Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 12, No. 3, 2024

Multi-scale image recognition strategy based on convolutional neural network

Huajun Zhang[1, a], Su Diao[2, b], Yining Yang[3, c], Jiachen Zhong[4, d], Yafeng Yan[5, e]

[1] Syracuse University, USA
[2] Auburn University, USA
[3] Carnegie Mellon University, USA
[4] University of Washington, USA
[5] Stevens Institute of Technology, USA

[a] jameszhang1023.java@gmail.com; [b] scottdiao33@gmail.com; [c] yiningya@alumni.cmu.edu; [d] mrjiachenzhong@gmail.com;
[e] yanyafeng0105@gmail.com

Abstract: The accurate recognition and interpretation of multi-scale visual information is a critical focus within contemporary computer vision research. To this end, this study explores and innovatively constructs a multi-scale image recognition strategy based on a Convolutional Neural Network (CNN) with a multi-level and multi-resolution perception domain. This strategy is embedded with an advanced multi-level convolutional operation mechanism, which enables the model to intelligently explore and learn the multi-scale feature representation space of images from tiny texture to grand structure, from shallow simple features to deep semantic abstraction. The core technology path of this paper is to design a deep separable convolutional architecture and combine pyramid pool technology to form a unique network module. This modular design not only ensures the computational efficiency of the model but also improves the ability of extracting and integrating multi-scale image features. Following intensive experimentation on an array of extensively recognized and substantial image datasets, the multi-scale image recognition approach introduced in our study has demonstrated marked enhancements in both recognition capability and stability, manifesting clear superiority compared to conventional, single-scale image recognition methodologies. This research not only enriches the theoretical framework of image recognition, but also provides a new and efficient solution for dealing with complex multi-scale image recognition challenges in practical applications, and further promotes the development of image understanding and recognition technology.

Keywords: Convolutional neural networks; Multi-scale feature; Image recognition strategies; Computer vision.

1. Introduction

In recent years, profound strides have been witnessed in the domain of image recognition, particularly due to the groundbreaking advancements in Deep Learning techniques, most prominently Convolutional Neural Networks (CNN) [1-3]. However, in the face of diverse image content in the real world, especially target objects with multi-scale characteristics, achieving accurate and efficient recognition is still a challenge. The traditional single-scale image recognition method is limited by the fixed feature extraction range, which can not fully explore and use the multi-scale image information. Particularly in scenarios with complex backgrounds and significant variations in target size, recognition performance is severely impacted.

In this paper, an innovative strategy based on a convolutional neural network is proposed for multi-scale image recognition. The key innovation of this strategy is to construct a multi-level, multi-receptive field convolutional network structure, which combines deep separable convolutional and pyramid pooling mechanisms. The depth wise separable convolution module diminishes the model's parameter count and computational intensity by segmenting the convolution procedure into two distinct phases: channel-wise spatial filtering followed by pointwise convolutions. The pyramid pooling module generates different scale pooling outputs at different levels, covering multi-scale feature representation from fine-grained to coarse-grained. These innovations enable models to simultaneously capture rich image features across different scales of space and fuse them efficiently.

We conducted a detailed experimental evaluation on several large representative publicly available image datasets. In the design of the experiment, we carefully selected datasets that included a variety of scenes, objects, and background variations, and ensured that they covered a wide range of scales. Through these experiments, we can comprehensively evaluate the applicability and performance of the proposed strategy in different scenarios. The experimental results show that the proposed multi-scale image recognition strategy based on convolutional neural network has achieved significant improvement in recognition accuracy and robustness. Compared with traditional single-scale recognition methods, our strategy shows higher recognition accuracy and stronger robustness when processing complex and multi-scale images. Specifically, we observe that the proposed strategies can identify target objects more accurately on image data at different scales, and have better resistance to background interference and scale changes. These empirical findings not merely substantiate the practicality of the proposed methodology, but also serve to reinforce its superior capability when dealing with a broad spectrum of intricate and varied real-world image data. Therefore, we believe that this strategy has a wide application prospect and can provide strong support for further research and application in the field of learning algorithms such as Semi-supervised learning [4-5], and federated learning [6-7]. In summary, this study not only proposes a novel multi-scale

image recognition method but also has important academic and practical significance for promoting the theoretical development and practical application of image recognition technology in the field of computer vision.

