International Core Journal of Engineering Volume 9 Issue 11, 2023
ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0027

Research on Fine-grained Bird Image Recognition based on Improved ResNet50

Zice Lu[1], Xiaofang Liu[2,*], Dewei Wang[2]

[1] School of Automation and Information Engineering, Sichuan University of Science & Engineering, Sichuan 643000, China

[2] School of Computer Science and Engineering, Sichuan University of Science & Engineering, Sichuan 643000, China

[*]Corresponding author: lxf1969@163.com

Abstract

Bird species are under great threat. In order to strengthen the protection of birds, this paper proposes an improved algorithm based on ResNet50 model to study the identification and classification of birds, in order to provide technical support for automatic bird image recognition system. In order to reduce the computational effort of the convolution process, the standard convolution with the convolution kernel size of 7×7 in the Input Backbone is replaced by a combination of 3×3 standard convolution and depth-separable convolution. In order to solve the problem of lost input feature mapping and information, the average pooling layer is added to the subsampling block of Stage4 in the backbone network to further improve the feature extraction capability. The improved ResNet50 model was compared with several classical classification models, and the Accuracy of the improved model reached 99.8% and the AUC reached 99.4%, which achieved excellent results in the task of bird classification and recognition.

Keywords

Image Classification; Resnet50; Depth-Separable Convolution; Average Pooling.

1. Introduction

The world's birds are in a critical situation, with many species on the brink of extinction. Habitat loss and degradation pose a great threat to the reproduction and survival of wild birds[1]. Therefore, it is urgent to strengthen the monitoring of birds and improve the protection measures of birds. Birds, as an important part of the ecological system, should not be allowed to get more damage than their ability to recover. However, with the rapid development of the global economy and the gradual improvement of the degree of industrialization, the ecological environment has suffered a lot of damage, so that many endangered species of birds have appeared. In addition, due to the wide variety of bird species and the complex diversity of bird characteristics, the protection of birds has caused great problems[2]. In order to protect birds, protect the diversity of species on earth, maintain ecological balance, the construction of automatic bird image recognition system has become an urgent thing to achieve. Therefore, image recognition and classification of bird species becomes a new research topic. In recent years, with the advent of the era of social information and big data and the substantial improvement of computer hardware computing capacity, artificial intelligence has gradually entered the public's vision, And the application of AI occurs in multiple academic fields and dimensions[3]. Related experts believe that artificial intelligence will lead the next industrial revolution.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0027

Computer vision is the "eyes" of artificial intelligence. It uses computers and related equipment to simulate the vision of organisms, accept the external image information for processing in the next step, and assist artificial intelligence to perceive the external world and make decisions and judgments[4]. Image recognition is one of the most fundamental and core problems in the field of computer vision. Its goal is to recognize and understand the content of the feature information in the image, and classify the image by distinguishing between each other. In terms of recognition content, Image recognition can be divided into two categories: universal image recognition and fine-grained image recognition. The task of universal image recognition is to distinguish large categories, such as people, trees, houses, etc., which is a coarse-grained image recognition task; Fine-grained image recognition, also known as subcategory image classification, is used to distinguish subcategories under large categories, such as different kinds of dogs such as corgi, Husky, collie, etc. Common fine-grained image recognition tasks include bird, flower, insect, plant, tea, food, car recognition, etc. Among all the fine-grained image recognition tasks, bird recognition has become one of the most typical and complex recognition tasks with its large intra-class differences and small inter-class differences, which has great academic research value and significance.

Bird recognition has been widely studied and applied in academia. Sharma[5] proposes a method for automatically identifying bird species. Image and audio classification models are built by recording bird images and audio using processing and classification techniques and built using pre-trained neural networks - ResNet50V2 and EfficientNetB0. The proposed model is also extended using multiple data sources in a dataset containing 137 bird species. The test and overall model accuracy performance was outstanding. Wang[6] proposes a method Based on Attention and Decoupled Knowledge Distillation, therefore, this model creates an efficient lightweight and fine-grained bird classification model, which achieves high accuracy in bird classification. Bin[7] summarized various research achievements of using artificial intelligence for bird recognition and analyzed the advantages and disadvantages of different types of application of this technology in airport bird prevention. Some scholars have also studied the classifier of bird species automatic recognition system based on image processing and support vector machine[8]. And some researchers have used artificial intelligence to identify birds by their sound[9,10]. Hai[11] present a novel invariant cues-aware feature concentration Transformer (TransIFC), which learns invariant and core information in bird images.

