Computers, Materials & Continua CMC, vol.65, no.2, pp.1691-1706, 2020

An Improved Deep Fusion CNN for Image Recognition

Rongyu Chen[1], Lili Pan[1,*], Cong Li[1], Yan Zhou[1], Aibin Chen[1] and Eric Beckman[2]

Abstract: With the development of Deep Convolutional Neural Networks (DCNNs), the extracted features for image recognition tasks have shifted from low-level features to the high-level semantic features of DCNNs. Previous studies have shown that the deeper the network is, the more abstract the features are. However, the recognition ability of deep features would be limited by insufficient training samples. To address this problem, this paper derives an improved Deep Fusion Convolutional Neural Network (DF-Net) which can make full use of the differences and complementarities during network learning and enhance feature expression under the condition of limited datasets. Specifically, DF-Net organizes two identical subnets to extract features from the input image in parallel, and then a well-designed fusion module is introduced to the deep layer of DF-Net to fuse the subnet's features in multi-scale. Thus, the more complex mappings are created and the more abundant and accurate fusion features can be extracted to improve recognition accuracy. Furthermore, a corresponding training strategy is also proposed to speed up the convergence and reduce the computation overhead of network training. Finally, DF-Nets based on the well-known ResNet, DenseNet and MobileNetV2 are evaluated on CIFAR100, Stanford Dogs, and UECFOOD-100. Theoretical analysis and experimental results strongly demonstrate that DF-Net enhances the performance of DCNNs and increases the accuracy of image recognition.

Keywords: Deep convolutional neural networks, deep features, image recognition, deep fusion, feature fusion.

1 Introduction

DCNNs [Ma, Li, Xia et al. (2020)] have made breakthrough progress in computer vision and have become the standard method of many visual object recognition algorithms in recent years. DCNN is a progressive structure, whose shallow neurons can sense semantic content such as structure, texture and location. On this foundation, the deep convolutional layers continue to learn more advanced and distinguishable features for

[1] College of Computer Science and Information Technology, Central South University of Forestry & Technology, Changsha, 410114, China.

[2] China Chaplin School of Hospitality of Hospitality and Tourism Management, Florida International University, North Miami, 33181, USA.

[*] Corresponding Author: Lili Pan. Email: lily_pan163.com.

Received: 25 May 2020; Accepted: 23 June 2020.

classification. The ability to automatically learn image features of DCNN has brought significant changes in the field of image recognition.

The basic image recognition network LeNet-5 [LeCun, Boser, Denker et al. (1989)] was composed of three operations: convolution, pooling and non-linear mapping. This kind of combination structure is also the foundation of mainstream DCNNs. With the update of neural network algorithms and the advance of hardware, the width and depth of DCNNs have been continuously improved.

In terms of network depth, He et al. [He, Zhang, Ren et al. (2016)] proposed a residual block structure that aimed at the training of networks. This research indicated that the residual networks were easier optimized and could obtain high accuracy by increasing network depth. Furthermore, expanding the network width has proved feasible. Inception module introduced by GoogLeNet [Szegedy, Liu, Jia et al. (2015)] implemented convolution and pooling operations simultaneously in a parallel manner to extract more potential information. The extension of different kernels means the fusion of different scale features, which effectively improves the expression of network.

It is important to design and develop accurate algorithms and systems for more accuracy of image recognition. However, creating an innovative network is a difficult task which requires abundant knowledge of DCNN as well as the stronger support from hardware. In practical application, the network structure, dataset scale and hardware computing power should be fully considered so that a designed DCNN architecture meets the needs of network performance and computing overhead simultaneously. Therefore, this paper proposes a novel deep fusion by using the latest development of deep learning. First, the fusion module is used to integrate two identical subnets in the deep layer to sufficiently improve network performance. Secondly, the training strategy of only fine-tuning the deep convolution layer is formulated to reduce the computational overhead. Consequently, a more in-depth and broader network DF-Net is constructed to extract more abundant image features and optimize the recognition effect.

2 Relative works

Image recognition is a fundamental problem in computer vision. The key factor to recognition accuracy is the performance of the extracted image features. Prior research is largely based on hand-crafted features, such as Histogram of Oriented Gradient (HOG) [Sugiarto, Prakasa, Wardoyo et al. (2017)], Harris [Qin, Li, Xiang et al. (2019)], YCbCr [Tan, Qin, Xiang et al. (2019)] and Scale-Invariant Feature Transform (SIFT) [Bharathidevi, Chennamsetty and Prasad (2017)]. However, these hand-crafted features have a strong reliance on expertise and task specificity. Despite of the feature selection and feature fusion [Qin, Sun, Xiang et al. (2009)] brought optimization for hand-crafted features, these may result in the cumbersome of design and the unsatisfactory of performance in practical applications.

