A Study of Image Recognition for Standard Convolution and Depthwise Separable Convolution

Fan-Hsun Tseng(✉) and Fan-Yi Kao

Department of Technology Application and Human Resource Development,
National Taiwan Normal University, Taipei 10610, Taiwan
skittles2567@gmail.com

Abstract. Artificial intelligence and deep learning techniques are all around our life. Image recognition and natural language processing are the two major topics. Through using TensorFlow-GPU as backend in convolutional neural network (CNN) and deep learning network, image recognition has been an extreme breakthrough in recent years. However, more and more model parameters result in overfitting problem and computation overhead. In the paper, the performance of image recognition between standard CNN and depthwise separable CNN is experimented and investigated. In addition, data augmentation technique is applied to both standard and depthwise separable CNNs to improve the image recognition accuracy. The experiments are implemented by an open source API called Keras with using CIFAR-10 dataset. Experimental results showed that the depthwise separable CNN improves validation accuracy compared with the standard CNN. Moreover, schemes with data augmentation achieve higher validation accuracy but training accuracy.

Keywords: Deep learning · Convolutional neural network · Depthwise separable convolution · Data augmentation

1 Introduction

The success of Google's AlphaGo man-machine war in 2016 made artificial intelligence (AI), which had been developed for more than six decades, once again attracted the attention of scientists with unprecedented acknowledgments. Machine Learning is a branch of AI that analyzes large amounts of data through algorithms, allowing machines to train and learn the rules on their own and generate models afterward. These rules can be applied to new data and make predictions. However, due to the development difficulties of the hardware environment at that time, the computing speed of the computer had not yet been improved. The storage space was small, and large amount of data was not readily available, the development of AI was at a bottleneck.

In 1986, scholars Rumelhar and Hinton proposed the Back Propagation algorithm [1]. It uses gradient descent optimization to update the gradient to solve the problem of

Y. Chen et al. (Eds.): BICT 2020, LNICST 329, pp. 189–198, 2020.
https://doi.org/10.1007/978-3-030-57115-3_16

complex computations in neural networks. Subsequently, Yann LeCun *et al.* presented LeNet [2] in 1998, which was the first model architecture uses Convolutional Neural Network (CNN). In 2006, Hinton proposed Deep Belief Network (DBN) [3], which successfully trained multi-layer perceptron. The network was renamed Deep Learning (DL) to officially open the field of deep learning. In 2012, two students under the supervision of Hinton used graphics processing unit (GPU) and coupled with a deep learning model of CNN, and then proposed an architecture called AlexNet [4] to win the ImageNet competition.

Deep learning is a sub-area of machine learning. Deep learning simulates the operation of human neural networks, and finds effective methods from data. Currently, common deep learning architectures include Multilayer perceptron (MLP), Deep Neural Network (DNN), Convolutional Neural Network, Recurrent Neural Network (RNN), etc. Deep learning accomplishes a great number of results and applications in the fields of speech recognition, visual recognition, natural language processing, biomedicine, autonomous driving, and so on.

This paper focuses on the difference between depthwise separable convolution and standard convolution in deep learning and the effect of using data augmentation. The rest of this paper is arranged as follows. Section 2 introduces the development background and related knowledge of convolutional neural networks. Section 3 describes the research method and model architecture of this experiment. Section 4 is the experimental results and analysis and Sect. 5 summarizes the paper and describes the future work and research direction.

2 Related Works

2.1 Convolutional Neural Network

Convolutional neural networks are the most basic models in deep learning. Convolutional neural networks use Feature Engineering to extract useful image features so it has a significant effect on image recognition. Convolutional neural networks are structured by Convolution Layer, Pooling Layer, and Fully Connected Layer. The convolution layer performs a convolution operation on the original image and a specific filter to extract a feature map. The backpropagation algorithm allows the neural network to continuously correct the value of the filter during training to reduce prediction errors [5].

The pooling layer is connected to the convolution layer, and then down sampling is performed on the picture. It is a method to reduce the amount of data in the picture and retain important features. It also reduces the pixels and computing resources that the neural network needs to process and speeds up the training of the model. There are two most common methods of pooling, Max Pooling [6] and Average Pooling [7].

Finally, the two-dimensional matrix after convolution and pooling operations is flattened into a one-dimensional vector, and then connected to the most basic neural network, called fully connected layer. It consists of three layers: flat layer, hidden layer, and output layer. Neurons in the fully connected layer are used as classifiers to represent the probability of classifying each class.

