Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/15/20230802

Cat classification based on improved ResNet50

Shipeng Sun

College of Mathematics and Statistics, Guangdong University of Technology, No. 161 Yinglong Road, 510000, Guangzhou, China

3120007004@mail2.gdut.edu.cn

Abstract. Cat species recognition holds significant potential in many fields. The primary objective of this research is to develop an automated algorithm for recognizing the presence of cats in images. The application prospects of this algorithm are diverse and include security, image search, and social media. Hence, this research has considerable practical value in various domains. In this study, we propose a cat image recognition algorithm based on the PyTorch, with ResNet50 as the foundational network architecture, and an attention mechanism (Efficient Channel Attention) integrated into the model for improved performance. We first introduced the Resnet network, and then introduced the combination of attention mechanism and Resnet in detail The proposed model achieved a 92.37% accuracy rate in classifying the 12 cat species, demonstrating its efficacy in accurately classifying and recognizing the collected images. The research conclusion of this paper has certain reference value.

Keywords: cat classification, attention, Resnet, efficient channel attention.

1. Introduction

With the rapid development of computer technology in the 21st century, artificial intelligence has pervaded various fields. Machine vision, an important branch of artificial intelligence, simulates human visual function to analyze and understand target images through feature extraction, ultimately achieving target classification and recognition. Machine vision technology has been applied to multi-font Chinese character recognition, Chinese medicine image recognition, car logo recognition algorithm research, and recently, cat species identification, which has broad application prospects in pet insurance, pet portrait, pet retail, pet health management, and other related fields. Researchers have conducted extensive research on the application of machine vision technology in cat classification and recognition.

Various classification methods have been studied, such as the eigenvalue recognition algorithm which is simple to operate, but with inadequate recognition rates [1]. The correlation coefficient recognition algorithm, which has high recognition accuracy but appears to have a narrow application range. The hierarchical classification algorithm has large calculations, and the support vector machine algorithm has high recognition rates but has difficulty in implementing large-scale training samples and requires manual selection of eigenvalues [2]. In comparison, the Convolutional Neural Network (CNN) has independent learning ability and good robustness. However, it requires a large amount of training data, and as the number of network layers deepens, the computational complexity and training model cycle increase. With the deepening of the network, the accuracy of the training set decreases, which affects the recognition effect and accuracy.

DOI: 10.54254/2755-2721/15/20230802

Based on the above analysis, this paper proposes a cat species recognition algorithm based on ResNet50, which fuse residual layer features and adds an attention mechanism to improve algorithm performance.

2. Mehtod

In the deep architecture of Convolutional Neural Networks (CNNs), the problem of gradient vanishing and explosion can impede training set accuracy as the network layers become increasingly deep. This issue can be resolved by employing residual networks, which allow for enhanced network performance despite increases in depth [3-4]. Specifically, a residual block structure, as depicted in Figure 1, is utilized, where ResNet includes both identity mappings and residual mappings. In this model, the image is convolved and processed through four residual modules to prevent overfitting, followed by one more convolutional layer to produce the output.

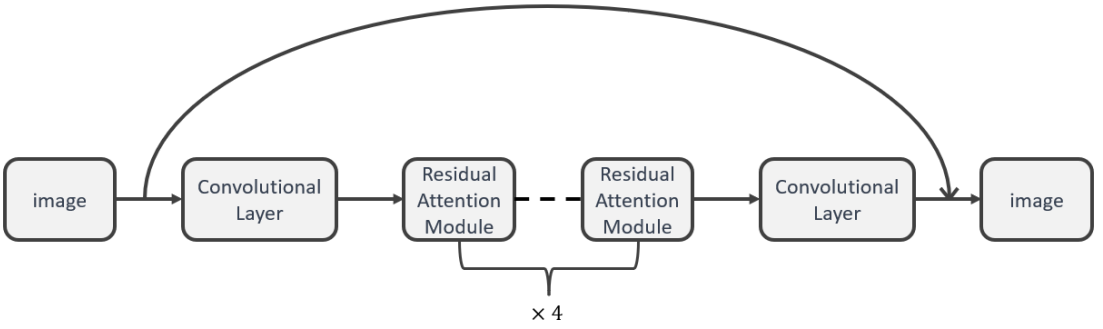

Figure 1. Main body framework diagram.