Recently, DCNN has performed well in computer vision. DCNN learns simple features from big data, and then gradually learns the more abstract deep features for image recognition. For instance, in the field of the classification of food ingredients, Pan et al. [Pan, Pouyanfar, Chen et al. (2017)] extracted rich and productive features from food ingredient images using DCNN, which improved the average accuracy of image

classification. Qin et al. [Qin, Pan, Xiang et al. (2020)] took advantage of DCNNs to classify the biological images effectively. In the field of target recognition, Nasrabadi [Nasrabadi (2019)] designed a high-performance system based on DCNN to detect the targets in forward looking infrared (FLIR). The advantage of DCNN is that it can learn the optimal feature representation from the target dataset and better express the information of the original image. Luo et al. [Luo, Qin, Xiang et al. (2020)] and Liu et al. [Liu, Xiang, Qin et al. (2020)] both used DCNNs to extract the high-level semantic features; Zhang et al. [Zhang, Wu, Feng et al. (2019)] used the attentional features extracted from the DCNN to localize the target accurately. These studies illustrated the effectiveness of DCNNs in the field of visual object recognition tasks.

As early as 1989, the classic DCNN LeNet-5 was proposed by LeCun et al. [LeCun, Boser, Denker et al. (1989)] for recognizing handwritten digits and machine-printed characters. LeNet-5 used a three-layer sequence combination: convolution, pooling and non-linear mapping, which formed the basis of current DCNNs. With the updating of algorithm and the improvement of deep structure, the accuracy of image recognition based on DCNNs has been continuously rising. AlexNet was the first DCNN applied to large-scale image classification by Krizhevsky et al. [Krizhevsky, Sutskever and Hinton (2012)]. The error rate won the first place in the 2012 ILSVRC (ImageNet Large Scale Visual Recognition Competition) and was far better than traditional methods. Its appearance proved the availability of DCNN in complex models and made DCNN popular in computer vision. Soon after a deeper convolutional network VGGNet was proposed by Simonyan et al. [Simonyan and Zisserman (2014)], which performed well in the 2014 ILSVRC classification task. The core idea of VGGNet is adopting the convolution combination of 3×3 instead of the large convolution and thus the depth of the network has been effectively increased. In conclusion, the performance of the basic network can be significantly improved by deepening the network.

However, the deeper the network, the more difficult the training. With the increase of network parameters, the network training will be more prone to the over-fitting problem, which would cause a low error rate on training set but a high error rate on test set. Researchers have proposed several advanced DCNNs to learn and simulate the complex data. Szegedy et al. [Szegedy, Liu, Jia et al. (2015)] proposed a new deep learning network GoogLeNet, which was the winner of the 2014 ILSVRC classification task. GoogLeNet decomposed the longitudinal connection into several block subnetworks, which were the inception module. In each module, the highly correlated features are connected and then transferred to the subsequent layers. Besides, different sizes of receptive fields are obtained using convolution kernels of different sizes, and the final concatenate means the fusion of various scale features. Simultaneously, the 1×1 convolution kernel was introduced for dimensionality reduction and eliminating the calculation bottleneck.

He et al. [He, Zhang, Ren et al. (2016)] presented the Residual Neural Network (ResNet) and achieved state-of-the-art performance in 2015 ILSVRC. ResNet introduced a creative architecture with “skip connections” and BN operation, which could train the deep network efficiently and accurately. DenseNet [Huang, Liu, Van Der Maaten et al. (2017)] gained the best paper of CVPR 2017 (Computer Vision and Pattern Recognition 2017),

which directly connected each layer with the same feature-map size in a feed-forward fashion. For each layer, the input was composed of the feature maps of all previous layers, and the feature maps of the current layer also became the input of all subsequent layers in the meanwhile. This operation alleviated the issue of vanishing-gradient, strengthened feature propagation and greatly reduced the network number of parameters.

It is quite formidable and complicated to design an innovative and excellent network. First, it is necessary to possess the relevant theoretical knowledge of DCNNs. In addition, it requires a lot of experimental accumulation and strong inspiration during the process of improving the basic model. Furthermore, creating a new model and training it on the large-scale datasets would consume a lot of time and computing resources, namely the powerful hardware support. Therefore, on the basis of the existing research, appropriately adjusting or improving the classical network is an effective measure for deep learning. Wang et al. [Wang, Qin, Xiang et al. (2019)] constructed a multi-classifier network based on DenseNet for CAPTCHA recognition. Amin-Naji et al. [Amin-Naji, Aghagolzadeh and Ezoji (2019)] constructed a new network with the support of the ensemble learning to decrease the overfitting on limited datasets. To learn more precise features, Hou et al. [Hou, Liu and Wang (2017)] provided a general framework DualNet to address image recognition by coordinating two parallel DCNNs. Zhang et al. [Zhang, Wang and Lu (2019)] proposed two novel lightweight networks that could obtain higher recognition precision of traffic sign images in a resource-limited setting. Xiang et al. [Xiang, Guo, Yu et al. (2020)] build a two-level cascaded DCNNs, which could automatically learn the steganographic features and improve the detection performance greatly. Pan et al. [Pan, Li, Pouyanfar et al. (2020)] proposed an up-to-date CBNet (Combinational Convolutional Neural Network) which combined two different DCNNs to extract complementary features for image classification.

Different from previous studies, this study innovatively designs a fusion module inspired by the inception module. The fusion module is created to implement the combination of deep features from two parallel subnets. It can give full play to the value of the basic network and enhance the expression of features. The latest constructed DF-Net brings double expression space and more complex mappings, which makes learning and representation easier. Compared with the DualNet framework which adopted the addition and combination, this paper presents the fusion concept that can effectively improve the performance of image recognition by extracted high-level semantic features with the network. At the same time, a training strategy of only fine-tuning the deep convolutional layers of the network is formulated to ensure the computational efficiency and the speedy constringency of network. Experimental results show that the proposed DF-Net framework achieves higher recognition accuracy on CIFAR100 [Krizhevsky and Hinton (2009)], Stanford Dogs [Khosla, Jayadevaprakash, Yao et al. (2011)] and UECFOOD-101[Matsuda, Hoashi and Yanai (2012)] datasets.

