ISSN 2959-6157

Impact of Pre-trained Weights on MobileNet Architectures for Animal Classification

Yusen Shi

International School, Beijing University of Posts and Telecommunications, Beijing, China
jp2019213423@qmul.ac.uk

Abstract:

This study aims to address the problem of improving animal image classification accuracy using different versions of MobileNet. Accurate animal classification plays a vital role in biodiversity protection, environmental monitoring, and agriculture. The research is significant because existing studies focus on specific models and datasets, leaving a gap in the comparative performance analysis of MobileNet versions. To address this issue, MobileNet V1, V2, and V3 models were utilized, both with and without ImageNet pre-trained weights. The models were trained on a dataset composed of 30,179 images from two sources, covering 13 animal categories. The experiment involved training the models over 10 epochs using a standard configuration of the TensorFlow framework, with accuracy serving as the primary evaluation metric.The results showed that MobileNet V3Large, with pre-trained weights, achieved the highest accuracy (97.43%), outperforming V1 and V2. Using pre-trained weights consistently enhanced performance, as models without pre-training exhibited lower accuracy and slower convergence. This study contributes by providing a comprehensive comparison of MobileNet versions in animal classification tasks, demonstrating the importance of pre-training and model architecture optimization for achieving high accuracy in image classification.

Keywords: Animal classification; MobileNet comparison; pre-trained weights

1. Introduction

Animals are one of the most important groups of life in nature. They not only play a key role in the ecosystem, but also have a profound impact on all aspects of human society. Animal classification and identification is a basic work in biological research. It not only helps to protect biodiversity, but also has important application value in agriculture, animal husbandry, environmental monitoring and other fields.

With the rapid development of Artificial Intelligence (AI) technology, AI has shown great potential in many fields, such as biomedicine [1], autonomous driving [2], etc. In image classification tasks, AI, especially deep learning technology, performs well. Typical algorithms include random forests [3], Convolutional Neural Networks (CNN) [4], etc. In the medical field, CNN has been successfully applied to disease classification [5], the authors designed a model to automatically identify and classify brain tumors from Magnetic Resonance Imaging (MRI) images, thereby helping doctors more accurately diagnose and treat brain tumor patients; in environmental science, CNN is used for garbage classification [6], the authors tried different models and methods and evaluated the performance. The results showed that deep learning technology performs well in clean backgrounds and can effectively solve the garbage classification problem, but still needs to be improved in more complex real-world scenarios. These successful cases prove the superiority of deep learning models in image classification tasks.

In animal classification, many studies have tried to use AI technology to improve classification accuracy. For example, Nguyen et al. [7] used CNN to classify wildlife images, significantly improving the classification accuracy. Wang et al. [8] used pre-trained models for transfer learning to improve the classification effect in the case of few samples. However, most of these studies focus on specific models or specific datasets, and there are few studies on the performance comparison of different versions of classic networks on animal classification tasks.

This study aims to compare the performance of MobileNet V1, V2, and V3 [9-11] on animal image classification tasks, and explore the impact of using ImageNet pre-trained weights on classification results. To solve the above problems, this paper merged two animal image datasets and selected classic MobileNet v1, v2, v3Large, and v3Small models for training and prediction. During the experiment, this study used pre-trained and non-pre-

trained weights, respectively, and compared the differences in training results of different models in these two cases.

Specifically, this study focuses on the following aspects: Model performance comparison: By comparing the classification accuracy of MobileNet v1, v2, and v3 on the same dataset, analyze their performance in animal image classification tasks. Pre-trained weight impact: Investigate the impact of using ImageNet pre-trained weights on model classification results, and analyze whether pre-training can significantly improve model performance.

2. Method

2.1 Dataset Preparation

The dataset comes from two datasets on Kaggle, namely Animal Classification [12] and Animals-10 [13]. There are 13 categories of images, a total of 30,179 images, the size of a single image is not uniform, and the images are 3-channel RGB images. Fig. 1 shows some image samples.

Fig. 1 Some sample images in the collected dataset [13].

Each model uses the corresponding method for data preprocessing. MobileNet V1 and V2 will scale the input pixel values between -1 and 1 according to the sample, while the method of MobileNet V3 will not process the input because the preprocessing logic is already included in this model implementation. The ratio of the training set to the test set is 80% to 20%.