2.1. Backbone ResNet50

ResNet50 is composed of 49 convolutional layers and 1 fully connected layer. Specifically, ID BLOCK x2 in the second to fifth stages denotes two residual blocks that maintain the spatial dimensions, whereas CONV BLOCK refers to a residual block that introduces spatial scaling. Each residual block consists of three convolutional layers, yielding a cumulative count of 1 + 3 x (3+4+6+3) = 49 convolutional layers. The ResNet50 architecture expects the input images to have a size of 256 × 256 × 3, which is not compatible with the required input size of 224 × 224 × 3. To resolve this incompatibility, a preprocessing step is employed to normalize the input images and crop them to the required size [5]. Specifically, the mean value of each channel across all images in the training set is subtracted to normalize the images, and then the images are cropped to the specified size. After passing through consecutive convolutional operations in the residual blocks, the depth of the pixel matrix channels of the image's increases. The Flatten layer is then applied to transform the pixel matrix into a 2D tensor of shape batch_size × 2048, which is then fed into the fully connected layer (FC).

The softmax classifier subsequently outputs the corresponding class probabilities. ResNet50 architecture utilizes skip connections, which are implemented through shortcut connections to propagate the input across layers and add it to the output that has undergone convolution. This process facilitates the effective training of the lower layers of the network, leading to a remarkable increase in accuracy as the depth of the network increases [6-7]. The shortcut connection can be viewed as performing an equivalent identity mapping, which does not introduce additional parameters or computational complexity to the network. In this case, the model is effectively reduced to a shallow network, and the key challenge is to accurately learn the identity mapping function $H(x)=x$. Directly fitting such a latent function can pose a significant challenge. Let $H(x)$ and $F(x)$ represent the output of the residual network and the output after convolution, respectively, such that

$$H(x)=F(x)+x \tag{1}$$

$$F(x) = (\omega_3\delta(\omega_2\delta(\omega_1 x+b))) \tag{2}$$

DOI: 10.54254/2755-2721/15/20230802

where ω represents convolutional operations and δ denotes activation functions. The problem of learning an identity mapping function can be reformulated as the task of training a residual function F(*x*)=H(*x*)-*x* that is easy to optimize [8].

2.2. Residual attention module based on ECA attention module

Distinctive variations in physical characteristics such as color, eye shape and size, ear morphology, coat type, and body proportions are readily apparent across various feline breeds. A number of feline breeds are typified by orbicular and voluminous ocular structures, whereas others exhibit slenderer and reduced ocular configurations. Likewise, some breeds are characterized by relatively diminutive aural appendages, while others possess comparatively larger and more erect auricular attributes. Variations in coat texture and length are also evident among feline breeds, with some breeds displaying velvety and extensive pelage, while others exhibit compact and abbreviated fur coverage. Additionally, certain breeds are distinguished by squat and plump body morphologies, while others are identified by elongated and slender body proportions [9]. This model improves upon the ResNet50 architecture by integrating an attention mechanism, which enhances the model's ability to selectively attend to highly discriminative features, thereby mitigating the effects of confounding information.

The Efficient Channel Attention Module (ECA) is an improvement over SENet, as the dimensionality reduction in SENet can have side effects on the channel attention mechanism, and capturing all interdependencies between channels is not always necessary or efficient (Figure 2). The ECA module uses a 1x1 convolutional layer directly after the global average pooling layer, and removes the fully connected layer. This module avoids dimension reduction and effectively captures cross-channel interactions. Additionally, ECA model achieves good results with few parameters. The Efficient Channel Attention Module is a channel attention mechanism frequently employed in visual models. Its plug-and-play design allows for simple integration into existing architectures, enabling it to enhance the channel features of input feature maps without altering their size. Consequently, the ECA module serves to strengthen channel features, while maintaining the original size of the input feature map. The ECA-Net addresses the potential drawbacks of channel attention prediction caused by dimension reduction in the SENet, and the inefficiencies of capturing dependencies among all channels.