2. Related work

The Convolutional Neural Network (CNN), a profoundly utilized deep learning architecture, has gained widespread adoption in various image recognition applications. Its development can be traced back to the 1980s. Among them, LeNet[8] is one of the earliest convolutional neural networks, proposed by Yann LeCun et al in 1998, which is mainly applied to handwritten digit recognition. Subsequently, AlexNet[9] achieved a breakthrough success at the ImageNet Image Recognition Challenge in 2012, ushering in a new era of deep learning in computer vision. Since then, classic CNN models such as VGG, GoogLeNet[10] and ResNet have been proposed successively, constantly improving the accuracy and generalization ability of image recognition. By simulating the information processing of the biological visual cortex [11], CNN uses convolutional layer to extract features from input images. The original convolutional layer mainly performed linear filtering operations, but with the advancement of technology, nonlinear activation functions such as ReLU were introduced, which greatly improved the learning ability and expressiveness of the model. In recent years, efficient designs such as deep separable convolution have been introduced to further optimize the number of model parameters and computational efficiency, making it more advantageous in processing large-scale image data.

Pooling technology, as an important operation in CNN, plays a key role in network design and optimization. The earliest pooling techniques can be traced back to the subsampling layer in the LeNet model. Presently, pooling techniques have bifurcated into two predominant types: maximum pooling and average pooling. The recently proposed pooling approach effectively simplifies computations, decreases parameter count, and fortifies the model's immunity to translation shifts and noise tolerance. Initially, prevalent pooling mechanisms involved sliding a fixed-size kernel over the feature map to either extract the maximum pixel value or compute the mean of pixels within each sub-region. With the deepening of research, adaptive Pooling technology comes into being. For example, RoI Pooling [12] and RoI Align [13], which are widely used in target detection, can carry out dynamic pooling according to the specific size of the target. In addition, pyramid pooling [14-17] (such as SPP-net and ASPP) allows multi-scale features to be extracted at the same level, thus strengthening the model's ability to identify objects at different scales.

With the continuous challenge and development of image recognition tasks, researchers have proposed a variety of multi-scale image recognition methods. Among them, the method based on image pyramid is a more common one in the early stage. In this method, multi-scale features are extracted and fused by constructing multi-scale pyramid of image. In addition, the method based on multi-scale network design has gradually attracted attention. These methods realize the capture and fusion of multi-scale features by introducing convolution kernel and pooling operations of different scales into the network. Multi-scale feature extraction by setting convolution kernels of different sizes in the CNN network or using hollow convolution technology, the network can capture image features from different scale levels, and realize both large and small target recognition. Multi-scale pooling and feature pyramid structure, the pyramid pooling technology can extract multi-scale feature mappings in the same layer, and the feature pyramid network (FPN) builds a multi-scale feature fusion structure across multiple levels to ensure that the model can maintain stable recognition performance at all scales.

To sum up, CNN, as the main model in the field of image recognition, has undergone continuous development and evolution in the past decades. Pooling technology, as an important part of CNN, is constantly improved with the optimization of network structure. The multi-scale image recognition method is a solution for the diversified image content in the real world, and its development is closely related to the complexity of image recognition tasks. Therefore, this paper aims to learn from the development experience of CNN and pooling technology, and propose a multi-scale image recognition strategy based on CNN to deal with the complex and changing real world image data.

3. Theoretical basis

3.1. Convolutional neural network

CNN is one of the important models in deep learning, and several classical models have emerged during its development, including LeNet5 model, AlexNet model, VGGNet model [18], GoogLeNet model and ResNet model [19]. The LeNet5 model was first proposed by American scientist Yann LeCun in 1998 and applied to convolutional neural networks for handwritten digit recognition. It is one of the most representative experimental systems in early convolutional neural networks. The LeNet5 architecture comprises several primary building blocks, including the input stage, convolutional layers, Rectified Linear Unit (ReLU) activation layers, subsampling or pooling layers. The features of the model include convolution, pooling and nonlinear processing of data, extraction of spatial features by convolution, introduction of multiple local regions, multi-scale decomposition by wavelet transform, etc. The LeNet5 model is simplified into a network consisting of four layers, each of which contains different training parameters. The structure of the model constructs a complete convolutional neural network by connecting each layer, in which the convolutional layer and the pooled layer alternately form a multi-layer neural network. The key principles of LeNet5 model include using each layer to learn the relationship between two adjacent pixels, using hyperbolic and S-type approximation methods to solve nonlinear problems, and introducing multi-layer neural networks into the classifier.

