ISSN: 2414-1895 DOI: 10.6919/ICJE.202402_10(2).0006

Improved Image Classification Algorithm based on Siamese Network

Mingyi Ma[1,2, a], Gongquan Tan[1,2, b], Weiguang Li[1,2, c], Xinyu Tang[1, d]

[1] School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin, Sichuan 644000, China

[2] Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science&Engineering, Yibin, Sichuan 644002, China

[a]1414756135@qq.com, [b]tgq11@126.com, [c]14465340@qq.com, [d]876078241@qq.com

Abstract

The current image classification algorithms need help classifying small sample data sets, including limited sample sizes, inadequate classification accuracy, and intricate model calculations. The objective of this paper is to suggest a siamese network-based algorithm for classifying images that is characterized by its lightweight nature. To begin with, the ResNext101 network is incorporated into the feature extraction backbone network. The purpose of the lightweight design is to minimize the model's parameters and calculations. The ECA attention mechanism is implemented to increase the model's focus on the image, and the feature image is combined to enhance the algorithm's classification accuracy. According to the research, the Omniglot data set yielded a 75.1% classification accuracy for the improved algorithm. There has been a 15% rise in comparison to Siamese-CNN. The algorithm ensures both accurate classification and a lightweight design.

Keywords

Image Classification; Siamese Network; Feature Fusion; Lightweight.

1. Introduction

Siamese networks have recently become increasingly popular for identifying and classifying datasets with limited data sets. Tan Jiachen and colleagues. [3] presented a classification model for retinal fundus images of diabetes utilizing siamese networks; a more intricate SSM model is employed to minimize the labor expenses associated with medical classification tasks. Additionally, the ResNext algorithm has numerous applications in image recognition detection [4][5][6]. The ResNet (Residual Network) structure was initially suggested by Kaiming He et al. in 2016. [7].The deep neural network model of the convolutional layer incorporates the residual block, resulting in outstanding performance on the ImageNet dataset. In the year 2017, Xie S and colleagues published a paper. [8] suggested the idea of uniformity utilizing the ResNext network. The ImageNet dataset yielded satisfactory outcomes. Zhaohui Yang and colleagues. [9] In 2019, a channel-based attention mechanism was suggested for ECA, and its utilization in deep convolutional neural networks is examined to enhance the model's performance and efficiency.

The ResNext network, which has a multi-layer feature extraction network, can extract the image features to be classified for image classification tasks due to its intense feature extraction and learning ability and its excellent performance in small sample data set classification tasks, as evidenced by the above analysis. The objective of this paper is to introduce SiRCNN, an enhanced image classification algorithm, which is built upon siamese networks. The ResNext101 neural network has been

ISSN: 2414-1895 DOI: 10.6919/ICJE.202402_10(2).0006

incorporated into the backbone feature extraction network, and minimal enhancements have been implemented. The feature extraction network incorporates the ECA attention mechanism and L2 regularization penalty to fuse image features. Not only does it enhance the feature extraction and generalization capability, but it also boosts the neural network model's anti-interference and classification accuracy. Using the L1 distance algorithm in regression prediction minimizes the model parameters and calculation amount, resulting in quicker model training and improved image classification accuracy.

2. Related Work

2.1 Algorithm in the Text

The ResNext101 residual network's parallel convolution mode is the primary concept of the siamese network [10], illustrated in Fig. 1 as the general structure of the algorithm proposed in this paper. The siamese network [11][12] is usually made up of two identical subnetworks, and the two subnetworks have the exact weights, which can minimize the mistakes caused by the neural network when extracting image features. This technology has significant implications for image categorization and facial identification. The ResNext101 network's residual parallel convolution mode enables the feature extraction network to apply multiple sets of convolution layers of varying sizes to separate images for feature extraction and then combine the layers to generate feature images with more details.

This paper's algorithm mainly comprises a siamese feature extraction network with shared weights, a channel attention module, and a regression prediction function. The weights of the two feature extraction network channels are the same. In order to classify a set of images, the algorithm initially employs the probability function to apply image enhancement to the target image randomly. Subsequently, the upgraded ResNext network, a lightweight siamese, is employed to identify the characteristics of 64 pairs of images and the image characteristics are augmented through ECA[13] channel attention. The L1 distance function [14] is then used to measure the similarity of the feature vector. The probability value between the feature vector and the label of the image to be classified is obtained, the relationship allocation between the input image and the extracted label is predicted, and the classification accuracy of the target image is obtained.

