**Design and Implementation of a Custom Computer Vision Model Using Classical Deep Learning Techniques with Explainable AI for Image Classification**

**Abstract**

Malaria, a persistent and life-threatening infectious disease, has long plagued humanity. Accurate diagnosis and treatment are crucial for its control, with cell image analysis being a simple yet effective diagnostic approach. To address this challenge, this study proposes a model that incorporates a convolutional neural network architecture and an advanced attention mechanism. By integrating spatial and channel attention and employing a residual-based approach to modulate feature maps, the model achieves superior classification performance. Meanwhile, the use of Grad-CAM, a cutting-edge Explainable AI (XAI) technique, effectively compensating for the lack of interpretability in this deep learning model. Evaluation results on the malaria cell dataset show that the proposed model achieves up to 96% accuracy, which greatly outperforms a host of fine-tuned pre-trained deep learning models. Quantitative results and qualitative results obtained by Grad-CAM show that the proposed model adequately learns effective classification guidelines.

**Keywords:** image classification; deep learning; attention mechanism; explainable AI

# **Introduction**

Image classification is a fundamental research task in computer vision, where the goal is to assign a given raw image to a set of predefined categories. This technique is already used in everyday life, for example, AI assistants on cell phones can tell us what kind of animal or plant a photo is. Many other advanced tasks in computer vision are also based on image categorization[1], in areas such as object detection and image segmentation. Early image classification methods were mainly implemented based on manually designed feature engineering and combined with classification algorithms in machine learning. For example, the SIFT or HOG feature descriptors in the image are extracted and after converting them into feature vectors, they are used as inputs to classification algorithms such as SVM, Random Forest, etc. and the corresponding categories are obtained. In this process, there is no unified solution for the design and selection of features, making the algorithms perform poorly in more complex real-world scenarios.

With the rise of deep learning technology, the field of image categorization has also made large breakthroughs. Deep learning methods are able to achieve end-to-end learning and classification[2] without having to be divided into two stages as in traditional methods. Large datasets such as ImageNet have been proposed, allowing deep neural network models to have more powerful learning and fitting capabilities, as well as more generalization capabilities in the face of new samples. Deep learning methods are also being updated and iterated, with a lot of work on network architecture, activation functions, regularization techniques, optimization methods, etc. to improve performance.

In addition to the widely used convolution, attention mechanisms are also important in visual tasks. The attentional mechanism in deep learning is inspired by the selective process in human cognition[3], whose core purpose is to select the more relevant and critical information to the current task goal from the redundant information, so as to process the information more efficiently and complete the task. The human visual attention mechanism is born with the ability to quickly identify the target area of visual focus and form the focus of attention. For example, when a photograph is in front of us, we tend to focus on the foreground objects in it, while easily ignoring the largely coherent background. Attention mechanisms first made groundbreaking progress in natural language processing, with one of the most important work being the self-attention and cross-attention proposed in Transformer, which achieve very good performance on sequence-type tasks. Inspired by this, many approaches have tried to bring the attention mechanism into image processing tasks, and one typical class of approaches is to compute the attention scores from the feature maps and adjust the feature maps. The introduction of the attention mechanism has led to further performance improvements in deep convolutional networks.

However, despite the good performance, the poor interpretability of deep learning methods has been an open question. Deep neural networks often have very complex structures and huge parameter counts, which makes the model's inference and decision-making process not well explained, and this is the reason why people refer to deep learning as a black box. This opacity not only reduces the possibility of applying and optimizing AI models in some specific domains, but also raises concerns about the safety and reliability of AI technology. XAI aims to solve this problem by making the inference process of deep learning methods easier to understand for humans. In addition to explaining the inner workings of deep learning models, some XAI techniques also aim to detect and eliminate incorrect inductive biases in models[4], as well as to meet specific requirements imposed on deep learning methods by some particular domains. In short, the goal of XAI is to provide reliable and efficient explanations for deep learning models without compromising the model performance and without incurring much extra overhead.