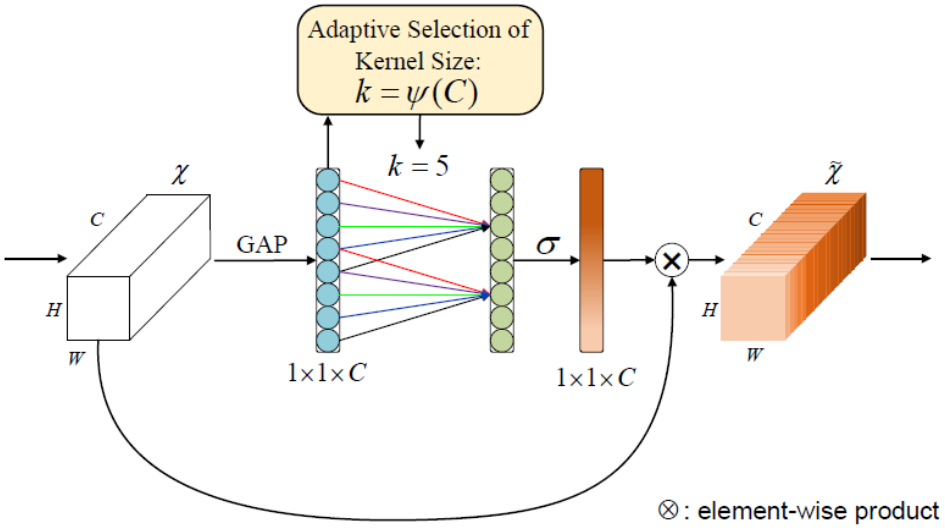

Figure 2. ECA model.

The ECA-Net achieves this by modifying the SENet architecture, replacing the fully connected layer (FC) responsible for learning channel attention information with 1x1 convolutional layers. This modification avoids the reduction of the channel dimension and reduces the overall number of parameters required for learning channel attention information [10-11]. Compared to the FC layer, the

DOI: 10.54254/2755-2721/15/20230802

1x1 convolutional layer has fewer parameters, making it a more efficient alternative. The ECA model follows a process in which an input feature map with dimensions of H * W * C undergoes spatial feature compression using global average pooling (GAP) on the spatial dimensions, yielding a feature map of 1 * 1 * C. Subsequently, a 1 * 1 convolution is employed to capture inter-channel dependencies and learn the relative importance of different channels, producing an output feature map with dimensions of 1 * 1 * C. Finally, a channel attention mechanism is integrated, whereby the channel attention feature map (1 * 1 * C) and the original input feature map (H * W * C) are element-wise multiplied on a channel-by-channel basis to produce a feature map with channel attention (Figure 3).

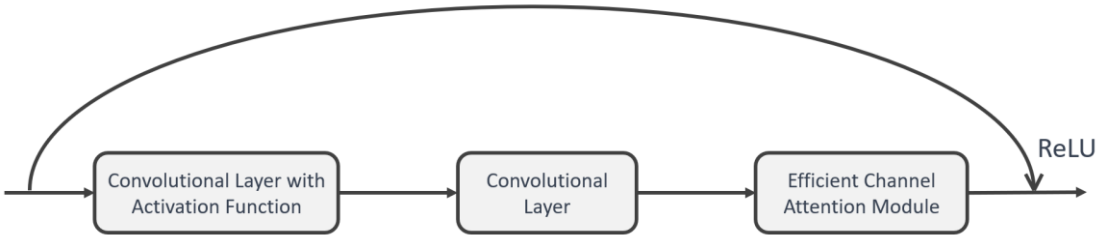

Figure 3. Residual block based on ECA attention.

3. Result

The aim of this experiment is to classify images of different cat breeds using deep learning algorithms. The dataset contains several thousand cat images, encompassing 12 different cat breeds. ResNet50 was selected as the base model and the ECA attention mechanism was added to improve the classification performance. The model was adjusted and optimized for hyperparameters during the training process.

3.1. Data set introduction

This paper employs the dataset provided by the Paddle platform's Cat Twelve-Classification Competition. The training dataset consists of 2,160 images of cats, divided into 12 categories. The testing dataset comprises 240 images of cats, without any annotation information (Figure 4).

In this study, the dataset was initially shuffled and partitioned into three distinct subsets: a training set, which accounted for 60% of the total dataset, a validation set (20%), and a testing set (20%). Uniform preprocessing techniques were implemented across all subsets to ensure consistency and accuracy of the validation and testing results. Firstly, the images were cropped to dimensions of 224x224 to ensure their compatibility with the input size of the model. The pixel values of the images were standardized, and finally, the dataset was divided into batches of size "32" to facilitate the training of the model in subsequent stages.

Figure 4. Data set diagram.

3.2. Hardware and software platform

The experimental hardware configuration used in this study consisted of a 64-bit Windows 10 operating system, a 2.60 GHz Intel i7 CPU, and an NVIDIA GeForce RTX 2060 graphics card based on the Pascal architecture. For software, PyCharm 2022.3.3 was used as the development platform, with the PyTorch

DOI: 10.54254/2755-2721/15/20230802

open-source deep learning framework selected as the programming framework, with a version of 1.10.2. The program was designed using Python 3.6.13.