2.2 MobileNet-based Classification

MobileNet is a class of lightweight CNN designed for mobile and embedded devices. Its main goal is to achieve efficient image classification in resource-constrained environments. By utilizing deep separable convolutions, MobileNet greatly reduces the number of model parameters and computational complexity, becoming a widely used neural network architecture on mobile devices.

The core idea of MobileNet is to decompose the standard convolution operation into depth convolution and point-by-point convolution to reduce computational complexity. Depth convolution performs convolution on each input channel separately, while point-by-point convolution uses 1x1 convolution to integrate these outputs in the channel dimension. This approach reduces the number of model parameters and computational requirements while still being able to capture effective feature representations.

For all models, an average pooling layer was added to the top of each model in the experiment and the dropout rate was set to 0.2 to prevent overfitting. Finally, the Fully Connected (FC) layer was used for classification prediction.

2.2.1 MobileNet V1

MobileNet V1 is the first version of the series, introducing the innovative concept of depth wise separable convolution. It has two main features. One is depth-separable convolution: by decomposing standard convolution into depth convolution and point-wise convolution, the computational complexity and number of parameters are significantly reduced. The second is model size adjustment: the introduction of Width Multiplier and Resolution Multiplier allows adjustment between calculation amount and accuracy.

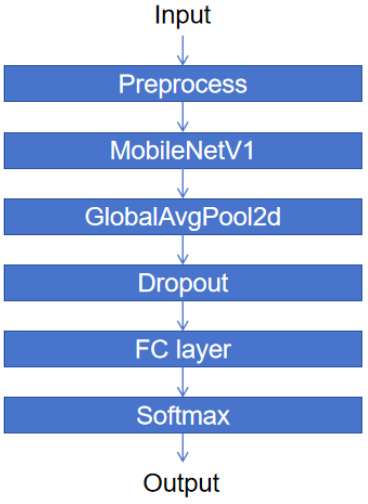

Fig. 2 The specific network structure of

MobileNetV1 used in the experiment (Photo/ Picture credit: Original)

The dropout rate in architecture shown in Fig. 2 is 0.2. The total number of parameters is 3242189. The number of trainable parameters of the model based on imagenet pre-training is 13325. The number of trainable parameters of the model without imagenet pre-training is 3220301.

2.2.2 MobileNet V2

MobileNet V2 has made important improvements on the basis of V1, especially in terms of network structure and feature expression ability. Its main improvement is the inverted residual structure. Compared with the traditional residual structure, the inverted residual structure reverses the number of input and output channels. Specifically, the input feature increases the number of channels through an expansion layer (i.e. 1x1 convolution), and after deep convolution processing, it is reduced by a compression layer (i.e. 1x1 convolution). This structure not only retains more feature information, but also reduces the amount of calculation. In addition, the linear bottleneck is also a major improvement. MobileNet V2 uses a linear activation function instead of Rectified Linear Unit (ReLU) at the output of the inverted residual block. The advantage of this design is that it reduces the loss of feature information, especially when the feature distribution is sparse.

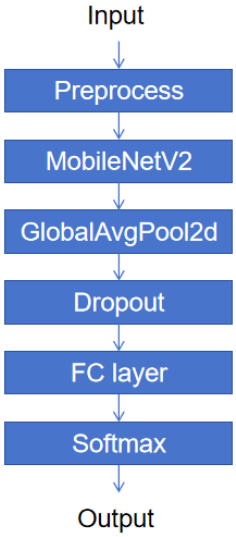

Fig. 3 The specific network structure of MobileNetV2 used in the experiment (Photo/ Picture credit: Original)

The dropout rate in architecture shown in Fig. 3 is 0.2. The total number of parameters is 2274637. The number of trainable parameters of the model based on imagenet pre-training is 16653. The number of trainable parameters of the model without imagenet pre-training is 2240525.

2.2.3 MobileNet V3

MobileNetV3 introduces multiple improvements and new technologies compared to previous versions. First, MobileNetV3 combines hardware-aware Neural Architecture Search (NAS) and NetAdapt algorithms. These technologies optimize the network architecture so that it can be fine-tuned for specific hardware (such as mobile phone CPUs). NAS is responsible for optimizing the overall architecture, while NetAdapt is used for layer-by-layer optimization to ensure optimal performance on specific hardware. Secondly, the hard swish activation function is introduced. This activation function is faster to calculate and more suitable for quantization, which is very suitable for mobile devices. In addition, compared with V1 and V2, V3 provides two models, Large and Small, which are optimized for high-resource and low-resource usage scenarios respectively.