In the year 2012, Alex Krizhevsky and colleagues presented the AlexNet model, which has since become a highly impactful reference in the realm of deep learning architectures. The model achieved great success in the ImageNet Challenge, introducing deep convolutional neural networks into the image classification task for the first time and achieving significant performance improvements. In addition, AlexNet uses techniques such as Local Response Normalization and data enhancement to further improve model performance.

VGGNet was proposed by the Visual Geometry Group at the University of Oxford in 2014 and is characterized by a very concise and deep network structure. The VGGNet model uses successive 3x3 convolution kernels for convolution operations. The hallmark design of the VGGNet model

involves a succession of convolutional and pooling operations, augmented by three dense layers. The central achievement of this model is substantiating the effectiveness of profoundly layered CNNs in image classification tasks, thereby facilitating further progress in deep learning frameworks.

GoogLeNet, also known as the Inception model, was proposed by a Google research team in 2014. The model uses a structure called Inception module to process input feature maps in parallel through convolution kernel and pooling operations of different sizes, and then concatenates the results to obtain a richer feature representation.

ResNet, or Deep Residual Network, was proposed by Microsoft Research in 2015 and its innovation is the introduction of a Residual Connection mechanism. This architectural feature facilitates the network's learning of residual mappings, hence easing the acquisition of identity mappings and addressing the issues of vanishing and exploding gradients that typically arise during the training of deep neural networks. At its core, the ResNet model is composed of a chain of residual units, each incorporating a duo of convolutional layers alongside an identity shortcut connection. The model demonstrated exceptional performance in the ImageNet Challenge, solidifying its status as a landmark milestone.

The above are the early models of CNN. When introducing the theoretical knowledge of CNN below, we need to start from its basic composition and principle, and its working mode is shown in Figure 1 for a deeper understanding. The convolutional layer is one of the core components of the CNN, its role is to slide over the input image through a filter (also known as the convolution kernel) and perform convolution operations to extract features in the image. The convolution operation can be mathematically expressed as:

$$\mathbf{Z}^{[l]} = \mathbf{W}^{[l]} * \mathbf{A}^{[l-1]} + \mathbf{b}^{[l]} \tag{1}$$

One, the $(Z^{[l]})$ said the first (l) convolution output layer, $(W^{[l]})$ is A convolution kernels (weight), and $(A^{[l-1]})$ is the value of the activation of A layer, $(b^{[l]})$ is a bias. By applying the convolution operation, the convolutional layer can glean local attributes across various positions within the image, thereby accomplishing the process of feature extraction.

Ordinarily situated subsequent to the convolutional layer, the pooling layer acts to truncate the dimensions of feature representations, thereby effectuating a reduction in parameter count and computational overhead, all while bolstering the model's robustness. The most common Pooling operations [20] are Max Pooling and Average Pooling. The mathematical expression of the pooling operation is:

$$\mathbf{A}^{[l]} = \text{Pooling}(\mathbf{Z}^{[l]}) \tag{2}$$

Where $(A^{[l]})$ represents the feature graph after the pooling operation.

Amidst the sequence of convolutional and fully connected layers, activation functions are often added to introduce nonlinearity. The ReLU function (Rectified Linear Unit) is one of the most commonly used activation functions, and its mathematical expression is:

$$\text{ReLU}(z) = \max(0, z) \tag{3}$$

ReLU function is easy to calculate and does not introduce the problem of disappearing gradient, so it is widely used in practical applications. The fully connected layer customarily resides towards the terminal portion of a CNN, functioning to project the features distilled by preceding convolutional layers onto the ultimate output classes. The mathematical representation of the fully connected layer is:

$$\mathbf{Z}^{[L]} = \mathbf{W}^{[L]} \cdot \mathbf{A}^{[L-1]} + \mathbf{b}^{[L]} \tag{4}$$

One, the $(Z^{[L]})$ said the final output, $(W^{[L]})$ is the weight matrix, $(A^{[L-1]})$is A former activation values, $(b^{[L]})$ is biased. Rights are still a challenge that requires further research and exploration.

The loss function [21-23], utilized to gauge the disparity between the model's output and actual labels, is a critical component in machine learning. Notable examples among common loss functions encompass Cross-Entropy Loss and Mean Squared Error Loss. Typically, the selection of a loss function hinges upon the unique requisites of the given task.

The backpropagation procedure computes the derivative of the function concerning the model parameters and employs the gradient descent technique to iteratively refine these parameters with the aim of minimizing the loss down to its lowest possible value. By implementing the chain rule, the gradients of every parameter concerning the loss function can be computed progressively layer-wise, facilitating the optimization of the model's parameters. In the convolution layer, the parameters of the convolution kernel are shared, that is, when sliding across the entire input image, the parameters of the convolution kernel are unchanged.

