Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/40/20230650

Performance of convolutional neural networks on CIFAR and Dog Breed classification dataset

Yang Bu[1,4], Weitong Xiong[2,5] and Lingfei Zhu[3,6]

[1]College of Network and Communication Engineering, Jinling Institute of Science and Technology, Nanjing, 211199, China
[2]Leicester International Institute, Dalian University of Technology, Dalian, 116081, China
[3]College of Chemistry, Chemical Engineering and Materials Science, Soochow University, Suzhou, 215000, China

[4]bu1486702069@gmail.com
[5]1787643613@mail.dlut.edu.cn
[6]17312779111@163.com

Abstract. This research aims to study the deep learning applications in image recognition and classification tasks. The advantages and limitations of deep learning are explored by analysing existing deep learning algorithms and their applications on tasks such as using images to do classification, image segmentation, and target detection. In the experimental part, this paper evaluates the performance of deep learning in image recognition by using classical neural network models, including training and testing models on a large size of image datasets. The results show that neural network has good classification and detection capabilities in image recognition tasks and also achieves good results for image segmentation tasks. However, the model training process of shallow neural network models is time-consuming and performs poorly for small-scale datasets. Based on these situations, this paper proposes some optimization strategies to achieve high performance and efficiency of neural networks in image recognition and analyze the performance differences between different strategies .

Keywords: Classification, Convolutional neural networks, Computer vision.

1. Introduction

With the rapid development of image technology and the widespread promotion of applications, image recognition plays a vital role in computer vision. Image recognition refers to the process of processing input digital images by computer, identifying and inferring image content. CNN, short for Convolutional Neural Network [1], is a powerful image recognition method, that has achieved great success in image recognition in recent years. Since Alex Krizhevsky [1] and others applied deep CNN to achieve extraordinary results in the ImageNet image recognition competition in 2012, CNN has become the mainstream model in the field of image recognition.

DOI: 10.54254/2755-2721/40/20230650

2. Related Work

The research results revealed the excellent performance of CNN on large-scale image datasets and also attracted extensive attention and research. In addition to the ImageNet competition, more and more fields and applications have begun to use CNN for image recognition, such as medical image diagnosis, autonomous driving, and face recognition. These studies show that CNN is robust and scalable for processing complex images, such as high-resolution images and large-scale datasets [2]. In addition to the above deep learning-based neural networks, they are also widely used in other fields: Object Detection and Localization [3], Multimodal Image Recognition [4], weakly supervised learning [5], Lightweight CNN [6], these works showed a good demonstration of the ability of neural networks to handle complex tasks. To further improve the accuracy and training efficiency of convolutional neural networks in image recognition, many researchers have also introduced methods such as attention mechanisms [7], multi-scale feature fusion [8], and transfer learning [9]. These techniques can enhance the robustness and generalization of the neural network model and have achieved outstanding results in many specific tasks.

3. Method

Compared to Artificial Neural Networks (ANNs), CNNs use kernels to extract relevant features from the input data through convolutional operations. The two core parts of a CNN are Convolution and Pooling [10]. Convolution mainly serves the purpose of extracting features and reducing dimensionality. As for Pooling, it plays a pivotal role in dimensionality reduction.

The various combination of Convolution and Pooling makes CNNs much more flexible and versatile. Thus, many famous networks were created [11]: AlexNet, VGGNet, Google InceptionNet, and ResNet.

3.1. Residual Block

The data go through some weight layers and activated functions and get the first result. The original input data is skipped over multiple convolution layers and conducted directly to the later layers, called shortcut or skip connection, and get the second result. The difference between the first and second results is what people call the residual [12].

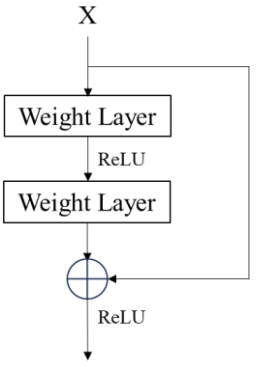

Figure 1. Image of Residual Block.

Compared with ordinary CNN, the most significant difference of ResNet is that it has many bypasses to build a connection of inputs directly to the layers behind the network. Therefore, the layers behind the network can also learn the residuals directly. This kind of network structure is called a shortcut or skip connection.

