Multi Scale Convolutional Fusion Network for Image Retrieval

Faraein Aeini*

Abstract–Image retrieval from large-scale databases poses a significant challenge in computer vision due to the limitations of traditional text-based and content-based methods in capturing the full spectrum of visual features, often resulting in a "semantic gap." A novel approach, called the Multi-Scale Convolutional Fusion Network (MSCFNet), has been proposed to enhance both accuracy and efficiency in image retrieval by utilizing multi-scale convolutional layers. MSCFNet leverages filters of various sizes to simultaneously extract fine, medium, and large-scale features, providing a richer and more comprehensive representation of images. This approach enables better detection of diverse patterns and visual details, significantly improving image matching and retrieval performance. Additionally, MSCFNet reduces model complexity by employing the "addition" operation for feature fusion, maintaining computational efficiency without increasing the dimensionality of the feature maps. MSCFNet is implemented in two variants, one with 2 multi-scale layers and another with 4 layers, and evaluated across three benchmark datasets: CIFAR-10, CIFAR-100, and Fashion-MNIST. The results show that MSCFNet consistently outperforms more complex models such as ResNet18 and ResNet50, achieving up to 74.43% accuracy on CIFAR-10, 38.87% on the more challenging CIFAR-100, and 92.47% on Fashion-MNIST. Furthermore, MSCFNet significantly reduces the number of parameters and training time, with the 2-layer version achieving a training time of just 113.1 seconds on CIFAR-10 while maintaining high accuracy. The 4-layer version demonstrates superior performance, improving accuracy and F-Score across all datasets. MSCFNet's ability to balance accuracy, computational efficiency, and reduced complexity makes it well-suited for deployment in resource-constrained environments.

Keywords: Image Retrieval, Feature Extraction, Deep Learning, Parallel Filters,

1. Introduction

Image retrieval is one of the most critical and widely used domains in artificial intelligence and image processing, having gained significant importance due to the rapid growth of visual data in recent decades [1]. This technology refers to the process where specific algorithms are utilized to retrieve similar or related images from a large database. Image retrieval has numerous applications across various fields, including medicine, security, identity recognition, and web image search [2-4]. However, challenges such as variations in lighting conditions, viewing angles, and image noise have necessitated the development of more advanced and precise methods in this area [5, 6].

Artificial Neural Networks, particularly Convolutional Neural Networks (CNNs), have emerged as one of the most powerful tools for addressing complex problems in image processing and retrieval. CNNs have rapidly established themselves among other machine learning methods due to their high capability in extracting complex and functional features from images. These networks extract local features of images, such as edges, patterns, and textures, through convolutional layers, thereby transforming images into compact and meaningful representations that can be utilized for image retrieval, object recognition, and classification [7, 8].

Feature extraction in CNNs is of paramount importance as these features directly determine the network's final performance in various tasks. The more accurate and comprehensive the extracted features, the more precisely the network can analyze images and deliver better results in image retrieval. Therefore, optimizing the feature extraction process through better design of convolutional layers and the use of image preprocessing techniques can have a significant impact on the network's efficiency and accuracy [9, 10].

In this paper, a novel approach named "Multi-Scale

*** Corresponding Author:** Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran. Email:aeini@iausari.ac.ir

Received: 2024.09.07; Accepted: 2024.10.06

Convolutional Fusion Network" (MSCFNet) is proposed to enhance the feature extraction process in CNNs. Instead of using filters of the same size in each layer, MSCFNet employs filters of different sizes, i.e., 3×3, 5×5, and 7×7, in parallel within a single layer. This approach allows the network to simultaneously extract features relevant to fine, medium, and large scales from images. Unlike traditional methods where uniform filters may limit the network's ability to recognize multi-scale features, MSCFNet enhances the network's capability in detecting and extracting richer and more diverse features by leveraging multi-scale filters.

The main innovation of this method lies in the simultaneous use of multi-scale filters within a single layer, enabling the network to respond more accurately to different image features. This approach not only increases the accuracy of the network in retrieving complex images but also reduces dependency on specific scales, enhancing the network's flexibility in processing images under various conditions. Additionally, this method can be used as a preprocessing stage alongside other advanced techniques to improve the overall performance of the network.

Compared to baseline methods, the proposed approach, due to the use of parallel filters with varying sizes, is capable of extracting features that might be overlooked by traditional methods. For example, in a complex image, patterns and details appearing at different scales can be simultaneously identified and extracted by multi-scale filters. This not only increases the accuracy of image retrieval but also enables the identification of more complex and multidimensional patterns.

In summary, the proposed method offers multiple advantages, including increased accuracy in image retrieval, reduced scale dependency, and enhanced network flexibility in various imaging conditions. These benefits, along with the introduced innovations, make this method a powerful tool for improving the performance of CNNs in image retrieval applications. The remainder of this paper will detail the proposed method, evaluate it on various datasets, and compare it with existing methods.

The second section of this paper explores related work, providing a comprehensive review of existing methods and their limitations. The third section introduces the objectives of the proposed method, detailing the structure and processes of MSCFNet. Section 4 focuses on the evaluation of the proposed model, while Section 5 presents the results of the experiments and simulations conducted on benchmark datasets. Finally, Section 6 concludes the paper, summarizing key findings and outlining potential avenues for future research.