The aim of this report is to utilize the above techniques to solve the malaria recognition problem, i.e., determining whether a subject is infected with malaria or not from a given cell photograph. By combining deep learning methods and attention mechanisms, I achieved a classification accuracy up to 96% in Malaria Cell Images Dataset. In addition, by using Grad-CAM, an XAI technique, I provide a better explanation of the decision-making process of the model proposed in this paper.

The remaining of the report is structured as follows. Section 2 provides a short introduction to the work related to the problem studied in this report. Section 3 describes the dataset used and the proposed model. Section 4 presents the experimental results of the model in detail and compares them with other models to show the robustness of the present model. Section 5 summarizes and explores possible future research directions.

# **Related work**

A large number of deep learning-based approaches have emerged for image classification tasks, many of which were compared for performance in the ILSVRC[5] (ImageNet Large Scale Visual Recognition Challenge) challenge on ImageNet. In 2012, AlexNet[6] made its debut in this challenge using a deep learning methods, and the Top-5 error rate dropped from the previous 26.2% to 15.3% compared to previous machine learning models, an advancement that shows the promise of deep learning in computer vision. In 2014, VGGNet[2] demonstrated the power of increasing the depth of the network and filtering using a combination of smaller convolution kernels by stacking convolutions. However, networks that are too deep are prone to stagnation problems in the learning process, and ResNet[6] in 2015 proposed residual connectivity to effectively alleviate the learning problem of deep networks. With this structure, the deeper the network, the better the performance achieved, given enough training data. With these techniques, the performance of deep networks on image classification tasks has even surpassed human recognition.

In 2017, with Transformer and self-attention[7] making a big splash in natural language processing tasks, research on introducing attention mechanisms into visual tasks has also received significant attention. Early attention mechanisms were mainly based on convolutional networks, such as CBAM[8] and SENet[9], by calculating spatial or channel attention from feature maps and reacting the attention result back to the feature maps as a way to enhance the useful information and suppress the ineffective information.ViT[10] was firstly proposed to use pure attention to complete the image classification without relying on convolution anymore. It divides the image into multiple 16x16 blocks, each patch is equivalent to words in natural language processing, and then applies Transformer to these sequences. However, this purely attentional structure is very dependent on large datasets and is prone to overfitting in case of insufficient data.

As deep learning techniques achieve excellent performance in various tasks, how to improve their interpretability has also sparked the research community's interest in XAI techniques. LIME[11] (Local Interpretable Model-agnostic Explanations) is a method for interpreting the predictive results of a model, which is done by in the vicinity of the input sample Perturbation is performed to generate new sample points. After obtaining the model predictions, the data is used to train an interpretable model (e.g., a linear model) in a local region. LIME can be applied to any type of model, and at the same time gives a very intuitive visual interpretation of the results. Grad-CAM[12] (Gradient-weighted Class Activation Mapping) is a visualization technique focused on interpreting deep learning models (especially convolutional neural networks). It uses gradient-based computation to locate the most important regions in the input for the model's predicted class, but may ignore fine-grained feature importance.

# **3. Material and Method**

## **3.1 Dataset**

In order to fully evaluate the validity of the constructed model, the malaria cell categorization dataset on Kaggle was used. This dataset consists of two main categories, pictures of infected cells and pictures of normal cells, and the pictures are neatly categorized into two folders whose names denote the label names. In this section, I will give a detailed description of the dataset.

This dataset was originally created to replace human labor in performing rapid cell categorization to determine whether a corresponding subject is infected with malaria. The images in the dataset have been organized in an orderly fashion, with infected and normal cell images divided into two separate folders. The dataset contains a total of 27,558 images, with infected and normal cells each accounting for half of the images, indicating that there is no annoying data imbalance. Below I have visualized some of the images in the dataset:

形状, 多边形

描述已自动生成

Figure 1. Random Samples from the dataset

***A. Data Preprocessing***

Then, I analyze the width and height of the images in the dataset:

图表, 直方图

描述已自动生成Figure 2. Image size analysis of the dataset

The results show that the vast majority of the images are within the range of [100, 150], while the height and width of an image directly does not vary much. Therefore, I standardized the image sizes to [128, 128] when loading the dataset. For the sake of robustness of gradient propagation during training, I performed a normalization operation on the images (pixel values divided by 255). Data augmentation is a common technique used in image processing tasks to enhance the performance of the model and the following data augmentation is used in this report:

1. RandomFlip: Randomly flips the image horizontally or vertically;
2. RandomRotation: randomly rotate the image at an angle of [-0.1, 0.1]\*2pi
3. RandomContrast: random contrast enhancement
4. RandomZoom: random zoom.