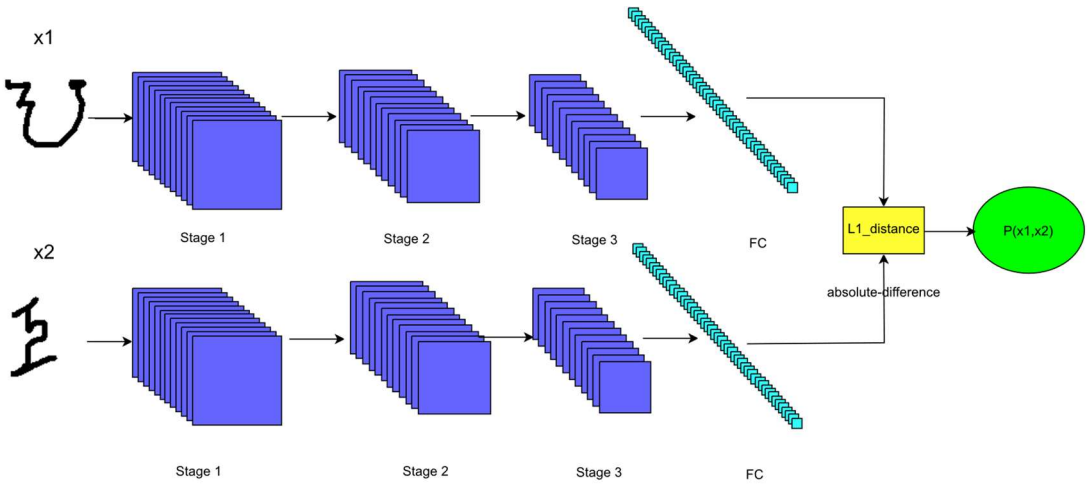

Fig. 1 Schematic diagram of the overall algorithm framework in this paper

This paper introduces the ResNext101 network model, which is based on the siamese network and reduces its complexity. The ResNext101 network model requires fewer layers than the lightened neural network. In Figure 1, Stage 1, Stage 2, and Stage 3 feature a Resnet_block module [15], which incorporates the ECA channel attention mechanism module to lessen the neural network's computational load and training parameters. The convolution layer's output is flattened, and the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202402_10(2).0006

extracted feature vector is used as the model output for similarity measurement. The probability value between the feature vector and the label of the image to be classified is obtained, the relationship distribution between the input image and the extracted label is predicted, and the classification accuracy of the target image is finally obtained.

Fig. 2 displays the residual block Stage 1 in the algorithm frame image, and the identical principle applies to Stage 2 and Stage 3.

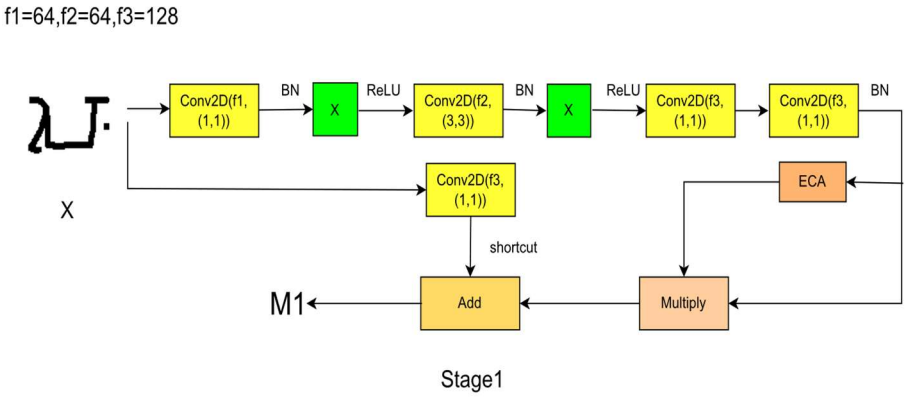

Fig. 2 Residual block Stage1

The sizes of the convolutional kernels are represented by f1, f2, and f3. The convolutional kernel group in a residual block comprises three distinct sizes of convolutional kernels. The residual block Stage1 has a convolutional kernel group of (64, 64, 128), Stage2 has a kernel group of (128, 128, 256), and Stage3 has a kernel group of (256, 256, 512).

