# The study of Advantages and Applications of Convolutional Neural Networks in Computer Vision Tasks

Zhonghao Xie

Aberdeen School of Data Science and Artificial Intelligence, South China Normal University, Foshan, 528200, China

z.xie1.22@abdn.ac.uk

**Abstract.** There are some deficiencies in the current Convolutional Neural Network (CNN) system. Some deep and complex CNN models take a long time to train, requiring a lot of computing resources and time for the training. At the same time, some CNN models may have overfitting problems, resulting in a decline in generalization ability. This paper provides an application case of CNN in computer vision tasks to help some people understand the actual application field of CNN. This paper takes Alexnet as an example, compares it with the Lenet-5 algorithm, and discusses the advantages of deep complex CNN models, especially in terms of transfer learning, which is very helpful for specific tasks in practical applications. In experimental results, through network architecture search and automated design methods, this paper finds a more suitable CNN architecture to improve model performance and generalization capabilities and explores how to improve the adaptability of the algorithm through structural updates and hyperparameter adjustments.

**Keywords**: Convolutional Neural Network, Image classification, Lenet-5, Alexnet

## 1 Introduction

Machine learning methods attempt to identify hidden patterns in a large amount of data. In several fields, it has been demonstrated that deep learning approaches surpass


P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_104

earlier learning methods. Undoubtedly, one of the most notable examples is computer vision [1, 2].

In the realm of computer vision, convolutional neural networks (CNN) are the most used deep learning technique [3]. From traditional machine learning methods for image classification to convolutional neural networks that use deep neural networks to identify images, CNN is becoming more diverse and its depth is constantly increasing. The emergence of datasets [4] coincides with the appearance of LeNet-5 Convolutional Neural Network (LENET-5) [5], Alexnet Convolutional Neural Network (Alexnet)[6], Visual Geometry Group(VGG)[7], Inception-v1 GoogLeNet(Googlenet) [8], Residual Neural Network(ResNet) [9], and Extreme Inception(Xception)[10]. However, the above algorithms lack horizontal comparison of dataset results. Just like what Zhang Xinche wrote in [11]: The goal of this study was not to improve accuracy but to use the CIFAR-10 database to apply Alexnet, LeNet-5, and VGG Net as well as to describe the characteristics and performance of these structures. Instead, this research analyzes how and why certain CNN structures can be better adapted to the database CIFAR-10 [12].

In image recognition, detecting image identification [13] is a well-known machine learning issue. Object or image recognition from digital photos or images to videos is an extremely difficult issue. The discipline of computer vision uses picture recognition in many different applications. Facial recognition, biometric systems, self-driving cars, emotion analysis, picture restoration, robotics, and more are a few examples. Some focus on new image feature extractors which can more effectively combine visual components into higher-level entities, which aids semantic inference. Some concentrate on Contrastive Predictive Coding, a non-supervised goal for learning these representations, and the features produced by this novel implementation support the ImageNet dataset's state-of-the-art linear classification accuracy.

In deep learning, some algorithms use deep learning to derive how to fight the COVID-19 epidemic and provide recommendations for ongoing COVID-19 research. There are some scholars who studied the principles of deep learning and machine learning to more fully grasp the technical basis of contemporary intelligent systems. They conceptually distinguished related terms and concepts and discussed the difficulties in putting such intelligent systems into practice in the context of electronic marketplaces and online businesses. Other scholars developed ZeroCostDL4Mic, a basic platform that makes deep learning access easier by utilizing Google Colab's

free, cloud-based computing resources, which give each network appropriate quantitative methods to assess model performance and enable model improvement.

In image classification, some researchers proposed the SpectralFormer backbone network, which is new and solves the problem that CNNs are difficult to mine and accurately represent the sequential properties of spectral features due to the limitations of their innate network backbones. Other researchers focused on modifications to the image recognition training process, such as tweaks to data augmentation and optimization techniques. The impact of a set of such modifications on the accuracy of the final model then was empirically assessed using an ablation study. Combining these enhancements, researchers can greatly enhance multiple CNN models. The study enhanced the functionality of the object and semantic recognition techniques utilized in a variety of application fields.

This research provides specific instructions for modernizing the structure to work with the new database and modifying the hyperparameters to better align the algorithm with the database process. At the same time, taking Alexnet as an example, by changing the hyperparameters in the structure, can be better applied to the data set CIFAR-10.

This essay is structured as follows: Section 2 reviews the state of the art. Our methodology is described in Section 3. The outcomes of the experiments are displayed in Section 4. The conclusions are reported in Section 5.

