Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/30/20230081

Application of CNN in computer vision

Jiamiao Yan

School of Computer Science and Technology, Tiangong University, Tianjin, 300000, China

2011630313@tiangong.edu.cn

Abstract. Today's deep learning continues to be hot, and the application of machine learning can be seen in more and more fields. A neural network model called a Convolutional Neural Network (CNN) was created to imitate the structure of the human brain. It is a convolution operation that maps the relationship between input features and output features to a two-dimensional in the vector space of , the network can effectively process the input data. CNN emerged to solve the computational bottleneck problem faced by traditional networks. This paper discusses the application of the deep learning model CNN in image classification, target detection and face recognition. In these fields, models are continuously proposed, and architectures in each field are constantly emerging. Among them will be the classic architecture of CNN in this field. These classic architectures have their advantages, but there will also be improvements brought about by the shortcomings of the classic architecture. Through the application of these different fields, we can see that CNN-based deep learning can help various fields, and the efficiency will be improved, but it is not perfect and needs continuous improvement.

Keywords: model CNN, image classification, target detection, face recognition.

1. Introduction

In the past decade or so, artificial intelligence has become very popular. Many people are curious about artificial intelligence, but most people still think that voice assistants are artificial intelligence. Artificial intelligence is everywhere in our lives. In recent years, artificial intelligence has developed very well, and this paper discusses machine learning in artificial intelligence. With the rise of machine learning, many fields are using machine learning to assist in processing work to achieve more efficient and smarter work while reducing labor costs. For example, facial recognition payments, search engines and recommendation systems, social media content recommendations, spam filtering, identification of identity information, and audio and image processing all use deep learning. There are still many such applications in our lives. This paper discusses the application of the deep learning model Convolutional Neural Network (CNN) in image classification, target detection and face recognition. Among them, there will be the classic architecture of CNN in this field. Although they are all classic models, they have their own usage scenarios and advantages. However, as demand increases, there will still be deficiencies. Therefore, this paper will also have improvement measures brought about by the lack of classic architecture. In image classification, the classic CNN architectures include LeNet, AlexNet and Visual Geometry Group Network (VGGNet) [1-3]. After development, some shortcomings have been improved, and new models such as ShuffleNet and EfficientNetV2 have been proposed [4,5]. In target

DOI: 10.54254/2755-2721/30/20230081

detection, the classic CNN architectures include Region-based Convolutional Neural Network(R-CNN), Fast Region-based Convolutional Neural Network (Fast R-CNN) and Faster Region-based Convolutional Neural Network (Faster R-CNN), which are sequentially improved and the accuracy is gradually improved [6-8]. In face recognition, there are DeepFace, FaceNet and ArcFace are also models applied in face recognition [10-12]. In face recognition, the accuracy can also be improved through hardware.

2. Research methodology

In our comparative study on image classification, object detection, and face recognition, we first collected datasets containing various types of images and preprocessed them, including resizing and data augmentation. Then, we selected several classic CNN models such as LeNet, AlexNet, VGGNe, ShuffleNet, and EfficientNetV2 as the comparison objects [1-5]. We trained and evaluated these models on the same training and testing sets. For image classification, we compared the differences in accuracy, speed, and model size among these models. For object detection, we compared the performance of models like R-CNN, Fast R-CNN and Faster R-CNN in terms of detection accuracy, speed, and localization precision [6-8]. For face recognition, we compared the differences in accuracy, robustness, and computational complexity among models like DeepFace, FaceNet, and ArcFace [10-12]. We also used appropriate evaluation metrics and statistical methods such as accuracy, recall, and F1 score to quantify and analyze the comparison results. Through these comparative studies, we can determine the strengths and weaknesses of different models in image classification, object detection, and face recognition, providing guidance and reference for practical applications. In summary, our research methodology primarily involves collecting and preprocessing datasets, selecting suitable comparison models, conducting training and evaluation, and analyzing the results using evaluation metrics and statistical methods. Through these methods, we can comprehensively compare the performance differences of different models in image classification, object detection, and face recognition.

3. Application of CNN in image classification

3.1. Application principle

As a deep learning model, CNN's unique structure and functions make it especially suitable for processing image data. CNN can automatically learn the features of images through convolutional layers and pooling layers for feature extraction. Among them, the convolution layer can filter the image through convolution, to extract the local features of different positions. The pooling layer compresses the obtained feature maps by downsampling and retains the necessary features to achieve better image classification results. When performing classification tasks, by adding activation functions and normalization operations in the convolutional layer and the fully connected layer, CNN can learn nonlinear feature representations and classify them in the output layer. Data augmentation techniques are used to expand the training set size. It is also possible to use the CNN model trained in other fields to perform fast training after fine-tuning in another field.