3.2. ResNet Model Architectures

Five different ResNet models are used in this study, namely ResNet18 [12], ResNet32 [12], ResNet50 [12], ResNet101 [12], and ResNet152 [12]. These deep residual learning models based on skip

DOI: 10.54254/2755-2721/40/20230650

connections aim to solve the problem of gradient disappearance and model degradation in deep networks. The structure diagrams of some models are shown in Figure 2 and Figure 3, respectively.

Different ResNet models have different layers and structures. For example, ResNet18 is a relatively shallow model with 18 layers, including eight basic blocks and one fully connected layer, while ResNet50 have 50 convolutional layers, which also includes multiple basic blocks and one fully connected layer.

The basic blocks in these models consist of multiple convolutional layers and skip connections. Introducing skip connections enables each basic block to learn a residual map, effectively alleviating the vanishing gradient problem. To further reduce the parameters that are calculated and decrease the complexity of the model structure, it also needs to introduce the batch normalization (Batch Normalization) technology and the residual learning strategy.

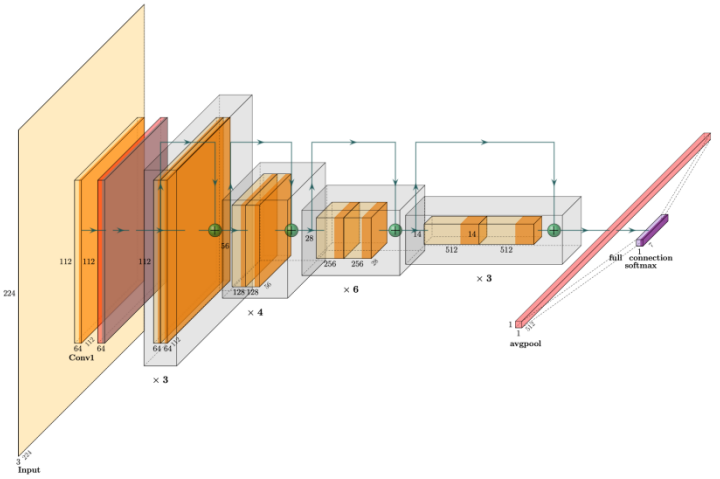

Figure 2. Image of ResNet34 Convolution layer structure.

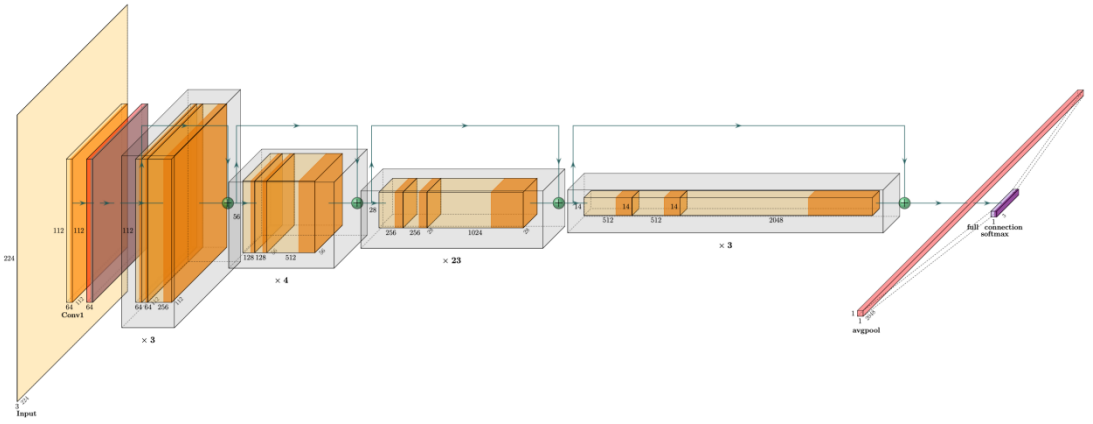

Figure 3. Image of ResNet34 Convolution layer structure.

a fine-tuning trick with the pre-trained model is used, there is a change in the layout of last full-connected layer. Since the tensor output of the avgpool layer of ResNet is a 1000-dimensional vector. There are also 1000 neural networks set up at the input layer. And finally the result needs to reduce the output dimension to 256 with a 120-neurons output layer, because the classification task ends up with 120 kinds of dogs. For the output layer, the researchers try to use the SoftMax function to turn the result into a more understandable probability distribution.

