Breast cancer X-ray image staging: based on efficient net with multi-scale fusion and cbam attention

Runyi Li\textsuperscript{1a*}, Sen Wang\textsuperscript{1b}, Zizhou Wang\textsuperscript{1c}, Lei Zhang\textsuperscript{1d}

\textsuperscript{1}Department of Computer Science, Sichuan University, Chengdu 610000, China
\textsuperscript{a}lirunyi@stu.scu.edu.cn, \textsuperscript{b}supremesam@stu.scu.edu.cn, \textsuperscript{c}wangzizhou@stu.scu.edu.cn, \textsuperscript{d}leizhang@scu.edu.cn

Abstract—With ever-progressing development period, image classification algorithms based on deep learning have shown good performance on some large datasets. In the development of classification algorithms, many proposals related to attention mechanism have greatly improved the accuracy of the model, and at the same time increased the interpretability of the network structure. However, on medical image data, the performance of the classification algorithm is not as expected, and the reason is that the fine-grained image data differs little among all classes, resulting that the knowledge domain is also hard to learn for models. We (1) proposed the EfficientNet model based on the cbam attention mechanism, and added a multi-scale fusion method; (2) applied the model to the breast cancer medical image data set, and completed the breast cancer classification task with high accuracy (Phase I, Phase II, Phase III, etc.); (3) Compared with other existing image classification algorithms, our method has the highest accuracy, thus the researchers conclude that EfficientNet with CBAM and multi-scale fusion will improve the classification performance. This result is helpful for deeper research on medical image processing and breast cancer staging.

1. Introduction
There has been a lot of demand regarding staging task of medical image, to diagnose and cure patients in an automatic and accurate way. Recent approaches are implemented by classification methods. Most of the image classification models are achieved through convolutional neural networks, starting from a simple convolution-fully connected structure, and constantly updated, and have achieved high-precision results on various natural image data sets. Related research has inspired researchers in the medical field and applied classification algorithms to medical image processing, especially the staging tasks of diseases. However, due to factors such as fine-graininess, narrow representation knowledge domain, and excessive noise of medical images, most of the results obtained by previous studies are not high in accuracy, so it is difficult to apply to clinical diagnosis. We apply the existing classification algorithm to the breast cancer image data set, combining the method of multi-scale fusion and cbam attention mechanism to improve the accuracy, generalization ability and interpretability of model on medical image classification tasks. Through the work proposed above, the accuracy and interpretability would be improved.

2. Related works
This part will introduce the related works including image classification and medical image classification.
**Image classification** The classification task is achieved using convolutional neural networks, and with the layer going deeper and more complex, the accuracy has risen rapidly in recent works.

Reference [1] proposed that when using identity mapping as skip connection and activation after adding, it is able to be propagated directly from one block to any other. ResNet has a better performance on training time and parameter amount, while compared to the recent work mentioned below, the accuracy of classification is still relatively low, which cannot meet the standard of medical data. Reference [2] studied the scaling of the algorithm, and it was ensured that a careful balance of network depth, width can lead to better accuracy and flexibility. This work achieved state of the art (simplified as SOTA) on ImageNet dataset challenge. Due to the small depth of network architecture compared to other proposals, it is hard to extract visual and semantic features on fine-grained tasks. Reference [3] showed that dependence of CNN is not necessary, and a transformer-only way directly applied to the image patch sequence can perform very good performance on image classification tasks. However, as ViT works on attention mechanism, it is hard for model to fetch features from data which have small difference among classes, like medical images. Reference [4] introduced MLP-Mixer, which is an architecture based entirely on Multilayer Perceptron (MLP). What makes it not capable for our goal is that MLP needs a huge amount of parameters and data to train, while medical image dataset is too small for that.

**Medical image classification** Reference [5] proposed a training procedure in two stages, allowing us to use high-capacity patch-level network to learn via pixel-level labels with a network learning from macroscopic breast-level labels. This work raised deep learning approach and multi-part detection method, which inspired the following work. However, this work is a positive-negative classification task, which lacks innovation and relation to our work. Reference [6] proposed three training methods: a baseline method for training the CNN architecture from scratch, a transfer learning method for further training the pre-trained VGG16 CNN architecture through ultrasound images, and a fine-tuning learning method. Fine-tuning method has been implemented, while it is still a two-class diagnosing task. Reference [7] classify and identify breast cancer images from the BreakHis dataset using CNN structure. This data set includes 7900+ breast cancer (BC) histopathological images, including 4 benign subcategories and 4 malignant subcategories. This approach is dealing with 8 classes, which is similar to ours, and our approach is inspired by this work.