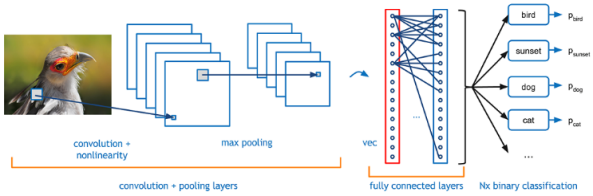

Figure 1. Workflow of CNN.

3.2. Pyramid pooingl technology

In our investigation, we introduce a novel approach to multi-scale image recognition grounded in convolutional neural network principles. One of the key innovations is the introduction of pyramid pool technology, whose structure is shown in Figure 2. The purpose of pyramid pooling methodology is to adeptly seize multi-scale image attributes and facilitate a versatile integration of these features, thereby augmenting the overall effectiveness and resilience in image recognition processes. Pyramid pooling is a multi-scale feature extraction method, which obtains multi-scale feature representation through pooling operations at different scales. The basic idea is to pool feature maps at different scales at different levels of the network, thus covering multi-scale information from fine-grained to coarse-grained.

Suppose we have an input feature graph (X) with dimensions $(H \times W)$, where (H) represents height and (W) represents width. Regarding pyramid pooling, we can execute pooling operations across various scales to derive an ensemble of feature representations with varying dimensions. The expression for the max pyramid pooling operation may be articulated as:

$$Y_{i,j}^{(k)} = \max_{(p,q)\in R_{i,j}^{(k)}} X_{p,q} \tag{5}$$

The $(Y_{i,j}^{(k)})$ said in the first (k) a scales characteristics after the pooling of coordinates for $((i,j))$ of pixels, $(R_{i,j}^{(k)})$said in pixels $((i,j))$ as the center of the pool area, $(X_{p,q})$ represents the pixel value in the input feature plot with coordinates $((p,q))$. The average pyramiding operation can be expressed as:

$$Y_{i,j}^{(k)} = \frac{1}{|R_{i,j}^{(k)}|} \sum_{(p,q)\in R_{i,j}^{(k)}} X_{p,q} \quad (6)$$

Herein, $(|R_{i,j}^{(k)}|)$ denotes the dimensionality of the pooling region, referring to the count of pixels involved.

In practical application, we can construct the feature pyramid by applying the pyramid pooling operation on the feature map of different levels. Through feature pyramid, we can obtain multi-scale feature representation of images at different scales, thus improving the accuracy and robustness of image recognition.

The optimization and application of pyramid pooling technology involves the selection and adjustment of parameters such as the size of the pooling area, the number of scales and the pooling method, as well as the collaborative design and training with other network components. In this paper, we optimize and apply the pyramid pooling technique according to the characteristics of actual data sets and tasks, and verify its effectiveness and performance in experiments. Through the above analysis, pyramid pool technology, as an important part of multi-scale image recognition strategy, is of great significance in extracting multi-scale image features, and has important application prospects for solving multi-scale image recognition problems in the real world.

In this research, we have implemented tactics akin to Spatial Pyramid Pooling and Atrous Spatial Pyramid Pooling methodologies. These techniques build a bottom-up and top-down feature pyramid structure in deep learning models, enabling the network to perform multi-scale, multi-level analysis of image content. Thus, by means of this method, the benefits of multi-scale pooling can be optimally harnessed without altering the native dimensions of the input image, ultimately enhancing the performance of multi-scale image recognition tasks.

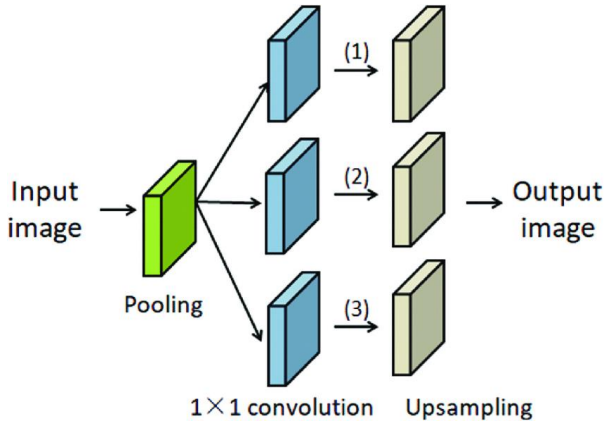

Figure 2. Pyramid pooling structure.