DOI: 10.54254/2755-2721/40/20230650

4. Experiments

4.1. Datasets Description

In this research, the datasets used are CIFAR 10 [1], CIFAR 100 [1], and a subset of ImageNet called Dog Breed.

There are 60000 color images in CIFAR 10 and CIFAR 100 datasets, whose size is 32 by 32. Among these, 50000 are for training, and 10000 are for testing.

Another dataset to be used is a subset of ImageNet called Dog Breed, which is more complex than the CIFAR 10 dataset, with 120 classes and an average image resolution of 256. There are 10,222 pictures for training and 10,357 pictures for testing. The main aim is to develop a convolutional neural network capable of accurately classifying 120 distinct breeds of dogs. To ensure the dataset was consistent, in the training set, all images needed to be cropped to 224 sizes, with a limited aspect ratio of random crops to either 3:4 or 4:3. Additionally, the probability of a random horizontal flip is 0.5, as well as randomly adjusting the image's brightness, contrast, and saturation to float within 40% of their original values. In the test set, the images were first resized to 256, and then 224 images were cropped using the center of the image as the initial point to ensure further accuracy. The purpose is to regularize the image size. In addition, in most pictures, the only thing that needs to be cared about is the dog in the picture's center, not the background information. Therefore, enlarging the image is in the hope that the network can learn more features of dogs.

4.2. Test to ResNet18 on CIFAR 10 and CIFAR 100

Only training ResNet18 on CIFAR 10 cannot be a good representation of the model's performance. Therefore, for CIFAR 10, to test the performance of CIFAR 10, comparing the ResNet18 model with a traditional CNN model is an effective method. The accuracy are shown in Figure 4.

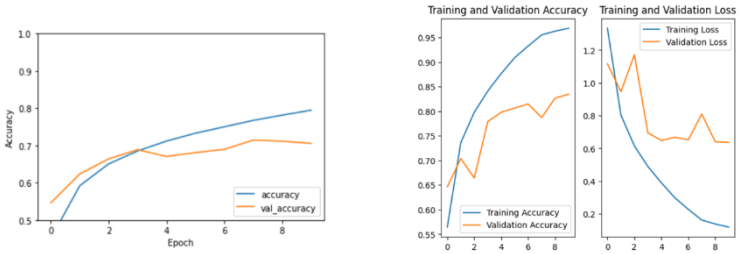

Figure 4. The left side is a traditional CNN and right side is ResNet.

The accuracy of the CNN is about 70-80%, while the accuracy of the ResNet on the training set is as high as 90% but is just over 80% on the test set. The latter is overfitting. After reviewing the data, it can be deduced that it might be because the images in CIFAR 10 are too small, the size is 32 by 32 pixels. However, the seven-by-seven convolutional kernel used in the first layer of the original ResNet18 is too large. It means that the kernel does not extract the features well. Therefore, there is a necessity to adjust the first convolutional kernel to a smaller size and change the step size and padding.

DOI: 10.54254/2755-2721/40/20230650

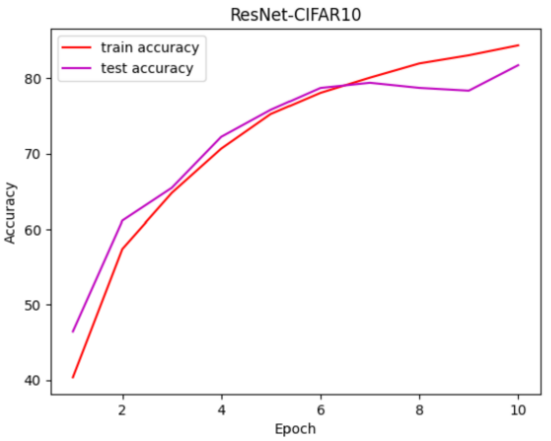

Figure 5. The Accuracy of ResNet (adjusted) on CIFAR10.

After adjusting, the result is shown in Figure 5, the accuracy is over 80% for two datasets.