Bird recognition is a very classic task in the field of fine-grained image recognition, which is full of challenges. The related techniques and algorithms of fine-grained image recognition are also loved and studied by many scholars. Zhang[12] proposes a fine-grained image recognition framework using CNN as the original feature extractor. Fu[13] propose a novel Recurrent Attention Convolutional Neural Network(RA-CNN) which recursively learns discriminative region attention and region-based feature representation at multiple scales in a mutually reinforced way. In order to solve the problem of lack of distinguishing ability of features, this paper propose a novel few-shot fine-grained image classification framework which enhances the Discriminative ability of Local structures utilizing class-aware Global structures (DLG)[14]. Sun[15] proposed attention-based Convolutional Neural Network (CNN) for fine-grained recognition based on dogs, which can modulate multiple target segments in different input images. The method first learns the multiple attention-area characteristics of each input image via the One-Squeeze Multi-Excitation (OSME) module. Then, Multi-Attention Multi-Class Constraint (MAMC) is applied in the metric learning framework. The method can be easily trained end-to-end and only requires one training session to be highly efficient. In addition, He[16] have proposed a novel converter architecture named "TransFG" for fine-grained recognition.

Through the rapidly developing computer vision technology, an image recognition system that can automatically identify bird images is successfully built. On the one hand, it can promote the protection of birds in the nature reserve, especially endangered birds, and on the other hand, the protection of birds can better maintain the ecological balance and protect the diversity of species.In conclusion, this paper is devoted to the research of bird image recognition based on deep learning.The research of bird image recognition technology based on convolutional neural network firstly has strong social

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0027

reality value and ecological environmental protection value, and secondly, the task itself also has very strong scientific and academic research value.

2. Materials and Methods

2.1 Algorithm Selection

There are many kinds of algorithms used for image classification, and the research on this kind of algorithm is a hot topic at present. Wang[17] proposes a convolutional neural network "residual attention network" using attention mechanism is proposed, which combines end-to-end training with the most advanced feedforward network architecture for image classification. Arco[18] proposes a multi-level integrated classification system based on Bayesian deep learning methods, which can achieve maximum performance and provide uncertainty for each classification decision. Kuswantori[19] proposes a fish recognition algorithm based on YOLOv4 was proposed and optimized with unique labeling technology. Optimize the automatic detection and classification of fish by using the video of real fish running on the conveyor belt as the experimental object. Gajula[20] proposes the ResNet50 model that was used to distinguish between tumor and non-tumor images, and the ability to find and classify them accurately was performed well.

Reading bird recognition papers and related image classification algorithms, this paper decided to use the improved ResNet50 network for bird recognition classification training. ResNet50 network is widely used for image recognition and classification tasks. Sharma[21] proposes the improved ResNet50 and enhanced watershed segmentation (EWS) algorithms are integrated into brain tumor classification and deep feature extraction to explain the construction of a new technique. The improved layer structure includes 5 convolution layers and 3 fully connected layers. The method can maintain the best computational efficiency under the high dimensional depth characteristics. Researcher extracted basic information from speech spectrogram through ResNet 50 training deep learning network for classification layer and gender detection[22]. Nijaguna[23] proposes that ResNet50 and VGG16 were used for feature extraction, and the results showed good performance. Selvaraj[24] proposes Resnet50 architecture was used for breed identification, which performed better than other models.

ResNet contains a series of network models with different layers[25], including 18 layers, 34 layers, 50 layers, 101 layers and so on. This paper selects the ResNet model of 50 layers, namely ResNet50. ResNet50 has a total of 50 layers, with 49 convolution layers and 1 full connection layer.

ResNet50 consists of an Input Backbone, a convolutional part and an output layer. The input to ResNet50 is a bird image of size 224×224. The Input Backbone consists of a 7×7 large convolution kernel with 2 steps and 64 channels, and a maximum pooling layer with 3×3 sizes and 2 steps. The convolution part consists of four stages, Stage1-Stage4. Stage1-Stage4 all start with a down-sampling module, followed by several general residual modules. The down-sampling module is a convolution layer on both the main network path and the branch path, while the general residual module only has a convolution layer on the main network path and no convolution layer on the branch path. The general residual module can be divided into Identity Block and Conv Block. The output layer includes an average pooling layer and a fully connected layer.

