IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 May 2022, accepted 4 June 2022, date of publication 8 June 2022, date of current version 25 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3181225

FixCaps: An Improved Capsules Network for Diagnosis of Skin Cancer

ZHANGLI LAN, SONGBAI CAI, XU HE, AND XINPENG WEN
Chongqing Jiaotong University, Chongqing 400074, China

Corresponding author: Songbai Cai (caisongbai@126.com)

This work was supported in part by the Chongqing Technology Innovation and Application Development Project.

ABSTRACT The early detection of skin cancer substantially improves the five-year survival rate of patients. It is often difficult to distinguish early malignant tumors from skin images, even by expert dermatologists. Therefore, several classification methods of dermatoscopic images have been proposed, but they have been found to be inadequate or defective for skin cancer detection, and often require a large amount of calculations. This study proposes an improved capsule network called FixCaps for dermoscopic image classification. FixCaps has a larger receptive field than CapsNets by applying a high-performance large-kernel at the bottom convolution layer whose kernel size is as large as 31×31, in contrast to commonly used 9×9. The convolutional block attention module was used to reduce the losses of spatial information caused by convolution and pooling. The group convolution was used to avoid model underfitting in the capsule layer. The network can improve the detection accuracy and reduce a great amount of calculations, compared with several existing methods. The experimental results showed that FixCaps is better than IRv2-SA for skin cancer diagnosis, which achieved an accuracy of 96.49% on the HAM10000 dataset.

INDEX TERMS Capsule network, CBAM, image classification, large-kernel convolution, skin cancer.

I. INTRODUCTION

The American Association for Cancer Research's Annual Cancer Report 2022 shows that cancer incidence and mortality in the United States continue to decline steadily [1]. The number of new cancer cases in China is approximately twice that in the United States, but nearly five times the number of deaths. It is helpful to reduce the cancer burden in China by comparing the latest cancer profiles, trends, and determinants between China and the United States, learning from the progress made in cancer prevention and care in the United State [2]. Skin cancer is one of the most common cancers diagnosed in the United States [3]. A report has shown that the five-year survival rate of localized malignant melanoma is 99% when diagnosed and treated early, whereas the survival rate of advanced melanoma is only 25% [4]. Hence, it is particularly important to detect and classify dermatoscopic images so that skin cancer can be diagnosed early. The traditional method is to first go through a doctor's visual inspection and then use dermoscopic imaging to aid in the diagnosis. However, a large number of skin cancer patients fail to receive early diagnosis and timely treatment due to the lack of professional doctors in China, the uneven level of doctors, and the pressure of doctors on repetitive reading work. With the development of artificial intelligence (AI) in the medical field, deep learning (DL) has been widely used for the detection and classification of medical images over the past few years [5]. The application of artificial intelligence to medical image-assisted diagnosis is called AI image diagnosis. It plays a pivotal role in the field of medical artificial intelligence, especially in intelligent image recognition, human-computer interaction-assisted diagnosis, precision treatment-assisted decision making, and other aspects [6]. Capsule networks (CapsNets) [7] have been widely applied in the medical field as an important research topic for deep learning. Afshar *et al.* [8] reduced the number of convolution kernels in the convolution layer of the capsule network. It has been successfully applied to the classification of brain tumors using magnetic resonance imaging (MRI), and its accuracy is superior to that of traditional convolutional neural networks (CNNs) [9]. Lin *et al.* [10]proposed a classification recognition algorithm for skin lesions based on ''Matrix Capsules with EM Routing'' [11], which achieved a high recognition accuracy in ISIC2017 dataset [12]. Mensah *et al.* [13] proposed Gabor CapsNets for tomato and citrus disease image recognition,

The associate editor coordinating the review of this manuscript and approving it for publication was Hossein Rahmani.

which could achieve 98.13% accuracy in the Plant-Village datasets [14], superior to AlexNet [15]and GoogLeNet [16] in terms of robustness and parameter amount. In addition, researchers have explored CapsNets from other aspects. Xiang *et al.* [17] proposed a multiscale capsule network and a capsule dropout. Robustness was achieved better than CapsNets on both Fashion MNIST dataset [18] and CIFAR10 dataset [19]. Rajasegaran *et al.* [20] constructed DeepCaps by using residual learning [21], which reduced the number of parameters by 68% compared with original CapsNets and was significantly superior to the existing capsule network architecture in benchmark datasets. This method provides a deep architecture for capsule networks. Other network models have been developed in the medical field. For example, Kawahara *et al.* proposed a fully convolutional network for skin-image classification. For the skin lesion datasets [22], the classification prediction accuracy was 81.8% [23]. Akram *et al.* [24] proposed a deep neural network based on integration and carried out a classification test in the ISIC2018 [25], [26].