3. **Methodology**
In this section, the details of methodology will be presented. Firstly, analysis of the problem and the structure of our method. Then present the whole architecture as shown in Fig.1. Next, the innovative modules used to our model are illustrated in the third part. Finally, the loss function of model was presented in the last part.

3.1. **Model overview**
Our model is based on the CNN structure of Efficientnet, and embeds the method of CBAM and multi-scale fusion to improve learning efficiency and generalization ability of the model. The visualization of the model is shown in the Fig.1.

The model first reads breast images through the input layer of Efficientnet, extracts spatial attention and channel attention through cbam during training, and learns semantic and geometric features under multi-scale features. The reasons for choosing cbam attention mechanism and multi-scale fusion as the optimization method will be explained in the following parts.
3.2. CBAM attention

CBAM is an effective attention module for CNN. The CBAM module sequentially infers the attention map along two independent dimensions with an intermediate feature map given, and then multiplies the map with the input feature map for adaptive optimization.

The channel attention module passes the input, and then respectively through MLP, performs element-wise summation operations on the features output by MLP to generate the final channel attention feature map. Perform a multiplication operation on it and input feature map to generate the input features required by the spatial attention module. The formula can be expressed as:

\[ M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \]

The spatial attention module uses the feature map output by attention mentioned above as the input feature map of this one. The feature and the input of the module are multiplied to generate the final generated feature map. The formula can be expressed as:

\[ M_s(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \]

which \( \sigma \) denotes activation function, and \( f^{7\times7} \) denotes \( 7\times7 \) convolution. The visualization of channel attention is shown as Fig.2 below.

3.3. Multi-scale feature fusion

The CNN gets the feature of the image in a layer-by-layer abstraction. One of the core concepts is the receptive field. This field of the high-level network is relatively huge, and the semantic information expressing ability is good, but the resolution of the feature map is low, and the expressing ability of geometric information is weak (lack of spatial geometric feature details); the receptive field of the low-level network is relatively small, and the representation ability of geometric details is strong. High-level semantic information can help us accurately detect and segment the target, but some geometric details
are often lost. These details are the characteristics that doctors will consider when processing medical images. Therefore, we apply the multi-scale fusion method to our model. The process of fusion is shown as Fig.3 below.

![Fig.3 Multi scale fusion](image)

3.4. Loss function
Since our image has four regions, namely the left, right, left, and right sides, the total loss function is equal to the sum of the loss functions of a single region. Considering the commonly used cross entropy for image classification tasks, our loss function is as follows:

$$L_{side}(y) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{m} y_j^i \log(y_j^i)$$

where $side = \{L, R, HL, HR\}$

The total loss function is as below:

$$L = L_L + L_R + L_{HL} + L_{HR}$$

4. Experiment

4.1. Dataset
The data of our experiment comes from breast cancer data provided by West China Hospital, which is composed of more than 300 breast cancer X-ray images and corresponding graded diagnosis results, as shown in the Fig.4. The steps of data processing are as follows:

**Region segmentation** We divide the X-ray image into four regions, namely the left side (L), the right side (R), the left side (HL), and the right side (HR). Each face is divided into levels 0-3 according to the corresponding diagnostic text of the image, corresponding to stage 0 (no cancer cells) to stage 3 (advanced stage) in the breast cancer grading standard.

**Data augmentation** The breast image of a single area is processed by rotation, partial cropping, and contrast adjustment. Considering that augmentation methods such as flipping and noise will affect the knowledge representation of image data, we did not choose them.
4.2. Implement details
Our work is based on PyTorch platform and the code is referred from GitHub repo [8]. The baseline Efficient Net’s version is b5 with appropriate size of parameters. To implement multi scale fusion, version b3 and b6 were chosen to get the shallow and deep feature map. Due to the restriction of computing resource, higher versions of Efficient Net are not chosen in our work.

The related hyperparameter settings and device information are shown in the Table1 below.

| Name                | Value         |
|---------------------|---------------|
| Initial learning rate | 7.5e-5       |
| Learning rate schedule | Cosine annealing |
| Optimizer           | RAdam         |
| Image size          | 456*456       |
| Epochs              | 100           |
| Batch size          | 2             |
| GPU information     | NVIDIA 3090   |

4.3. Quantitative results
We conducted a comparative experiment and compared our model with the existing SOTA model Vision Transformer and the baseline model Efficient Net. At the same time, in order to prove the effectiveness of improvement, the researchers conducted ablation experiments and compared the results of only Efficient Net, Efficient Net+CBAM, and Efficient Net+ multi-scale fusion.