2.2 Improved ResNet50

2.2.1 Improvment Introduction

There are two improvements in this paper. First, the standard convolution with the convolution kernel size of 7×7 in the Input Backbone is replaced by a combination of 3×3 standard convolution and depth-separable convolution, so as to reduce the amount of calculation in the convolution process. Secondly, in order to solve the problem of easy loss of input feature mapping and information loss, average pooling layer was added to the subsampling block of Stage4 in the backbone network to further improve the capability of network feature extraction.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0027

2.2.2 Preparatory Knowledge

To understand the improvements in the Input Backbone, first we need to learn what deep separable convolution is. Here is how it differs from the general convolution operation. General Convolution content is shown in the Fig. 1.

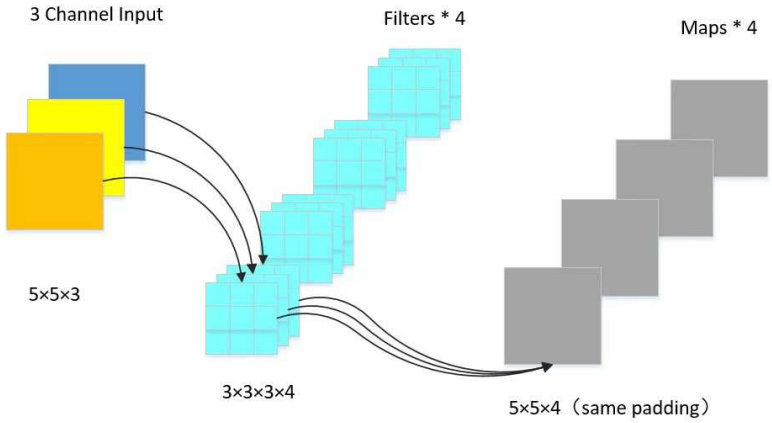

Fig. 1 General Convolution

For an input picture, there are three channels, each of which is 5 by 5 pixels. After four convolution kernels with shape of 3×3×3, four Feature maps with size of 5×5 are finally output. For easy understanding, assuming the image convolution has the same padding, the size is the same as that of the input layer. The number of convolution kernel parameters of general convolution can be calculated by the following formula:

$$SUM\_N = 4 \times 3 \times 3 \times 3 = 108 \tag{1}$$

As shown in Fig. 2, Depthwise Separable Convolution can be divided into depthwise (DW) and pointwise (PW) two parts, used to extract feature map, compared with general Convolution, it has the advantage of less number of parameters and operation cost.

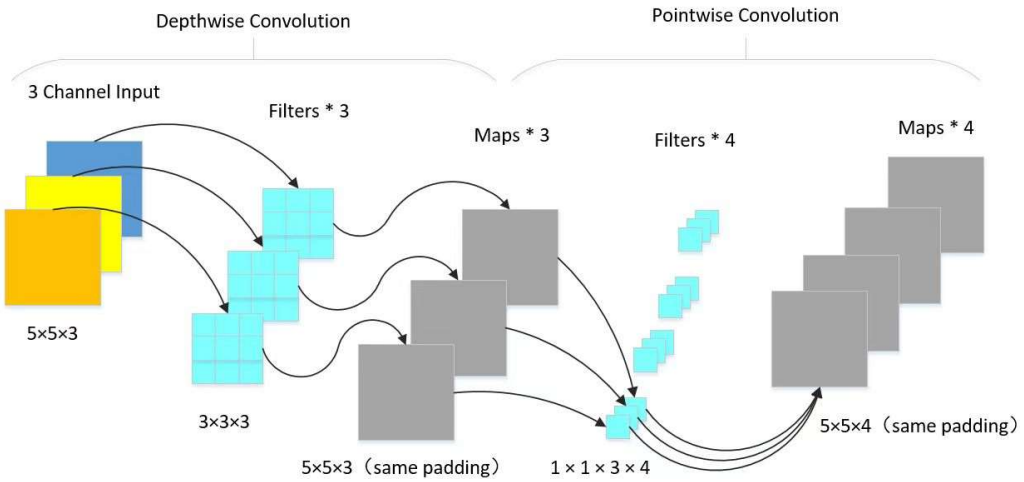

Fig. 2 Depthwise Separable Convolution

A convolution kernel and channel of depthwise convolution have one-to-one correspondence. The same input image. There are three channels, and each channel is 5 by 5 pixels. Depthwise Convolution