The loss curve demonstrates that with an increase in the number of training iterations, the loss function progressively decreases until it converges to a minimum value (Figure 5). The experimental results demonstrate that the proposed model has high classification performance and can effectively classify different cat breeds. Ultimately, the accuracy on the test set reaches 92.37%. Figure 6 shows the classification results for different cats. The numbers in "[]" are confidence (%), followed by the classification results for cats.

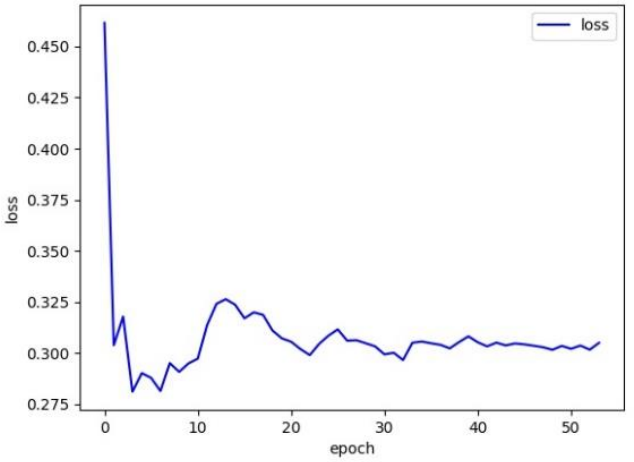

Figure 5. loss curve.

Figure 6. Classification result.

This study presents a novel deep learning network model based on ResNet50 for accurate recognition and classification of cat images, using the latest advancements in image classification techniques. The residual blocks incorporated in the model address the issue of network degradation, ensuring that the model's performance does not deteriorate with the increase in network depth. This proposed model outperforms the existing models in terms of its increased depth, faster convergence rate, higher precision, and superior generalizability. The feasibility of the proposed model has been demonstrated in practical applications of cat image classification, providing an effective solution for animal recognition.

4. Conclusion

The present study introduces an ECA attention mechanism to the ResNet50 architecture, resulting in a faster convergence rate, shorter training time, and improved performance. The residual network effectively addresses the issue of difficult training in deep networks, leading to a high accuracy rate of

DOI: 10.54254/2755-2721/15/20230802

92.37%. This model not only demonstrates its potential in cat breed image classification but also holds applicability in classifying other animals such as dogs. Although the ResNet50 model achieved the expected classification performance, the limited size of the dataset used in this study necessitates future work with a larger dataset to further improve the recognition accuracy.

References

[1] Edgar Solomonik, Grey Ballard. A Communication-Avoiding Parallel Algorithm for the Symmetric Eigenvalue Problem. 2017 *Sym. Para. Alg. Arch*. 111–121.

[2] An Zeng, Qi-Gang Gao, and Dan Pan. A global unsupervised data discretization algorithm based on collective correlation coefficient. 2011 *Conf. Ind. Eng. Appl. Intel. Sys.* 146–155.

[3] Susan Dumais and Hao Chen. Hierarchical classification of Web content. 2000 *Conf. Res. Deve. Infor. Ret*. 256–263.

[4] Glenn Fung and Olvi L. Mangasarian. Proximal support vector machine classifiers. 2001 *Conf. Knowl. Disc. Data Min.*, 77–86.

[5] Nestler E G, Osqui M M, Bernstein J G. Convolutional Neural Network, 2017, *Wire. Net.*, **201 (74)**,137-151.

[6] Weijie Liu, Weiwei Chen, and Xinmiao Dai. Capsule Embedded ResNet for Image Classification. 2021 *Conf. Com. Sci. Arti. Intel.*, 143–149.

[7] Kaiming He, Xiangyu, and Jian Sun. Deep residual learning for image recognition. 2016 *Computer Vision and Pattern Recognition*, 1-10.

[8] Wu Z, Shen C, Hengel A. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. Pattern Recognition, 2016 *Comp. Vis. Pat. Rec.*,1-12.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. 2018, *Comp. Vis. Pat. Rec*, 201-211.

[10] Santosh Kumar Mishra, Gaurav Rai, Sriparna Saha, and Pushpak Bhattacharyya. 2021. *Effic. Cha. Att. En.*, **21 3**, Article 49, 17

[11] Chen P. Efficient Channel Allocation Tree Generation for Data Broadcasting in a Mobile Computing Environment. 2003, *Wire. Net.*, 200-213.