## 2 Methods

First of all, the data need to be pre-processed. The initial data source is used to construct the training and test sets. Once the model has been trained, the appropriate hyper-parameters are selected, and then some metrics are used to evaluate the accuracy and accuracy of the output results to judge the effectiveness of the two functions in the CIFAR-10 dataset. The space of the output volume in the convolution layer is a function of the input volume, the rod of the nuclear domain convolutional layer neuron, and the zero-filling charging at the boundary.

### 2.1 Alexnet

In Alex, it contains 8 learning layers, including 3 full-connected layers and 5 convolutional layers, and uses Local Response Normalization (LRN) to achieve a

local response. LRN mimics neuromas, with a function called lateral restraint, which refers to active neurons' suppression of nearby neurons. The notion of borrowing side suppression to produce local suppression is normalized by LRN's local response, making the value of a relatively big response, which enhances the model's capacity for generalization. LRN does not alter the size or dimensions of the data; it just processes the adjacent areas. The feature map from the preceding layer is learned and twisted in the convolutional layer. In order to create output feature diagrams, a linear or nonlinear activation function, such as dual-curve cut, soft magnetic, rectifier linear, or identity function, is created by the kernel. Multiple input functions can be paired with each output function map. The algorithm is generally expressed as

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{1}$$

Here, $x_j^l$ the output of the current layer refers to the output of the layer preceding it, $k_{ij}^l$ signifies the current layer's kernel, and $b_j^l$ indicates the biases of the current layer. The given input map array is represented by $M_j$ . A cumulative bias b is applied to each output map. To create equal output maps, the input maps will be convolved using different kernels. The final models are then subjected to an activation function that is either nonlinear or linear. The collecting layering is carried out underneath the input diagram in the sub-sampling layer. This is usually called the merger layer. Down sampling can be written as

$$x_j^l = down(x_j^{l-1}) \tag{2}$$

where a sub-sampling function is represented by down(.). Then dropouts are used to prevent excessive neural networks.

### 2.2 LeNet-5

Convolutional layers, pooling layers, and complete connection layers make up LENET-5. It has a convolutional neural network structure commonly used in deep learning. After convolution, the output is symbolized by $a_{i,j}$ , and the position of

each item for every convolution is denoted by b, $F(\bullet)$ which expresses the triggering procedure. Then the convolution layer can be expressed as

$$y_{k,l} = \{f(\sum_{a=0}^{D-1}\sum_{b=0}^{D-1} m_{a,b} n_{k+a,l+b} + b)\}_{k=1,2.....,K;l=1,2.....L} \quad (3)$$

where the feature extraction method is mostly carried out using the convolutional layer. There are many convolutional kernels in each layer. A compressed input matrix is used, utilizing the convolution kernel at this layer. Assume the entry table equals $N = \{n_{k,l} \mid k = 1,2,...K, l = 1,2,...L\}$ where $K$ =32 and $L$ = 32. K is the number of sensors and L is the amount of data for gas received for each sensor. And $M = \{m_{k,l} \mid k = 0,1,...D-1, L = 0,1,...D-1\}$ is the notation for the convolution kernel, where F stands for the convolutional kernel's size, which is equal to its width or height. The activation function is depicted by $f(\bullet)$, the offset term for each convolution is displayed by b, and $y_{k,l}$ depicts the result of convolution.

The first pooling layer employs the largest 2x2 pooling procedure to condense the size of the characteristic diagram. The size of the feature diagram is once again reduced by the second pooling layer, which employs the greatest pooling operation of 2x2. The pooling layer's expression is

$$a_n^l = pool(a_n^{l-1}) \quad (4)$$

Where $pool(\bullet)$ denotes the largest operation of pooling. The result of the $lth$ layer is often shown as a $lth$. Additionally, the preceding layer's output is represented by the following: $a_n^l$ and $a_n^{l-1}$, where n corresponds to the nth sample. In the full connection layer, the LENET-5 training process usually uses the random gradient decreased (SGD) to optimize the parameters, and the network weight and bias are adjusted by minimizing the loss function.

Applying LENet-5 and Alexnet to CIFAR-100 needs to be appropriately adjusted and optimized to adapt to the characteristics and complexity of the CIFAR-100 data set. Through training and testing, they can evaluate their classification performance on CIFAR-100 and compare it with other models to select the best model architecture and parameter settings.

## 3 Results

The 60,000 images that make up the CIFAR-10 dataset are color images of 10 different 32x32 kinds, with 6000 images for each type. 10,000 of them are utilized as test sets, while 50,000 of them are used as training sets. Each of the five training batches and the one test batch in the CIFAR-10 dataset has 10,000 images. The test set batch's image is made up of 1,000 images that were chosen at random from each category. The remaining 50,000 images are included in the training collection batch in a random sequence. However, certain training sets could include more of a particular sort of image than other types of image.