4. Multi-scale image recognition model based on convolutional neural network and pyramid pool

In this research endeavor, we put forth a novel Multi-Scale Image Recognition Model referred to as "Multi-Scale Convolutional Neural Network incorporating Pyramid Pooling" (abbreviated as MSCNN-PP), built upon the foundations of Convolutional Neural Networks and Pyramid Pooling mechanisms. The purpose of this model lies in proficiently capturing and leveraging the multi-scale image information to elevate the precision and robustness of image recognition. The procedural workflow of this model can be dissected into several essential stages as follows:

First, the input image passes through a series of convolution layers, each of which includes multiple convolution cores, responsible for extracting features at different levels of the image. Each convolutional layer is then processed by activation function, which increases the nonlinear fitting ability of the network. The outputs of these convolutional layers will serve as inputs to the pyramidal pools. The convolutional layer is responsible for extracting the features of the image and converting the input image into a series of feature maps through the convolution operation. Set the first l layer of convolution kernels to $(W^{(l)})$, bias for $(b^{(l)})$, is the first l the output characteristics of layer figure $(X^{(l)})$ can be represented as:

$$X^{(l)} = \sigma(W^{(l)} * X^{(l-1)} + b^{(l)}) \quad (7)$$

Wherein, (σ) signifies the activation function, while $(*)$ symbolizes the convolutional operation.

Secondly, the MSCNN-PP architecture incorporates multiple convolutional tiers, pyramid pooling stages, and fully-connected layers, with the pyramid pooling component representing a distinctive novelty introduced in this study, and pyramid pooling layer is the key component of the MSCNN-PP model, which carries out pyramid pooling operations on the input feature maps at different scales. More precisely, the pyramid pooling module performs pooling operations at varying scales on the input feature map, thereby extracting feature information across different levels and skillfully integrating these diverse scale feature details. The pyramid pooling layer is used for multi-scale feature extraction to capture multi-scale information of images by applying pooling operations at different scales. Set the first (l) layer pyramid pooling operation to $(P^{(l)})$, the resultant feature representation stemming from the pyramid pooling layer, denoted as $(Y^{(l)})$, can mathematically be expressed as:

$$Y^{(l)} = P^{(l)}(X^{(l)}) \quad (8)$$

Afterward, the merged feature map progresses to the fully connected layer, where it undergoes classification or regression based on the extracted features. This layer utilizes a softmax function to transform the feature map into a final output suitable for classification predictions, or it may directly yield the forecasted result. Ultimately, the fully connected layer serves to classify or regress the extracted features, converting them into conclusive output. Set the weight for (L) layer $(W^{(L)})$, bias for $(b^{(L)})$, is the final output results (O) can be represented as:

$$O = \text{softmax}(W^{(L)} \cdot Y^{(L)} + b^{(L)}) \quad (9)$$

In the course of the model training process, this study utilizes the backpropagation algorithm in tandem with gradient descent optimization methods to fine-tune the model parameters. This refinement involves minimizing the loss function, thereby aligning the model outputs increasingly closer to their corresponding ground truth labels. Common optimizers can be used in the optimization process. Finally, after model training is completed, we can use MSCNN-PP model to recognize new images. The image to be recognized is input into the model, and after a series of convolution and pyramid pool processing, the model will output the classification result of the image, thus completing the image recognition task. To sum up, the operation steps of the MSCNN-PP model include convolutional feature extraction, pyramid pooling, full connection classification, and model training and optimization. Through these steps, the model can

effectively identify images and output corresponding prediction results. As a novel multi-scale image recognition model, the MSCNN-PP model is a novel multi-scale image recognition model. Combined with convolutional neural network and pyramid pool technology, it has a good theoretical basis and practical application prospect.

5. Experimental analysis

5.1. Data set

In the empirical phase of our investigation, we opted for the MSCOCO dataset, known colloquially as Microsoft Common Objects in Context, as the core platform to examine the effectiveness of our innovated MSCNN-PP model when confronted with multi-scale image recognition tasks. Widely adopted for image comprehension and various computer vision applications, the MSCOCO dataset [24] is extensive in scope. Created by Microsoft Research, the dataset contains a rich variety of images of real-world scenes, each accompanied by detailed annotated information. The MSCOCO dataset encompasses over a hundred thousand snapshots of genuine settings, featuring a vast array of scenes and objects such as humans, fauna, transportation modes, interior environments, and an assortment of additional subjects. These images have high quality and rich visual information and are suitable for various image understanding tasks. Each image is equipped with detailed annotation information, including the location, category, and attributes of the objects in the image. In addition, the MSCOCO dataset also provides the annotation of image caption, that is, the natural language description of the image content. This annotation information is important for image understanding and natural language processing tasks. The MSCOCO dataset is prevalently employed for assessing and investigating a diverse range of computer vision undertakings, which includes tasks like visual comprehension, image synthesis, object localization, and caption generation. Due to its rich and varied image content and detailed annotation information, MSCOCO data integration serves as an important benchmark for evaluating model performance and driving research progress.