Although these studies have promoted the development of AI image diagnosis, they have been found to be either ineffective or defective to the prediction of skin lesions, and often need a great amount of calculations. Hence, we propose an improved capsule network called FixCaps for dermoscopic image classification in this study. It can obtain a larger receptive field than CapsNets by applying a large-kernel convolution at the bottom layer whose kernel size is as large as 31×31, in contrast to commonly used 9×9. And the convolutional block attention module (CBAM) [27] is used to reduce the losses of spatial information caused by convolution and pooling. Meanwhile, the group convolution (GP) [15] is used to avoid model underfitting in the capsule layer. The network can improve the detection accuracy and reduce a large number of calculations, compared with the several existing methods. This research has verified the effectiveness of FixCaps for diagnosis (classification) of skin cancer. We address the problems of the limited amount of annotated data and the imbalance of class distributions. To ensure the validity of our perspectives, we make a large number of experiments on the HAM10000 dataset [26].

II. RELATED WORK

With the increasing incidence of skin cancers, a growing population, a lack of adequate clinical expertise and services, there is an immediate necessity for AI image diagnosis to assist clinicians in this field. Before 2016, most research adopted the traditional machine learning progress of preprocessing (augmentation), segmentation, feature extraction, and classification [28]. Nowadays, various types of skin lesion datasets are publicly accessible. Researchers have developed AI solutions, notably deep learning algorithms, to distinguish malignant skin lesions and benign lesions in different image modalities, such as dermoscopic, clinical and histopathology images [29]. For instance, Datta *et al.* [30] combined soft-attention (SA) and Inception ResNet-V2 (IRv2) [31] to construct IRV2-SA for dermoscopic image classification, which reached an accuracy of 93.47% on the HAM10000 dataset.

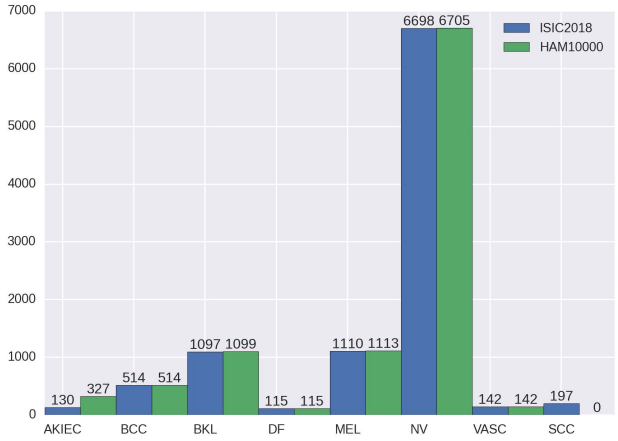


FIGURE 1. Sample distribution and comparison of different skin lesion images in ISIC2018 and HAM10000. The task3 of ISIC2018 included training data(HAM10000), validation data and test data. They are all organized into seven types of skin lesions. Zhao *et al.* added some images of squamous cell carcinoma(SCC) to ISIC2018 and deleted some samples, so that the distribution of ISIC2018 dataset is similar to that of ISIC2019 dataset. More details of the HAM10000 dataset are shown in the "DATASET" subsection.

Zhao *et al.* [32] proposed a classification framework based on skin lesion augmentation style-based GAN (SLA-StyleGAN), which achieved an accuracy of 93.64% on ISIC2018 dataset and ISIC2019 dataset [33]. In their work, the distribution of the ISIC2018 dataset differs considerably from the HAM10000 dataset, as shown in Fig. 1. Hence, IRv2-SA is the state-of-the-art performance in the dermatoscopic image classification on HAM10000 dataset, to the best of our knowledge.