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0027

first generates 3 Feature maps with a size of 5×5 (assuming the same padding).This process is shown on the left side of Fig. 2.

Pointwise Convolution is very similar to general convolution. The size of its convolution kernel is 1×1×M, where M is the number of channels in the previous layer. Therefore, Pointwise Convolution weights maps of the previous step in depth direction to generate new Feature maps. After Pointwise Convolution, four Feature maps are also output, with the same output dimension as that of General Convolution.

The Convolution kernel parameters contained Depthwise Separable Convolution is two parts together to get:

$$SUM\_DW = 3\times3\times3 = 27 \tag{2}$$

$$SUM\_PW = 1\times1\times3\times4 = 12 \tag{3}$$

$$SUM\_SEP = SUM\_DW + SUM\_PW = 39 \tag{4}$$

The same input, is also to get four Feature map, Depthwise Separable Convolution number of parameters is about one-third of the General Convolution.

2.2.3 Improvements to The Input Backbone

The Input Backbone of the ResNet50 network is mainly composed of a 7×7 convolution kernel and a maximum pooling layer. The calculated cost is the square of the convolution kernel width or height. Therefore, this paper replaces the standard convolution whose convolution kernel size is 7×7 in the Input Backbone with a combination of 3×3 standard convolution and depth-separable convolution, so as to reduce the computation in the convolution process and deepen the depth of the network. Use Depthwise Separable Convolution improved Input Backbone process is shown in fig. 3.