2. Related Work

Image retrieval has emerged as one of the key areas in image processing and artificial intelligence, attracting significant attention. This technology, which involves searching for related or similar images from large databases, faces challenges such as variations in lighting conditions, viewing angles, and noise present in the images. Therefore, improving existing methods to enhance the accuracy and efficiency of this technology has always been a primary goal for researchers.

Convolutional Neural Networks (CNNs) have become the gold standard for solving various image retrieval problems due to their remarkable ability to extract complex features from images. CNNs are composed of multiple layers, each responsible for extracting different features from the input image. Each layer typically includes filters of specific sizes that are employed to recognize certain patterns and features within the image. Early architectures, such as LeNet [11] and AlexNet [12], achieved considerable accuracy in image recognition and classification by utilizing multiple convolutional layers with fixed filters. These methods were particularly effective in identifying simple features such as edges and basic textures, but they encountered limitations when faced with more complex images and the need for multi-scale feature extraction.

With advancements in deep learning techniques, researchers have found that utilizing filters of different sizes in various network layers can facilitate the extraction of more diverse and precise features. In this regard, the VGGNet architecture [13] marked a significant advancement. By using fine filters (such as 3x3) in deeper layers, VGGNet achieved notable results in various image processing tasks. However, while VGGNet succeeded in recognizing complex features, it still required sequential processing across layers, which did not fully support simultaneous multi-scale feature extraction.

One of the most important developments in multi-scale feature extraction was the introduction of the Inception architecture by Google. For the first time in the GoogLeNet network, this architecture employed multiple filters of varying sizes in parallel within a single layer [14]. This approach allowed the network to extract both local and global features from the image simultaneously, resulting in higher accuracy and efficiency in tasks such as object detection and image retrieval. The Inception architecture, through the combination of different-sized filters, significantly improved the performance of CNNs. However, this architecture also introduced challenges, such as

structural complexity and the need for precise parameter tuning [15]. The complexity of these structures made their implementation and optimization more difficult, particularly in large-scale and practical applications.

As CNNs continued to evolve, researchers sought ways to increase the depth of networks and improve their performance in image recognition and classification. This led to the introduction of the ResNet architecture, which, by adding skip connections, enabled the training of much deeper layers [16]. These skip connections allowed the network to mitigate issues such as gradient instability and saturation, ensuring improved stability and accuracy. However, the primary focus of ResNet was on increasing network depth, without specifically addressing the challenge of simultaneous multi-scale feature extraction.

To address the limitations of traditional image retrieval methods, new approaches have been proposed to reduce the semantic gap between image features and human perception. For example, deep neural network (DNN)-based approaches for saliency prediction and the use of image quality assessment (IQA) algorithms have been introduced to select high-quality and salient regions. These methods aim to mimic human visual perception by combining these regions and applying constraint-based metrics [17].

Moreover, deep learning techniques in content-based image retrieval (CBIR) have been extensively explored [18]. These techniques automatically extract more complex features from images, resulting in significant improvements in image retrieval performance. Recently, there has been growing interest in the development of image retrieval methods using hybrid models that combine visual and textual features [19]. These approaches have proven particularly effective in specialized applications such as person re-identification, product retrieval in e-commerce, and image retrieval in scientific and historical domains [20].

To further address the challenges posed by the semantic gap and improve image retrieval efficiency, deep learning and hashing-based methods have been introduced. Hashing, as an efficient method for nearest-neighbor search in large-scale data, transforms high-dimensional descriptive features into low-dimensional Hamming space. While this method allows for faster retrieval on a large scale, it may result in reduced retrieval accuracy compared to traditional methods. Consequently, novel methods, such as multi-view hashing models leveraging deep neural networks, have been proposed to preserve diverse data features and enhance retrieval accuracy [21].

Other approaches, employing deep learning and adaptive weight fusion, have been developed to preserve privacy in encrypted image retrieval. By combining low-level and high-level features, and using dimensionality reduction and local hashing techniques, these methods improve the efficiency and security of image retrieval in cloud environments [22].

In the field of remote sensing images, content-based image retrieval (CBRSIR) has also attracted significant attention. In this context, deep hashing models using convolutional neural networks and adversarial learning have been employed to learn compact features and map them to compressed hash codes, improving the accuracy and efficiency of large-scale image retrieval [23, 24].

Despite the significant advancements introduced by models like VGGNet, Inception, and ResNet, each of these architectures presents certain limitations, particularly in terms of multi-scale feature extraction, computational efficiency, and model complexity. VGGNet, for instance, relies heavily on sequential processing, with fixed filter sizes across deeper layers. While effective in recognizing basic features, VGGNet's sequential nature limits its ability to simultaneously capture multi-scale features within a single layer. As a result, VGGNet may miss important visual information, especially in images where critical features appear at different scales. In contrast, MSCFNet leverages multi-scale convolutional layers, enabling the extraction of diverse features (fine, medium, and large) at the same time, which significantly improves its ability to handle more complex visual patterns. By applying parallel filters of varying sizes within each layer, MSCFNet provides a more comprehensive feature set without the need for a deeper network.

The Inception architecture introduced a similar multi-scale strategy by using filters of different sizes in parallel, allowing for both local and global feature extraction. However, the structural complexity of Inception increases the demand for parameter tuning and results in a heavier computational load, particularly due to its multi-branch design. This makes its implementation challenging, especially in environments with limited computational resources. MSCFNet improves upon Inception by using a simplified architecture that achieves comparable or better results through parallel filter application combined with the addition operation to merge features. This streamlined design reduces the need for tuning complex branches while maintaining multi-scale feature extraction, resulting in a model that is lighter and more computationally efficient.