III. FixCaps

As an important research direction in deep learning, the capsule network has the greatest advantage of being able to encode the pose and spatial relations of features, which significantly improves the shortcomings of deep learning in image classification. However, CapsNets [7] often exhibit poor performance in complex images, such as dermatoscopic images. Studies have shown that multiple routing layers lead to higher training costs and reasoning time in large initial layers [34]. Therefore, an improved capsule network called FixCaps for dermoscopic image classification was proposed. The main components and architecture are described below.

A. LARGE-KERNEL CONVOLUTION

FixCaps, compared with paper [8], not only reduces the number of convolution kernels at the bottom convolution layer, but also increases fractional max-pooling (FMP) [35] to reduce the size of the initial layer and the cost of dynamic routing in the capsule layer. The commonly used convolution kernels are as follows: 3×3, 5×5, or 7×7 in the convolutional layer

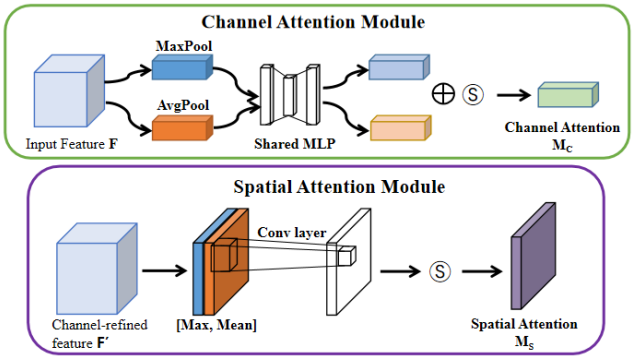


FIGURE 2. Diagram of each attention sub-module [27]: CBAM consists of two independent submodules the part in the green box is the channel attention module (CAM), the purple box is the spatial attention module (SAM), and Ⓢ denotes the sigmoid function.

of the neural network. In this study, the convolution layer with convolution kernels larger than 9×9 is called large-kernel convolution (LKC). The experimental result shows that the larger the convolution kernel is, the more picture information is "seen" and the better the features are learned [36]. In this study, the LKC has a larger receptive field compared to the small convolution kernel used in the literature compared with the small-kernel convolution used in the literature [7], [8], [17], [20]. The features available for CBAM screening are better, which improves the ability of the capsule layer to deal with the long-term relationship of feature vectors.

B. CONVOLUTIONAL BLOCK ATTENTION MODULE

An CBAM (see Fig. 2) is added between the convolution layer at the bottom and the capsule layer to make FixCaps pay more attention to the object and reduce the losees of spatial information caused by convolution and pooling. The feature maps are output from the convolution layer through CAM and SAM to strengthen the connection of each feature in the channel and space. This enabled the network to effectively avoid overfitting without dropout [37].The overall attention process can be summarized as Formula (1).

$$F' = M_s(M_c(F) \otimes F) \otimes F, \tag{1}$$

where $\otimes$ denotes the element-wise multiplication. During multiplication, the attention values are broadcasted (copied) accordingly: channel attention values are broadcasted along the spatial dimension and vice versa [27]. Here $M_c(R^{C\times1\times1})$ is the channel attention map, and $M_s(R^{1\times1\times H\times W})$ is the spatial attention map. F is the feature map output of the convolutional layer. F′ denotes the final refined output.

C. CAPSULE LAYER

The capsule layer is divided into two parts: the primary capsule and the digit capsule. FixCaps uses convolution with an inner size of nine and stride size of two in the primary capsule, which is consistent with CapsNets. In addition, the group convolution in the primary capsule was used to avoid underfitting of the model and reduce the amount of calculations while improving the accuracy of classification prediction. And the "Squashing" function is used to process the input vector, so that the modulus of the vector can represent the probability of this feature [7]. Its expression is shown in Equation (2).