To begin with, the given filters parameter yields three distinct convolution kernels. First, go through the three convolutional feature extraction separately. Then, the feature map X_1 is obtained by normalisation and ReLU activation function, as shown in the formula below.

$$X_1 = Conv3(\delta_1(Conv2(\delta_1(Conv(X1))))) \tag{1}$$

Subsequently, employ the ECA attention mechanism to acquire an improved feature map Y_1 for the X_1 image features. This is the formula:

$$Y_1 = ECA \cdot X_1 \tag{2}$$

Subsequently, utilize a Conv3 convolutional feature extraction on the input image to acquire a expedited shallow convolutional feature map. This is the formula:

$$shortcut = Conv3(X1) \tag{3}$$

Finally, the feature maps are summed and processed through the ReLU activation function to obtain the feature map M1 after the residual block Stage1 processing. The formula is shown in the following equation.

$$M1 = \delta_1(shortcut +) \ Y_1 \tag{4}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202402_10(2).0006

The input image is denoted by X1, the ReLU activation function is denoted by δ_1, the feature map after the residual block convolution operation is represented by X_1, the image convolution operation is denoted by Conv, the parallel layer of convolution operation in the residual block is denoted by shortcut, and the feature map of the input image after M1 denotes Stage1.

The paper optimizes the network structure by removing the convolutional layers with minimal input, eliminating redundant and repetitive feature extraction convolutional layers, and altering some of the parameters of the backbone network. By reducing the size of the model and the computational complexity of model training, the risk of model overfitting is minimized. Table 1 displays the specifications of the enhanced backbone network.

Table 1. Improved backbone network parameters

	Layer	Kernel	Stride	Image
	Conv1	3	2	(105,105,1)
Stage1	Conv2	3	1	(53,53,64)
	Conv3	1	1	(54,54,64)
	Conv4	3	2	(54,54,128)
Stage2	Conv5	1	1	(27,27,128)
	Conv6	1	1	(27,27,128)
	Conv7	3	2	(27,27,256)
Stage3	Conv8	1	1	(14,14,256)
	Conv9	1	1	(14,14,256)
	Conv10	3	2	(14,14,512)

The table displays the redesigned backbone network's residual blocks and convolutional parameters. At regular intervals of three convolutional layers, a core convolutional group of parallel residual convolutions is created to extract and combine features from the input image, thereby enhancing the neural network model's accuracy in image classification tasks.

2.2 ECA Attention Mechanism

The feature channel attention operation (ECA) is a mechanism used to enhance the information interaction between feature channels, which enhances the model's ability to represent the input features. The purpose of the fully connected layer in the ECA module is to aggregate the feature channels by assigning varying weights to the features between them, thereby enhancing their interaction. This mechanism can augment the model's capacity to depict the input characteristics and enhance the model's generalization capability.

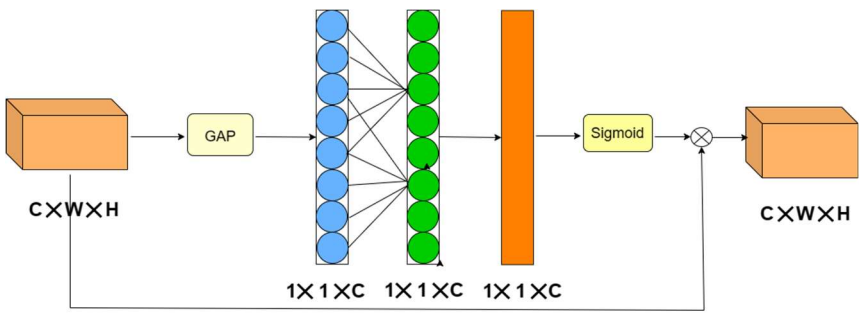

Fig. 3 Schematic diagram of the ECA attention mechanism

ISSN: 2414-1895 DOI: 10.6919/ICJE.202402_10(2).0006

First, the graph input image is subjected to global leveling pooling to obtain the global information of the image. Subsequently, a dimensionality reduction is done to decrease the feature map dimensions and the computational and parametric quantities. The feature map dimensions are then further reduced by A convolution and sigmoid activation function processing. Finally, the Multiply function combines the resulting low- and high-dimensional features. The following equation displays the formula.

$$Y = Mutiply(x, \delta_2(Con1D(Rs(GAP(x))))) \tag{5}$$

The input in the formula above is x, and the resulting feature fusion map is Y. The multiplication function is denoted by multiplication, the sigmoid activation function by δ_2, the one-dimensional convolution operation by Conv1D, the Reshape dimensionality reduction operation by Rs, and the global average pooling operation by GAP.