Data preprocessing is a very important part in the experiment, it is very important to ensure that the training and evaluation of the model has a good foundation and reliability. Therefore, in the data preprocessing phase, we take the following steps to process the MSCOCO dataset: First, we load and resize the images in the MSCOCO dataset. Loading the original image is to obtain the pixel information and annotation information of the image for subsequent processing and analysis. Resizing images is to adjust all images to a uniform size for easy model input. This step ensures that the model can handle images of different sizes and increases training efficiency. Secondly, we process the annotation information of the image. This includes parsing the annotation information of the image, extracting the location bounding box and category information of the object, and further processing as needed, such as screening specific categories of objects or expanding the annotation information. The purpose of processing labeling information is to extract useful information for our research tasks and provide accurate labeling for model training and evaluation. Finally, we use data enhancement techniques to enhance the data set.

5.2. Evaluation indicators

The classification task evaluation index used in this paper is calculated based on confusion matrix. The confusion matrix divides the sample into four categories: Correct Positives (CP/TP), Correct Negatives (CN/TN), Misclassified Positives (MP/FP), and Missed Positives (NP/FN) – these metrics play a pivotal role in assessing and refining the performance of the Government Information Hybrid Classification Model (GIHCM). CP indicates the count of positively-labeled instances correctly identified, CN reflects the accurate classification of negatively-labeled instances, MP signifies the number of negatives misjudged as positives, and NP represents the positive samples mistakenly categorized as negatives. Collectively, these metrics construct the confusion matrix for the classification task, offering insight into the model's predictive accuracy across different classes. Thorough analysis of the confusion matrix enables a more holistic evaluation of the classification model's capabilities, with the corresponding category classifications detailed in Table 1.

Table 1. Confusion Matrix Of Samples

	Positive sample	**Negative sample**
Positive	TP	FN
Negative	FP	TN

When gauging model proficiency, evaluative metrics like Accuracy and F1 Score frequently serve as barometers for measuring model competence. Accuracy, essentially, measures the ratio of accurately predicted instances relative to the entire sample population, serving as a broadly accepted benchmark in evaluating classification models' performance. Its computation typically adheres to the following formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{10}$$

The F1 Score constitutes a balanced measure that integrates both Precision and Recall, reflecting the dual aspects of model exactness and inclusiveness. This metric is computed through the application of the harmonic mean, employing the following formula:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

Precision reflects the percentage of instances labeled as positive by the model that truly are positive cases, whereas Recall calculates the extent to which the model accurately identifies actual positive cases within the dataset. The F1 Score varies from 0 to 1, with higher values signifying better model performance, ideally approaching 1. These dual metrics offer a holistic perspective when assessing the model's classification performance. Accuracy focuses on the proportion of the model's overall predictions that are correct, while the F1 score takes into account both the accuracy and recall of the model, which better reflects the model's performance in different categories.

5.3. Experimental setup

In the experimental setting part, this paper chooses the internationally recognized Microsoft Common Objects in Context (MS COCO) dataset as the benchmark platform. The efficacy of the abbreviated MSCNN-PP model, which is rooted in a convolutional neural network and pyramid pooling, has been thoroughly scrutinized for its multi-scale image recognition capabilities. According to the standard partitioning rules of MS COCO, to guarantee the autonomy

and fairness of model training, calibration, and assessment, the dataset has been judiciously partitioned into three subsets consisting of 80% for training, 10% dedicated to validation, and another 10% reserved for testing purposes. In the process of model construction, we designed a network structure integrating deep separable convolution and pyramid pool technology, and adopted SGD optimization algorithm and learning rate attenuation strategy for training, while taking into account key factors. Regarding the assessment of efficiency and effectiveness, accuracy and F1 score are taken as the core indexes to measure the comprehensive performance and robustness of the model on multi-scale and multi-state image recognition tasks. In order to further confirm the superiority of MSCNN-PP model, we arranged a comparison experiment with the current mainstream image recognition methods and single-scale convolutional neural network, and obtained stable and reliable result data through repeated experiments to ensure the reliability and universality of the conclusion. Through this series of rigorous experimental setup and implementation, the outstanding performance and technological innovation value of MSCNN-PP model in solving multi-scale image recognition problems are demonstrated scientifically and intuitively.