ResNet, another important advancement, introduced skip connections that allowed for the training of extremely deep networks, which solved issues like the vanishing gradient problem. While ResNet is highly effective for deep hierarchical feature extraction, its reliance on depth comes

at the cost of increased training time and a higher risk of overfitting on smaller datasets. Furthermore, ResNet's architecture does not explicitly address multi-scale feature extraction, meaning its performance might not be optimal in cases where diverse feature scales are important. MSCFNet, on the other hand, mitigates this limitation by focusing on multi-scale feature extraction with fewer layers. By using filters of different sizes within a shallower architecture, MSCFNet achieves higher accuracy and faster training times compared to deeper networks like ResNet50, as evidenced by the experimental results. This balance between depth and scale makes MSCFNet more suitable for resource-constrained environments without sacrificing performance.

The choice to use the addition operation instead of concatenation in MSCFNet also contributes to its computational efficiency. While concatenation increases the dimensionality of feature maps, leading to higher memory and processing requirements, addition maintains a constant feature map size. This not only preserves the integrity of multi-scale features but also reduces the overall computational load, which is critical in real-time or large-scale applications. Compared to Inception, which uses a more complex merging mechanism, MSCFNet's addition operation simplifies the integration of features across different scales, resulting in a more streamlined and faster model.

In summary, MSCFNet directly addresses the limitations of prior models by introducing a multi-scale, lightweight architecture that balances feature richness, computational efficiency, and model simplicity. It outperforms more complex architectures like ResNet and Inception, particularly in scenarios where computational resources are limited or real-time performance is essential. These innovations in MSCFNet provide a more adaptable and scalable solution for modern image retrieval tasks, positioning it as a strong alternative to existing deep learning models. In the next section, we will describe the architecture and methodology of MSCFNet in detail, demonstrating how this model leverages its multi-scale capabilities to outperform traditional CNN architectures in accuracy, training time, and computational cost.

3. Proposed Method

In this section, a novel method for multi-scale feature extraction in Convolutional Neural Networks (CNNs) is introduced, aiming to enhance the accuracy and efficiency of image retrieval. Traditional methods typically use filters of the same size in each layer of the network, which may result in the loss of valuable information at different image scales. Therefore, a new approach is proposed in which MSCFNet utilizes multi-scale filters to improve the network's ability to extract comprehensive and meaningful features from images. Figure 1 illustrates the flowchart of the proposed MSCFNet architecture. The diagram demonstrates the process of feature extraction using multiple convolutional layers with varying filter sizes (3×3, 5×5, and 7×7), which operate in parallel. The outputs from these layers are then merged through an addition operation to create a unified feature map. This feature map is subsequently passed through a Max Pooling layer to reduce dimensionality, resulting in a final feature map that can be utilized for image classification or retrieval tasks. The multi-scale approach employed by MSCFNet allows for the simultaneous capture of both fine and coarse features, improving accuracy and computational efficiency.

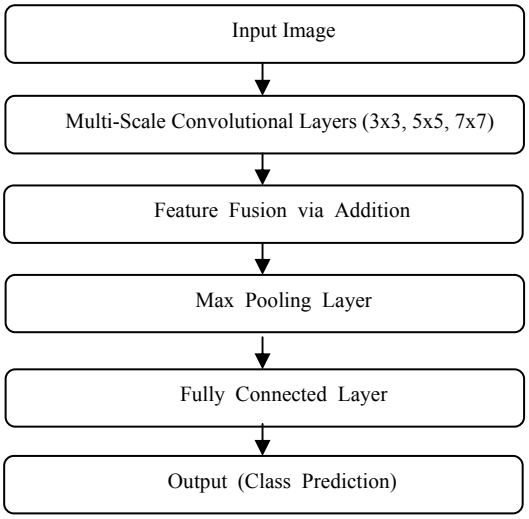

Fig. 1.A diagram showing the architecture of MSCFNet.

The proposed neural network, MSCFNet, consists of multiple convolutional layers, where filters of varying sizes are employed in each layer. These filters are arranged in parallel within a single layer and are applied simultaneously to the input image. In this way, as shown in Figure 2, each filter extracts features corresponding to different image scales, and the results from these filters are subsequently combined to create a comprehensive and multi-scale representation of the image.

- Fine filters (3x3): These filters are suitable for extracting fine details and small edges within the image, such as sharp boundaries and precise details. Fine filters are particularly effective for recognizing small local patterns, which may be more critical in the initial layers of the network.

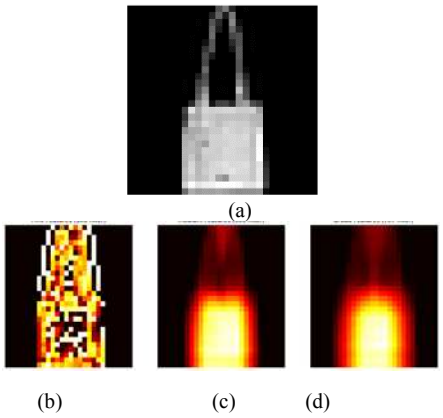

Fig. 2.a) Original Image from Fashion MNIST. b) Fine Features (3×3). C) Medium Features (5x5 filter). D) Large Features (7x7 filter)

• Medium filters (5x5): Medium-sized filters are used to capture larger patterns and more complex textures within the image. These filters allow the network to identify mid-scale features that may arise from the combination of multiple smaller patterns.