The next step is testing the ResNet18 model on CIFAR 100. Similarly, comparing the result on CIFAR 100 with the result on CIFAR 10 can be more convincing. The end results are pretented in Figure 6 below.

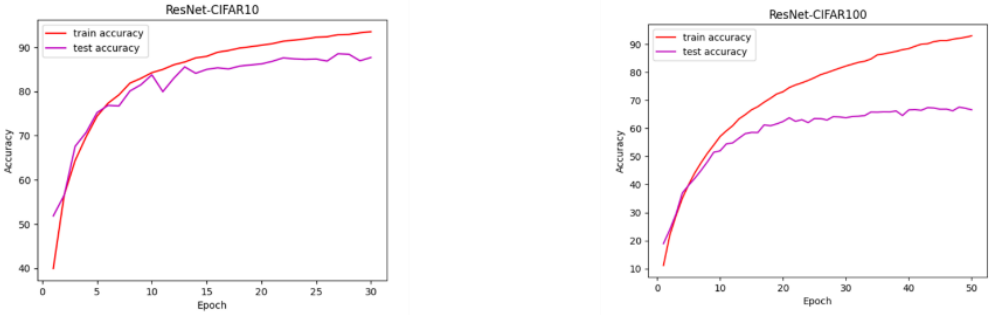

Figure 6. The Accuracy of ResNet18 on CIFAR10 (left) and CIFAR100 (right).

The accuracy is more than ninety percent. However, although the accuracy on CIFAR 100 is over ninety percent, the model does not perform well in the test set, the accuracy is just over sixty percent. Moreover, as the Figure shows, at 30 iterations, the model in CIFAR10 performs well, while that in CIFAR100 is overfitting.

4.3. Test to ResNet and MobileNet on Dog Breed

Due to the limited hardware computing power, adjusting the model that has been pre-trained is a better choice to train the dog category classification task in this research. The dataset, Dog Breed, include the training set, validation set and test set. Through training the model iteratively on the training set, the performance indicators on the test set are used to select the model and perform hyperparameter tuning. Using the SGD algorithm which is short from Stochastic Gradient Descent, the Adam algorithm (comparing algorithm performance with SGD), and the loss function of cross entropy to update the weights of the model. Among the SGD, using a version with momentum optimization to accelerate the convergence of gradient descent (GD).

To better observe the state of the dataset, Mobilenetv2 [2] is used as the backbone network for the first training to classify dog species. MobileNetv2 is a lightweight network. The number of parameters is 3.5m, which is 1/10 of ResNet50. Therefore, its training time is much less than ResNet, which means that people can get results in a short time. The Mobilenetv2 was directly used to train the dataset without any data enhancement, and the results are shown in Figure 7. Although the loss function in training set

DOI: 10.54254/2755-2721/40/20230650

gradually approaches 0, the loss value does not decrease very well in the test set, which is a typical overfitting phenomenon. However, the good news is that by adjusting different batch sizes, the latter is gradually approaching the loss value in the training set. The results about using ResNet50 without changing any hyperparameters are shown in Figure 8. The test set results are better, so they can be used more in the next training.