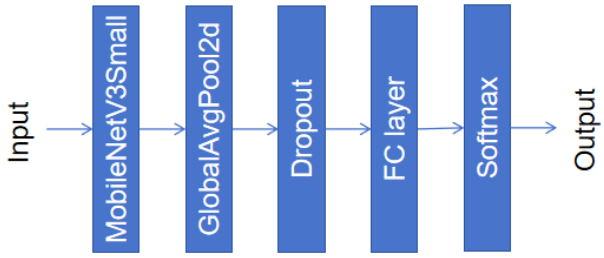

Fig. 4 The specific network structure of MobileNetV3Small used in the experiment (Photo/Picture credit: Original)

The dropout rate in architecture shown in Fig. 4 is 0.2. The total number of parameters is 946621. The number of trainable parameters of the model based on imagenet pre-training is 7501. The number of trainable parameters of the model without imagenet pre-training is 934509.

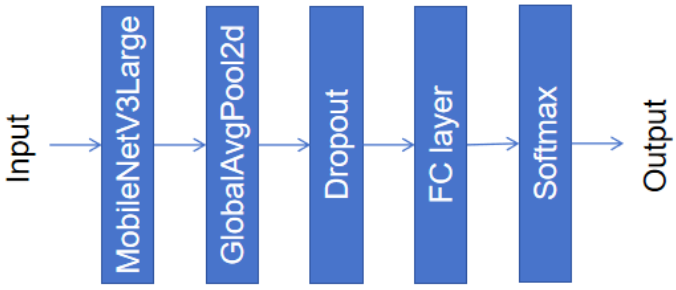

Fig. 5 The specific network structure of MobileNetV3Large used in the experiment (Photo/ Picture credit: Original)

The dropout rate in architecture shown in Fig. 5 is 0.2. The total number of parameters is 3008845. The number of trainable parameters of the model based on imagenet pre-training is 12493. The number of trainable parameters of the model without imagenet pre-training is 2984445.

2.3 Implementation Details

The experiment conducted in this study uses the TensorFlow framework, version 2.10.0, and the program runs on a GPU with CUDA version 11.8. The GPU model is an RTX 2070 laptop. The base learning rate is set to 0.001, and all optimizers use the Adam optimizer. The loss function employed is cross-entropy loss. The batch size is 32, and each model is trained for 10 epochs. The evaluation metric is accuracy.

3. Results and Discussion

3.1 The Performance of Various Models

Table 1. Model validation accuracy and loss

Model name	Validation Accuracy	Validation Loss	Recall	Precision
MobileNetv1-imagenet	0.9604	0.1453	0.96	0.96
MobileNetv1-none	0.7791	0.7906	0.78	0.79
MobileNetv2-imagenet	0.9637	0.1356	0.96	0.96
MobileNetv2-none	0.7586	0.7843	0.76	0.77
MobileNetv3Small-imagenet	0.9428	0.1680	0.94	0.94
MobileNetv3Small-none	0.7511	0.8252	0.75	0.75
MobileNetv3Large-imagenet	0.9743	0.0936	0.97	0.97
MobileNetv3Large-none	0.7901	0.7225	0.79	0.80

In this experiment, MobileNet V1, V2, and V3 models were used to compare the performance of animal image classification tasks with and without ImageNet pre-trained weights. By observing the validation accuracy and validation loss of different models, the following conclusions shown in Table 1, Fig. 6 and Fig. 7 can be drawn:

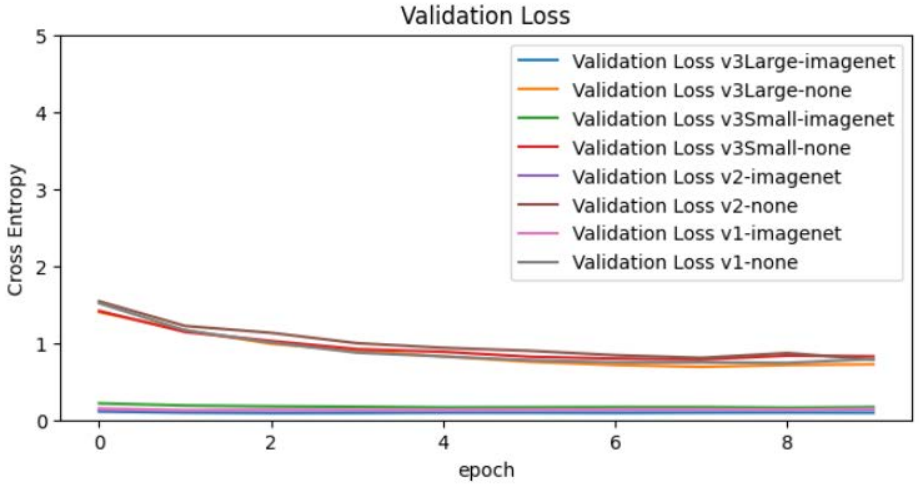

Fig. 6 Validation loss for all models during the training processes

(Photo/Picture credit: Original)

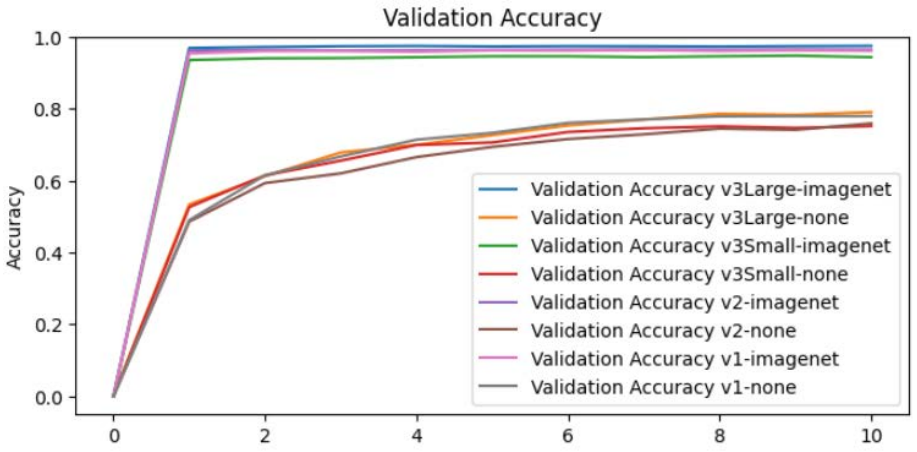

Fig. 7 Validation accuracy for all models during the training processes

(Photo/Picture credit: Original)

It can be observed that MobileNet V3Large performed best with pre-trained weights. In contrast, MobileNet V1 and V2 also had high accuracy, but they were slightly lower than V3. Models that did not use pre-trained weights generally performed poorly, with validation accuracies ranging from 0.75 to 0.8.

From the perspective of validation accuracy, models that used pre-trained weights performed significantly better than models that did not use pre-trained weights at the beginning of training. In particular, the V3Large model quickly reached an accuracy close to 1.0 after the initial first epoch, performing the best. The accuracy of models using pre-trained weights was relatively stable and fluctuated slightly within a higher accuracy range. In contrast, the models that do not use pre-trained weights also have a high accuracy increase in the early stage of the training process, but the accuracy increases slowly in the later stage of training, and the final accuracy is low. From the perspective of validation loss, the models that use pre-trained weights maintain a low and very stable loss value throughout the training process, especially V3Large and V2, whose loss values always remain in a very low range.

Under the same pre-training conditions, MobileNet V3Large performs better than other versions, especially the difference with V1 is particularly obvious, with an accuracy increase of about 1.39 percentage points. Although V3Small has a slightly lower accuracy than other models, the total number of parameters is one-third of V1, V2, and V3Large, which shows that the V3Small model can also achieve good classification results at a smaller scale.

Models that do not use pre-trained weights, such as V1 and V2, have high validation losses and slow convergence throughout the training process. This further proves the importance of pre-trained weights in improving model performance.

3.2 Discussion

From the experimental results, it can be found that MobileNet V3Large performs best in the animal image classification task, which may be closely related to its structural improvements. MobileNet V3 combines hardware-aware Neural Architecture Search (NAS) and NetAdapt algorithms to further optimize the network structure and achieve optimal performance on specific hardware. In addition, the hard swish activation function introduced by MobileNet V3 is not only faster in calculation, but also more suitable for quantization. These improvements may have a positive impact on the improvement of model performance.