• Large filters (7x7): Large filters are employed to identify broader features and larger structures within the image. These filters are capable of recognizing large-scale patterns and shapes, which are typically used in the later stages of the network for final classification tasks.

In the next stage, instead of processing the features extracted by each filter separately, these features are combined using an "addition" operation. By using the "addition" method to merge the features extracted from different filters, MSCFNet can achieve an integrated representation of the image that simultaneously covers all scales. This process is carried out by summing the values at each pixel position from the feature maps generated by each filter. The result of this summation is a unified feature map that encompasses information from all image scales. This final feature map is then passed to the subsequent layers of the network for further processing. This process is repeated in each layer that utilizes multi-scale filters, ensuring that the network retains multi-scale information throughout the entire processing pipeline.

The addition method in convolutional neural networks offers several key advantages.

• Preservation of Information Integrity: By summing the features extracted from different filters, the network can retain critical information from all scales in the final feature map. This allows MSCFNet to simultaneously process and analyze features at different scales.

• Reduction in Computational Complexity: Unlike the concatenation method, which increases the dimensionality of the feature map, the addition method keeps the feature map's dimensions constant. This reduces computational complexity and memory usage, resulting in improved overall network performance.

• Enhancement of Key Features: During the summation process, features that are consistent across different scales and recognized as important patterns become more prominent. This can enhance the network's accuracy in identifying and retrieving images.

• Ease of Implementation: The addition method is simple and fast, making it easy to implement in various CNN architectures. This method also simplifies parameter optimization, as the dimensions of the feature maps remain unchanged during processing.

By employing this method, MSCFNet can more effectively handle variations in images and identify the key features that appear at different scales. Consequently, the accuracy and efficiency of the network in image retrieval are improved.

After extracting features using multi-scale filters and merging them through the addition method, the next step in the processing pipeline is dimensionality reduction and the selection of important and prominent features. For this purpose, Max Pooling layers are utilized.

Max Pooling is one of the most popular methods in CNNs, used to reduce the dimensionality of feature maps and select the most salient information. Max Pooling is a downsampling operation that divides the feature maps into smaller patches and selects the maximum value within each patch. This process helps reduce the size of the feature maps and highlight the stronger and more important features.

• Creation of a Reduced Feature Map: The output of Max Pooling is a reduced feature map, where only the most prominent values are retained, and unnecessary information and noise are discarded.

The Max pooling method in convolutional neural networks offers several key advantages.

• Reduction of Dimensionality and Computational Complexity: Max Pooling reduces the dimensionality of the feature map, which in turn decreases computational complexity in the subsequent layers of the network. This reduction in dimensions allows the network to operate faster and with fewer computational resources.

• Preservation of Prominent Features: One of the main advantages of Max Pooling is its ability to retain the most prominent and important features. These features, which typically have larger values, represent key information within the image and help improve the model's accuracy in image recognition and retrieval.

• Resistance to Small Variations: Max Pooling helps the

network remain robust to small variations in the image (such as minor noise or changes in lighting). This allows MSCFNet to handle a greater diversity of images while maintaining high accuracy.

The combination of Max Pooling with multi-scale filters and feature merging allows MSCFNet to more accurately extract important features from images, leading to significant improvements in image retrieval. This method also helps reduce the complexity of the network and increase processing speed, which is especially important for real-world, large-scale applications.

In summary, the proposed method, MSCFNet, integrates multi-scale filters, feature merging using the addition method, and Max Pooling to enable the CNN to extract and process key features from images with higher accuracy and efficiency. This method not only improves the accuracy of image retrieval but also reduces computational complexity, making it an ideal choice for real-world, large-scale applications.

4. Performance Evaluation

In this section, the performance of the proposed model, MSCFNet, is evaluated using three well-known datasets: CIFAR-10, CIFAR-100, and Fashion-MNIST. These datasets are widely utilized in deep learning and computer vision to assess the effectiveness of different models.

1. CIFAR-10 Dataset: CIFAR-10 consists of 60,000 32x32 colored images distributed across 10 different categories. This dataset is considered a standard for evaluating models in the field of image classification. The proposed model was first trained on this dataset, and its performance was subsequently evaluated. Metrics such as accuracy, F-Score, the number of parameters, and training time were employed for the assessment.

2. CIFAR-100 Dataset: CIFAR-100 is similar to CIFAR-10 but comprises 100 different categories, with each category containing 600 images. The higher number of classes poses a greater challenge for models. The proposed model was also trained on this dataset, and its performance was evaluated using the same metrics as mentioned above.

3. Fashion-MNIST Dataset: Fashion-MNIST contains 70,000 28x28 grayscale images of clothing items divided into 10 categories. This dataset serves as a more complex alternative to the MNIST dataset and is suitable for evaluating models designed for object recognition. The proposed model was also evaluated on this dataset.

