Progress in Artificial Intelligence
https://doi.org/10.1007/s13748-019-00203-0

REVIEW

Convolutional neural network: a review of models, methodologies and applications to object detection

Anamika Dhillon[1] · **Gyanendra K. Verma**[1]

Received: 28 May 2019 / Accepted: 25 November 2019

Abstract
Deep learning has developed as an effective machine learning method that takes in numerous layers of features or representation of the data and provides state-of-the-art results. The application of deep learning has shown impressive performance in various application areas, particularly in image classification, segmentation and object detection. Recent advances of deep learning techniques bring encouraging performance to fine-grained image classification which aims to distinguish subordinate-level categories. This task is extremely challenging due to high intra-class and low inter-class variance. In this paper, we provide a detailed review of various deep architectures and model highlighting characteristics of particular model. Firstly, we described the functioning of CNN architectures and its components followed by detailed description of various CNN models starting with classical LeNet model to AlexNet, ZFNet, GoogleNet, VGGNet, ResNet, ResNeXt, SENet, DenseNet, Xception, PNAS/ENAS. We mainly focus on the application of deep learning architectures to three major applications, namely (i) wild animal detection, (ii) small arm detection and (iii) human being detection. A detailed review summary including the systems, database, application and accuracy claimed is also provided for each model to serve as guidelines for future work in the above application areas.

Keywords Deep learning · CNN architectures · Transfer learning · Object detection

1 Introduction

Due to the success and rapid development of deep learning (DL) [1], a number of fields including robotics, medicine, biology, commerce, etc., have achieved considerable results. The fundamental concept of DL comes from artificial neural network (ANN) [2] research, and the idea is to learn the data representations by enlarging the abstraction levels. It endeavors to model the hierarchal representations beyond the data and classify the patterns by stacking numerous layers of information modules in the hierarchal structure. This type of hierarchal learning model is very robust as it enables the system to learn the complex representations right from the input data.

DL and deep convolution neural network (DCNN) have dramatically upgraded the performance beyond the state of the art in the above fields, compared to the conventional machine learning (ML) techniques, such as support vector machine (SVM) [3] and Naive Bayes [4]. It takes advantage of extracting higher-level features directly from the raw data. There are several advancements of DL that make this model more reliable and adaptive. For instance, in computer vision, a new optical character recognition (OCR) engine [5] is introduced in maps, through which we can identify the street as well as the store signs. Another one is generative adversarial networks (GAN) [6] which enable to tackle the problem of unsupervised learning. Furthermore, there is a task known as visual reasoning, where neural network (NN) is used to answer a question, with the help of a photograph, and so on.

There are many variants of DL such as autoencoder [7, 8], restricted Boltzmann machines (RBM) [9] and CNN [10]. An autoencoder is a deep neural network approach, utilized for unsupervised learning and contains the following layers: an input layer, the hidden layer and the output layer. This whole network is prepared to reconstruct its input data so

✉ Anamika Dhillon
dhillon.anamika2390@gmail.com

Gyanendra K. Verma
gyanendra@nitkkr.ac.in

[1] Department of Computer Engineering, National Institute of Technology Kurukshetra, Kurukshetra 136119, India

Published online: 20 December 2019

that it can enforce the hidden layer to learn better presentations of the input. Autoencoders can be divided into various categories: stacked autoencoder (SAE) [11], denoising autoencoder (DAE) [12], correspondence autoencoder (Corr-AE) [13], sparse autoencoder [14]. RBM is a feature generation model, which has two-layer structure: The first layer is known as a visible layer, and the second layer is identified as hidden layer. RBM characterizes the likelihood distribution on the arrangement of its visible layers, and its main aim is to magnify the probability of the training data. This model is utilized to demonstrate those temporal sequences which have high dimensionality, for example, distinct sorts of data including unlabeled or labeled images, client rating of videos or movies, etc. RBM can be employed as a part of deep learning models, such as convolutional neural systems [15], pre-preparing and DBN [16].