$$V_j = \frac{\|S_j\|^2}{1+\|S_j\|^2} \cdot \frac{S_j}{\|S_j\|}, \tag{2}$$

where V_j is the vector output of capsule j and S_j is the total input. FixCaps uses the marginal loss as a loss function to enhance the class probability of the correct class [7]. Its expression is as Equation (3):

$$L_k = T_k \cdot max(0, m^+ - \|V_k\|)^2 + \lambda(1-T_k) \cdot max(0, \|V_k\| - m^-)^2, \tag{3}$$

where T_k=1, λ= 0.5, m^+=0.9 and m^-=0.1. The total loss is simply the sum of the losses of all the digit capsules. It was found in the experiment that the reconstruction cost of CapsNets was too high for large-size and high-resolution images such as dermoscopic imaging. Run FixCaps with and without the reconstitution module on server A. The results showed that reconstruction is not helpful in the classification of dermatoscopic images. Hence, the aim was to reduce the run time of FixCaps by deleting the reconstitution module. In the eval stage, the L2 norm of the output vector of the digit capsule layer is calculated, and the index number of the longest layer is taken as the predicted classification label. The calculation formula is given by Equation (4).

$$\|V_j\|_2 = \sqrt{a_1^2 + a_2^2 + \cdots + a_i^2}, \tag{4}$$

where V_j is the output vector of the digit capsule and j ∈[1,7]. The a_i is the value of the V_j, and i∈[1,16]. Here i and j are the positive integers.

D. FixCaps-DS

Increasing the depth of the model is an important research direction for deep learning, but the deeper the model is, the faster the gradient vanishes, so that back propagation is difficult to train the shallow network, and the network performance deteriorates instead. Residual learning makes it easier for gradients to shallow networks, and skip connections improve the performance of deep models [21]. However, the consequent problem is a rapid increase in the number of model parameters. In recent years, models such as GoogLeNet, Inception, and ResNet have used convolution with an inner size of 1 [38] to lightweight the model, but they still fail to solve the problem in which the weight parameters are too large to be applied in mobile terminals. CapsNets and their improved models strive for a balance between network depth and performance, as do FixCaps, the architecture of which is shown in Fig. 3. In this study, FixCaps-DS is a deep-wise separable convolution (DS) [39] combined with FixCaps. It only has approximately 35% parameter and 50% computation amount combined with FixCaps and is more

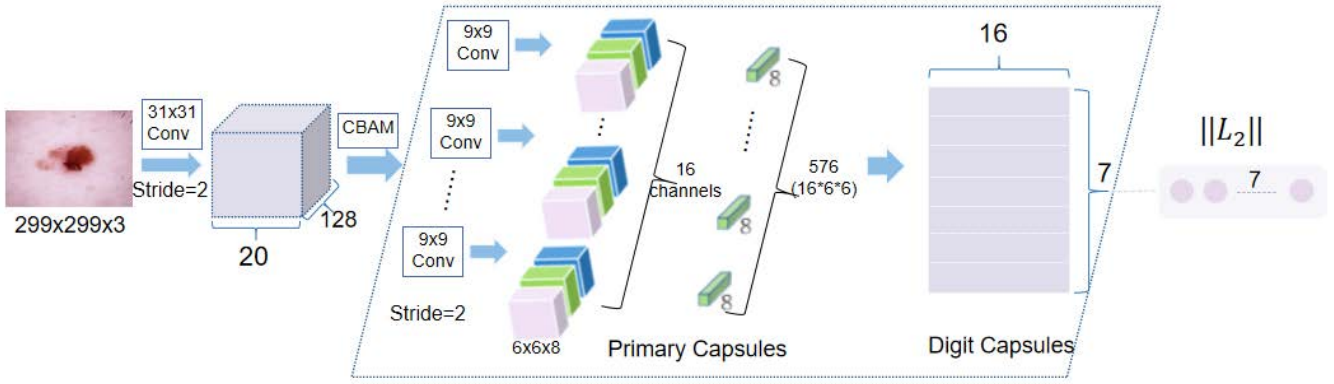


FIGURE 3. Diagram of FixCaps: The main components include convolution layer with large-kernel, CBAM, and capsule layer.

suitable for mobile terminal deployment. The experimental results showed that FixCaps-DS was also better than IRv2-SA in classification prediction of dermatoscopic images, which is achieved an accuracy of 96.13% on the HAM10000 dataset.