2.3 Similarity Measures the Loss Function

The Manhattan distance (L1 distance) formula is used in this paper to measure the similarity of the extracted features. This formula calculates the sum of the absolute values of the difference between the two vectors in each dimension.

$$L_1 = |x_1 - y_1| + |x_2 - y_2| + ... + |x_n - y_n| \tag{6}$$

The feature vectors obtained after feature extraction for the input image are denoted by x and y in the equation above.

The features extracted by the network undergo a similarity measure using L1 distance through degree, enabling the prediction of the classification of two input images.

3. Experimental Analysis

3.1 Experimental Parameters

The paper specifies that the input image size is $105 \times 105 \times 1$. The siamese network receives 64 pairs of images simultaneously as input, and a randomly generated matrix is utilized to determine whether or not to incorporate image enhancements like random rotations, cuts, offsets, and scaling on the training images, thereby improving the model's generalization capability.The values assigned to each parameter of the network are outlined in Table 2.

Table 2. Network parameter settings

lr	momentum	momentum_slope	L2_Conv1	L2_Conv2	L2_Conv3	Batch_size
0.01	0.5	0.01	0.01	0.01	0.01	32

The initial learning rate is set to 0.01, and the learning rate is updated every 500 iterations with the updated formula shown in the following equation:

$$lr^* = lr \times 0.99 \tag{7}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202402_10(2).0006

In the above formula, 0.99 is the learning rate update coefficient, and lr^* represents the updated learning rate.

The algorithm increases the momentum value dynamically, with the momentum update formula shown as follows:

$$new\_momentum = momentum + momentum\_slope \tag{8}$$

3.2 Datasets and Evaluation Metrics

3.2.1 Datasets

For training and testing, the Omniglot handwriting dataset [16], which consists of 1623 classes with 20 distinct characters each, was selected for the paper. Due to its attributes as a multilingual multi-character system, it can be utilized for investigating issues like character recognition, cross-language image classification, and small-sample learning, making it an optimal dataset for validating image classification algorithms for small-sample datasets. The diagram illustrates the representation of AN-C1 using the handwritten letters of the initial letter in Angelic, AN-C1, N-C, AN-C1, AN-C15 from various letters in Angelic, AN-C1, AU-C1, AMN-C1, UL-C1 from different alphabetic writing systems within the validation set. A selection of images from the Omniglot dataset is depicted in Fig. 4.

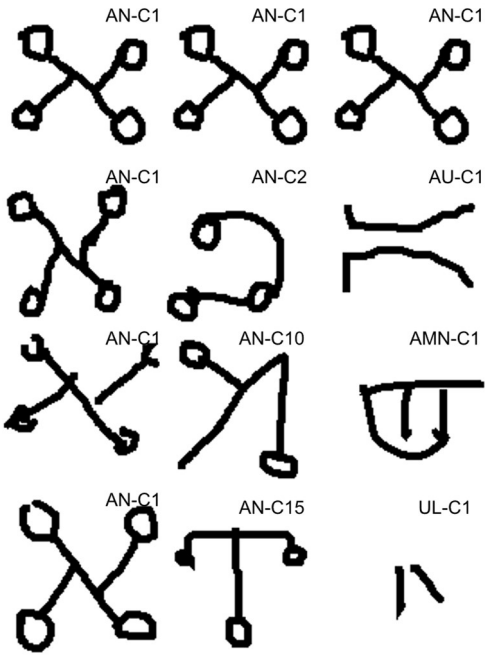

Fig. 4 Omniglot dataset partial validation set

The folder where the extracted images are located is assigned to the image labels. For example, the label of AN-C1 is Angelic-character01, which eliminates the need for human labeling and simplifies the image classification process.