2.2 Depthwise Separable Convolution

In 2012, Alex Krizhevsky *et al.* presented AlexNet [4] convolutional neural network model, and in that year's massive visual recognition contest (ImageNet Large Scale Visual Recognition Challenge 2012, ILSVRC2012) won the championship. AlexNet used Rectified Linear Unit (ReLU) for the first time as an activation function and local response in convolutional neural networks. He also used technologies such as Local Response Normalization (LRN) and Dropout. Future convolutional neural network models have developed their own neural networks with reference to the architecture of AlexNet. AlexNet is an important indicator of neural network models. Then, models such as Xception, VGG16, VGG19 [8], ResNet50, Google Net, MobileNets [9] were successively proposed, which improved the training accuracy dramatically.

In MobileNets [9], the construction concept of depthwise separable convolution is mentioned in detail. Compared with the general standard convolutional neural network, a depthwise separable convolution that can disassemble the convolutional layer operations is proposed to construct a lightweight deep neural network. Depthwise separable convolution divides the convolution into two parts for operation, Depthwise Convolution and Pointwise Convolution. Deep convolution is to create a $k \times k$ convolution core for each channel of the input data. Then each channel performs convolution for the corresponding convolution kernel independently; point by point convolution is to perform 1×1 convolution core for each completed channel; 1×1 convolution core helps decomposition channel feature learning and spatial feature learning. The premise is that each channel is highly correlated with spatial information, but different channels are not highly correlated with each other.

Depthwise separable convolution reduces the parameters and calculation costs used in convolution operations, resulting in a better model than the previous one for a given number of parameters [10]. Compared with the standard convolution parameter, the footprint is very small, and the kernel does not change over time, showing competitive performance [11].

2.3 Keras

Using python syntax, Keras is an open-source, advanced deep learning library that uses minimal code and takes the least amount of time to build deep learning models. Keras is a Model-level library that provides advanced modules needed by developers, but Keras only handles model creation, training, forecasting, etc. As a result, it cannot handle underlying operations such as tensor and matrix operations and differential. It relies on a specially crafted and optimized tensor library as a backend engine for Keras' backend engine for underlying operations such as tensor computing.

3 Proposed Analysis Approach

3.1 System Architecture

This paper is based on standard CNN and depthwise separable convolution to build training models. The hardware environment uses GPUs to speed up training. Regarding to the standard CNN, the architecture and operations are introduced as follows. The image size of the input data set in standard CNN is 32 × 32, and performs one convolution operation. The filter is set to 3 × 3 to generate 32 sets of filters, followed by a maximum pooling operation. The filter is set to 2 × 2, and the size after image down sampling is 16 × 16, and there are 32 groups of 16 × 16 images. Then, one convolution operation is performed again. The filter is set to 3 × 3, the original 32 sets of filters are converted to 64 sets of filters, and the second maximum pooling operation is performed. The filter is set to 2 × 2, and the image is reduced. After sampling, the size is 8 × 8 to generate 64 groups of 8 × 8 images. Finally, the third convolution operation is performed. The filter is set to 3 × 3, the original 64 sets of filters are converted to 128 sets of filters, and the third maximum pooling operation is performed. After down sampling, the size is 4 × 4, resulting in 128 sets of 4 × 4 images. The overall architecture is shown in Table 1.

Table 1. The architecture of standard convolutional neural network.

Type/stride	Filter shape	Input size
Conv/s1	3 × 3 × 3 × 32	32 × 32 × 3
Max Pool/s1	Pool 2 × 2	32 × 32 × 32
Conv/s1	3 × 3 × 64	16 × 16 × 32
Max Pool/s1	Pool 2 × 2	16 × 16 × 64
Conv/s1	3 × 3 × 128	8 × 8 × 64
Max Pool/s1	Pool 2 × 2	8 × 8 × 128
FC/s1	–	4 × 4 × 128
FC/s1	1024 × 10	1 × 1 × 1024
Softmax/s1	Classifier	1 × 1 × 10

Each convolution layer uses ReLU as the activation function, followed by a fully connected layer behind the convolution neural network. The first layer is a flat layer. A hidden layer and a total of 1024 neurons are set, and a Dropout layer is added to randomly discard 50% of the neurons. Finally, an output layer is established to output a total of 10 neurons, corresponding to 10 categories in the data set. After that, the Softmax activation function is used for conversion, which can convert the output of the neuron into the probability of predicting each image category.