The enhancement process defined in this report will only implement data enhancement during training and are automatically turned off during test. For the images in the training set, there is a 30% probability of implementing data enhancement, and each enhancement will randomly select one of the above four algorithms, a practice that greatly enhances the diversity of the training data.

***B. Data Split***

In order for the model to be adequately trained and effectively evaluated, the dataset is divided into training set, validation set, and test set according to the ratio of 75%, 12.5%, 12.5% in this report. After the division, the training set contains 20668 images, and the validation and test sets contain 3444, 3446 images, respectively (the number of images in the validation and test sets are not strictly equal due to the rounding operation in training).

## **3.2 Proposed Model**

Describe your own proposed medel. An example is provided below:

This report uses a deep learning based approach to solve the malaria cell classification problem, the basic architecture of the proposed model is a convolutional neural network combined with an attention module.

图示

描述已自动生成Figure 3. Overview of the proposed model

The proposed model uses the common components and processing order in convolutional neural network. During preprocessing, I have resized the inputs all to the size of [128,128], which is a relatively small input in visual tasks. Hence the convolution size used in the proposed model is 3x3. After convolution Batch Normalization process is performed.The BN layer looks at each batch of data, calculates its mean and variance and adjusts the data accordingly to make its distribution more standardized.The use of BN layer is a popular technique in CNNs to make the gradient back propagation process more stable and helps to improve the performance of the network.

Activation function enables a neural network to have the ability to fit complex nonlinearities and is an essential part of the network structure. I studied commonly used activation functions, starting with the S-type Sigmoid and tanh activation functions, which are commonly used in some machine learning algorithms, but both involve calculating the exponential term, which is a relatively complex computational process. Then there are ReLU and its variants, such as Leaky ReLU, ELU and so on. Among these functions, ReLU is the simplest to compute and can make the forward and backward propagation process of the network more efficient. At the same time, ReLU sets the negative value to 0, which can make the network with sparse characteristics and avoid the appearance of overfitting. Therefore, ReLU is used as the activation function in this model.

Pooling can reduce the size of the feature map and increase the sensory field in the depth of the network, and MaxPooling is used in this model.In the final part of the classification header, Global Average Pooling is used to aggregate the feature maps into feature vectors, and a two-layer MLP is used to get the classification results.The number of neurons in the first layer of the MLP is not easy to be determined, so I set it to 64 for the time being, and try to adjust it in the subsequent layers to get the classification results. The number of neurons in the first layer of the MLP is not easy to determine, so I tentatively set it to 64 and try to optimize it in the subsequent tuning to find the optimal value.

***Attention Module***: in addition to the basic convolutional architecture, an attention mechanism is incorporated in this model. Most of the previous approaches deal with spatial attention and channel attention separately, which increases the model complexity to some extent. Meanwhile, some high-level semantic information in the feature map often cannot be simply interpreted separately as spatial and channel roles. Therefore, inspired by the GCNet[13], this paper implements a mechanism with both channel and spatial attention. First, I do not simply use GlobalAveragePooling to extract the initial value of attention for [C, 1,1] from the feature map. Instead, I used convolution with a kernel size of 1 in conjunction with Softmax to preserve spatial dimensionality while adaptively assigning weights to different spatial locations. Matrix multiplication is then used to interact with the corresponding view of the input to obtain the initial result of [C, 1,1].