3 Algorithm implementation

3.1 The deep fusion convolutional neural network (DF-Net)

So far DCNNs have made great progress in the field of image recognition, and the deep feature is the most competitive visual features. The key factor in the success of DCNNs is

the training samples which consist of numerous tagged data. Through the training on the samples of known correct answers, DCNNs can learn deep semantic information from a machine perspective. However, DCNNs will easily occur over-fitting when the training samples are limited. In other words, a model with excellent extensive ability is challenging to achieve directly from the limited training samples. Although the model can perfectly perform on training set during training, the performance of this model will be relatively weak for unknown data. One effective approach for this problem is to prepare more data, but the cost of collecting enough training samples in a reasonable amount of time is enormous. To solve the problem, this paper designs a novel DF-Net architecture (see Fig. 1), which fuses two parallel models in the deep layer. Our goal is to enable the new architecture to learn more complementary features under constrained training samples. As shown in Fig. 1, the proposed architecture includes Subnet Component (SC, left of Fig. 1) and Fusion Component (FC, right of Fig. 1), which together form the DF-Net for automatic image recognition.

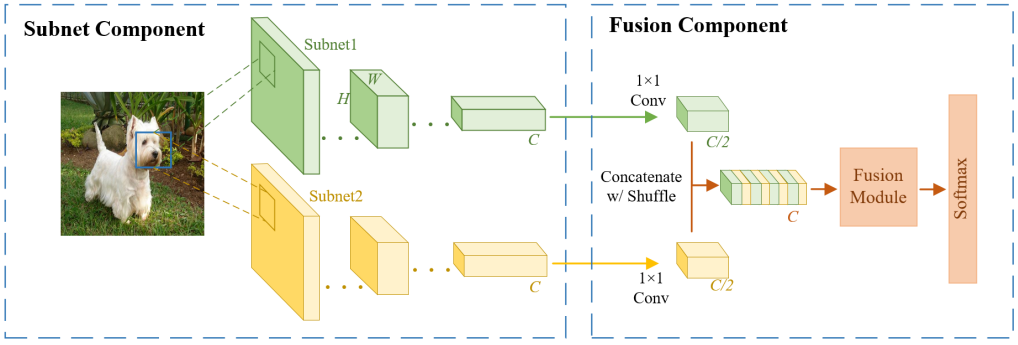

Figure 1: DF-Net architecture

3.1.1 Subnets component

Through the multiple convolution and pooling, DCNNs can map the raw data to a multi-level representation and abstraction. However, the complexity of a single network is limited, which would restrict the ability of learning complex function mapping, particularly for small-scale datasets. It is acknowledged that if the features extracted from subnets are the same, the concatenated features are linear, similar to a single network. Therefore, it is impossible to enhance the complexity of the network by concatenating two groups of identical features. Overall, instead of extending dimension on a single network, this work deploys two independent networks to catch more potential information from input images simultaneously in SC.

3.1.2 Fusion component

In our architecture, the subnets' features are extracted before the final global pooling layer of network. It means that features fusion implements in 3D feature space, not 1D feature space. In Fig. 1, after features extracted from SC, they will be further analyzed and processed in FC. Here double-stream features are bound to be different, which exactly is the premise of information complementarity. In order to make full use of the

complementarity, based on the most advanced achievements of DCNNs, an innovative fusion strategy consisting of 1×1 Convolution, Channel Shuffle [Zhang, Zhou, Lin et al. (2018)] and fusion module is proposed. By doing so, the network complexity is effectively improved, and thus the ability of learning and simulating more complex types of data is strengthened. Furthermore, the corresponding training strategy is proposed to speed up the convergence and reduce the computation overhead of network training. The fusion strategy and training strategy will be outlined in Sections 3.2 and 3.3.

3.2 Fusion strategy

The fusion strategy is an important strategy in which the proposed DF-Net architecture is utilized to improve the non-linear mapping relations of DCNNs and heighten the network expression.

The subnets of DF-Nets can be employed for most of the well-known DCNNs, such as ResNet, GoogLeNet, DenseNet, etc. These DCNNs have one thing in common: They all obtain the final 1D features by using global average pooling instead of the fully connected layer. Therefore，a well-designed fusion module is constructed to fuse 3D features of two subnets. Although the 3D features learned more abundant information from subnets than 1D features, the learning process will produce excessive parameters, which is unfavorable for fusion module to calculate and realize the transformation of spatial information. In view of this, one essential method is that the 3D features are preprocessed to reduce the network parameters to make feature abstraction effective in fusion module.