The depthwise separable convolution in this work is based on the architecture of MobileNet. The depthwise separable convolution's architecture consists of depth convolution, batch normalization, ReLU activation function, and 1 × 1 point by point convolution. It is also connected to batch normalization and ReLU activation function. The overall architecture of depthwise separable convolution in this work is captured in Table 2.

The first layer is a standard convolution layer, the input image size is 32 × 32, the filter is set to 3 × 3, and 32 sets of filters are generated. The batch normalization and ReLU are connected as activation functions. Next, a layer of 64 groups of filters

Table 2. The architecture of depthwise separable convolution.

Type/stride		Filter shape	Input size
Conv/s1		3 × 3 × 3 × 32	32 × 32 × 3
Conv dw/s1		3 × 3 × 32 dw	32 × 32 × 32
Conv/s1		1 × 1 × 32 × 64	32 × 32 × 32
Conv dw/s2		3 × 3 × 64 dw	32 × 32 × 64
Conv/s1		1 × 1 × 64 × 128	16 × 16 × 64
Conv dw/s1		3 × 3 × 128 dw	16 × 16 × 128
Conv/s1		1 × 1 × 128 × 128	16 × 16 × 128
Conv dw/s2		3 × 3 × 128 dw	16 × 16 × 128
Conv/s1		1 × 1 × 128 × 256	8 × 8 × 128
Conv dw/s1		3 × 3 × 256 dw	8 × 8 × 256
Conv/s1		1 × 1 × 256 × 256	8 × 8 × 256
Conv dw/s2		3 × 3 × 256 dw	8 × 8 × 256
Conv/s1		1 × 1 × 256 × 512	4 × 4 × 256
5×	Conv dw/s1	3 × 3 × 512 dw	4 × 4 × 512
	Conv/s1	1 × 1 × 512 × 512	4 × 4 × 512
Conv dw/s1		3 × 3 × 512 dw	4 × 4 × 512
Conv/s1		1 × 1 × 512 × 1024	4 × 4 × 512
Avg Pool/s1		Pool 4 × 4	4 × 4 × 1024
FC/s1		1024 × 10	1 × 1 × 1024
Softmax/s1		Classifier	1 × 1 × 10

of depthwise separable convolution with a step size of 1 and a layer of 128 groups of filters of depthwise separable convolution with a step size of 2 are connected to generate 128 groups of 16 × 16 images. A layer of 128 groups of filters of depthwise separable convolution with a step size of 1 and a layer of 256 groups of filters of depthwise separable convolution with a step size of 2 are connected to generate 256 groups of 8 × 8 images. A layer of 256 groups of filters of depthwise separable convolution with a step size of 1 and a layer of 512 groups of filters of depthwise separable convolution with a step size of 2 are connected to generate 512 groups of 4 × 4 images. Five layers of 512 groups of filters of depthwise separable convolution with a step size of 1 and a layer of 1024 groups of filters of depthwise separable convolution with a step size of 1 are connected. The global average pooling layer is connected to solve the problem of too many parameters in the fully connected layer. Finally, an output layer is established to output a total of 10 neurons, and the Softmax activation function is also used for conversion.

This paper explores the impact of comparing standard CNN and deep separable architectures on accuracy in training and validation sets before model training. In addition, the paper explores the impact of applying data augmentation techniques to both architectures on training and validation accuracy.

3.2 Data Augmentation

Deep learning has certain requirements for the number of data sets. If the original number of data sets is small, it is not possible to effectively train neural network models, which can affect the model's performance. Therefore, while encountering such a situation, you can use the data augmentation techniques. Data augmentation is the processing of the data set of the original image to expand the number of data sets, generate more data for machine learning, and improve the problem of insufficient raw data.

Data augmentation technology produces a new image by rotating, resizing, scaling, or changing the brightness color temperature, flipping, and so on. Although for humans, the two images can still be identified as the same image, but for the machine, the new image is a brand-new image, so the image recognition model can improve the performance to a certain degree. In addition, data augmentation technology can also prevent over-training of the model, improve the recognition accuracy of the model, and obtain a network with stronger generalization ability. Due to privacy issues in the medical industry, access to data is strictly protected. Therefore, data enhancement uses effective methods to improve the accuracy of classification tasks and [12] explores the effectiveness of different data enhancement techniques in image classification tasks.