3.2.2 Assessment of Indicators

The article evaluates the algorithm's performance using test set image classification accuracy, average classification accuracy, and model training convergence round time. The total number of samples is denoted by P_n, and the number of correctly predicted samples is represented by p_n. This is the formula for the calculation:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202402_10(2).0006

$$P = (P_n / p_n) \times 100\% \tag{9}$$

For training, a validation set is formed by randomly selecting five alphabets and one fixed alphabet, and the resulting classification accuracy and average classification accuracy are calculated for every 1000 iterations. If the average classification accuracy P_{m+1} is higher than P_m after the successive 1000 training iterations, then P_{m+1} will be adjusted to the highest validation accuracy; however, if P_{m+1} is lower than P_m, it will not be modified. Suppose the validation accuracy has not improved in the last 10000 training iterations. In that case, the training will end with the 10001st training iteration, and each alphabet's classification accuracy and average classification accuracy in the test set will be outputted. The number of times the model is trained is used as an evaluation metric for the time used for training.

3.3 Different Algorithms are Used to Classify and Verify the Omniglot Dataset

The algorithms mentioned in the paper, as well as different neural networks including VGG16, VGG19, RES18, RES50, EfficientNet-B2, RES101, Siamese (CNN), and the algorithm SiRCNN neural network models in the paper, are compared experimentally on the image classification benchmarks on Omniglot dataset. As part of the training process, the dataset is validated in tandem with the iterations of model training to enhance the efficiency of the model's training. The average validation accuracy varies, as shown in Fig. 5.

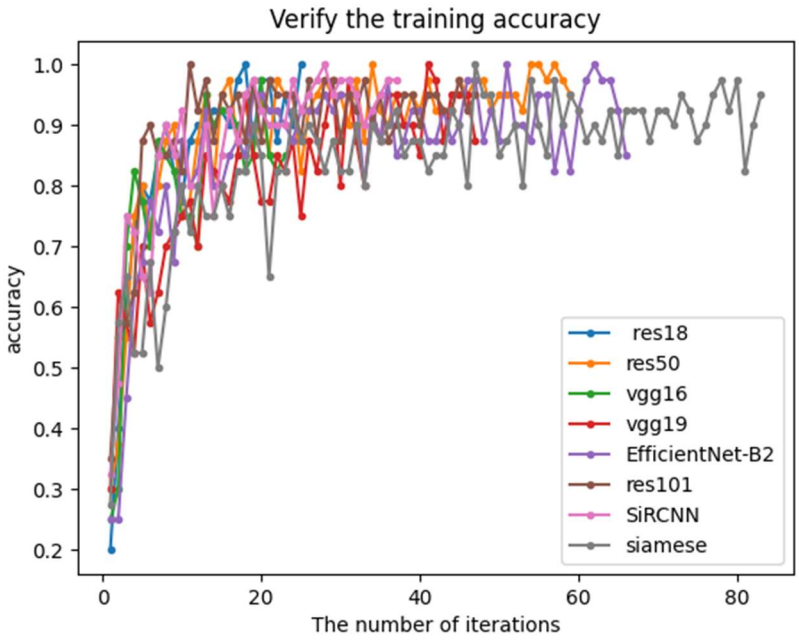

Fig. 5 Change in average verification accuracy

The average classification accuracy of the validated alphabet output by the different models during the training process is depicted in the figure above, and this accuracy increases as the models are trained until the model training is finished.

In training a neural network model, the maximum threshold is set at 100000 times, while the minimum threshold is set at 10000 times. The accuracy of the validation set's classification accuracy change curve during the model training process is illustrated in Fig. 6.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202402_10(2).0006