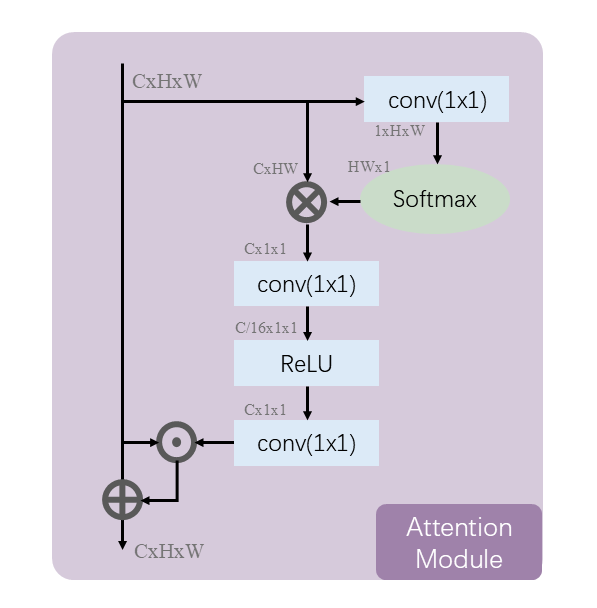


Figure 4. Attentiont module in the model

Next, I constructed the bottomleneck structure in the attention module using a convolution with a kernel size of 1. The dimensionality was first reduced to C//16, and then re-upsized to C after processing by the activation function. this approach, while efficiently reducing the number of parameters, enables the model to reconsider the correlations between the channels and construct effective attention weights. Finally, in applying the attention weights to the inputs, instead of simply using additive or multiplicative operations, I weighted the inputs on the channels according to the attention weights and superimposed them on the original inputs. This processing is somewhat analogous to residual concatenation, which is intended to back propagate gradients more efficiently in deep networks. The entire attention module can be plug-and-play added to the CNN, thus introducing the advantages of the attention mechanism in visual tasks.

***XAI Integration***: In order to provide a reasonable explanation for the model's predictions, I integrated Grad-CAM into the model. Grad-CAM is particularly suitable for convolutional architecture based classification models, which can generate heat maps of images based on the trained model, showing which regions the model focuses on when making decisions. The principle is to calculate the gradient of the target category score relative to the last convolutional layer feature map, perform GAP on the gradient in the width and height dimensions, obtain importance weights, and weight and sum them.

## **3.3 Evaluation Strategy**

For the classification task, the evaluation metrics used in this report include accuracy, precision, recall, F1, and AUC, and their calculation methods are explained below:

* Accuracy(A, B) =
* Precision =
* Recall:
* F1:
* AUC: area under ROC curve

Each of these metrics has its own focus, and by using them together, they can complement each other and provide a comprehensive performance evaluation of the model.

## **3.4. Environment Execution**

The training of models in deep learning is often done on servers, and the use of GPUs can greatly accelerate the training process of models. The graphics card model I used in my experiments was an RTX 2080Ti with 12GB of video memory.The CPU model in the system was an Intel(R) Core(TM) i9-9820X CPU with a base frequency of GHz.The operating system was Ubuntu 20.04.4 LTS, which is a commonly used system on servers. The construction of the model is done under the tensorflow framework.

## **3.5. Model Training**

After the model was constructed, I completed training as well as evaluation of the model on the dataset. The losses recorded during the training process are shown below:

图表, 折线图

描述已自动生成

Figure 5. Training record before parameter tuning

However, there are some hyperparameters in the model that affect the structure of the model or the training process and they are difficult to determine directly. Therefore, I use keras tuner to perform a search on these hyperparameters to determine the relatively optimal values. In order to improve the efficiency of the tuning process, I set the tuner's objectives as the learning rate, and the number of neurons in the first fully connected layer. After 10 random search trials, the values of these two parameters were determined: 96, 0.003.

The model was constructed using the optimal parameters, retrained, and recorded as follows:

图表, 折线图

描述已自动生成

Figure 6. Training record after parameter tuning

# **4. Experimental Results**

In this section I evaluate the proposed model on the malaria cell dataset and compare it to other AI models.

**4.1. Performance Results on the Malaria Cell Dataset**

The results obtained by each method are shown in **Table 1**.