IV. EXPERIMENTS AND RESULTS

In this study, all the experiments were implemented using PyTorch 1.8, except augmentation of the training set. Run FixCaps and FixCaps-DS on two servers. The difference is that DS is used in convolution, and everything else is the same. Run FixCaps on server A, which was configured with an Intel I9 CPU and an RTX 3090 GPU. FixCaps-DS was run on server B, which was configured with an Intel I5 CPU and an RTX 3070 GPU.

A. DATASET

The dataset used in this study was HAM10000 [26], which consisted of 10015 dermatoscopic images with a size of 450×600. There are seven types of skin lesions: actinic keratosis/intraepithelial carcinoma (AKIEC), basal cell carcinoma (BCC), benign keratosis (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevi (NV), and vascular lesions (VASC), as shown in Fig. 4. To make a fair comparison with IRV2-SA, 828 images were extracted from the dataset as the test set in the same manner as IRV2-SA in the dataset division and data augmentation of the training set [30]. Subsequently, translation and other methods were used to increase the number of samples in the training set, and the processed data were saved as 299×299 JPG images.

It was found that the pixels in the center of the image had a high correlation with the prediction, whereas the pixels in the edge had a low correlation with the prediction in the experiment. So the original image of the dataset is decomposed into R, G, and B channels before training; subsequently, they are considered as input matrix A. Three matrices U, Σ, and V are obtained after singular value decomposition (SVD) according to Formula (5). In the experiment, $K = 90$ was used to obtain R*, G*, and B*, which were subsequently fused and stored the

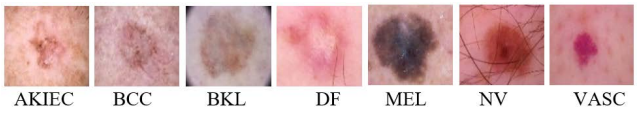


FIGURE 4. Example of Skin lesions in HAM10000 dataset. Among them, BKL, DF, NV, and VASC are benign tumors, whereas AKIEC, BCC, and MEL are malignant tumors.

images in the PNG format.

$$A_{(m,n)} = U_{(m,k)} \cdot \Sigma_k \cdot V_{(k,n)} \quad , \tag{5}$$

where U and V are orthogonal matrices called the left and right singular values, respectively, and Σ is the singular value. In sigma, the singular values are arranged from the largest to the smallest, with the latter values closer to zero and retaining less image information.

B. EVALUATION METRICS

In this study, the model was evaluated using Recall, Accuracy and F_1-score. The calculation formula is given by Equation (6). In the confusion matrix, TP samples were distributed on the diagonals (in this study, the diagonals refer to the diagonals from the upper left to the lower right). Accuracy was defined as the ratio of the number of correctly classified samples (on the diagonal) to the total number of samples.

For multi-classification problems, the accuracy measures the prediction of global samples, whereas F_1-score and recall represent the prediction of a certain category. Therefore, the F_1-score and recall of each skin lesion must be calculated separately, but the accuracy does not. The confusion matrix in multi-classification has a special case: micro-precision, micro-recall, and the accuracy is always the same. Because the FP is in one class of samples, to the others must be FN.

$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN}$$

TABLE 1. Classification accuracy (%) on the HAM10000 test set. FixCaps-DS has 0.08 billion FLOPs, which is only 66% of MobileNet V3 (0.12 billion FLOPs).

Method	Accuracy[%]	Params(M)	FLOPs(G)
GoogLeNet	83.94	5.98	1.58
Inception V3	86.82	22.8	5.73
MobileNet V3	89.97	1.53	0.12
IRv2-SA	93.47	47.5	25.46
FixCaps−DS	**96.13**	0.14	0.08
FixCaps	**96.49**	0.50	6.74

$$Accuracy = \frac{TP + TN}{T}$$
$$F_1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \quad (6)$$

where TP, FP, FN, and T are the true negative, true positive, false positive, false negative, and the total number of samples, respectively.