The common metric for image classification is accuracy. Taking into account the false positive situation in medical practice, we choose F1-score as our model metric. The result is shown in Table2 and 3.

| Model                | Accuracy | Recall | F1-score |
|----------------------|----------|--------|----------|
| ViT                  | 63.81    | 66.94  | 65.34    |
| EfficientNet(baseline) | 64.71    | 50.23  | 56.56    |
| Ours                 | 76.47    | 65.73  | 70.69    |

| Model                               | Accuracy | Recall | F1-score |
|-------------------------------------|----------|--------|----------|
| EfficientNet(baseline)              | 64.71    | 50.23  | 56.56    |
| EfficientNet+cbam                   | 70.58    | 56.79  | 62.94    |
| EfficientNet+cbam+multi-scale-blending(ours) | 76.47    | 65.73  | 70.69    |
According to the results of the above-mentioned comparative experiments and ablation experiments, we can conclude that the EfficientNet with cbam and multi-scale fusion in breast cancer imaging data surpasses the baseline model and the optimal model for classification tasks, which proved that the model is effective in the task.

4.4. Analysis and visualization

According to the results in the previous section, we conclude that the combination of cbam and multi-scale network structure can effectively solve the problem of breast cancer staging and treatment. The following is an in-depth analysis and display of the experimental results.

In the result of ours, its staging accuracy reached a high 76.47%, surpassing other existing models, and it is also in a good effect in the same type of research. However, its recall rate is only 65.73%, which is relatively low for medical practice. In addition, we have counted the accuracy rates of the four categories, and the results are as follows:

| Class   | Accuracy |
|---------|----------|
| Stage0  | 75.52    |
| Stage1  | 80.91    |
| Stage2  | 74.21    |
| Stage3  | 68.95    |

It can be seen that the recognition accuracy of breast cancer in stage 0-2 is the highest, but the recognition effect of breast cancer in stage 3 is poor. In fact, according to the distribution of the data set we use, the data of the third period is the least, and even though the data has been expanded moderately, it is not possible to fully express the knowledge. If the content diversity and representation diversity of the data set can be enhanced, we believe that the overall effect of the model will be better.

The training curve of epoch-accuracy is as follows, from Fig5 to Fig7:

![Fig.5 Baseline training procedure](image)
![Fig.6 Training procedure with CBAM module](image)
![Fig.7 Training procedure with CBAM module and multi scale fusion](image)

It can be seen that the baseline model and our work have a certain degree of overfitting during the later training. If the early stopping mechanism is set, the model performance will be better.
In addition, different from the classification task of natural images, the staging task of breast cancer has connections and distance differences between classes, and the four periods have temporal order. Inspired by the task of age estimation, the ordinary regression method may help solve the task, and further research is needed and encouraged.

5. Conclusion
Based on the results and discussions presented above, the conclusion is shown as below:

We proposed an Efficientnet classification model that combines cbam and multi-scale fusion, and applied it to the staging task of breast cancer, achieving an accuracy of about 76%, surpassing the existing related research. Cbam and multi-scale fusion improve the accuracy of Efficientnet in terms of interpretability and performance. This method can be extended to other classification tasks.

However, such accuracy can only be used for auxiliary diagnosis, and its actual effect still needs to be tested by professional doctors. In addition, algorithms that deal with the gap between classes, such as ordered regression, can be applied to this task, and we encourage more in-depth research.

Acknowledgements
This work was financially supported by the Innovation and Startup Project of Sichuan University (C2021117404).

References
[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016, October). Identity mappings in deep residual networks. In European conference on computer vision (pp. 630-645). Springer, Cham.
[2] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (pp. 6105-6114). PMLR.
[3] Gillespie, L., Ruffley, M., & Expósito-Alonso, M. (2021). An Image is Worth a Thousand Species: Scaling high-resolution plant biodiversity prediction to biome-level using citizen science data and remote sensing imagery. Biodiversity Information Science and Standards, 5, e74052.
[4] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., ... & Dosovitskiy, A. (2021). Mlp-mixer: An all-mlp architecture for vision. arXiv preprint arXiv:2105.01601.
[5] Jush, F. K., Biele, M., Dueppenbecker, P. M., Schmidt, O., & Maier, A. (2020, September). DNN-based Speed-of-Sound Reconstruction for Automated Breast Ultrasound. In 2020 IEEE International Ultrasonics Symposium (IUS) (pp. 1-7). IEEE.
[6] Samala, R. K., Chan, H. P., Hadjiiski, L., Helvie, M. A., Richter, C. D., & Cha, K. H. (2018). Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. IEEE transactions on medical imaging, 38(3), 686-696.
[7] Rakhlin, A., Shvets, A., Iglovikov, V., & Kalinin, A. A. (2018, June). Deep convolutional neural networks for breast cancer histology image analysis. In International conference image analysis and recognition (pp. 737-744). Springer, Cham.
[8] https://github.com/whisperLiang/Efficientnet_pytorch_cbam_gui