The proposed Multi-Scale Convolutional Fusion Network (MSCFNet) was implemented in two architectures: one with two multi-scale convolutional layers and another with four such layers. Each layer comprised 3x3, 5x5, and 7x7 filters, applied in parallel and then merged using the addition operation. In the two-layer architecture, after each convolutional layer, Max Pooling was applied to reduce the dimensionality, and the resulting features were fed into fully connected layers. In the four-layer architecture, two additional convolutional layers were introduced, enabling the extraction of more complex features and enhancing the model's accuracy. This flexible design allowed the evaluation of the model at different levels of depth and complexity.

The proposed MSCFNet was compared with three other models: a CNN with five convolutional layers, ResNet18, and ResNet50. In this comparison, the models were evaluated in terms of accuracy, F-Score, the number of parameters, and training time. The results demonstrated that MSCFNet, in both architectures (two and four layers), achieved a balanced trade-off between accuracy and model complexity by utilizing its multi-scale structure. Compared to the standard CNN, MSCFNet provided higher accuracy with fewer parameters and more efficient training times. Additionally, when compared to ResNet18 and ResNet50, the MSCFNet models with fewer layers and a simpler structure offered competitive performance, making them particularly suitable for environments with limited resources or a need for faster training. The results of this evaluation are presented and analyzed below.

4.1. Analysis of Experimental Results

The experimental results are presented in three sections for each dataset. In this section, the model's performance is analyzed and compared based on the mentioned metrics. Specifically, the model's accuracy in recognizing different categories, its efficiency under varying conditions, as well as its complexity and the time required for training, will be evaluated. This analysis will help identify the strengths and weaknesses of the proposed model, providing insights into potential improvements for future development.

- **Evaluation and Analysis of Results on the CIFAR-10 Dataset**

The results of the comparison between five different models, including CNN (5 layers), ResNet18, ResNet50, MSCFNet (2 layers), and MSCFNet (4 layers), are presented in Table 1. These models were evaluated based on the metrics of accuracy, F-Score, number of parameters, and training time. The analysis of these results is provided below.

Table 1. The caption must be followed by the table

Method	Accuracy	F-Score	Parameters#	Training time (seconds)
CNN (5 layers)	51.03	49.38	122570	40.965
ResNet18	67.98	63.34	11184778	646.67
ResNet50	36.00	28.91	23608202	394.49
MSCFNet (٢ layers)	70.75	70.60	881,994	113.1
MSCFNet (٤ layers)	74.43	74.42	1070794	159.76

Performance of CNN (5 layers): The CNN model with five convolutional layers is a simple and standard model that achieved an accuracy of 51.03% and an F-Score of 49.38% in this comparison. Given its relatively low number of parameters (122,570) and short training time (40.965 seconds), this model is a suitable option for applications with limited computational resources. However, due to its shallow depth and lack of advanced mechanisms such as skip connections or multi-scale feature extraction, this model is unable to compete with more complex models and exhibits comparatively lower performance.

Performance of ResNet18 and ResNet50: The ResNet18 model achieved an accuracy of 67.98% and an F-Score of 63.34%, outperforming the CNN model. This improvement is attributed to the deeper architecture and the use of skip connections, which enable the extraction of more complex features from images. However, the large number of parameters (11,184,778) and long training time (646.67 seconds) increase the model's complexity. The ResNet50 model, on the other hand, exhibited weaker performance with an accuracy of 36.00% and an F-Score of 28.91%. This decrease in performance may be due to the model's higher complexity and the need for more precise hyperparameter tuning. Furthermore, ResNet50, with its very high number of parameters (23,608,202) and long training time (394.49 seconds), faced issues with convergence and extracting appropriate features.

*Performance of MSCFNet (2 layers and 4 layers):*The proposed MSCFNet method, in both its 2-layer and 4-layer versions, demonstrated significantly better performance compared to the other models. The 2-layer MSCFNet achieved an accuracy of 70.75% and an F-Score of 70.60%, while the 4-layer MSCFNet obtained an accuracy of 74.43% and an F-Score of 74.42%. These results show that the use of multi-scale convolutional layers and the fusion of various features improves model performance. Additionally, the reduced number of parameters (881,994 for 2 layers and 1,070,794 for 4 layers) and shorter training time (113.1 and 159.76 seconds, respectively) are further advantages of MSCFNet. These results demonstrate that MSCFNet, with its simpler and more efficient architecture, can achieve higher accuracy compared to more complex models.

- **Evaluation and Analysis of Results on the CIFAR-100 Dataset**

In this section, the results of the evaluation of various models, including CNN (5 layers), ResNet18, ResNet50, and the two versions of MSCFNet (2 layers and 4 layers), on the CIFAR-100 dataset are analyzed. Due to the presence of 100 different classes and the high diversity of images, this dataset poses greater challenges for the models. The results obtained based on the metrics of accuracy, F-Score, number of parameters, and training time are presented in Table 2.

Table.2. Performance Comparison: Results on CIFAR-100

Method	Accuracy	F-Score	Number of parameters	Training time (seconds)
CNN (5 layers)	28.79	27.26	128420	84.80
ResNet18	24.8	15.78	11230948	679.17
ResNet50	27.45	26.43	23792612	737.81
MSCFNet (٢ layers)	38.51	37.83	893604	107.98
MSCFNet (٤ layers)	38.87	38.74	1082404	192.86

Performance of CNN (5 layers): The CNN model with five layers is the simplest model in this comparison, achieving an accuracy of 28.79% and an F-Score of 27.26%. The number of parameters in this model is relatively low (128,420), and the training time is also relatively short (84.80 seconds). These results indicate that in resource-constrained environments, this model may be a suitable choice. However, in terms of accuracy, it lags behind the other models.