3.2.1 Feature pre-processing

Once a network is adopted as the subnet of DF-Net, its corresponding pre-trained network will be applied and fine-tuned on the target dataset. During various training, two of the top fine-tuned models are deployed in SC to obtain the output of the final feature maps (3D feature space). The information gathered by double-stream features has a high complexity, which may lead to dimensionality disaster. In some networks (e.g., AlexNet, VGGNet), researchers employed the fully connected layer to reduce feature dimension. However, it is not suitable for our architecture because the fully connected layer is used for the 1D features, not the 3D features. For other networks (e.g., ResNet, GoogLeNet), the 1×1 convolution kernel provides help for dimensionality reduction. Inspired by the scheme, the 1×1 convolution layer is employed to subnets' features to alleviate the dimensionality disaster. Specifically, the 1×1 convolution layer with a compression ratio of 50% is appended behind the subnets to avoid the dimensional explosion. As shown in FC, the feature maps of subnets are compressed from c-channel to (c/2)-channel by the 1x1 convolution layer. Then the two compression features are concatenated so that the channels of the concatenated features are consistent with a single network. It is worth noting that RELU activation is not used after 1×1 convolution as suggested by Chollet [Chollet (2017)].

Since the rudimentary way of combination makes the data distribution highly dispersed, it may cause one of the double-stream features to be eccentrically selected during training, which is not beneficial to the mutual learning and cooperation between subnets.

Therefore, the channel shuffle technology is adopted to achieve an even data distribution. The channel shuffle is similar to ShuffleNet [Zhang, Zhou, Lin et al. (2018)]. In our module, the number of groups is set as 2 (corresponding to two subnets). As shown in FC of Fig. 1, the final feature maps are gained with a staggered concatenation of two subnets.

In summary, the feature pre-processing includes a 1×1 convolution layer for the feature reduction and a channel shuffle operation for information flow from two subnets. Feature pre-processing can substantially decrease the computation overhead of the fusion module and enhance the information exchange and complementarity between subnets.

3.2.2 Fusion module

In the inference of DF-Net, two subnets with the same architecture capture the visual information from the input images respectively, but their weights of neurons are not shared. In other words, these two subnets are independent of each other before channel concatenation, so the acquired features are complementary for the two subnets. In order to make full use of this complementarity, the approach is absorbed from Inception-ResNet-v2 [Szegedy, Ioffe, Vanhoucke et al. (2017)], which offers effective insights to improve the network learning by non-linear mapping, the fusion of different scale features, skip connection, etc. Inception module is an excellent local topology that performs multiple convolutions and pooling operations for the input data in parallel, and concatenates all the output results together into a multi-channel feature map.

As for Inception-ResNet-v2, the residual connection is introduced into the inception module to improve recognition performance as much as possible. Enlightened by the inception module, a novel inception module is created as the fusion module that the particular design is favorable to analyze the concatenated features. In theory, the fusion module can extract more informative features from subnets for image recognition. The structure of the fusion module is described in Fig. 2.

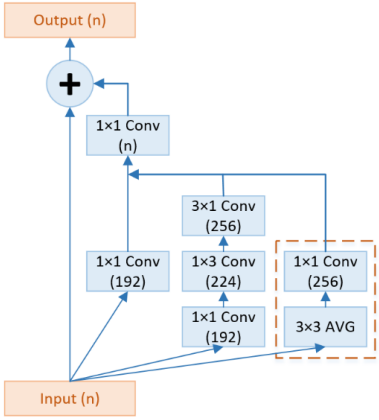

Figure 2: The structure of the novel fusion module

In Fig. 2, an additional branch (dotted box) is added to the original Inception-ResNet-C module. After feature pre-processing, the concatenated features are used as input of the fusion module. Through the calculation of three branches on the right, the different scale

features are generated and concatenated to form a 704-channel feature map. To keep the residual addition operating normal, the 704-channel feature map is followed by a 1×1 convolution layer, which is used to match the channels of the initial input by dimensionality expansion. In the residual branch, the residuals are added to the final output and a scaling factor is set as 0.3.

As for the case that the fusion module is applied to network fusion, the advantages include as followed: (1) The usage of small-scale kernel size can learn and perceive high-level semantic features with more details; (2) The multi-branch design realizes the fusion of different scale features, which can enhance the adaptability to different scales and improve the expression ability of the network; (3) The additional branch composed of a 3×3 average pooling and a 1×1 convolution is conducive to increasing the feature's diversity. More importantly, the features extracted by average pooling operations inherit the ability of classification from subnets; (4) The residual connection maps the module input to the output, which may guide the convolution and pooling operations from other branches to ensure the classification performance. In conclusion, the fusion module provides powerful abstraction abilities for reintegrating the concatenated features, as well as the fusion features are more generalizable compared with other features.

3.3 Training strategy

Logically, the DF-Net is a quite large network, which comprises 4 modules, including the two subnets (remove the global average pooling and classifier), the feature pre-processing module, the fusion module and the newly appended classifier. Considering the training time and GPU memory, this paper drafts a staged training scheme to update global parameters.