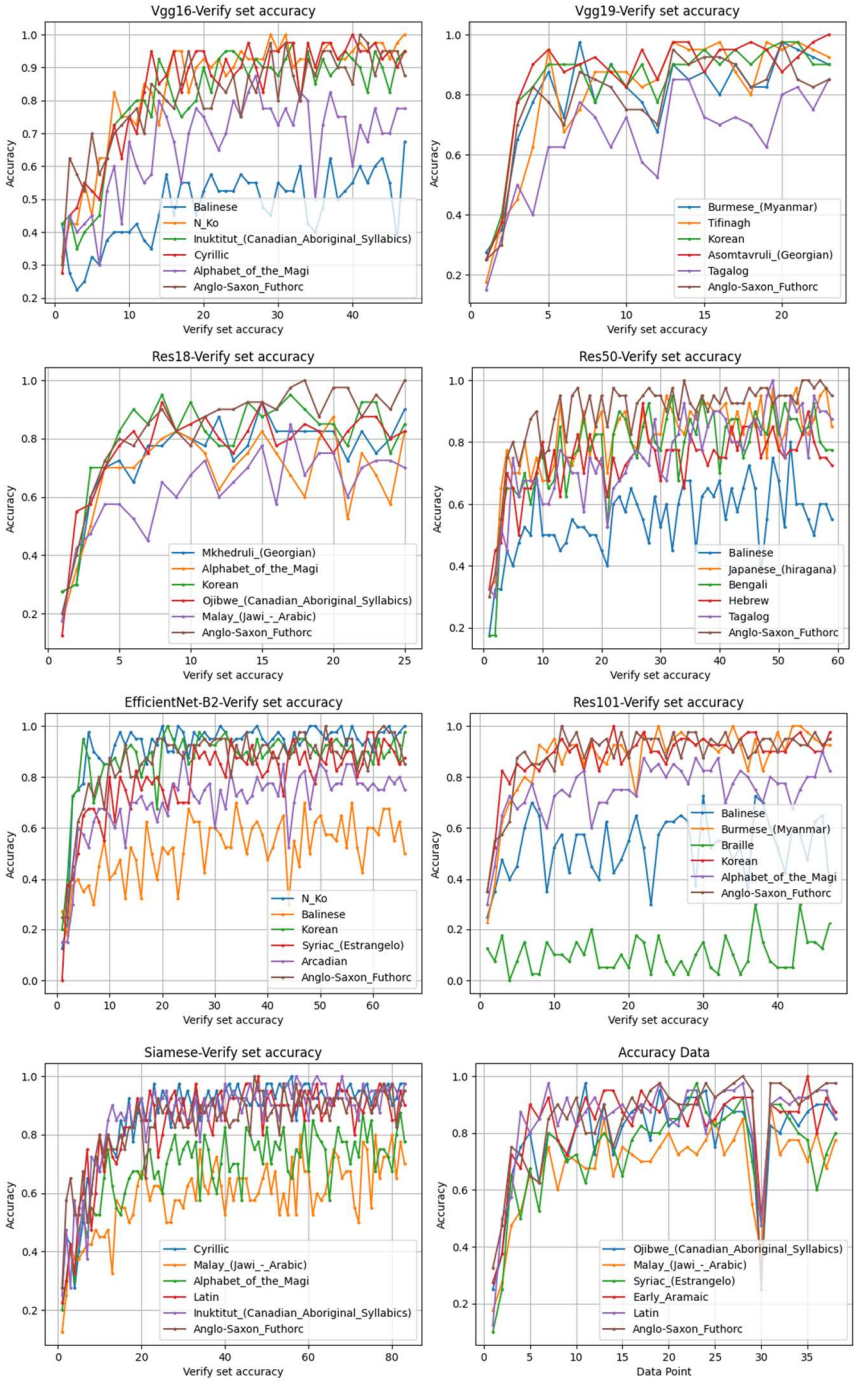

Fig. 6 The classification accuracy of different models on random alphabets in the Omniglot dataset

The results of Fig. 6 demonstrate that the random 6-validation alphabet's classification accuracy is on the rise with training, suggesting that the model's performance is improving. The evidence suggests that the model could better understand the features and patterns in the dataset, implying that it was trained correctly and could classify the data more precisely.

3.4 Lightweight SiRCNN Algorithm Experimental Validation

In this paper, the lightweight algorithm training parameters have been drastically decreased compared to the original network and ResNext network model, following the initial enhancement of the algorithm and its feature extraction network being incorporated into the optimized computational model. Table 3 displays the parameters used to train the lightweight model.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202402_10(2).0006

Table 3. Model Parameter Comparison

	Total_params	Trainable_params	Non_trainable_params
Siamese	38,951,745	38,951,745	0
Resnext101	90,486,736	90,377,984	90,752
SiRCNN	1,569,792	1,562,496	7,296

The SiRCNN model, as demonstrated in the table, drastically cuts down the algorithm's total parameters and training parameters, decreasing the computational complexity and workload of model training and optimizing the algorithm's calculation process. Table 4 displays the classification accuracy of 20 distinct classes in the test set, which were trained with Siamese, ResNext101, and SiRCNN models.