3.2. Classical CNN architectures for image classification

The LeNet model introduces the concept of convolutional layer and pooling layer into the CNN model so that local features can be extracted, and downsampling can also be performed to better process image data. It laid the foundation for the subsequent CNN model, but it was limited by resources and performance at that time, and there were relatively few parameters [1].

AlexNet is a new CNN model proposed in 2012, which has a deeper network structure than previous CNN models. The increase in network depth allows AlexNet to learn richer and abstract feature representations to improve image classification accuracy. In addition to the deep structure, the nonlinear activation function ReLU, local response normalization, Dropout regularization, and parallel computing methods are introduced to significantly reduce the error rate of image classification and also lay the foundation for the field of computer vision [2].

DOI: 10.54254/2755-2721/30/20230081

The VGGNet model was proposed in 2014. The salient features of this model are the concise and regular network structure and the increased depth. VGGNet deepens the network structure by increasing the number of convolutional layers, which is conducive to extracting more levels of abstract features, thereby improving the performance of image classification. A small-size convolution kernel is beneficial to reduce the risk of overfitting. This model also uses max pooling layers and multiple fully connected layers.Although the VGGNet model has a larger computational cost compared to later and more complex models, its simple structure and improved performance have made it one of the classic CNN architectures [3].

3.3. New CNN models for image classification

The proposal of ShuffleNet solves the problem that some advanced basic architectures become inefficient due to the high computational complexity of dense 1 × 1 convolution in very small networks. ShuffleNet employs a unique channel rearrangement technique to facilitate information flow between feature channels and get around the drawbacks of group convolution, as well as point-by-point group convolution to lessen the computational cost of 1 × 1 convolution. This model can allow more feature map channels under a given computational complexity budget, which can help encode more information and optimize the performance of very small networks [4].

EfficientNetV2 is a new type of convolutional network. Compared to earlier models, this one has a higher parameter efficiency and a quicker training rate. Compared with EfficientNet, .Scaling and a mix of training-aware neural architecture search (NAS) and neural architecture search (NAS) are used to improve EfficientNetV2.This brings a 4 times improvement in training speed for the model, and the size of the parameters is also reduced by 6.8 times. During training, the training speed can be further accelerated by increasing the image size. By dynamically adjusting the regularization method, the training speed can be accelerated without reducing the accuracy [5].

The advantages, disadvantages and improvements of the image classification models are listed in table 1.

Table 1. Analysis of the advantages and disadvantages of CNN in image classification.

Image Classification	LeNet	AlexNet	VGGNet	ShuffleNet	EfficientNetV2
Advantages	The introduction of convolutional layers and pooling layers lays the foundation for the subsequent CNN model	Deeper network structure, using parallel computing to reduce error rate	There is a concise and regular network structure and an increase in depth	Fixed some advanced basic architectures being less efficient in very small networks due to the high computational complexity of dense 1 × 1 convolutions	Adding an extra operation to the search space and scaling for optimization combined with training-aware Neural Architecture Search (NAS) improves parameter efficiency and speeds up training

DOI: 10.54254/2755-2721/30/20230081

Table 1. (continued).

Disadvantages	relatively few parameters	high demand for computing resources, large restrictions on input size, large amount of parameters	large amount of calculation	A certain amount of computational overhead will be introduced, and the local perception ability will be reduced to a certain extent.	More complex ,increased complexity of the training process, more memory resources to store network parameters and intermediate feature maps

4. Application of CNN in object detection

4.1. Application principle

CNN's hierarchical feature extraction can extract multi-level and multi-scale features from the original image. These features capture some key information about the target and can distinguish and locate the target. Shared weights provide great robustness and generalization capabilities and may reduce the number of network parameters. Expanding the receptive field allows CNN to obtain a larger receptive field for more accurate target positioning. The CNN model is also invariant to spatial transformation, so that even if the target moves, the model can still correctly identify and locate the target, enhancing the robustness of the target pose and changes. At the same time, the use of large-scale data sets to train the CNN model allows it to learn richer feature representations.