Performance of ResNet18 and ResNet50: The ResNet18 and ResNet50 models achieved accuracies of 24.8% and 27.45%, and F-Scores of 15.78% and 26.43%, respectively, which are lower than expected. Although these models have potential due to their deeper layers and the use of skip connections, the weak results may be attributed to the overly complex architecture and the need for more precise parameter tuning. The large number of parameters in these models (11,230,948 for ResNet18 and 23,792,612 for ResNet50) also results in longer training times (679.17 and 737.81 seconds, respectively), which could pose challenges in resource-limited environments.

Performance of MSCFNet (2 layers and 4 layers): The proposed MSCFNet method, in both its 2-layer and 4-layer

architectures, demonstrated superior performance compared to the other models. The 2-layer MSCFNet achieved an accuracy of 38.51% and an F-Score of 37.83%, while the 4-layer MSCFNet achieved an accuracy of 38.87% and an F-Score of 38.74%. These results indicate that MSCFNet effectively extracts various image features and outperforms more complex models such as ResNet. The number of parameters in these models (893,604 for the 2-layer MSCFNet and 1,082,404 for the 4-layer MSCFNet) is relatively low, and the training time is also shorter compared to ResNet models (107.98 and 192.86 seconds, respectively).

- **Analysis of Results on the Fashion-MNIST Dataset**

Table 3 presents the comparative results of five different models, including CNN (5 layers), ResNet18, ResNet50, and the two versions of MSCFNet (2 layers and 4 layers), based on the metrics of accuracy, F-Score, number of parameters, and training time on the Fashion-MNIST dataset after 10 training epochs.

Table.3. Performance Comparison Results on Fashion-MNIST (10 epochs)

Method	Accuracy	F-Score	Number of parameters	Training time (seconds)
CNN (5 layers)	88.69	88.65	93322	83.59
ResNet18	91.3	91.54	11183626	641.38
ResNet50	87.18	87.03	23601930	682.06
MSCFNet (2 layers)	91.92	9195	748490	109.06
MSCFNet (4 layers)	92.47	92.45	1035594	173.87

Performance of CNN (5 layers): The CNN model with five convolutional layers achieved an accuracy of 88.69% and an F-Score of 88.65%. With a relatively low number of parameters (93,322) and a short training time (83.59 seconds), this model demonstrates that it can be utilized in applications where training speed and model simplicity are prioritized. However, the accuracy and F-Score of this model are lower compared to more complex models like ResNet18 and MSCFNet, which can be attributed to the model's limitations in extracting more complex features from images.

Performance of ResNet18 and ResNet50: The ResNet18 model showed better performance than CNN, with an accuracy of 91.3% and an F-Score of 91.54%. With a higher number of parameters (11,183,626) and a longer training time (641.38 seconds), ResNet18 benefits from a deeper architecture and the use of skip connections, enabling it to extract more complex features and achieve higher accuracy. The ResNet50 model, on the other hand, demonstrated weaker performance with an accuracy of 87.18% and an F-Score of 87.03%. This decrease in performance could be due to the model's higher complexity and the need for more precise tuning to better adapt to the Fashion-MNIST data. Additionally, ResNet50, with a significantly higher number of parameters (23,601,930) and a longer training time (682.06 seconds), may have encountered convergence issues, leading to reduced accuracy.

Performance of MSCFNet (2 layers and 4 layers): The proposed MSCFNet method, in both its 2-layer and 4-layer versions, outperformed the other models. The 2-layer MSCFNet achieved an accuracy of 91.92% and an F-Score of 91.95%, while the 4-layer MSCFNet reached an accuracy of 92.47% and an F-Score of 92.45%. These results indicate that the use of multi-scale convolutional layers has been successful in extracting more effective features from images. Furthermore, the lower number of parameters (748,490 for 2 layers and 1,035,594 for 4 layers) and shorter training times (109.06 and 173.87 seconds) demonstrate the high efficiency of MSCFNet. These models, with their simpler and more efficient structure compared to the ResNets, have achieved better accuracy and performance.

4.2. Discussion

This section analyzes the results obtained from the three datasets: CIFAR-10, CIFAR-100, and Fashion-MNIST. Due to the varying levels of complexity and data diversity, these datasets serve as suitable benchmarks for evaluating the performance of different models. CIFAR-10, with its 10 distinct classes, provides a relatively straightforward classification task, while CIFAR-100, with 100 classes, presents a more challenging scenario. Fashion-MNIST, containing grayscale images of clothing items, offers a unique perspective with its focus on texture and shape features.

In Figure 3, we observe a line chart comparing the accuracy of different models across these three datasets. As shown, MSCFNet (both the 2-layer and 4-layer versions) consistently outperforms other models in terms of accuracy, especially in CIFAR-10 and Fashion-MNIST. This demonstrates the effectiveness of MSCFNet's multi-scale feature extraction in capturing relevant features at various image scales, leading to better overall performance. The

trends in the chart also highlight the difficulty posed by CIFAR-100, where even the best-performing models show lower accuracy due to the dataset's inherent complexity.