4 Experimental Results

4.1 Experiment Setting

In this paper, the data sets used are shown in Table 3. This paper uses the CIFAR-10 dataset, which is composed of scholars Alex Krizhevsky, Vinod Nair and Geoffrey Hinton collected a collection of 60,000 images in size 32 × 32. Among them, there are 60,000 images with a size of 32 × 32, of which 50,000 are used as the training sets and 10,000 are used as the test sets. From the 50,000 training sets, another 10,000 are used as the verification sets. There are ten categories in total including airplane (0), automobile (1), bird (2), cat (3), deer (4), dog (5), frog (6), horse (7), ship (8), and truck (9). There are 6,000 images in each category, for a total of 60,000 images.

The software and hardware system information used in this experiment is shown in Table 4. The operating system is Windows 10 64-bit version. The main deep learning tool is the establishment and training of code processing models developed by Keras. The programming language environment is Python version 3.6, TensorFlow accesses the GPU through CUDA and cuDNN provided by NVIDIA, and Keras is a high-level API of TensorFlow, so it must access GPU through TensorFlow.

Table 3. Data sets for experiments.

Data set	CIFAR-10
Size	32 × 32
Category	10 categories
Training set	40,000 pics
Validation set	10,000 pics
Testing set	10,000 pics

Table 4. System architecture.

Operating system	The Windows 10 64-bit version
Graphics card	NVIDIA TITAN RTX
Development environment	TensorFlow Keras
Programming language	Python 3.6
CUDA	Version 9.0
cuDNN	Version 7.6

4.2 Experiment Results

Batch count size for each architecture is set to 64 and Epoch number is set to 100. Conv represents standard convolutional neural network, DepthSepConv represents depthwise separable convolution, DA-Conv represents a standard convolutional neural network that adds data enhancement techniques, and DA-DepthSepConv represents a depthwise separable convolution that adds data enhancement techniques. Data enhancement technology is mainly targeted at training sets, as shown in Fig. 1.

The experimental results of the training concentration accuracy show that the depthwise separable convolution architecture can obtain a higher accuracy than the standard convolution neural network. The accuracy of the depthwise separable convolution is 30, already very close to 1. Therefore, this experiment further explored the impact of adding two architectures to data enhancement techniques on model accuracy. It was found that if data enhancement techniques were included, the recognition accuracy rate would be reduced in both architectures.

From the experimental results to verify the centrality accuracy of the concentration we found that the depthwise separable convolution and standard convolution neural networks are lower than that of the training set. However, the depthwise separable convolution architecture is still lower than that of the standard convolution neural network. Higher accuracy can be obtained, as shown in Fig. 2. In addition, with the addition of data enhancement technology, the accuracy of recognition results for both architectures have improved.

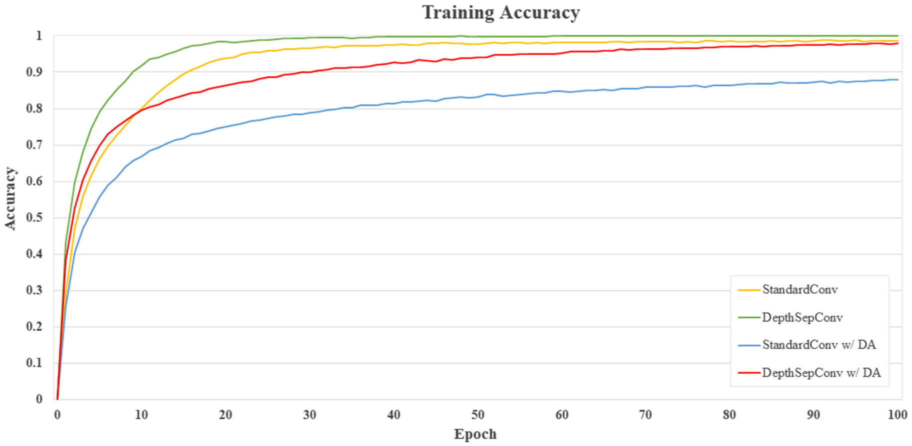

Fig. 1. Training accuracy of standard CNN and depthwise separable CNN with and without data augmentation.