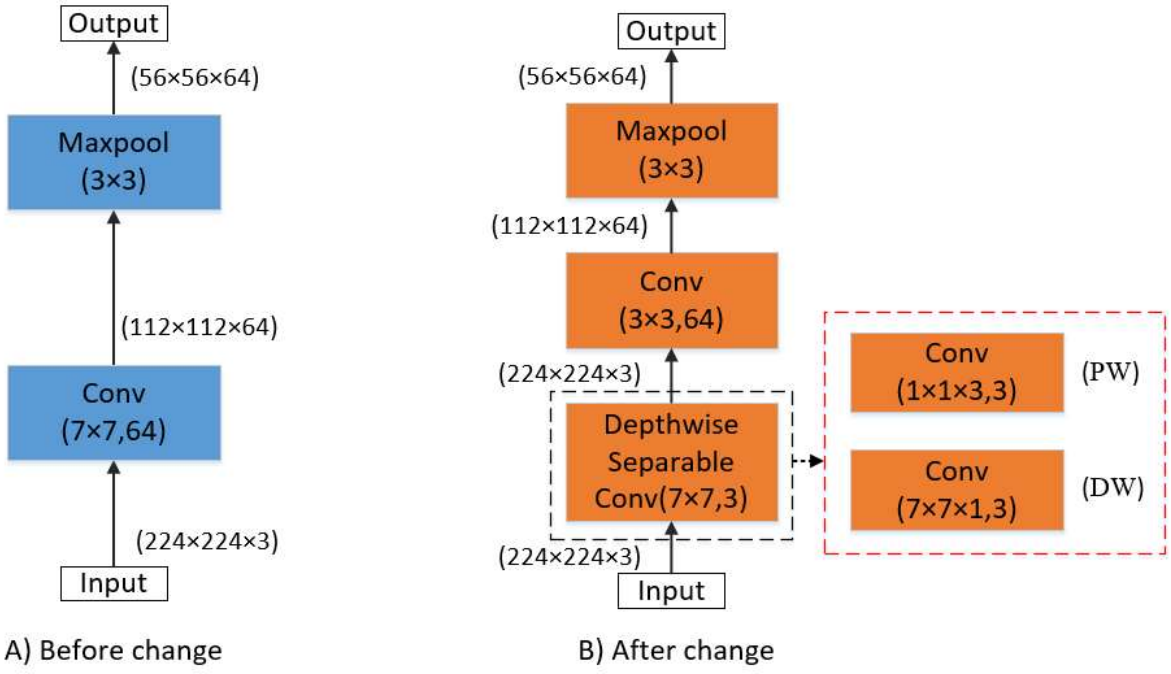

Fig. 3 Improved Input Backbone Process

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0027

First, the image is entered into DW, which uses three convolution kernels of size 7×7×1, instead of a single convolution kernel of size 7×7×3. Each convolution kernel convolves just one of the channels in the input layer, resulting in an output of size 224×224×3. PW uses a convolution kernel with a size of 1×1×3 for point-by-point convolution. After convolution of each convolution kernel to the input image, a mapping with a size of 224×224×1 can be obtained. After three point-by-point convolution, an output image with a size of 224×224×3 can be obtained.Then input the 3x3 standard convolution with channel size of 3, its stride size is 2, and using the same padding convolution, then the output will become 112×112×3 picture. Finally, after a 3×3×64 maximum pooling layer, the output size of 56×56×64 is the same as before the improvement. Depth-separable convolution not only has the same receptive field as a standard convolution with a convolution kernel size of 7×7. And with the Depthwise Separable Convolution improved part model parameters was only about 20% of the original Convolution kernels. It greatly reduces the calculation cost of classification network, deepens the depth of network and improves the accuracy of classification model.

2.2.4 Improvements in The Stage4 Down-Sampling Module

As mentioned above, Stage4 starts with a down-sampling module, followed by several general residual modules. The down-sampling module has convolution layers on both the main network path and the branch path, while the general residual module only has convolution layer on the main network path and no convolution layer on the branch path. In this paper, the model is improved in the Down-Sampling Module of Stage4.

In the Main Path of Down-Sampling, when the input data passes through the convolution layer with the convolution kernel size of 1×1 and step size of 2, the input feature mapping is easy to be lost, leading to information loss. Based on this weakness, this paper proposes to modify the step size to 1 when data passes through the first convolution layer in the Main Path. After the second convolution layer, its step size is changed to 2. AvgPooling is commonly used when it is required to combine all information on the feature map to make a decision. For example, in the field of image segmentation, Global AvgPooling is used to obtain global context information. Therefore, in Branch Path, an average pooled layer with step size 2 and convolution kernel size 2×2 is added before the convolutional layer. Fig. 4 shows the comparison between before and after improvement.