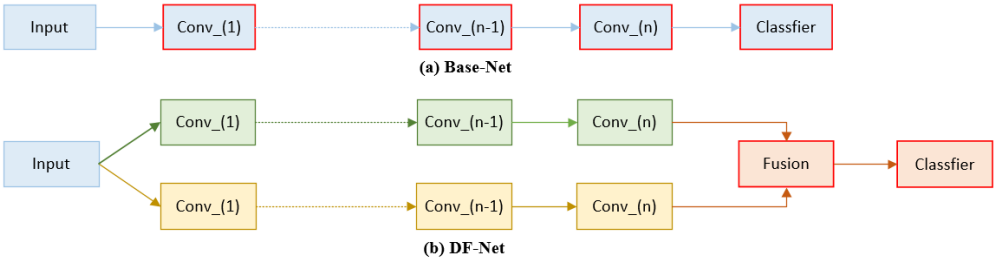

Figure 3: The training process of DF-Net

The process of training is usually called fine-tuning that the pre-trained parameters are updated on the target dataset and adjusted to the target dataset. In this work, once a network is adopted as the subnet of DF-Net, this pre-trained network by ImageNet will be utilized and firstly fine-tuned on the target dataset. During various tests, two of the top fine-tuned models are deployed to the DF-Net as subnets, and the fine-tuned weights are reused to initialize DF-Net. Accordingly, just fine-tuning the fusion module and the newly appended classifier can achieve a prediction model. According to the training strategy, the computation overhead of network training can be reduced substantially. It also makes the DF-Net convergence faster during training. Moreover, since the fine-tuned weights of the subnets are preserved, the classification performance of DF-Net is ensured effectively.

Fig. 3 shows the training process of DF-Net, which is orderly divided into two stages, i.e., Base-Net training and DF-Net training. Training with a large amount of tagged data is a key factor of deep learning. In practice, it is hard to achieve high performance on limited training samples even if how excellent the network is. Nevertheless, the fine-tuning technology can solve this problem well. At the first stage shown in Fig. 3(a), the Base-Net (such as ResNet50) is fine-tuned on the target dataset with the pre-trained network on ImageNet. The main purpose of this stage is to adapt the parameters to the target dataset and improve the recognition performance of Base-Nets as much as possible. The next step is employing the fine-tuned Base-Nets as subnets of the DF-Net. Then, at the second stage (Fig. 3(b)), after freezing all the parameters of subnets, the fusion module and the newly appended classifier are only fine-tuned to make the parameters adapt to the subnets and the target dataset. Correspondingly, the number of total parameters and trainable parameters is contrasted between Base-Nets and DF-Nets in Tab. 1.

Table 1: The comparison of network parameters between Base-Nets and DF-Nets (million)

Network	Total parameters	Trainable parameters
Base-MobileNetV2	2.26	2.22
DF-MobileNetV2	8.18	3.66
ResNet50	23.66	23.61
DF-ResNet50	54.51	7.33
DenseNet121	7.01	6.93
DF-DenseNet121	16.78	2.70

As shown in Tab. 1, although the number of total parameters of DF-ResNet50 and DF-DenseNet121 is about 2-3 times more than the corresponding Base-Nets, the number of total trainable parameters is only equal to 30-40% of the Base-Net. That means the training of these DF-Nets needn't additional hardware requirements. Besides, the number of trainable parameters of DF-Net constructed by the lightweight network MobileNetV2 is equivalent to 165% of the Base-Net, even higher than the total parameters of the Base-MobileNetV2. Essentially, the substantial growth of parameters implies that the DF-Net gains the larger capacity of expression space and the more complex mappings than Base-Net. Subsequent experiments show that DF-Nets achieve a markable accuracy improvement for the lightweight network.

4 Experimental analysis

In this section, the proposed DF-Nets which are built based on ResNet50 (DF-ResNet50), DenseNet121 (DF-DenseNet121), MobileNetV2 (DF-MobileNetV2) are tested on multiple widely-used datasets, including CIFAR-100, Stanford Dogs and UEC FOOD-100. Firstly, the model (e.g., ResNet50) trained on ImageNet is loaded and fine-tuned on the target dataset several times. Secondly, two models with the best performance are selected as the subnets of DF-Net (e.g., DF-ResNet50). Finally, the fusion module and classifier of the DF-Net are fine-tuned on target dataset so that the network parameters

adapt to subnets and datasets. Most importantly, to measure the effectiveness and stability of novel DF-Nets architecture, this work uses the average accuracy of the DF-Net compared with the highest accuracy of the corresponding subnets.

Although these classic DCNNs are perfect, the fine-tuning process is very complicated and time-consuming. In order to train all the networks well, a serial of methods is utilized for the optimizer, learning rate, parameters, etc. Here the Stochastic Gradient Descent (SGD) is used for optimization. The mini-batch size is set to 16 to balance the memory utilization and capacity. All the networks are fine-tuned 30 epochs with an initial learning rate of 0.003. In addition, to achieve a better convergence, the learning rate is decayed with a rate of 0.94 per epoch. Keras is used and it is a popular deep learning tool which provides many advanced DCNNs with pre-trained weights and supports fast network design and experimentation. The hardware is the NVIDIA RTX2080Ti GPU with 11 GB of memory, 4352 CUDA cores and 544 Tensor cores.

4.1 CIFAR-100

In order to evaluate the effectiveness of the DF-Net framework, first experiment is operated on the publicly available and challenging CIFAR-100 dataset, which contains 60,000 32×32 color images for 100 categories. Commonly, 50,000 images are used for training and 10,000 images for test. During the training, images are resized to 224×224 as the network input. Tab. 2 shows the comparison of accuracy on CIFAR-100, where the (Δ) represents the improvement of the DF-Nets compared with the corresponding Base-Nets.