Table 4. Accuracy of 20 Alphabets in the Test Set

	Siamese	ResNext101	SiRCNN		Siamese	ResNext101	SiRCNN
OC	0.8	0.8725	0.9	AN	0.6	0.675	0.75
OR	0.5	0.675	0.475	AV	0.525	0.725	0.675
AB	0.8	0.85	0.875	SY	0.4	0.675	0.575
MO	0.675	0.75	0.825	MA	0.625	0.725	0.725
SS	0.55	0.725	0.575	ATL	0.525	**0.725**	**0.825**
TI	0.675	0.8	0.9	TE	0.425	0.725	0.75
GU	0.6	0.675	0.725	MAN	0.8	0.925	0.8
AT	0.55	0.65	0.75	UL	0.6	0.4	0.675
GL	**0.55**	**0.875**	**0.9**	GE	0.625	0.875	0.75
KE	0.825	0.825	0.9	KN	0.525	0.725	0.675
Mean global accuracy	0.60875	0.74375	0.75125	Number of training iterations	83000	47000	38000

The differences in individualization may be due to the different picture content of the alphabet data for the different classes. The test set yielded a classification accuracy of 20 distinct alphabets.

Table 4 demonstrates that utilizing ResNext101 and SiRCNN algorithms for image classification tasks leads to an enhancement in the classification accuracy of 20 test set alphabets compared to the Siamese model. The GL alphabet had the highest classification accuracy, increasing by 32.5% and 35%, respectively. The average classification accuracy of the test set also increased by 14% and 15%, respectively. The ATL alphabet exhibited a 10% enhancement in classification accuracy compared to the ResNext model. Furthermore, the siameseCNN model underwent 83,000 training iterations, followed by the ResNext101 network model with 48,000 iterations, and the SiRCNN model with 37,000 iterations, surpassing Siamese by 45,000 times. The completion of the training occurred 9000 times ahead of the ResNext101 model, leading to a substantial reduction in the model's training duration and a significant saving in operational time.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202402_10(2).0006

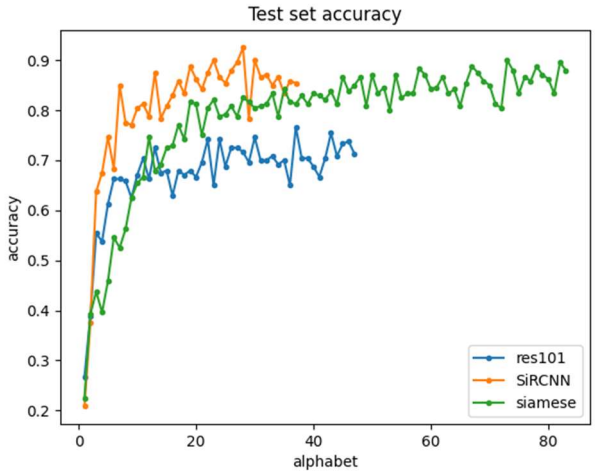

Fig. 7 Average Classification Accuracy in Validation

As the number of iterations in the model training process increases, the average accuracy of the random letter table validation in the graph also increases. The SiRCNN model surpasses the Siamese and ResNext101 models regarding classification accuracy and converges sooner than the other two. The SiRCNN model outperforms the other two neural network models in terms of both classification accuracy and training time.

Once the training is finished, various neural network models will provide data regarding the amount of training iterations, the optimal training loss rate, and the optimal learning rate throughout the training procedure. The most effective learning and loss rates are used as test parameters for the test set. The test set's most accurate validation and average classification accuracy are then obtained for different neural network models, serving as the final evaluation indicator of the overall performance of different neural network models. Table 5 displays the information on the various models.

Table 5. Different Model Overall Evaluation Data

	Average classification accuracy	Training time	Optimal loss	Optimal learning rate	Optimal verification accuracy
Siamese(CNN)	0.60875	83000	0.554829	0.00186	0.8875
VGG16	0.6275	47000	0.843890	0.00388	0.870833
VGG19	0.65125	23000	0.253298	0.006298	0.925
Res18	0.6625	25000	0.344828	0.006050	0.875
Res50	0.6995	59000	0.420847	0.003055	0.904166
EfficientNet-B2	0.72125	66000	0.381858	0.002654	0.8875
ResNext101	0.74375	47000	0.314022	0.003888	0.7666
SiRCNN	0.75125	38000	0.391897	0.004659	0.925

The SiRCNN model attains an average classification accuracy of 75.1% in this instance, which then converges at the 38,000th iteration, enabling a quicker completion of the training process.