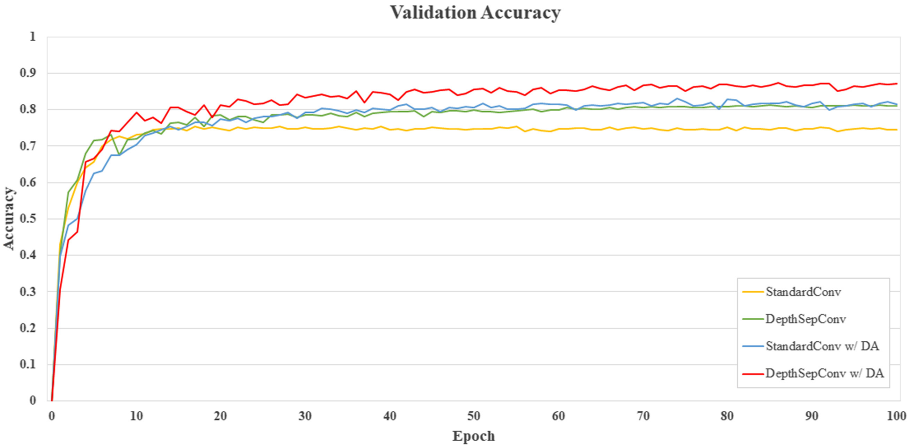

Fig. 2. Validation accuracy of standard CNN and depthwise separable CNN with and without data augmentation.

From the experimental results in Fig. 1 and Fig. 2, it can be observed that the depthwise separable convolution architecture achieves higher accuracy than standard convolution neural networks. Since depthwise separable convolution is a lightweight model, the main characteristics are to reduce the number of parameters and reduce the amount of memory per parameter. Although the model is lightweight, the depthwise separable convolution architecture is composed of depth convolution and point by point convolution which leads to an increase in the number of layers. However, although the number of layers increased and the accuracy increased, it did not increase the training time significantly. In our experimental architecture, the number of model parameters for standard convolution and depthwise separable convolution was 2,201,674 and 1,094,538, respectively. Even though there are many layers of depthwise separable convolution, the number of parameters is still lower than the standard convolution. Therefore, the memory usage can be reduced during the model training process.

In addition, take further advantage of data enhancement techniques, even if the accuracy rate is reduced on the training set, it can be found on the validation set that no data enhancement will result in a difference between the accuracy rate and the recognition accuracy of the training set. Although the addition of data enhancement techniques may lead to a decrease in the accuracy of the training set, it can improve the accuracy of the validation set and improve the generalization of the model.

5 Conclusions and Future Works

In this paper, the different architecture and performance between the standard convolutional and the depthwise separable convolution neural network are compared and analyzed. In addition, the data augmentation technique is applied to both standard and depthwise separable convolution to prevent overfitting problem and to enhance the generalization ability of the model. The experimental results showed that although the depthwise separable convolution architecture can reduce the number of model parameters, it will increase the number of layers in neural network, thereby it will not speed up model's training time. In general, computing capability is limited by hardware specification, the architecture of depthwise separable convolution is able to reduce memory usage and maintain recognition accuracy. In the future, we expect to optimize the architecture of depthwise separable convolution such as type and stride, filter shape, and input size for pursuing the higher recognition accuracy.

Acknowledgements. This work was financially supported from the Young Scholar Fellowship Program by Ministry of Science and Technology (MOST) in Taiwan, under Grant MOST109-2636-E-003-001 and MOST108-2636-E-003-001. The authors would like to thank Crimson Interactive Pvt. Ltd. (Ulatus) for their assistance in language polishing.

References

1. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**, 533–536 (1986)
2. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
3. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
5. Bouvrie, J.: Notes on convolutional neural networks (2006), http://cogprints.org/5869/1/cnn_tutorial.pdf. Accessed 25 Dec 2019
6. Nagi, J., et al.: Max-pooling convolutional neural networks for vision-based hand gesture recognition. In: Proc. 2011 IEEE International Conference Signal Image Processing Applications (ICSIPA), pp. 342–347, IEEE, Kuala Lumpur (2011)

7. Liu, L., Shen, C., Hengel, A.: The treasure beneath convolutional layers: cross-convolutional-layer pooling for image classification. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition (CVPR), pp. 4749–4757. IEEE, Boston (2015)
8. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 6th International Conference Learning Representations (ICLR), San Diego, California, pp. 1–14 (2015)
9. Howard, A.G., et al.: MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861v1, pp. 1–9, arXiv (2017)
10. Kaiser, L., Gomez, A.N., Chollet, F.: Depthwise separable convolutions for neural machine translation. arXiv:1706.03059, pp. 1–10, arXiv (2017)
11. Wu, F., Fan, A., Baevski, A., Dauphin, Y.N., Auli, M.: Pay less attention with lightweight and dynamic convolutions. arXiv:1901.10430v2, pp. 1–14, arXiv (2019)
12. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Sig. Process. Lett. **24**(3), 279–283 (2017)