Convolutional neural networks (CNN), first introduced by Fukushima [17] in 1998, have wide applications in activity recognition [18, 19], sentence classification [20], text recognition [21], face recognition [22], object detection and localization [23, 24], image characterization [25], etc. They are made up of neurons, where each neuron has a learnable weight and bias. It contains an input layer, an output layer and multiple hidden layers, where hidden layer consists of a convolutional layer, pooling layer, fully connected layer (FC) and various normalization layers. Convolutional layer applies a convolution operation, to merge two sets of information. It imitates the feedback of an individual neuron to visual stimuli. Pooling layer is used to reduce the dimensionality, by associating the output of neuron cluster at one layer with the single neuron. FC layer connects every neuron in one layer to every neuron in another layer. Its primary purpose is to classify the input images into several classes, based on the training datasets.

Deep neural network (DNN) [26] is applied in several areas, where machine intelligence would be useful. This segment presents some of those instances where DNNs are presently having a considerable effect:

- Human activity recognition [27, 28].
- Facial expression recognition [29].
- Image classification [30].
- Feature modeling [31].
- Object detection [32] etc.

Though DL has advancements over traditional machine learning (ML), it faces various challenges [33] including:

1. Lots of data. A large database is essential to get the desired accuracy in DL algorithms.
2. Facial expression recognition.
3. Energy-efficient techniques and high-performance hardware. To ensure great accuracy and less time consumption, high performing GPU´s should be taken into account.
4. Big data analytics using DL. Dealing with various criteria including variety, volume and velocity of big data problem.
5. Lack of flexibility and multi-task learning. To solve multiple tasks in various application domains, without the need for reworking on the whole architecture.

The name “convolutional neural system” shows that the system utilizes a mathematical linear operation called convolution, instead of general matrix multiplication in at least one of their layers. A CNN contained at least one convolutional layer and after that pursued by at least one fully connected layer as in a standard multi-layer neural network. They are one of the more prevalent strategies utilized in image recognition and computer vision frameworks, and they are verifiably noteworthy in the field of data science.

In this paper, we discuss the advances made in applications of DL. We focus on the architecture of CNN, which includes its various layers, like convolutional layer, nonlinearity layer, pooling layer, FC layer and the classification layer. Then, we discuss the multiple models of CNN like, AlexNet, ResNet, SENet, etc. In transfer learning, a develop model and pre-trained model approaches are described. Develop model approach uses four steps, including selection of source task, development of source task, reusing the model and finally tuning or refinement of the model. Pre-trained model approach has three steps, including selection of source model, reuse of the model and refinement of model. Finally, some areas of applications are discussed, which mainly embrace wild animal detection [34, 35], arm detection [36, 37] and miscellaneous object detection.

The primary purpose of this survey is to present the comprehensive body of knowledge of DL and CNN architecture and its various pre-trained models. Our contributions are outlined as follows:

1. A comprehensive perspective of DL.
2. Identification of key challenges.
3. A resolute discussion on the architecture of CNN.
4. Description of various CNN’s pre-trained models.
5. Pointing the open applications in the areas of object detection.

We organize the remainder of this paper as follows: In Sect. 2, we provide the architecture of CNN, which discusses its different layers and their explanations. Section 3 describes various pre-trained models of CNN, including LeNet, AlexNet, ZFNet, GoogleNet, VGGNet, ResNet, ResNeXt and SENet. In Sect. 4, transfer learning is described, and in Sect. 5 various applications of CNN in the area of object

detection are explained. We conclude in Sect. 6 with a brief overview of our contributions.

2 A brief history of CNN in object detection

In this section, we present a brief history of CNN algorithms and related applications.

2.1 Object detection

Object detection is proposed for classifying and localizing every object in the bounding box. Time complexities are reduced in the detection techniques by utilizing the **sliding window method** [38] which is the very basic technique for detecting the objects. In this method, a window of reasonable size, say, $M \times N$, is picked to search over the objective image. Initially, a classifier is prepared on a set of training images, spreading over the object of interest for detection as one class and irregular objects as different classes. Tests that belong to objects of interest for detection are alluded to as positive images, while irregular samples are alluded to as negative images.