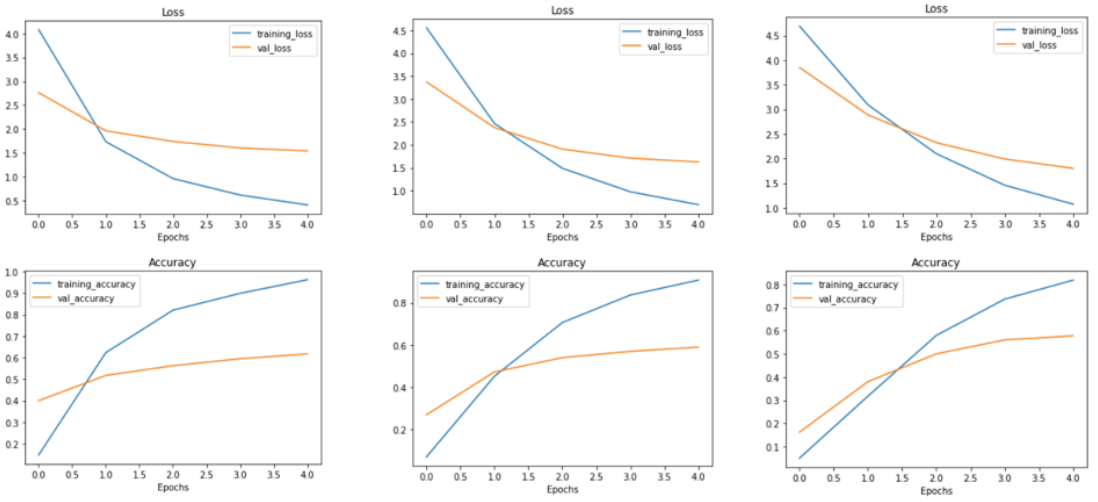

Figure 7. Mobilenet result. The training results and the image from left to right are the loss function values under different batch sizes, 32, 64, and 128, respectively.

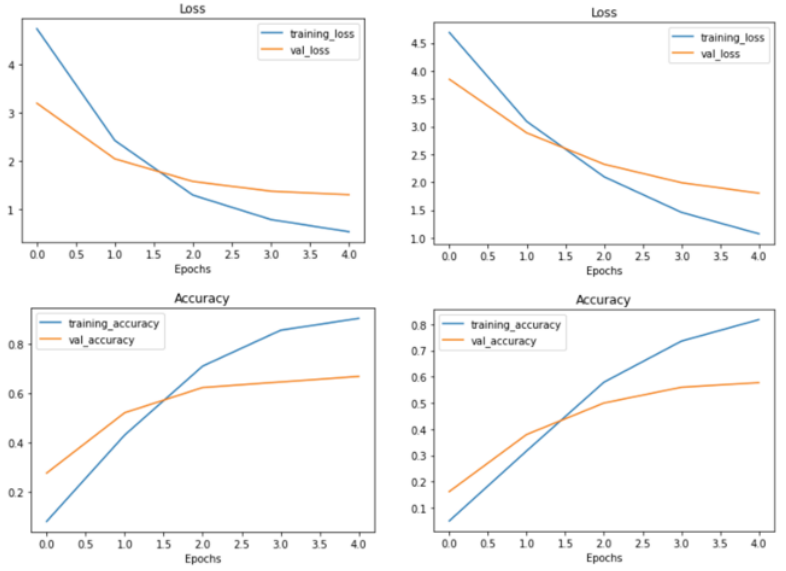

Figure 8. ResNet vs Mobilenet.The left side of the image is ResNet50, and the right is Mobilenetv2.

4.4. Performing a classification task using ResNet.

In this training, ResNet34 is used in the Pytorch library. At the same time, to reduce the training time, the researchers adjust full connection layer and set to a hidden layer and an input layer. In the first hidden layer, there are 1000 to 256 neurons set up, activation function ReLU, and the neurons in the output layer are 256 to 120. Freeze all convolutional layer parameters at the same time. In the first training, the hyper-parameter of batch size is 128 and used Mini-batch SGD with a momentum value of 0.9, the

DOI: 10.54254/2755-2721/40/20230650

learning rate of GD is 0.001, and its decay is 0.1 in each epoch. The performance of the loss function in mini-batch descent at different scales is tested. The result is shown in Figure 9.

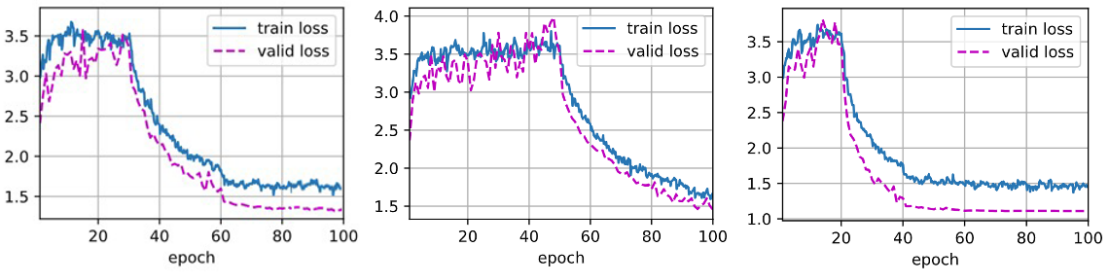

Figure 9. Gradient Descent Performance for Different Batch Sizes. The decrease rate of the loss function is the fastest when the batch size is 20.