4.2. Classical CNN architectures for object detection

The R-CNN model was proposed in 2013 for object detection. This model combines deep learning and target detection methods to improve the efficiency of target detection. The R-CNN model uses the candidate area generation algorithm to select candidate areas that are likely to be include targets and then performs feature extraction in each candidate area through a pre-trained convolutional neural network model. The support vector machine classifier classifies the retrieved characteristics, and the regressor is utilized to modify the bounding box of the candidate area in order to better frame the target location. However, R-CNN also has the disadvantages of a large amount of calculation and slow training and reasoning speed [6].

Fast R-CNN was proposed in 2015. This model is an optimization and improvement of the R-CNN model. This model replaces R-CNN's need for independent proposal generation and object classification by using end-to-end training. In this way, training the process of target detection as a single model can improve training efficiency and model performance. To feed candidate regions of any size to the fully connected layer for target classification, Fast R-CNN introduces the ROI pooling layer. Any size candidate region can be mapped using this layer to a fixed-size feature map. The feature sharing of fast R-CNN can enable feature calling to be shared among multiple candidate regions by Extraditing Only One Feature on the Image, Reducing the Amount of Computation and Memory Overhead. A softmax classifier performs object classification, and a regressor is used to fine-tune the candidate regions. In this way, accurate classification and location positioning of objects can be achieved. But the disadvantage is that Fast R-CNN also needs to use the candidate area generation algorithm to extract the candidate area [6,7].

4.3. Model improvements for object detection

Faster R-CNN is an optimized improvement of R-CNN and Fast R-CNN. It introduces a region proposal network to automatically generate region proposals. RPN is a specially designed neural network that generates proposals by sliding a small window on the feature map and provides an objective score for

DOI: 10.54254/2755-2721/30/20230081

each proposal. RPN utilizes the settings of anchor boxes to generate candidate regions of different sizes and aspect ratios. This eliminates the need to independently train the candidate region generator like R-CNN and Fast R-CNN, thereby further improving speed and accuracy [6-8].

Due to the use of CNN, object detection has advanced significantly in recent years. Faster R-CNN, R-FCN, Multibox, SSD and YOLO are a few of these detectors that are rapid enough to be employed in consumer products. Operators, however, may find it difficult to decide which architecture is ideal for their applications [9].

The advantages, disadvantages and improvements of the object detection model are listed in the table 2.

Table 2. Analysis of the advantages and disadvantages of CNN in object detection.

Object Detection	R-CNN	Fast R-CNN	Faster R-CNN
Advantages	Integrating deep learning and target detection methods	fast detection speed, end-to-end training, high accuracy	A region proposal network is introduced to automatically generate candidate regions
Disadvantages	too much calculation	It is also necessary to use the candidate area generation algorithm to extract the candidate area	The understanding and implementation of the algorithm is relatively difficult, and it is not easy for beginners to get started. The calculation complexity is high and requires certain computing resources.Difficulty detecting small objects

5. Application of CNN in face recognition

5.1. Application principle

Train a CNN model to extract characteristics related to faces and categorize photos based on the presence or absence of faces. In terms of face feature extraction, through large-scale data sets, CNN can extract discriminative feature vectors for subsequent face recognition tasks. In face recognition, the face is input into the trained CNN model, and the similarity is judged by the feature vector, to judge whether it is the same person.By contrasting the input face picture with the face image of a known identity, face recognition may also determine whether the identity matches.It can also be used to train a CNN model to judge the expression of input face images.

5.2. Classical CNN architectures for face recognition

DeepFace uses a deep network with multiple convolutional layers, pooling layers, and fully connected layers, and it is pretrained on large-scale datasets.After using the traditional face detection algorithm to detect the face area in the image, face alignment technology is used to reduce the influence of posture on recognition. The feature vector is obtained by feature extraction to distinguish faces. Identify the identity by comparing the input face with the known face feature vector. But DeepFace can only be used to detect known faces, not real-time face recognition [10].

FaceNet uses a CNN architecture and uses a triplet loss function to train the model. This makes it possible to make the same person, the face eigenvectors are closer, and the eigenvectors of different people are farther away. If the distance between the eigenvectors of two people is smaller than the preset value, then it is judged that the two people are the same person. At the same time, FaceNet [9] also employs an end-to-end approach, including feature extraction, simultaneous training of distance

DOI: 10.54254/2755-2721/30/20230081

measures, and automated learning of the feature representation and similarity metric of facial picture data. FaceNet can be used for real-time face recognition [11].