Table 2: The accuracy achieved by Base-Nets and DF-Nets on CIFAR-100 (%)

Model	Base-Net	DF-Net	Δ
MobileNetV2	79.91	82.27	2.36
ResNet50	83.31	84.57	1.26
DenseNet121	84.86	85.81	0.95

Tab. 2 compares the accuracy between DF-Nets and the corresponding Base-Nets on the CIFAR-100 dataset. Obviously, as can be seen from Tab. 2, DF-Nets have better accuracy than Base-Nets. It can be concluded that the DF-Net can effectively improve the complexity of network and learn a model with fine generalization. Specifically, the DF-Net based on MobileNetV2 (DF-MobileNetV2) achieves the highest promotion, and the accuracy reaches 82.27%, which is 2.36% higher than its Base-Net. Referring to Tab. 1, the results show that the greater the parameters increase by the deep fusion, the more remarkable the performance improvement will be.

As shown in Tab. 3, the DF-ResNet50 is compared to other state-of-the-art methods with ResNet architecture. From Tab. 3, the DualNet based on ResNet56 (DNR56) [Hou, Liu and Wang (2017)] obtains the accuracy of 75.57%, which is 2.76% higher than its basic network. With the improved residual network, RoR-3-WRN58-4 [Zhang, Sun, Han et al. (2017)] and WRN [Zagoruyko and Komodakis (2016)] get the accuracy of 80.27% and 81.15% respectively. In this paper, the DF-ResNet50 achieves 84.57% accuracy, which improves the performance of Base-ResNet50 by more than 1.26%, and much higher

accuracy (almost 3.42%) than WRN. The growth rate of the DF-Net is slightly lower than Dual-Net's. The reasons are that both the subnets are optimized perfectly, and the highest accuracy of subnets is set as baseline in our experiments. For the DF-Net, only the fusion module and the final classifier need be trained during the fine-tuning so that the DF-Net is approximately equal to subnet in time cost. The experimental results demonstrate the effectiveness of the proposed DF-Net framework which improves network performance and has a very high image recognition accuracy compared to other existing methods.

Table 3: The comparison of performance with the existing methods on CIFAR-100 (%)

Method	Accuracy	Δ
ResNet56	72.81	2.76
DNR56	75.57	
RoR-3-WRN58-4	80.27	-
WRN	81.15	-
Base-ResNet50 (ours)	83.31	1.26
DF-ResNet50 (ours)	84.57	

4.2 Stanford Dogs

In this section, the DF-Net is further analyzed on a Fine-Grained image Visual Classification (FGVC) Stanford Dogs dataset, which consists of 20580 images and 120 categories of dogs. The experiment applies 100 images for each category for training and the rest of the dataset for test. The Stanford Dogs dataset has an extremely high similarity that can be used for the FGVC task. Besides, compared to the image size of CIFAR-100 (32×32), the Stanford Dogs is proper to DCNNs' training.

Table 4: The accuracy achieved by Base-Nets and DF-Nets on Stanford Dogs (%)

Model	Base-Net	DF-Net	Δ
MobileNetV2	73.93	76.01	2.08
ResNet50	76.12	77.63	1.51
DenseNet121	78.21	79.34	1.13

In practice, the complexity of a single network is limited, which would affect the ability of learning complex function mapping. However, the proposed DF-Net can extract more abundant and accurate fusion features from two parallel subnets for image recognition even with small-scale datasets. Tab. 4 shows the accuracy comparison between Base-Nets and DF-Nets on the Stanford Dogs dataset. All the DF-Nets beat the corresponding Base-Nets and gain better accuracy. The Base-DenseNet121 obtains the highest accuracy of 78.21% among all the Base-Nets, but DF-DenseNet121 gets an average accuracy of 79.34%, which achieves an improvement of 1.13%. Additionally, the DF-Nets increase the accuracy of 2.08% and 1.51% comparing to Base-MobileNetV2 and Base-ResNet50, respectively.

In Tab. 5, the DF-ResNet50 is compared with other techniques which uses ResNet50 as basic network. The PC-ResNet50 [Dubey, Gupta, Guo et al. (2018)] obtains an accuracy of 73.35%, which is better accuracy of 3.43% than ResNet50. In this paper, the DF-ResNet50 achieves 77.63% accuracy, which improves the almost 1.51% accuracy of Base-ResNet50, and its accuracy is 4.28% higher than PC-ResNet50. In summary, FGVC problems benefit from the DF-Net framework, and the DF-Net shows more prominent advantages for the lightweight network.

Table 5: The comparison of performance with the existing methods on Stanford Dogs (%)

Method	Accuracy	Δ
ResNet50	68.92	3.43
PC-ResNet50	73.35	
Base-ResNet50 (ours)	76.12	1.51
DF-ResNet50 (ours)	77.63	

4.3 UECFOOD-100

In this section, another dataset is used that is FGVC dataset (with bounding box), i.e., UECFOOD-100, which includes 100 food categories with 8643 images. Each image is annotated with a label and a bounding box that indicates the food location. In the experiments, the raw images are cropped from the given bounding boxes. The dataset is divided into 5 folds. The 3 folds are used for training and the rest for test.

