheduling alleviates the local minimum trap. Label smoothing reduces the risk of overfitting through soft one-hot encoding [4]. In the regularization techniques, Dropout randomly blocks neurons to force the online learning to express the redundant feature, and DropPath randomly drops branch paths in the residual network to enhance the robustness of model. The latest development, such as RandAugment, simplifies the search space to achieve the automatic generation of non-parametric enhancement strategies, and significantly reduce the calculation cost. Knowledge distillation transfers large model knowledge through the teacher-student framework to maintain high performance of lightweight models. The systematic combination of these techniques forms the core pillar of modern image classification training processes.

2 Optimization Method of Image Classification Algorithm

2.1 Optimization Strategy Based on Data Enhancement

Data enhancement improves the generalization ability of the model by expanding the diversity of training samples. Traditional methods rely on manual work to design geometric transformation (such as flipping and cropping) and color distortion (such as brightness adjustment), but it is difficult to cover semantic changes in complex scenes. The adaptive enhancement strategy uses automated learning to optimize directions, such as dynamic selection of enhanced operation combinations by reinforcement learning, or adjustment of distortion intensity based on image semantic features. Hybrid enhancement techniques (such as Mixup and CutMix) generate virtual samples through the linear interpolation or region substitution to force the model to learn more robust feature representations. Recent research focuses on the generative adversarial networks (GAN) and diffusion models, which use generative enhancement to synthesizes high-quality realistic samples, and effectively alleviate the category imbalance especially in small sample scenarios. These strategies significantly improve the adaptability of the model to real environmental interference such as occlusion and noise through multi-level and multi-modal data reconstruction.

2.2 Optimization Strategy Based on Model Compression

Model compression aims to reduce computational complexity to meet the needs of edge device deployment, and its core lies in the balance between accuracy and efficiency. Structured pruning dynamically removes redundant channels or layers by assessing neuronal importance, so as to preserve the ability to extract key features while reducing the scale of parameters. Quantitative technology converts the high-precision floating-point arithmetic to the low-bit integer arithmetic, and reduces the memory footprint and computation delay with dynamic range calibration. Knowledge distillation uses the soft label of teacher network to supervise the lightweight student network to realize the knowledge transfer and performance retention. The hybrid compression framework integrates pruning, quantification, and architecture search. For example, neural architecture search (NAS) automatically design the efficient network topology

and achieve the real-time reasoning ability in mobile terminals. These methods can greatly reduce the consumption of computing resources while maintaining the classification accuracy through multi-level collaborative optimization.

2.3 Optimization Strategy Based on Attention Mechanism

The attention mechanism enhances the semantic understanding ability of the model through dynamic feature weight allocation. Channel attention (such as SE module) learns the importance weights of different feature channels to highlight task-related feature expression. Spatial attention focuses on key areas to suppress background noise interference. Temporal attention models the frame dependency in video classification. Self-attention (such as Transformer) breaks through the local perception of convolution to establish global context associations, and is particularly good at capturing long-distance feature interactions. The hybrid attention network combines the inductive bias of convolution with the overall modeling benefit of attention, such as embedding cross-attention layers in the CNN backbone or introducing convolutional location coding in Transformer. Multi-scale attention enhances the model's collaborative perception of details and overall structures through parallel processing of feature maps with different resolutions. These mechanisms significantly improve the classification accuracy and interpretability in complex scenarios through collaborative optimization of feature selection and information integration.

3 Application Analysis of Image Classification Algorithm Optimization

3.1 Application in Medical Image Classification

The accuracy, robustness and interpretability of algorithms are strictly required by medical image classification. For multi-modal data such as pathological sections, X-rays and MRI, the optimization algorithm needs to solve the challenges of data scarcity, category imbalance and high labeling cost. Transfer learning reuses the low-level feature extraction ability of ImageNet pre-training models (such as ResNet-50 and DenseNet-121) and fine-tune the top-level network combined with medical data to significantly improve the generalization performance in small sample scenarios. The multiple instance learning (MIL) framework can effectively reduce local noise interference by dividing whole slide images (WSI) into multiple regional instances, using attention mechanism to identify key lesion areas (such as cancer cell clusters), and combining the graph neural network to model spatial relationships among regions. Interpretability enhancement techniques such as Grad-CAM generate heat maps, visually display the model's areas of concern, and assist doctors to verify the clinical rationality of classification results. To solve the problem of insufficient data, the generative adversarial network based on StyleGAN can generate virtual samples with pathology, adjust the shape and distribution of lesions by controlling latent variables, and expand the training data of rare cases.