These methods are classified into two categories, namely region proposal-based methods and classification-based methods. Region proposal-based methods include R-CNN, Faster R-CNN [39], etc., and classification-based methods include You Only Look Once (YOLO) [40], single shot detector (SSD) [41], etc.

2.2 Faster R-CNN

Faster R-CNN consists of two networks: region proposal network, which is used to generate the regions of interest, i.e., region of proposals and the network utilizing these proposals for detecting the object. So in Faster R-CNN, region proposal generation and detection of object are occurred by the similar network. It proposed the idea of anchor boxes. These are placed at sliding window in question and accomplished with three scales (128×128, 256×256 and 512×512) and three aspect ratios (1:1, 2:1 and 1:2). So, total 9 boxes are there on RPN to forecast the likelihood of it being background or foreground. So, Faster R-CNN is considered as one of the most accurate object detection techniques.

2.3 YOLO—You Only Look Once

YOLO is a technique where distinct components of detected objects are unified into a single neural network. It takes the input image and utilizes its features for predicting each bounding box. It partitions every image into $S \times S$ grid, and every grid anticipates N bounding boxes and confidence score for these boxes. The confidence reveals the accuracy of the bounding box, whether it encloses the accurate image or not. YOLO is superfast, and it can be run for real-time environment, but its drawback is that it only concludes 1 type of class in 1 grid.

2.4 Single shot detector (SSD)

SSD, as the name predicts, for detecting numerous objects in an image requires only one single shot, meaning that the convolutional network will run only once and predict the feature map. It also utilizes the anchor boxes with various aspect ratios and scales. For predicting variable size objects, the network combines various feature map predictions with distinct resolutions. It is much faster than those methods which are based on two-shot RPN.

2.5 RetinaNet

RetinaNet [42, 43] is a technique which is assembled by composing two advancements over techniques like YOLO and SSD: (1) feature pyramid networks for object detection and (2) focal loss for dense object detection.

A feature pyramid network (FPN) [44] utilizes the characteristic multi-scale pyramidal order of deep CNNs to make feature pyramids. It joins low-resolution, semantically solid features with high-resolution, semantically weak features through a top-down pathway and sidelong associations. The Focal loss is intended to address the single-stage object detection issues with the irregularity where there are an exceptionally enormous number of conceivable background classes and only a couple of foreground classes. Focal loss is an enhancement for cross-entropy loss which decreases the corresponding loss for well-arranged models and establishing extra spotlight on hard, misclassified models.

RetinaNet is a solitary, unified system made up of a backbone network and two task specific subnetworks. The backbone is in charge of figuring a conv feature map over a whole input image. The first subnet performs object classification by utilizing backbone's output; the second subnet performs bounding box regression by utilizing the backbone's output.

3 Convolutional neural networks: related work

CNNs have demonstrated viable in the areas of image recognition and classification. They are DL algorithms, which take input in the form of video/images, assign weights and bias to several aspects in that image and then differentiate them from each other. They aim to utilize the spatial information among the pixels of an image. Therefore, they are based on discrete convolution. This section gives an

overview of some existing works of CNN and its applications in object detection and recognition.

Zhiqiang [45] propounded a technique for extracting the features of CNN and also proposed a detecting algorithm, based on regional proposal and regression. This paper compensates those drawbacks which exist in hand-crafted features but also contains some drawbacks like occlusion and low resolution.

Ranjan [22] propounded a deep learning technique for understanding the faces. The paper presents deep convolutional neural network (DCNN)-based techniques for detection of faces, and it can be partitioned into two categories: sliding window approach and the region-based approach. The main issues which are addressed in this paper are as follows: minimizing the reliance on huge training database, controlling data bias and mortification in training data and reducing the time of training when the network is broader and more profound.