When applying LENET-5 to the data set CIFAR-10, adjusting the input size of the network to 5 × 5, and it is necessary to increase the input channel count to three. Through this method. At the same time, the quantity of filters across the convolutional layer also needs to be adjusted, set the quantity of first convolutional layer filters to 6, and set the quantity of second spherical harmonic filters to 16. As well as the initial batch of neurons in the whole connection layer is set to 120, the second set is 84, and the third settings are 10 (10 categories corresponding to CIFAR-10 data sets). Only in this way, can the class in CIFAR-10 be matched in this way.

In order to apply Alexnet to the CIFAR-10 of the data set, adjusting the input size and modifying is needed. The first convolutional layer contains 64 filters, 3 input channels, 64 output channels, and convolution kernels. 11 × 11, the second convolutional layer is set to have 192 filters, 64 input channels, 192 output channels, convolution kernel 5 × 5, and convolution cores of 5 × 5. Convolution layer settings include setting the initial convolution layer to 6 and equipping each layer with the max-pool 2D (MaxPool2D) that adjusts to RGB pictures. The second layer is modified to contain 16 5X5 filters. In order to accommodate the CIFAR-10 figure, the last complete connection layer's number of neurons has been adjusted to 10. The retrieved characteristics are classified using the classifier pieces, which contain a sequence module with numerous complete connection layers and Dropout layers. The

following layers are included: A RELU activation function comes after the first complete connection layer (Linear), which includes 9216 input features and 4096 output features. A dropout method is then applied to avoid overfitting. The RELU activation functions and various complete connection layers are comparable. The final complete connection layer's output characteristics, which are 10 in number, are utilized to forecast 10 different CIFAR-10 data set categories. Table 1 shows the two algorithms' short structures.

**Table 1.** Brief structure of two algorithm

| Structure | Number of convolution layers | Full connection layer number | Placement method | Convolution core size |
|---|---|---|---|---|
| **Lenet-5** | 2 | 3 | Follow a pooling layer after each convolutional layer | 5×5 |
| **Alexnet** | 5 | 3 | Layer interval uses the pooling layer and the first layer of the layer | 11×11 and 5×5 |

There are 3072 input nodes when LENET-5 is applied to the CIFAR-10 dataset. A color picture of 32x32 pixels featuring three RGB color channels makes up the CIFAR-10 data set's image. Therefore, each image has 3072 pixels values. LENET-5's input layer accepts the image data after the flat and uses these 3072-pixel values as the input of the network. So on the CIFAR-10 dataset, LENT-5 has 3072 input nodes in total.

When Alexnet is applied to the CIFAR-10 dataset with the input feature graph scaled to 6x6, there are 6 x 6 x 256 = 9216 input nodes. Following Alexnet's average pooling layer (AVGPOOL), the feature diagram was resized to a 6x6 size with 256 channels. The whole connection layer's (Classifier's) input was the updated feature

diagram, which was shown as a vector. The length of the vector after the flattening is 6x6x256 = 9216 since the feature diagram is 6x6 in size and has 256 channels. This is the number of input nodes of Alexnet on the CIFAR-10 dataset. Each neuron corresponds to a category of the CIFAR-10 dataset. The value of each output node depicts the likelihood or score for the relevant category of the network, so the output nodes of the two algorithms are 10. For LeNet-5, the original design is for processing 32x32 pixel images, so when applied to the CIFAR-10 dataset, the image needs to be adjusted. The image size of the CIFAR-10 dataset is 32x32, which matches the original design of LeNet-5. Consequently, it is unnecessary to modify The total amount of neurons in the layer with 100% connectivity, which is still 120 and 84. For Alexnet, the original design was to process images of 224x224 pixels, which did not match the image dimensions of the CIFAR-10 dataset. In general, for the application of CIFAR-10, appropriate modifications can be made to accommodate smaller image sizes. Reducing the number of neurons within the completely interconnected layer of Alexnet is a frequent adjustment. The hyperparameter's value is indicated in Table 2.

**Table 2.** The value of the hyperparameter

| Hyperparameters | Lenet-5 | Alexnet |
|---|---|---|
| **Input image size** | 32x32 | 224x224 |
| **Activation function** | Sigmoid | ReLU |
| **Number of neurons in fully connected layer 1** | 120 | 256 |
| **Number of neurons in the fully connected layer 2** | 84 | 256 |
| **Input layer** | 3072 | 9216 |
| **Output layer** | 10 | 10 |

The results of the two models are shown in Table 3.