To conclude, the SiRCNN algorithm discussed in this article provides superior image classification capabilities and immunity to interference, thus making it a better choice for classifying small-sample datasets.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202402_10(2).0006

4. Conclusion

This paper suggests a SiRCNN-based image classification algorithm for Siamese networks to tackle the limited data samples in small-sample datasets, ensuring algorithm precision while minimizing model intricacy. The Omniglot dataset is used to assess performance. After 38,000 training iterations, the improved algorithm converges and achieves a final image classification accuracy of 75.1%, accomplishing a faster and more accurate small-sample image classification task.

References

[1] Hinton G E,Salakhutdinov R R.Reducing the Dimensionality of Data with Neural Networks[J].Science, 313[2023-11-03].DOI:10.1126/science.1127647.

[2] Deng J,Guo J,Zafeiriou S.ArcFace: Additive Angular Margin Loss for Deep Face Recognition[J]. 2018. DOI:10.48550/arXiv.1801.07698.

[3] Tan Jiachen, Dong Yongquan, Zhang Guoxi. SSM: Diabetic Retinopathy Fundus Image Classification Model Based on siamese Networks [J]. Journal of Nanjing University (Natural Sciences),2023,59(03): 425-434.DOI:10.13232/j.cnki.jnju.2023.03.006.

[4] Zhang,Zhang L,Du B.Deep Learning for Remote Sensing Data:A Technical Tutorial on the Stateof the Art[J].IEEE Geoscience & Remote Sensing Magazine,2016,4(2):22-40.DOI:10.1109/MGRS.2016. 2540 798.

[5] Sumbul G , Charfuelan M ,Demir, Begüm,et al.BigEarthNet: A Large-Scale Benchmark Archive For Remote Sensing Image Understanding[J].IEEE, 2019.DOI:10.1109/IGARSS.2019.8900532.

[6] Ullah A , Ahmad J , Muhammad K ,et al.Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features[J].IEEE Access, 2018, 6(99):1155-1166.DOI:10.1109/ ACCESS. 2017.2778011.

[7] He K,Zhang X,Ren S,et al.Deep Residual Learning for Image Recognition[J].IEEE, 2016.DOI:10.1109/ CVPR.2016.90.

[8] Xie S,Girshick R,Dollar P,et al.Aggregated Residual Transformations for Deep Neural Networks[C]// Computer Vision and Pattern Recognition.IEEE,2017.DOI:10.1109/CVPR.2017.634.

[9] Wang Q,Wu B,Zhu P,et al.ECA-Net:Efficient Channel Attention for Deep Convolutional Neural Networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.DOI:10.1109/CVPR42600.2020.01155.

[10] Chopra S , Hadsell R , Lecun Y .Learning a similarity metric discriminatively, with application to face verification[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).IEEE, 2005.DOI:10.1109/CVPR.2005.202.

[11] Gregory Koch,Richard Zemel,Ruslan Salakhutdinov.Siamese Neural Networks for One-shot Image Recognition[J].[2023-11-28].

[12] Yi D,Lei Z,Liao S,et al.Learning Face Representation from Scratch[J].Computer Science,2014.DOI:10. 48550/arXiv.1411.7923.

[13] Fu J,Liu J,Tian H,et al.Dual Attention Network for Scene Segmentation[J]. 2018.DOI:10.48550/arXiv. 1809.02983.

[14] Comaniciu D,Meer P.Mean shift: a robust approach toward feature space analysis[J].IEEE Trans Pattern Analysis & Machine Intelligence,2002,24(5):603-619.DOI:10.1109/34.1000236.

[15] Huang G,Liu Z,Laurens V D M,et al.Densely Connected Convolutional Networks[J].IEEE Computer Society, 2016.DOI:10.1109/CVPR.2017.243.

[16] Lecun Y,Bottou L.Gradient-based learning applied to document recognition[J].Proceedings of the IEEE, 1998, 86(11):2278-2324.DOI:10.1109/5.726791.