**Table 1**. Quantitative comparison on malaria cell dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **Total Params** |
| MobileNet | 0.946 | 0.932 | 0.965 | 0.949 | 3,327,458 |
| MobileNetV2 | 0.942 | 0.946 | 0.941 | 0.944 | 2,381,154 |
| ResNet50 | 0.853 | 0.857 | 0.856 | 0.857 | 23,784,610 |
| VGG16 | 0.873 | 0.876 | 0.877 | 0.877 | 14,764,130 |
| VGG19 | 0.882 | 0.883 | 0.888 | 0.885 | 20,073,826 |
| **Ours** | **0.966** | **0.952** | **0.981** | **0.966** | **87,760** |

To further compare the performance performance between the models, I show their training records in Figure 7. The results show that the proposed model achieves the best on multiple metrics on this cell classification dataset.

图表, 折线图

描述已自动生成

Figure 7. Training records of different models

## **4.2. XAI Results**

After the model training is completed, I use Grad-CAM to explain the decision-making process:

图片包含 图示

描述已自动生成

Figure 8. Grad-CAM interpretation results

From left to right in the Figure 8 are the attention heatmap obtained by Grad-CAM, the original parasitized image, and the original image overlay attention heatmap. The results indicate that the model focuses on the cell boundary and a small abnormal area inside the cell when classifying infected cells. This is similar to the basis for human classification, so this technology makes the model's predictions more transparent. This attention pattern is very similar to the diagnostic methods of human experts, which indicates that the model has learned effective classification criteria.

## **4.3 Fair comparison with other Deep Learning Models**

I selected five deep learning models to compare in my experiment, as in Table 1. Each of these models freezes the weights of the backbone component and fine-tunes the redesigned classification header on the malaria cell dataset. Among them, MobileNet and MobileNetV2 performs better, ResNet50 performs worse. The reason for this may be that this is a small dataset, and models with too many parameters tend to overfit, while lighter models tend to perform better. Also, the pre-training weights of these models are on ImageNet, which is quite different from the cell images used in this experiment.

## **4.4 Discussion**

Maintaining the simplicity of the model structure is crucial for relatively straightforward tasks. I fully considered this point when designing the model. Instead of using complex structures with a large number of parameters, I completed the task using small convolutions and efficient attention mechanisms. The sufficient experimental results indicate that the proposed model accomplishes the classification task with the least number of parameters and the best performance.

# **5. Conclusion, Limitation, Future work**

In this report, I proposed a method combining deep learning techniques with advanced attention mechanisms for the malaria cell classification task. The proposed method is primarily based on convolutional neural networks, integrating both spatial and channel attention modules to achieve precise and effective modulation of feature maps. The motivation for this work is to fully leverage the strengths of convolutional networks in feature extraction while enhancing key features through attention mechanisms. This study is expected to assist or even replace medical personnel in determining whether patients are infected with malaria. Subsequent research can focus on developing more lightweight versions suitable for computationally constrained scenarios such as mobile devices, thus expanding the applicability of this research.

**References**

[1] W. Rawat and Z. Wang, “Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review,” *Neural Computation*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017.

[2] Z. Yan, W. Liu, S. Wen, and Y. Yang, “Multi-Label Image Classification by Feature Attention Network,” *IEEE Access*, vol. 7, pp. 98005–98013, 2019.

[3] Y. Peng, X. He, and J. Zhao, “Object-Part Attention Model for Fine-Grained Image Classification,” *IEEE Trans. on Image Process.*, vol. 27, no. 3, pp. 1487–1500, Mar. 2018.

[4] R. Dwivedi *et al.*, “Explainable AI (XAI): Core Ideas, Techniques, and Solutions,” *ACM Comput. Surv.*, vol. 55, no. 9, p. 194:1-194:33, Jan. 2023.

[5] D. Lu and Q. Weng, “A survey of image classification methods and techniques for improving classification performance,” *International Journal of Remote Sensing*, vol. 28, no. 5, pp. 823–870, Mar. 2007.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[7] H. Guo, K. Zheng, X. Fan, H. Yu, and S. Wang, “Visual Attention Consistency Under Image Transforms for Multi-Label Image Classification,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 729–739.

[8] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” presented at the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[9] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[10] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, “The Application of Two-Level Attention Models in Deep Convolutional Neural Network for Fine-Grained Image Classification,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 842–850.

[11] A. Barredo Arrieta *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization,” presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[13] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond,” presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.