5.4. Experimental result

The tabular presentation of contrasting foundational models' experimental findings is depicted in Table 2, and the MSCNN-PP model shows significant advantages in image recognition tasks. Specifically, the accuracy rate reached 97.15% and the F1 score reached 95.69%, indicating the model has a higher degree of accuracy and precision in image classification and recognition. In contrast, other mainstream image recognition methods, including MobileNet, CNN and DenseNet, have also achieved good results, but are slightly inferior to the MSCNN-PP model in accuracy and F1 scores. The accuracy of MobileNet model is 95.12% and F1 score is 92.99%. The accuracy of CNN model is 96.38%, F1 score is 93.12%; The accuracy of DenseNet model is 90.24%, and the F1 score is 90.87%. In particular, while DenseNet was slightly ahead of the MSCNN-PP model in F1 scores, its accuracy was significantly lower than MSCNN-PP and other models. This result fully verifies the effectiveness and practicability of MSCNN-PP model in multi-scale image recognition tasks. Consequently, the MSCNN-PP model presents a promising and expansive application potential within the realm of image recognition, furnishing a potent remedy for addressing image recognition challenges across diverse real-world contexts.

Table 2. Experimental results under different sets

MODEL	ACC	F1
MSCNN-PP	97.15	95.69
MobileNet	95.12	92.99
CNN	96.38	93.12
DenseNet	90.24	90.87

5.5. Ablation experiment

The outcomes from the ablation study are presented in Table 3, which provides valuable information on each component of the MSCNN-PP model. In the experiment, two variations were considered: the MSCNN variant devoid of pyramid pooling (referred to simply as MSCNN), and another configuration excluding the CNN structure (termed MS-PP). Specifically, in the MSCNN-PP model, the pyramid pooling component was removed to create the MSCNN version, preserving solely the convolutional neural network (CNN) framework. The empirical findings revealed that the MSCNN model achieved an accuracy rate of 90.24% and an F1 score of 91.85%. This shows that the performance of the model in the image recognition task decreases after the pyramid pooling technique is removed. Employing the pyramid pooling technique plays a pivotal role in the procurement and integration of features across multiple scales, thereby substantially enhancing the model's precision and resilience. The MS-PP model removes the CNN structure in the MSCNN-PP model and retains the pyramid pool technology. According to the experimental outcomes, the MS-PP model exhibits an accuracy rate of 94.32%, coupled with an F1 score amounting to 88.21%. This shows that the performance of the model in image recognition tasks also decreases after the CNN structure is removed. As a core component of image recognition, CNN can effectively extract local and global features of images, and provide powerful feature representation capabilities for models.

In summary, the results of ablation experiments further verify the importance and effectiveness of pyramid pooling technology and CNN structure in MSCNN-PP model. Pyramid pool technology can improve the model's perception ability of multi-scale features, and CNN structure can extract rich image features. Only by working together can MSCNN-PP model achieve superior performance in image recognition tasks. These findings provide important references for further understanding and improving multi-scale image recognition models.

Table 3. Ablation results

MODEL	ACC	F1
MSCNN-PP	97.15	95.69
MSCNN	90.24	91.85
MS-PP	94.32	88.21

6. Conclusion

After experimental evaluation of several large representative public image datasets, we found that the MSCNN-PP model achieved significant improvements in recognition accuracy and robustness. Compared with traditional single-scale image recognition methods, the MSCNN-PP model shows better performance in processing complex and multi-scale images. Specifically, we observe that the model achieves excellent recognition accuracy on the data set and is robust against various complex situations. Through the analysis of experimental results, we find that the superiority of the MSCNN-PP model is mainly due to the application of pyramid pool technology. The pyramid-pooling module efficiently retrieves and harnesses the multi-scale information inherent in images, thus empowering the model to perform a comprehensive and adaptable extraction of image features across a diverse range of scales. Compared with traditional single-scale pooling, pyramid pooling can better adapt to the multi-scale changes of images, thus improving the recognition performance and generalization ability of the model. To put it concisely, when employing the MSCNN-PP architecture for visual recognition tasks, we deduce that this approach significantly enhances both the precision and resilience of the recognition process, especially in the processing of complex and multi-scale images. This conclusion holds significant value in promoting the development of image recognition technology for practical applications. It also provides robust support for further

research and application in the field of multi-scale image recognition.