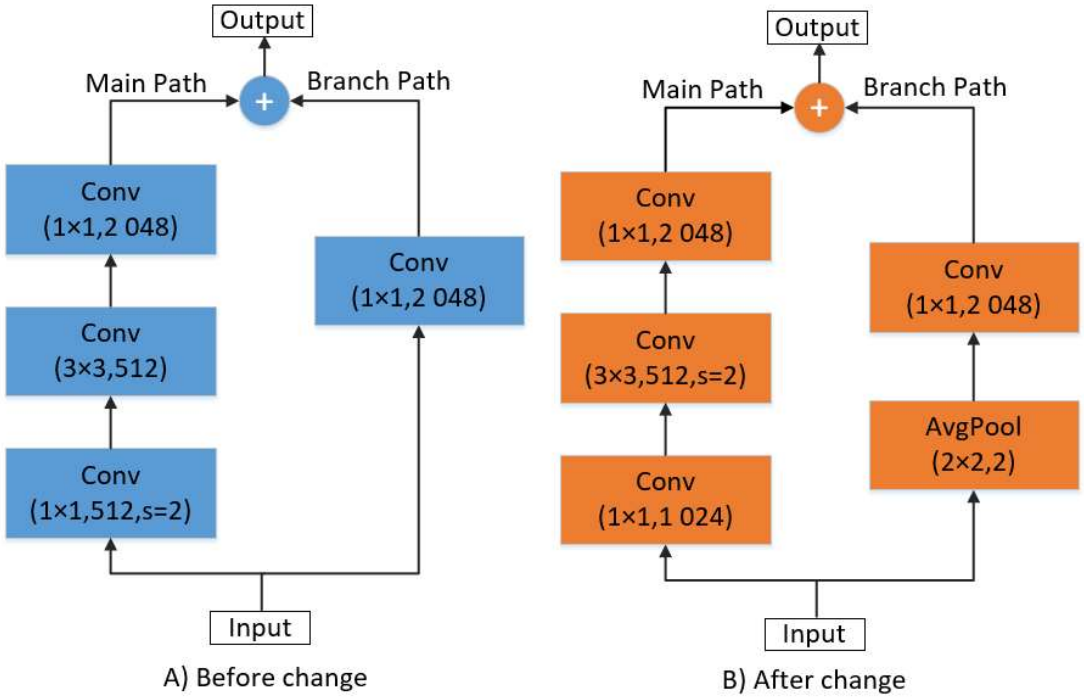

Fig. 4 Improved Stage4 Down-Sampling Module

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0027

3. Experiment

The experimental data set for this study is 62388 photos of 400 bird species. Fig. 5 shows part of the data set. The training set contains 58,388 images, the verification set contains 1,997 images, and the test set contains 2,003 images. The computer operating system is Windows 10 Professional Edition, the GPU is NVIDIA GeForce RTX 3060 with 8GB of video memory, the CPU is i7-12700, and the computer memory is 32GB. The deep learning framework adopted is PyTorch2.0.1, and the epoch number is 120 times. The improved ResNet50 model and ResNet50, VGG19, AlexNet and GoogLeNet were used for bird classification recognition and comparison experiments. Fig. 6 shows the results of the improved ResNet50 classification. Table 1 shows the performance comparison of the classification results of different models.

Fig. 5 Part of the data set

Fig. 6 Improved ResNet50 classification result graph

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0027

Table 1. Performance comparison of classification results of different models

Model	Accuracy	Scheme 2
AlexNet	98.9	98.2
VGG19	98.4	97.6
GoogLeNet	97.2	98.0
ResNet50	98.7	97.4
improved ResNet50	99.8	99.4

Accuracy and AUC (Area Under the Curve) are common indicators used to evaluate the performance of classification models. Accuracy is used to evaluate classification accuracy, representing the proportion of samples predicted correctly by the model in the total number of samples. AUC represents the area under the Receiver Operating Characteristic (ROC) curve. The AUC can comprehensively evaluate the classification capability of the model, regardless of the selection of classification thresholds. The value ranges from 0.5 to 1. A larger value indicates better model performance.

As can be seen from Table 1, Accuracy and AUC improved by ResNet50 in this study are both the best, with Accuracy reaching 99.8% and AUC reaching 99.4%, respectively. Compared with AlexNet, VGG19, GoogLeNet and ResNet50, Accuracy increased by 0.9%, 1.4%, 2.6% and 1.1%, and AUC increased by 1.2%, 1.8%, 1.4% and 2%.

4. Summary

In this paper, an improved ResNet50 model is proposed for bird recognition and classification. Based on ResNet50, the algorithm replaces the standard convolution of the input trunk with a combination of standard convolution and dept_separable convolution, thus reducing the computational load in the convolution process. The average pooling layer is added to the subsample block of Stage4 in the backbone network to further improve the feature extraction capability of the network. The improved ResNet50 model was compared with several classical classification models, and the Accuracy reached 99.8% and AUC reached 99.4%, which effectively realized the classification effect of bird identification and provided strong technical support for bird protection.

Acknowledgments

Funding: Supported by The Innovation Fund of Postgraduate, Sichuan University of Science & Engineering under Grant Y2022158.

References

[1] Nugent D T, Baker-Gabb D J, Green P, et al. Multi-scale habitat selection by a cryptic, critically endangered grassland bird-The Plains-wanderer (Pedionomus torquatus): Implications for habitat management and conservation[J]. Austral Ecology, 2022, 47(3): 698-712.

[2] Lin C W, Hong S, Lin M, et al. Bird posture recognition based on target keypoints estimation in dual-task convolutional neural networks[J]. Ecological Indicators, 2022, 135: 108506.

[3] Tavares, Diana, et al. The Intersection of Artificial Intelligence, Telemedicine, and Neurophysiology: Opportunities and Challenges[J]. Handbook of Research on Instructional Technologies in Health Education and Allied Disciplines, 2023(7): 130-152.