3.2 Application in the Autonomous Driving Scenario

Image classification in autonomous driving scenarios requires real-time and robust scene understanding in dynamic environments. Multi-task learning frameworks (such as Multi-Task CNN) share backbone network parameters to simultaneously complete road type classification, traffic sign recognition, and obstacle detection, reducing computational redundancy and improving system efficiency. Time series modeling techniques (such as 3D convolution or Transformer encoders) capture the spatiotemporal correlation of consecutive frames, and combine optical flow estimation to predict the trajectory changes of moving objects, enhancing response to emergencies (such as pedestrians crossing the road). In view of complex conditions such as light change and rain and fog interference, the self-supervised pre-training strategy learns robust feature representation by designing proxy tasks such as rotation prediction and jigsaw restoration, and using massive unlabeled driving data. In terms of model lightweight, structured pruning removes redundant convolution kernels, combines 8-bit quantization technology to compress the model to a scale suitable for deployment of on-board edge devices such as NVIDIA Jetson, and increases inference speed to more than 30 frames per second. The scene adaptive module dynamically adjusts the classification threshold through online learning, such as reducing the false detection rate for distant vehicles in the highway scene, and increasing the sensitivity to small targets such as bicycles and scooters in the urban environment. In actual deployment, the improved hybrid architecture based on YOLOv5 realizes multi-target cooperative classification on KITTI data set, and its spatial pyramid pool module effectively extracts multi-scale features to ensure classification accuracy of targets at different distances.

3.3 Application in Security Monitoring Scenarios

Security monitoring scenarios require algorithms to achieve accurate target recognition and real-time early warning under complex lighting, occlusion and multi-view conditions. The spatial features of RGB frame and the motion feature of

optical flow sequence are extracted respectively by the spatio-temporal dual-flow network, and the spatio-temporal information is fused through the cross-modal attention mechanism to effectively capture abnormal behaviors (such as climbing over fences and leaving suspicious objects behind). In view of the high redundancy of video data, the sparse sampling strategy selects key frames (such as extracting 1 frame every 10 frames) and combines the bidirectional LSTM to complete timing context to reduce the computing load and maintain the continuity of action. Multi-modal fusion technology integrates visible, infrared, and depth sensor data to build an anti-interference classification system. Infrared data is used for human detection at night or in low light conditions, and depth information helps distinguish real targets from projection deception. By maximizing cross-view feature consistency of the same object, self-supervised contrast learning improves robustness to degraded scenes such as occlusion and blurring. In terms of privacy protection, the edge anonymization module performs Gaussian blur or regional Mosaic processing on sensitive information such as human faces and license plates before feature extraction to ensure data compliance. In actual deployment, the improved model based on SlowFast network can realize the high-precision anomaly detection on UCF-Crime data set. Its time-resolved branches capture fast motion changes, and the spatial branches analyze static scene semantics. The dual-path collaboration improves the classification reliability.

4 Conclusion

This paper studies the optimization of image classification algorithm based on deep learning, systematically analyzes the feature extraction mechanism of mainstream models such as convolutional neural network (CNN) and Transformer, and proposes optimization strategies based on data enhancement, model compression and attention mechanism. Through data enhancement, the generalization ability of the model is significantly improved and the overfitting phenomenon is reduced. Through the model compression technique, the calculation cost is reduced effectively and the practicability of the algorithm is enhanced. By introducing the attention mechanism, the feature focusing ability of the model is further improved, and the classification accuracy is improved. These optimization methods show good adaptability in different application scenarios. It is expected that future studies can further explore multi-modal data fusion, new model architecture design and more efficient learning paradigms to cope with more complex image classification tasks, and promote the widespread implementation and in-depth development of image classification technology in practical applications.

About the Author

Dou Wang, Master, Teaching Assistant, Research Direction: Information Management and Deep Learning.

References

[1] Liu Hongda, Sun Xuhui, Li Yibin, et al. Review of Deep Learning Models for Image Classification Based on Convolutional Neural Networks [J/OL]. Computer Engineering and Applications,1-29[2025-02-24].

[2] Cui Yuchen, Xie Yuandong, Wu Yumiao, et al. Research on Classification of Benign and Malignant Oral Lesions Using ViT-B Deep Learning Model [J]. Journal of Oral Science Research, 2019,41(01):16-20.

[3] Wei Zongyue, Qiu Dawei, Liu Jing, et al. Research and Progress of Deep Learning in the Diagnosis of Upper Limb Fractures [J/OL]. Journal of Frontiers of Computer Science and Technology,1-27[2025-02-24].

[4] Wang Feifei, He Xinyi. Research on Deep Learning Method of Garbage Classification Images Recognition [J]. Journal of Sanming University, 2024,41(06):48-58.