References

[1] O'shea, Keiron, and Ryan Nash. "An introduction to convolutional neural networks." arXiv preprint arXiv:1511.08458 (2015).

[2] He, W., Vu, M. N., Jiang, Z., & Thai, M. T. (2022, December). An explainer for temporal graph neural networks. In GLOBECOM 2022-2022 IEEE Global Communications Conference (pp. 6384-6389). IEEE.

[3] Li, K., Zhu, A., Zhou, W., Zhao, P., Song, J., & Liu, J. (2024). Utilizing Deep Learning to Optimize Software Development Processes. arXiv preprint arXiv:2404.13630."

[4] He, W., & Jiang, Z. (2020). Semi-supervised learning with the em algorithm: A comparative study between unstructured and structured prediction. IEEE Transactions on Knowledge and Data Engineering, 34(6), 2912-2920.

[5] Ning, Q., Zheng, W., Xu, H., Zhu, A., Li, T., Cheng, Y., ... & Wang, K. (2022). Rapid segmentation and sensitive analysis of CRP with paper-based microfluidic device using machine learning. Analytical and Bioanalytical Chemistry, 414(13), 3959-3970.

[6] Lan, G., Liu, X. Y., Zhang, Y., & Wang, X. (2023). Communication-efficient federated learning for resource-constrained edge devices. IEEE Transactions on Machine Learning in Communications and Networking.

[7] Lan, G., Han, D. J., Hashemi, A., Aggarwal, V., & Brinton, C. G. (2024). Asynchronous Federated Reinforcement Learning with Policy Gradient Updates: Algorithm Design and Convergence Analysis. arXiv preprint arXiv:2404.08003.

[8] Al-Jawfi, Rashad. "Handwriting Arabic character recognition LeNet using neural network." Int. Arab J. Inf. Technol. 6.3 (2009): 304-309.

[9] Iandola, Forrest N., et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size." arXiv preprint arXiv:1602.07360 (2016).

[10] Khan, Riaz Ullah, Xiaosong Zhang, and Rajesh Kumar. "Analysis of ResNet and GoogleNet models for malware detection." Journal of Computer Virology and Hacking Techniques 15 (2019): 29-37.

[11] Zhu, A., Li, J., & Lu, C. (2021). Pseudo view representation learning for monocular RGB-D human pose and shape estimation. IEEE Signal Processing Letters, 29, 712-716.

[12] Sun, Yuxuan, et al. "Roi pooled correlation filters for visual tracking." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[13] Gong, Tao, et al. "Temporal ROI align for video object recognition." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 2. 2021.

[14] Martinel, Niki, Gian Luca Foresti, and Christian Micheloni. "Deep pyramidal pooling with attention for person re-identification." IEEE Transactions on Image Processing 29 (2020): 7306-7316.

[15] He, W., & Jiang, Z. (2023). A survey on uncertainty quantification methods for deep neural networks: An uncertainty source perspective. arXiv preprint arXiv:2302.13425.

[16] Lan, G., Wang, H., Anderson, J., Brinton, C., & Aggarwal, V. (2024). Improved Communication Efficiency in Federated Natural Policy Gradient via ADMM-based Gradient Updates. Advances in Neural Information Processing Systems, 36.

[17] Zhu, A., Li, K., Wu, T., Zhao, P., Zhou, W., & Hong, B. (2024). Cross-Task Multi-Branch Vision Transformer for Facial Expression and Mask Wearing Classification. arXiv preprint arXiv:2404.14606.

[18] Wang, Limin, et al. "Places205-vggnet models for scene recognition." arXiv preprint arXiv:1508.01667 (2015).

[19] Targ, Sasha, Diogo Almeida, and Kevin Lyman. "Resnet in resnet: Generalizing residual architectures." arXiv preprint arXiv:1603.08029 (2016).

[20] Chen, Jiacheng, et al. "Learning the best pooling strategy for visual semantic embedding." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[21] Barron, Jonathan T. "A general and adaptive robust loss function." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[22] Christoffersen, Peter, and Kris Jacobs. "The importance of the loss function in option valuation." Journal of Financial Economics 72.2 (2004): 291-318.

[23] Spiring, Fred A. "The reflected normal loss function." Canadian journal of statistics 21.3 (1993): 321-330.

[24] Vinyals, Oriol, et al. "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge." IEEE transactions on pattern analysis and machine intelligence 39.4 (2016): 652-663.