C. RESULTS

We compare the performance of FixCaps, IRv2-SA and other methods while classifying skin lesions, as shown in Table 1. FixCaps outperforms GoogleNet, IRv2, and other methods in terms of accuracy on the HAM10000 dataset. It is worth noting that FixCaps has fewer parameters and lower complexity than IRv2-SA. And FixCaps has 6.74 billion FLOPs, which is only 26% of IRv2-SA (25.46 billion FLOPs). Moreover, FixCaps-DS has 0.14 million parameters, which is approximately 10 percent of MobileNet V3 [40] (1.53 million parameters).

V. DISCUSSION

In the clinical, the first task needs a correct specific diagnosis out of multiple classes on skin lesion classification. F_1-score and AUC (area under ROC curve) both hope that the actual true samples can be detected. The difference between the two is that AUC aims for a model with as few false positives as possible, while F_1-score aims for a model that does not miss any possibilities. We hope to diagnose as many suspected cases as possible, so we prefer F1-score as the evaluation index of the model. In fact, FixCaps has a good performance of ''Recall'' on skin cancer classification. The details of the performance of different methods on the test set are listed in Table 2. The last column shows the number of samples of each skin lesion type in the test set. Clearly, FixCaps outperformed IRv2-SA on diagnosis (classification) of skin cancer, except for the VASC type. There are two main reasons for this consequence. One is that capsule networks not only learn excellent weights for feature extraction and image classification, but also learn how to encode the pose and spatial relations of features [7]. The other is FixCaps can obtain a larger receptive field than IRV2-SA by applying a high-performance LKC whose kernel size is as large as 31×31, in contrast to IRV2-SA used 3×3.

As shown in Table 3 and Fig. 5, FixCaps' performance on the test was evaluated by using different LKC. Finally, it was found that the convolution with kernel size of 31×31 showed the best performance. In conclusion, this work has proven the effectiveness of FixCaps in the classification prediction of dermatoscopic images, and demonstrated that using the large convolutional kernels instead of a stack of small kernels could be a more powerful paradigm.

TABLE 2. Evaluation metrics of FixCaps and IRV2-SA for each skin lesion type on the test set.

Dis.	Recall			F1-score			#
	Fix Caps-DS	Fix Caps	IRv2-SA [30]	Fix Caps-DS	Fix Caps	IRv2-SA [30]	828
AKIEC	0.913	**0.957**	0.520	0.875	**0.917**	0.690	23
BCC	0.769	0.846	**0.880**	0.769	**0.898**	0.880	26
BKL	0.803	**0.864**	0.830	0.869	**0.881**	0.770	66
DF	**0.833**	0.667	0.170	**0.769**	0.615	0.290	6
MEL	0.853	**0.912**	0.650	0.879	**0.925**	0.660	34
NV	**0.992**	0.986	0.980	**0.986**	0.985	0.980	663
VASC	**1**	0.700	**1**	**1**	0.824	**1**	10

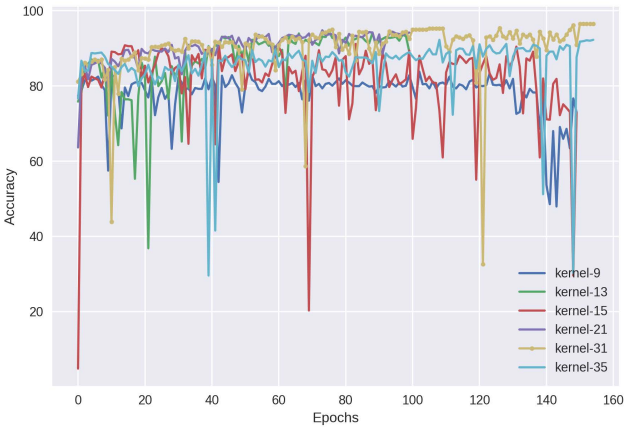


FIGURE 5. The accuracy is evaluated on the test set by using different LKC. The "kernel-N" denotes a convolution kernel of N × N in the FixCaps.

TABLE 3. The F_1-score is evaluated on the test set by using different LKC.