[4] Xiao S, Wang Y, Perkes A, et al. Multi-view Tracking, Re-ID, and Social Network Analysis of a Flock of Visually Similar Birds in an Outdoor Aviary[J]. International Journal of Computer Vision, 2023: 1-18.

[5] Sharma N, Vijayeendra A, Gopakumar V, et al. Automatic identification of bird species using audio/video processing[C]. In: 2022 International Conference for Advancement in Technology (ICONAT). IEEE, 2022: 1-6.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0027

[6] Wang K, Yang F, Chen Z, et al. A Fine-Grained Bird Classification Method Based on Attention and Decoupled Knowledge Distillation[J]. Animals, 2023, 13(2): 264.

[7] Guo B, Du W, Cheng L, et al. Application of artificial intelligence bird recognition technology in airport bird strike prevention safety management[C]. In: IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2020, 565(1): 012092.

[8] Chandra B, Raja S K S, Gujjar R V, et al. Automated bird species recognition system based on image processing and svm classifier[J]. Turkish Journal of Computer and Mathematics Education, 2021, 12(2): 351-356.

[9] Qiu Z, Zhu X, Liao C, et al. Detection of bird species related to transmission line faults based on lightweight convolutional neural network[J]. IET Generation, Transmission & Distribution, 2022, 16(5): 869-881.

[10] Yang F, Jiang Y, Xu Y. Design of Bird Sound Recognition Model Based on Lightweight[J]. IEEE Access, 2022, 10: 85189-85198.

[11] Liu H, Zhang C, Deng Y, et al. TransIFC: Invariant Cues-aware Feature Concentration Learning for Efficient Fine-grained Bird Image Classification[J]. IEEE Transactions on Multimedia, 2023: 1-14.

[12] Zhang W, Yan J, Shi W, et al. Refining deep convolutional features for improving fine-grained image recognition[J]. EURASIP Journal on Image and Video Processing, 2017, 2017: 1-10.

[13] Fu J, Zheng H, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4438-4446.

[14] Cao S, Wang W, Zhang J, et al. A few-shot fine-grained image classification method leveraging global and local structures[J]. International Journal of Machine Learning and Cybernetics, 2022, 13(8): 2273-2281.

[15] Sun M, Yuan Y, Zhou F, et al. Multi-attention multi-class constraint for fine-grained image recognition[C]. In: Proceedings of the european conference on computer vision (ECCV). 2018: 805-821.

[16] He J, Chen J N, Liu S, et al. Transfg: A transformer architecture for fine-grained recognition[C]. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(1): 852-860.

[17] Wang F, Jiang M, Qian C, et al. Residual attention network for image classification[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3156-3164.

[18] Arco J E, Ortiz A, Ramírez J, et al. Uncertainty-driven ensembles of multi-scale deep architectures for image classification[J]. Information Fusion, 2023, 89: 53-65.

[19] Kuswantori A, Suesut T, Tangsrirat W, et al. Fish Detection and Classification for Automatic Sorting System with an Optimized YOLO Algorithm[J]. Applied Sciences, 2023, 13(6): 3812.

[20] Gajula S, Rajesh V. Performance Analysis of Classification and Detection of Brain Tumor MRI Images Using Resnet50 Deep Residual U-Net[J]. Journal of Iranian Medical Council, 2023, 6(1): 101-110.

[21] Sharma A K, Nandal A, Dhaka A, et al. Enhanced watershed segmentation algorithm-based modified ResNet50 model for brain tumor detection[J]. BioMed Research International, 2022, 2022.

[22] Alnuaim A A, Zakariah M, Shashidhar C, et al. Speaker gender recognition based on deep neural networks and ResNet50[J]. Wireless Communications and Mobile Computing, 2022, 2022: 1-13.

[23] Nijaguna G S, Babu J A, Parameshachari B D, et al. Quantum Fruit Fly algorithm and ResNet50-VGG16 for medical diagnosis[J]. Applied Soft Computing, 2023, 136: 110055.

[24] Selvaraj S, Thangarajan R, Anbukkarasi S. ResNet50 Architecture Based Dog Breed Identification Using Deep Learning[J]. Applied and Computational Engineering, 2023: 251-259.

[25] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.