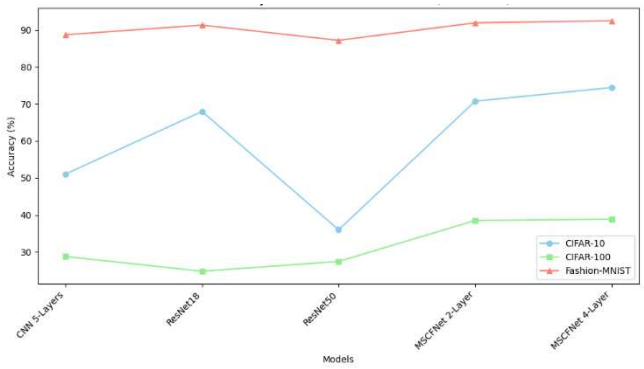

Fig. 3. Model Accuracy across Different Datasets

The Fashion-MNIST dataset, being a relatively simple dataset with images of clothing items across 10 different categories, is well-suited for revealing small differences between models. The results show that MSCFNet (4 layers) achieved the best performance with an accuracy of 92.47% and an F-Score of 92.45%. Despite being less complex than ResNet18 and ResNet50, this model was able to extract more effective features due to its multi-scale structure, leading to higher accuracy. The CIFAR-10 dataset, with more diverse images and 10 different categories, presents a greater challenge compared to Fashion-MNIST. Here, MSCFNet, in both the 2-layer and 4-layer versions, outperformed the other models. MSCFNet (4 layers) achieved an accuracy of 74.43%, demonstrating its ability to extract more complex features from the images and surpass more complex models like ResNet50. The CIFAR-100 dataset, with 100 different categories, is one of the most complex datasets, requiring a high capability in extracting complex and multi-scale features. MSCFNet (4 layers) also showed superior performance on this dataset with an accuracy of 38.87%, outperforming other models. These results suggest that even when faced with highly challenging datasets, the proposed method can deliver better performance than more complex models with a greater number of parameters.

The advantages of the proposed MSCFNet include a balance between accuracy and complexity, shorter training time, efficiency in feature extraction, flexibility, and a lower number of parameters. By combining multi-scale convolutional layers and using the addition operation to fuse features, MSCFNet has managed to achieve a good balance between high accuracy and model complexity. This characteristic has allowed the model to perform well across various datasets without requiring a large number of parameters. Another important advantage of MSCFNet over more complex models like ResNet18 and ResNet50 is the reduced training time. MSCFNet, with its more efficient structure, requires less time to train, which is especially important in environments with time or computational resource constraints. The use of multi-scale convolutional layers in MSCFNet has enabled the model to extract a broader range of features from the images. This capability has proven particularly beneficial when dealing with complex datasets like CIFAR-100, where MSCFNet has outperformed other models. Due to its modular structure, MSCFNet allows for flexibility in the number of layers. This flexibility enables users to adjust the number of convolutional layers based on specific needs and data complexity, optimizing the model for the best performance. Compared to ResNet18 and ResNet50, MSCFNet has shown competitive or superior performance in terms of accuracy and F-Score, despite having fewer parameters. This reduction in the number of parameters results in lower memory and computational resource requirements, which is a significant advantage in many practical applications.

The performance of ResNet50 was consistently lower than MSCFNet across all datasets, as shown in the results Figure.3. ResNet50's poor performance and convergence issues can be attributed to several factors. First, its high complexity with 50 layers leads to significant challenges in convergence, especially when hyperparameters like the learning rate are not finely tuned. In contrast, MSCFNet's simpler architecture, with fewer layers, allows for faster and more reliable convergence. Additionally, ResNet50's large number of parameters increases the risk of overfitting, particularly on datasets with limited or less complex data, while MSCFNet, with fewer parameters, shows better generalization across different datasets. ResNet50 is also highly sensitive to hyperparameter tuning, requiring precise adjustments to perform optimally. If not properly tuned, it may experience slow or suboptimal convergence. In comparison, MSCFNet’s design, which uses simpler operations such as "addition" for feature fusion, requires less fine-tuning and provides more consistent results. Furthermore, ResNet50’s deep architecture and high parameter count demand more computational resources, leading to longer training times and higher memory usage. MSCFNet, with its lightweight and efficient design, is better suited for resource-constrained environments, requiring significantly less memory and training time while still achieving superior performance.

Based on the results from the three datasets with varying levels of complexity, the proposed MSCFNet has

demonstrated that by maintaining a proper balance between accuracy and model complexity, it can outperform more complex models like ResNet. The benefits of this method include high accuracy, shorter training times, efficient feature extraction, flexibility, and fewer parameters, making MSCFNet an efficient and suitable model for a wide range of applications.

5. Conclusion

In this study, the proposed MSCFNet method was evaluated on the CIFAR-10, CIFAR-100, and Fashion-MNIST datasets, using a multi-scale structure. The results demonstrated that MSCFNet, in both its 2-layer and 4-layer versions, outperformed more complex models such as ResNet18 and ResNet50 by achieving a proper balance between accuracy, model complexity, and training time. The multi-scale structure of MSCFNet enabled the extraction of richer features from the images, resulting in higher classification accuracy. Additionally, the lower number of parameters and shorter training time made this model a suitable option for various applications, especially in environments with limited computational resources.
Several avenues for improving and extending MSCFNet can be explored in future research. One possible direction is to investigate advanced techniques for optimizing model performance, such as leveraging transfer learning methods. These techniques could help MSCFNet improve its efficiency and accuracy in new and challenging domains by utilizing prior knowledge from other domains. Moreover, developing deeper and more advanced models while maintaining the balance between accuracy and complexity, as well as enhancing training time reduction and computational resource optimization, could lead to broader applications and greater success for MSCFNet in various fields.