Dis.	F_1-score								
	9	11	13	15	18	21	27	31	35
AKIEC	0.323	0.488	0.545	0.634	0.818	0.880	0.809	**0.917**	0.800
BCC	0.410	0.557	0.767	0.738	0.774	0.815	0.772	**0.898**	0.772
BKL	0.426	0.516	0.811	0.806	0.824	0.832	0.770	**0.881**	0.739
DF	0	0	0.444	0.333	0.400	0.545	0.533	0.615	**0.625**
MEL	0.261	0.246	0.853	0.600	0.774	0.829	0.722	**0.925**	0.687
NV	0.923	0.929	**0.989**	0.975	0.989	0.980	0.972	0.985	0.966
VASC	0.588	0.667	**0.947**	0.778	0.842	0.889	0.737	0.824	0.889

VI. CONCLUSION

In this work, we introduced an improved capsule network called FixCaps for dermoscopic image classification. FixCaps can obtain a larger receptive field than CapsNets by applying a high-performance large-kernel at the bottom convolution layer whose kernel size is as large as 31×31, in contrast to commonly used 9×9. Moreover, the CBAM was used to reduce the losses of spatial information and

the GP was used to avoid model underfitting in the capsule layer. We evaluate FixCaps on HAM10000 dataset, and the experiment results show that FixCaps achieves an accuracy of 96.49%, while achieving a 92% reduction in the number of parameters. FixCap can improve the detection accuracy with less calculations, compared with several existing methods. Hence, FixCaps will be helpful to doctors (especially those with little experience) by providing valid auxiliary diagnosis. Moreover, it will promote the perfection and popularization of skin cancer screening technologies. In this work, however, the generalization performance of FixCaps has not been adequately studied, which we will study in the future.

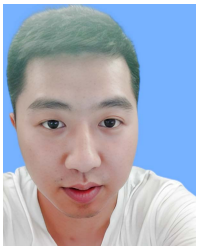
ACKNOWLEDGMENT

The authors would like to thank Yitian Li and Hanzhong Zhang for the helpful discussions.