Table 6: The accuracy achieved by Base-Nets and DF-Nets on UECFOOD-100 (%)

Model	Base-Net	DF-Net	Δ
MobileNetV2	80.63	82.29	1.66
ResNet50	82.87	84.00	1.13
DenseNet121	84.20	85.35	1.15

Unlike other datasets, this dataset is pre-processed before training. The input images are cropped with the provided object bounding boxes so that the dataset has more significant interclass similarity and intra-class variation than the original data. Even so, the proposed DF-Net still makes remarkable advancement. Here, the average accuracy of DF-Nets is above 1.1% than their subnets. Among all the DF-Nets, DF-DenseNet121 is best and beats other networks, and its classification accuracy reaches 85.35%. The experimental results strongly demonstrate that DF-Net achieves excellent performance for high-granularity classification tasks.

The performance comparison between DF-Nets and the existing techniques on UECFOOD-100 is shown in Tab. 7. All the evaluated methods use the same dividing and bounding box during experiments. Liu et al. [Liu, Cao, Luo et al. (2017)] proposed a practical deep learning-based food recognition system and reported the accuracy of 77.5% on

UECFOOD-100. Yanai et al. [Yanai and Kawano (2015)] fine-tuned DCNN which was pre-trained with a large food-related dataset and achieved the classification accuracy of 78.77%. Using DF-Net with DenseNet121 as the Base-Net, our framework obtains best accuracy on this food dataset, improving the accuracy over the published methods by 6.58% in Tab. 7. For computing time, DF-DenseNet121 takes 0.03 seconds per image for training using our proposed architecture. When the model is applied to image recognition, it only takes 0.01 seconds per image on average. As a comparison, the training time for Liu's method is usually around 2~3 seconds. The experimental results indicate that the DF-Net architecture is a significant improvement both computing time and recognition accuracy comparing to other existing methods.

Table 7: The comparison of performance with the existing methods on UECFOOD-100 (%)

Method	Accuracy
Liu	77.5
DCNN-FOOD (ft2)	78.77
DF-MobileNetV2 (ours)	82.29
DF-ResNet50 (ours)	84.00
DF-DenseNet121 (ours)	85.35

In summary, the above experimental results illustrate the superiority of the novel DF-Net architecture by exhibiting extensive improvement for image recognition compared to other published technologies. Most important, the DF-Net model improves the accuracy for image recognition without additional hardware overhead.

5 Conclusion

In this paper, an up-to-date automatic classification architecture Deep Fusion Convolutional Neural Networks (DF-Net) is proposed, where the model has a strong generalization ability for image recognition with the limited dataset and without additional hardware overhead. Specifically, DF-Net firstly organizes two identical subnets to catch more potential information from images in parallel. Then the extracted subnets' features are pre-processed with a 1×1 convolution kernel to reduce the features redundancy, and a channel shuffle operation to adopt information flow each other. Next, the fusion module is introduced to the end of subnets to reintegrate the subnets' features and generate more abundant and accurate fusion features for image recognition. Furthermore, the corresponding training strategy is proposed to speed up the convergence and reduce the computation overhead of network training. Finally, DF-Nets constructed based on well-known ResNet50, DenseNet121 and MobileNetV2 are evaluated on public datasets CIFAR100, Stanford Dogs, and UECFOOD-100 using accuracy measurement. Theoretical analysis and experimental results strongly demonstrate that DF-Nets achieve more advanced recognition performance than the existing research results. Additionally, the DF-Net framework is beneficial for the fine-grained visual classification tasks of small and medium datasets. Future studies will

acknowledge the importance of the above research and continue to explore more advanced fusion strategy to optimize the DF-Net.

Funding Statement: This work is partially supported by National Natural Foundation of China (Grant No. 61772561), the Key Research & Development Plan of Hunan Province (Grant No. 2018NK2012), the Degree & Postgraduate Education Reform Project of Hunan Province (Grant No. 2019JGYB154), the Postgraduate Excellent teaching team Project of Hunan Province (Grant [2019]370-133) and Teaching Reform Project of Central South University of Forestry and Technology (Grant No. 20180682).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Amin-Naji, M.; Aghagolzadeh, A.; Ezoji, M. (2019): Ensemble of CNN for multi-focus image fusion. *Information Fusion*, vol. 51, pp. 201-214.

Bharathidevi, B.; Chennamsetty, L. P.; Prasad, A. R.; Balijepalli, A. K. (2017): Logo matching for document image retrieval using SIFT descriptors. *International Journal of Engineering Research and Application*, vol. 7, no. 2, pp. 55-60.

Chollet, F. (2017): Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251-1258.

Dubey, A.; Gupta, O.; Guo, P.; Raskar, R.; Farrell, R. et al. (2018): Pairwise confusion for fine-grained visual classification. *Proceedings of the European Conference on Computer Vision*, pp. 70-86.

He, K.; Zhang, X.; Ren, S.; Sun, J. (2016): Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.

Hou, S.; Liu, X.; Wang, Z. (2017): DualNet: learn complementary features for image recognition. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 502-510.

Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Q. (2017): Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700-4708.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; Li, F. F. (2011): Novel dataset for fine-grained image categorization: Stanford Dogs. *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, no. 1.

Krizhevsky, A.; Hinton, G. (2009): Learning multiple layers of features from tiny images. *Technical Report.* University of Toronto.