In contrast, although MobileNet V2 introduces an inverted residual structure and a linear bottleneck, its performance is still not as good as V3, which may be because V3 combines more optimization techniques to make it perform better in complex tasks.

Without the use of pre-trained weights, the performance of all models dropped significantly. This shows that the role of pre-trained weights in image classification tasks cannot be ignored. It can provide better initialization for the model, thereby accelerating convergence and improving the final classification effect.

However, some shortcomings were also exposed during the experiment. For example, the generalization ability of these models may need further examination in more complex real-world scenarios. In addition, due to the limitations of the data set, the experimental results may be affected by the data distribution to a certain extent. Therefore, future research can consider introducing more diverse data sets or trying more model structures to further improve the classification effect.

Through the above analysis, it can be seen that different versions of MobileNet have their own advantages and disadvantages in the animal image classification task, but in

general, with the upgrade of the version, the performance of the model has improved, especially when using pre-trained weights, the effect is particularly significant.

4. Conclusion

In this paper, a comprehensive comparison of different MobileNet models (V1, V2, V3Large, and V3Small) was conducted for the task of animal image classification. By leveraging both pre-trained and non-pre-trained versions of these models, the performance and the impact of using ImageNet pre-trained weights on classification accuracy were explored.

Three main versions of MobileNet V1, V2, and V3 (both Large and Small)—were employed to perform the classification tasks using a dataset comprising 13 categories of animal images. The models were evaluated under two conditions: with and without ImageNet pre-trained weights. The methodology involved training these models over 10 epochs and comparing their validation accuracy and loss.

The experimental results demonstrated that MobileNet V3Large with pre-trained weights achieved the highest classification accuracy, closely followed by V2 and V1. The V3Small model, although slightly less accurate than its larger counterpart, proved to be an effective and efficient alternative with fewer parameters. The use of pre-trained weights significantly improved model performance, confirming the importance of transfer learning in enhancing model accuracy and reducing loss. Without pre-training, all models exhibited lower performance, with higher loss and slower convergence.

Despite the promising results, this study has some limitations. The dataset used was relatively small and may not fully represent real-world complexity, which could affect the generalization of the results. The future work could involve testing these models on larger and more diverse datasets, as well as exploring additional architectures or optimization techniques to further improve classification performance.

References

[1] Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Medical image analysis. 2017 Dec 1;42:60-88.

[2] Grigorescu S, Trasnea B, Cocias T, Macesanu G. A survey of deep learning techniques for autonomous driving. Journal of field robotics. 2020 Apr;37(3):362-86.

[3] Bosch A, Zisserman A, Munoz X. Image classification using random forests and ferns. In2007 IEEE 11th international conference on computer vision 2007 Oct 14 (pp. 1-8). Ieee.

[4] Rawat W, Wang Z. Deep convolutional neural networks for image classification: A comprehensive review. Neural computation. 2017 Aug 24;29(9):2352-449.

[5] Khasim S, Basha SS. Brain Tumor Identification and Classification System Using Convolutional Neural Network. International Journal of Health Sciences.(II):7264-75.

[6] Meng S, Chu WT. A study of garbage classification with convolutional neural networks. In2020 indo–taiwan 2nd international conference on computing, analytics and networks (indo-taiwan ican) 2020 Feb 7 (pp. 152-157). IEEE.

[7] Nguyen H, Maclagan SJ, Nguyen TD, Nguyen T, Flemons P, Andrews K, Ritchie EG, Phung D. Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring. In2017 IEEE international conference on data science and advanced Analytics (DSAA) 2017 Oct 19 (pp. 40-49). IEEE.

[8] Wang X, Li P, Zhu C. Classification of wildlife based on transfer learning. InProceedings of the 2020 4th International Conference on Video and Image Processing 2020 Dec 25 (pp. 236-240).

[9] Howard AG. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. 2017.

[10] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted residuals and linear bottlenecks. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 4510-4520).

[11] Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, Le QV. Searching for mobilenetv3. InProceedings of the IEEE/CVF international conference on computer vision 2019 (pp. 1314-1324).

[12] Kaggle, Animal Classification, 2022, https://www.kaggle.com/datasets/ayushv322/animal-classification

[13] Kaggle, Animals10, 2019, https://www.kaggle.com/datasets/alessiocorrado99/animals10