REFERENCES

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *CA Cancer J. Clin.*, vol. 72, no. 1, pp. 7–33, Jan. 2022.
[2] C. Xia, X. Dong, H. Li, M. Cao, D. Sun, S. He, F. Yang, X. Yan, S. Zhang, N. Li, and W. Chen, "Cancer statistics in China and United States, 2022: Profiles, trends, and determinants," *Chin. Med. J.*, vol. 135, no. 5, pp. 584–590, 2022.
[3] *Cancer Facts & Figures 2018*, Amer. Cancer Soc., Atlanta, GA, USA, 2018.
[4] H. K. Koh, "Melanoma screening," *Arch. Dermatol.*, vol. 143, no. 1, pp. 101–103, Jan. 2007, doi: 10.1001/archderm.143.1.101.
[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
[6] C. H. Qiu, C. F. Huang, S. R. Xia, and D. X. Kong, "Application review of artificial intelligence in medical images aided diagnosis," *Space Med. Med. Eng.*, vol. 34, no. 5, pp. 407–414, 2021, doi: 10.16289/j.cnki.1002-0837.2021.05.009.
[7] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," 2017, *arXiv:1710.09829*.
[8] P. Afshar, A. Mohammadi, and K. N. Plataniotis, "Brain tumor type classification via capsule networks," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3129–3133, doi: 10.1109/ICIP.2018.8451379.
[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.
[10] K. Lin, H. Du, H. Wang, and L. J. Zhu, "Research on classification and recognition algorithm of melanoma in dermoscopic image based on matrix capsule network," *J. Hubei Minzu Univ. Natural Sci. Ed.*, vol. 39, no. 2, pp. 175–179 and 240, 2021, doi: 10.13501/j.cnki.42-1908/n.2021.06.010.
[11] G. Hinton, S. Sara, and N. Frosst, "Matrix capsules with EM routing," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–15.
[12] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172, doi: 10.1109/ISBI.2018.8363547.
[13] P. M. Kwabena, B. Asubam, and A. Abra, "Gabor capsule network for plant disease detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 10, pp. 388–395, 2020, doi: 10.14569/IJACSA.2020.0111048.
[14] D. P. Hughes and M. Salathe, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," 2015, *arXiv:1511.08060*.
[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2012, vol. 25, no. 2, pp. 1097–1105.
[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[17] C. Xiang, L. Zhang, Y. Tang, W. Zou, and C. Xu, "MS-CapsNet: A novel multi-scale capsule network," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1850–1854, Dec. 2018, doi: 10.1109/LSP.2018.2873892.
[18] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.
[19] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," in *Handbook of Systemic Autoimmune Diseases*, vol. 1, no. 4, 2009.
[20] J. Rajasegaran, V. Jayasundara, S. Jayasekara, H. Jayasekara, S. Seneviratne, and R. Rodrigo, "DeepCaps: Going deeper with capsule networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10717–10725.
[21] K. He, X. Zhang, and S. Ren, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.
[22] C. D. Leo, V. Bevilacqua, L. Ballerini, R. Fisher, B. Aldridge, and J. Rees, "Hierarchical classification of ten skin lesion classes," in *Proc. SICSA Dundee Med. Image Anal. Workshop*, 2015.
[23] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 1397–1400, doi: 10.1109/ISBI.2016.7493528.
[24] T. Akram, H. M. J. Lodhi, S. R. Naqvi, S. Naeem, M. Alhaisoni, M. Ali, S. A. Haider, and N. N. Qadri, "A multilevel features selection framework for skin lesion classification," *Hum.-Centric Comput. Inf. Sci.*, vol. 10, no. 1, pp. 1–26, Dec. 2020.
[25] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.
[26] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, Dec. 2018, doi: 10.1038/sdata.2018.161.
[27] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[28] M. A. Kassem, K. M. Hosny, R. Damaševičius, and M. M. Eltoukhy, "Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review," *Diagnostics*, vol. 11, no. 8, p. 1390, Jul. 2021.
[29] M. Goyal, T. Knackstedt, S. Yan, and S. Hassanpour, "Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities," *Comput. Biol. Med.*, vol. 127, Dec. 2020, Art. no. 104065, doi: 10.1016/j.compbiomed.2020.104065.
[30] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, "Soft-attention improves skin cancer classification performance," *Comput. Sci.*, vol. 12929, 2021.
[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*.
[32] C. Zhao, R. Shuai, L. Ma, W. Liu, D. Hu, and M. Wu, "Dermoscopy image classification based on StyleGAN and DenseNet201," *IEEE Access*, vol. 9, pp. 8659–8679, 2021, doi: 10.1109/ACCESS.2021.3049600.
[33] M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, "BCN20000: Dermoscopic lesions in the wild," 2019, *arXiv:1908.02288*.
[34] E. Xi, S. Bing, and Y. Jin, "Capsule network performance on complex data," 2017, *arXiv:1712.03480*.
[35] B. Graham, "Fractional max-pooling," 2014, *arXiv:1412.6071*.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
[38] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv.1312.4400*.
[39] F. Mamalet and G. Christophe, "Simplifying convnets for fast learning," in *Proc. Int. Conf. Artif. Neural Netw.*, 2012, pp. 58–65.
[40] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324, doi: 10.1109/ICCV.2019.00140.

ZHANGLI LAN received the Ph.D. degree in software and theory from Chongqing University, in 2008. He is currently pursuing the Doctor of Engineering degree with Chongqing Jiaotong University. He is also a Professor and the Young Backbone Teacher with Chongqing Jiaotong University. He focuses on the undergraduate teaching of analog circuit, digital image processing, and other courses. His research interests include traffic information and intelligence, solar energy, and image processing. He is an ACM Member, a CCF Senior Member, a CCF University Liaison Officer, a YOCSEF AC Member, and a Chongqing Artificial Intelligence Society Member.



SONGBAI CAI was born in Fujian, China, in 1989. He graduated in computer science and technology from Chongqing University, in 2016. He is currently pursuing the master's degree in computer science and technology with Chongqing Jiaotong University. His research interests include image classification, dermatoscopic image, deep learning, and large-kernel convolution.

XU HE was born in 1995. He received the bachelor's degree in agronomy from Jiangxi Agricultural University, in 2017. He is currently pursuing the master's degree in computer science and technology with Chongqing Jiaotong University. His research interest includes computer image processing.

XINPENG WEN was born in Chongqing, China, in 1997. He is currently a Graduate Student with Chongqing Jiaotong University, Chongqing. His research interests include image processing and machine vision.