Krizhevsky, A.; Sutskever, I.; Hinton, G. E. (2012): ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pp. 1097-1105.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E. et al. (1989): Backpropagation applied to handwritten zip code recognition. *Neural Computation*, vol. 1, no. 4, pp. 541-551.

Liu, C.; Cao, Y.; Luo, Y.; Chen, G.; Vokkarane, V. et al. (2017): A new deep learning-based food recognition system for dietary assessment on an edge computing service infrastructure. *IEEE Transactions on Services Computing*, vol. 11, no. 2, pp. 249-261.

Liu, Q.; Xiang, X.; Qin, J.; Tan, Y.; Tan, J. et al. (2020): Coverless steganography based on image retrieval of DenseNet features and DWT sequence mapping. *Knowledge-Based Systems*, vol. 92, no. 2020, pp. 105375-105389.

Luo, Y.; Qin, J.; Xiang, X.; Tan, Y.; Liu, Q. et al. (2020): Coverless real-time image information hiding based on image block matching and dense convolutional network. *Journal of Real-Time Image Processing*, vol. 17, no. 1, pp. 125-135.

Ma, B.; Li, X.; Xia, Y.; Zhang, Y. (2020): Autonomous deep learning: a genetic DCNN designer for image classification. *Neurocomputing*, pp. 152-161.

Matsuda, Y.; Hoashi, H.; Yanai, K. (2012): Recognition of multiple-food images by detecting candidate regions. *IEEE International Conference on Multimedia and Expo*, pp. 25-30.

Nasrabadi, N. M. (2019): Deeptarget: an automatic target recognition using deep convolutional neural networks. *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 6, pp. 2687-2697.

Pan, L.; Li, C.; Pouyanfar, S.; Chen, R.; Zhou, Y. (2020): A novel combinational convolutional neural network for automatic Food-Ingredient classification. *Computers, Materials & Continua*, vol. 62, no. 2, pp. 731-746.

Pan, L.; Pouyanfar, S.; Chen, H.; Qin, J.; Chen, S. C. (2017): DeepFood: automatic multi-class classification of food ingredients using deep learning. *IEEE 3rd International Conference on Collaboration and Internet Computing*, pp. 181-189.

Qin, J.; Li, H.; Xiang, X.; Tan, Y.; Pan, W. et al. (2019): An encrypted image retrieval method based on Harris Corner Optimization and LSH in cloud computing. *IEEE Access*, vol. 7, no. 1, pp. 24626-24633.

Qin, J.; Pan, W.; Xiang, X.; Tan, Y.; Hou, G. (2020): A biological image classification method based on improved CNN. *Ecological Informatics*, vol. 58.

Qin, J.; Sun, X.; Xiang, X.; Niu, C. (2009): Principal feature selection and fusion method for image steganalysis. *Journal of Electronic Imaging*, vol. 18, no. 3, pp. 1-14.

Simonyan, K.; Zisserman, A. (2014): Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Sugiarto, B.; Prakasa, E.; Wardoyo, R.; Damayanti, R.; Dewi, L. M. et al. (2017): Wood identification based on histogram of oriented gradient (HOG) feature and support vector machine (SVM) classifier. *2nd International Conferences on Information Technology*, pp. 337-341.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. A. (2017): Inception-v4, Inception-ResNet and the impact of residual connections on learning. *Thirty-first AAAI Conference on Artificial Intelligence.*

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S. et al. (2015): Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9.

Tan, Y.; Qin, J.; Xiang, X.; Ma, W.; Pan, W. et al. (2019): A robust watermarking scheme in YCbCr color space based on channel coding. *IEEE Access*, vol. 7, no. 1, pp. 25026-25036.

Wang, J.; Qin, J. H.; Xiang, X. Y.; Tan, Y.; Pan, N. (2019): CAPTCHA recognition based on deep convolutional neural network. *Mathematical Biosciences and Engineering*, vol. 16, no. 5, pp. 5851-5861.

Xiang, L.; Guo, G.; Yu, J.; Sheng, V. S.; Yang, P. (2020): A convolutional neural network-based linguistic steganalysis for synonym substitution steganography. *Mathematical Biosciences and Engineering*, vol. 17, no. 2, pp. 1041-1058.

Yanai, K.; Kawano, Y. (2015): Food image recognition using deep convolutional network with pre-training and fine-tuning. *IEEE International Conference on Multimedia & Expo Workshops*, pp. 1-6.

Zagoruyko, S.; Komodakis, N. (2016): Wide residual networks. arXiv:1605.07146.

Zhang, J.; Wang, W.; Lu, C.; Wang, J.; Sangaiah, A. K. (2019): Lightweight deep network for traffic sign classification. *Annals of Telecommunications*. https://doi.org/10.1007/s12243-019-00731-9.

Zhang, J.; Wu, Y.; Feng, W.; Wang, J. (2019): Spatially attentive visual tracking using multi-model adaptive response fusion. *IEEE Access*, vol. 7, no. 1, pp. 83873-83887.

Zhang, K.; Sun, M.; Han, T. X.; Yuan, X.; Guo, L. et al. (2017): Residual networks of residual networks: Multilevel residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1303-1314.

Zhang, X.; Zhou, X.; Lin, M.; Sun, J. (2018): ShuffleNet: an extremely efficient convolutional neural network for mobile devices. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848-6856.