ArcFace uses CNN to extract the feature vectors of the input face images. The separation between the feature vectors of the same individual is minimal, but the separation between the feature vectors of different individuals is big. This model introduces a special loss function, the angle cosine interval loss (ArcFace loss), which is used to train the network and increase the difference between classes to make the distinction more accurate. Afterwards, the eigenvector is normalized, and the modulus length of the eigenvector is fixed to 1 to reduce the error of scale change. Introducing non-linear mapping can further increase feature discriminability and robustness [12].

5.3. Improvements to face recognition models

Liveness detection is a technology created to distinguish real faces from fake faces (such as photos, videos, etc.). The current technology judges whether a face is a real face by using various information sources such as dynamic features. And living body detection can be carried out through the cooperation of various sensors.

Face key point detection can accurately locate important positions in face images, such as facial features. The introduction of multi-task learning can associate face key point detection with other tasks (such as face pose estimation, face expression recognition, etc.) for joint training. The advantages, disadvantages and improvements of the face recognition model are listed in table 3.

Table 3. Analysis of the advantages and disadvantages of CNN in face recognition.

Face Recognition	DeepFace	FaceNet	ArcFace
Advantages	High accuracy, using large-scale training set training	The triplet loss function is introduced to optimize the face feature embedding space	Use angle cosine loss function, easy to train and implement
Disadvantages	Face data needs to be labeled, and the demand for computing resources is high	High computational complexity	Too dependent on data sets, large computational requirements

6. Conclusion

According to the discussion of CNN, the many models put forward have a wide range of applications in numerous disciplines, which is advantageous for understanding CNN in various domains. Following the ease of labor, many models based on CNN were created to aid the task. This is conducive to improving accuracy in work and avoiding some errors caused by manual work. Although various models have been proposed one after another, it is very difficult and expensive to obtain large-scale labeled data in some fields, so it is still necessary to improve more accurate machine learning in the case of insufficient data. The noise and disturbance of the input data will affect the machine learning and make it misjudgment. Therefore, it is necessary to improve the robustness of the model in the future so that it has better generalization ability. At the same time, machine learning requires a lot of computing resources and time. Therefore, a major difficulty remains in figuring out how to increase the model's learning efficiency while decreasing the application of resources. While there are still many problems to be solved, this is exactly why the field of machine learning needs to evolve. In development, the accuracy of machine learning will definitely become higher and higher, and at the same time, the consumption of resources will become less and less. I hope that in the future, machine learning will be easier to use, so that more ordinary people can get started, so as to reduce work pressure and make work more efficient.

DOI: 10.54254/2755-2721/30/20230081

References

[1] Lecun Y and Bottou L 1998 Gradient-based learning applied to document recognition. *Proc. IEEE* **86(11)** p 2278-2324 doi:10.1109/5.726791.

[2] Krizhevsky A, Sutskever I and Hinton G 2012 ImageNet classification with deep convolutional neural networks *NeurIPS* **25(2)** doi:10.1145/3065386.

[3] Simonyan K and Zisserman A 2014 Very Deep Convolutional Networks for Large-Scale Image Recognition *CS* doi:10.48550/arXiv.1409.1556.

[4] Zhang X, Zhou X and Lin M ,et al 2017 ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices doi:10.48550/arXiv.1707.01083.

[5] Tan M and Le Q V 2021 EfficientNetV2: Smaller Models and Faster Training doi:10.48550/arXiv.2104.00298.

[6] Girshick R, Donahue J and Darrell T ,et al 2014 Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation *CVPR.IEEE* doi:10.1109/CVPR.2014.81.

[7] Girshick R 2015 Fast R-CNN *ICCV* doi:10.1109/ICCV.2015.169.

[8] Ren S, He K and Girshick R 2016 Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks *NIPS* doi:10.1109/tpami.2016.2577031.

[9] Huang J, Rathod V and Sun C 2016 Speed/accuracy trade-offs for modern convolutional object detectors. *CVPR.IEEE* doi:10.1109/CVPR.2017.351.

[10] Taigman Y, Yang M and Ranzato M 2014 DeepFace: Closing the Gap to Human-Level Performance in Face Verification *CVPR.IEEE CS* doi:10.1109/CVPR.2014.220.

[11] Schroff F, Kalenichenko D and Philbin J 2015 FaceNet: A Unified Embedding for Face Recognition and Clustering *IEEE* doi:10.1109/CVPR.2015.7298682.

[12] Deng J, Guo J and Xue N 2019 Arcface: Additive angular margin loss for deep face recognition *Proc. CVPR.IEEE* p 4690-4699