References

[1] a, G. G., & Khanna, A. (2024). Content Based Image Retrieval System Using CNN based Deep Learning Models. Procedia Computer Science, 235, 3131-3141. doi:https://doi.org/10.1016/j.procs.2024.04.296

[2] a, L. C., & Liu, M. (2024). An intelligent deep hash coding network for content-based medical image retrieval for healthcare applications. Egyptian Informatics Journal,27(100499).doi:https://doi.org/10.1016/j.eij.2024.100499

[3] a, Y. L., b, J. M., & Zhang, Y. (2021). Image retrieval from remote sensing big data: A survey. Information Fusion, 67, 94-115. doi:https://doi.org/10.1016/j.inffus.2020.10.008

[4] b, Y. Y. a., a, S. J., b, J. H., b, B. X., a, J. L., & Xiao, R. (2020). Image retrieval via learning content-based deep quality model towards big data. Future Generation Computer Systems, 112, 243-249.

[5] Chen, Y., Ling, M., Liu, Y., Chen, X., Li, Y., & Tong, B. (2024). Enhancing MRI image retrieval using autoencoder-based deep learning: A solution for efficient clinical and teaching applications. 17(3). doi:https://doi.org/10.1016/j.jrras.2024.100932

[6] Ciocca, G., Napoletano, P., & Schettini, R. (2018). CNN-based features for retrieval and classification of food images. Computer Vision and Image Understanding, 176, 70-77. doi:https://doi.org/10.1016/j.cviu.2018.09.001

[7] Dubey, S. R. (2022). A Decade Survey of Content Based Image Retrieval Using Deep Learning. EEE Transactions on Circuits and Systems for Video Technology, 32(5), 2687-2704. doi:0.1109/TCSVT.2021.3080920

[8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Paper presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA.

[9] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Paper presented at the In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA.

[10] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

[11] Kumar, M., Singh, R., & Mukherjee, P. (2024). VTHSC-MIR: Vision Transformer Hashing with Supervised Contrastive learning based medical image retrieval. 184, 28-36. doi:https://doi.org/10.1016/j.patrec.2024.06.003

[12] Lecun, Y., Bottou, L., Bengio, Y., &Haffne, P.

(1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324. doi:10.1109/5.726791

[13] Leticio, G. R., Kawai, V. S., Valem, L. P., Pedronette, D. C. G., & Torres, R. d. S. (2024). Manifold information through neighbor embedding projection for image retrieval. Pattern Recognition Letters, 183, 17-25. doi:https://doi.org/10.1016/j.patrec.2024.04.022

[14] Li, X., Yang, J., & Ma, J. (2021). Recent developments of content-based image retrieval (CBIR). Neurocomputing, 452, 675-689.

[15] Liu, C., Ma, J., Tang, X., Liu, F., Zhang, X., & Jiao, L. (2021). Deep Hash Learning for Remote Sensing Image Retrieval. IEEE Transactions on Geoscience and Remote Sensing, 59(4), 3420-3443. doi:10.1109/TGRS.2020.3007533

[16] Qin, J., Chen, J., Xiang, X., Tan, Y., Ma, W., & Wang, J. (2020). A privacy-preserving image retrieval method based on deep learning and adaptive weighted fusion. J Real-Time Image Proc, 17, 161–173. doi:https://doi.org/10.1007/s11554-019-00909-3

[17] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[18] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). Going deeper with convolutions. Paper presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA.

[19] Wang, S., Xia, Y., Xiang, N., Qian, K., Yang, X., You, L., & Zhang, J. (2024). Multi-colour sketch-based image retrieval with an explicable feature embedding. Engineering Applications of Artificial Intelligence, 135. doi: https://doi.org/10.1016/j.engappai.2024.108757.

[20] Yan, C., Gong, B., Wei, Y., & Gao, Y. (2021). Deep Multi-View Enhancement Hashing for Image Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(4), 1445-1451. doi:10.1109/TPAMI.2020.297579

[21] Yu, W., Yang, K., Yao, H., Sun, X., & Xu, P. (2017). Exploiting the complementary strengths of multi-layer CNN features for image retrieval. Neurocomputing, 237,235-241. doi:https://doi.org/10.1016/j.neucom.2016.12.002

[22] Zhang, Z., Cheng, S., & Wang, L. (2023). Combined query image retrieval based on hybrid coding of CNN and Mix-Transformer. Expert Systems With Applications,234. doi:https://doi.org/10.1016/j.eswa.2023.121060

[23] Zhao, D., Qiu, Z., Jiang, Y., Zhu, X., Zhang, X., & Tao, Z. (2024). A depthwise separable CNN-based interpretable feature extraction network for automatic pathological voice detection. Biomedical Signal Processing and Control, 88.

doi:https://doi.org/10.1016/j.bspc.2023.105624

[24] Zhao, K., Xiao, J., Li, C., Xu, Z., & Yue, M. (2023). Fault diagnosis of rolling bearing using CNN and PCA fractal based feature extraction. Measurement, 223. doi:https://doi.org/10.1016/j.measurement.2023.113754