For a deep learning problem, it is usual to define a loss function first. Once having the loss function, an optimization algorithm can be used in an attempt to minimize the loss. In the subsequent training, the algorithm is switched from the Mini-batch SGD algorithm to the Adam optimization algorithm. As shown in Figure 10, replacing the Adam optimization algorithm allows the learning rate to be better modified, making the loss function approach 0 faster.

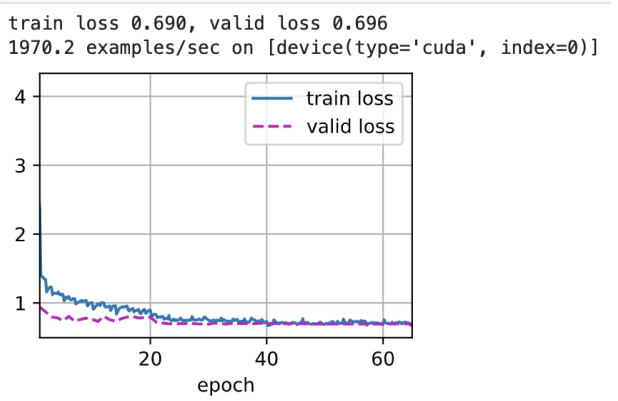

Figure 10. The Adam algorithm replaced the SGD algorithm without changing hyperparameters.

The convolutional neural network can learn the detailed features of the target through the convolutional layer. As the convolutional layer deepens, the network can learn more details. Many pre-trained models are provided in the Pytorch library, such as ResNet34, ResNet101, ResNet152, etc. The result is in Figure 11. And Figure 12. Show that the more convolutional layers, the smaller the loss value. However, Through the data graph, it is intuitive to find that both the test set loss and the training set loss have a good fit without overfitting and underfitting. In the data enhancement part, every 128 pictures are cropped as a batch, which effectively improves the quality of the dataset.

DOI: 10.54254/2755-2721/40/20230650

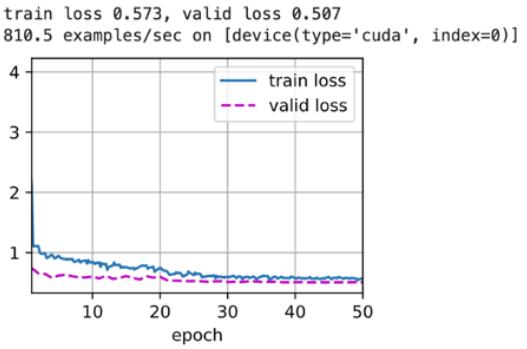

Figure 11. Results of the ResNet101 network.

Figure 12. Results of the ResNet152 network.

5. Conclusion

The results demonstrate the accuracy of object classification in complex datasets with convolutional neural networks of different depths. It is worth noting that in most cases, as the convolutional layer deepens, the final classification accuracy is higher. Still, in the case of average dataset quality, it does not achieve the expected performance. Through the analysis of the final results, the full-connected layer in the last can still be optimized.

To briefly explain the work, a pre-trained ResNet network is used for fine-tuning learning. In the last full connection layer, there is only one single hidden-layer and a output-layer with SoftMax algorithm is set for the probability prediction in the final result. Since ResNet itself is pre-trained on the ImageNet and has reached 95% accuracy, it deduces that the convolutional layer of the network itself has a fairly good fitting result, which means that the data has been well-learned related features of the set. In the end, the oversimplified fully connected layer leads to the loss of some results and the correct rate is slightly lower than expected, it will be deeply researched in future work.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

[2] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[4] He, D., Li, F., Zhao, Q., Long, X., Fu, Y., & Wen, S. (2018). Exploiting spatial-temporal modelling and multi-modal fusion for human action recognition. arXiv preprint arXiv:1806.10319.

[5] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2921-2929).

[6] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

DOI: 10.54254/2755-2721/40/20230650

[7] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.

[8] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

[9] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[10] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

[11] Wang, W., Yang, Y., Wang, X., Wang, W., & Li, J. (2019). Development of convolutional neural network and its application in image classification: a survey. Optical Engineering, 58(4), 040901-040901.

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).