**Table 3.** The output of the two models

| Model | Validation set accuracy | Training time |
|---|---|---|
| **Lenet-5** | 38% | 1minute26seconds |

Table 3 indicates that Alexnet's accuracy is plainly considerably greater than Lenet-5's, but the time spent is also many times that of Lenet-5. According to the experimental findings, LENET-5 and Alexnet differ significantly in terms of training duration, output accuracy, and migration learning effects when the identical hardware configuration and testing environment are used. LENET-5 is a relatively simple CNN algorithm. It was originally applied to MNIST, but because the CIFAR-10 is more complicated, the accuracy is unsatisfactory and not as good as Alexnet. This result is because Alexnet has a deeper and more complicated convolutional neural network model. Alexnet may be able to better capture the characteristics in the image, thereby achieving higher accuracy. Lenet-5's validation and test accuracy on CIFAR-10 is shown in Figure 1, and Alexnet's validation and test accuracy is shown in Figure 2.

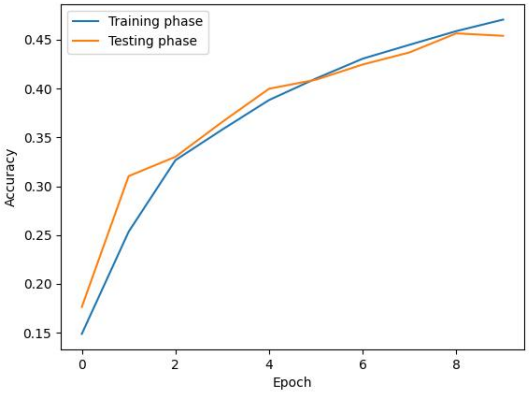


**Fig. 1.** Training and testing accuracy of Lenet-5 on CIFAR-10 (Photo/Picture credit:Original)

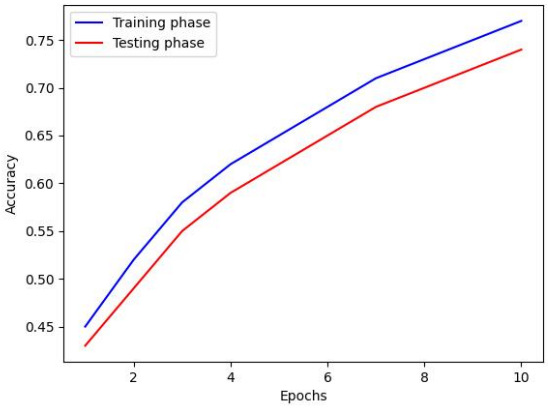


**Fig. 2.** Training and testing accuracy of Alexnet on CIFAR-10 (Photo/Picture credit: Original )

In Figure 1 and Figure 2, both the training set and testing set of Alexnet is more accurate than Lenet-5. The accuracy of Alexnet is often greater than the accuracy of Lenet-5, regardless of how many times the training is done simultaneously.

## 4 Conclusion

This research focuses mostly on the use of deep learning in computer vision, particularly in picture identification. The study compares and analyzes the performance of several convolutional neural network topologies on the CIFAR-10 database as well as investigates how to make the algorithm more adaptable by changing the structure and modifying hyperparameters. Experimental results show that different CNN structures exhibit different performances on the CIFAR-10 dataset.

## References

1. Batta, M.: Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR), Volume 9, pp:381-386(2020).

2. Diego, A.: Deep Learning for Computer Vision: A Brief Review. Hindawi Computational Intelligence and Neuroscience Volume 2018,13 pages(2018).
3. Li, Z. , et al.: A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions on neural networks and learning systems* (2021).
4. Ring, M., et al.: A survey of network-based intrusion detection data sets. *Computers & Security* 86, 147-167(2019).
5. Wei, G., et al.: Development of a LeNet-5 gas identification CNN structure for electronic noses.*Sensors* 19.1, 217 (2019).
6. Alom, Md Z., et al.: The history began from Alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv, 1803.01164* (2018).
7. Sengupta, A., et al.: Going deeper in spiking neural networks: VGG and residual architectures. *Frontiers in Neuroscience* 13, 95(2019).
8. Anand, R., et al.: Face recognition and classification using GoogleNET architecture. *Soft Computing for Problem Solving: SocProS 2018, Volume 1* (2020).
9. Wightman, R., Hugo T. and Hervé J.: Resnet strikes back: An improved training procedure in time. *arXiv preprint arXiv, 2110.00476* (2021).
10. Rahimzadeh, M. and Abolfazl A.: A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2. *Informatics in medicine unlocked* 19, 100360(2020).
11. Zhang, X.: The AlexNet, LeNet-5, and VGG NET applied to CIFAR-10. In: 2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering, ICBASE(2021).
12. Smith, R J., Ryan A., and Malcolm I. Heywood.: Evolving simple solutions to the CIFAR-10 benchmark using tangled program graphs. *2021 IEEE Congress on Evolutionary Computation (CEC)*(2021).
13. Chauhan, R., Kamal Kumar G., and Joshi R. C.: Convolutional neural network (CNN) for image detection and recognition. In: *2018 first international conference on secure cyber computing and communication, ICSCCC*( 2018).