Park [27] proposed human activity recognition technique by using recurrent neural network (RNN). Firstly, they create an input feature matrix from MSRC-12 activity dataset, and then, RNN is trained. Finally, the trained RNN is evaluated by utilizing the test datasets. They have gained 99.55% of the mean recognition accuracy for the actions.

Zhao [46] discussed a method for classification of fine-grained objects and semantic segmentation based on deep learning. This method distinguishes the subordinate-level groups, such as breeds of dogs and bird species. They have achieved a 3.57% error rate on the ImageNet dataset. Vinyals [47] proposed a method, namely a neural image caption generator. This method describes a framework for the neural and probabilistic model, for creating the representations from the images. Their further work includes improvement in image description approaches by using unsupervised data.

Dai [48] propounded a network of multi-task cascades, for instance-aware semantic segmentation. Their framework contains three networks, which are as follows: distinguishing the instances, approximation of the masks and classification of objects. They took 360 ms for testing the image by using VGG-16 and achieved convincing object detection outputs which exceed the Fast/Faster R-CNN model.

Zhou [24] discussed a paper which describes the deep learning applications in object detection. This paper includes various neural networks like SPPNet, R-CNN, Faster R-CNN [49] and Fast R-CNN [50] and their applications on a football game image dataset. In the future, some synthetic data will be considered, so that the amount of data can be increased further.

Girshick [50] proposed Fast R-CNN, and Ren [49] propounded Faster R-CNN for object detection. They present the latest training algorithm which helps in fixing the disadvantages of R-CNN and SPPNet and the experiments are mainly performed on VOC 2007 and 2010 datasets.

Xu [51] discussed a method for detection of the RGB-D object, based on multimodal deep feature learning. This method contains two phases: estimation of objectness and the recognition of the region-wise object. Experiments are carried out on two datasets, namely NYU Depth v2 and SUN RGB-D, and they have shown a great performance over other methods like RPN-RGB, RPN-RGB-D, etc. In future, a study for a particular influence of depth statistics on several object classes will be carried out.

Girshick [52] propounded a technique for object detection and semantic segmentation on the PASCAL VOC dataset. Their system mainly contains three modules: the first module initiates the category-independent region proposals, the other one is the large CNN which extricates rigid length feature vector from each area and the third module is a set of class definite linear SVMs. This paper propounded a straightforward and adaptable detection calculation that enhances mAP by over 30% with respect to the past outstanding outcome on VOC 2012 accomplishing a mAP of 53.3%.

Abousaleh [53] proposed a method known as comparative region convolutional neural network (CR-CNN), for estimation of facial age. Moreover, they include the technique of auxiliary coordinates (MAC) for training that diminishes the ill-conditioning issue of deep neural network and manages a proficient and appropriate performance.

Fang [54] propounded a technique to solve the problems of image classification. In their work, they pick up a superior comprehension of deep learning by investigating the misclassified cases of emotions and facial recognition. In their future work, they will place the data augmentation model on the training data.

4 CNN architecture

CNNs have widely used deep learning algorithms and the most prominent category of neural networks, mainly in high dimensional data, like images and videos. It is a multi-layer neural network (NN) architecture, stimulated by the neurobiology of visual cortex, which contains convolutional layer(s) pursued by fully connected (FC) layer(s). Subsampling layers can exist between these two layers. They achieve the best of DNNs, which have a complication in scaling well along with multidimensional locally correlated input data. Thus, the primary application of CNN exists in databases, where the number of nodes and parameters required to be trained is comparatively large (e.g., image processing) (Tables 1, 2, 3).

4.1 Convolutional layer

This layer is the primary building block of a convolutional neural network, which determines the output of associated

inputs, in the receptive field. This output is achieved through kernels, which are convolved over the height and width of the information data, computing the dot product between the input and filter values, therefore building a 2-D activation map of that filter. With this, CNN quickly learns those filters, which activate when a particular type of feature at some spatial position of the input is observed.

4.2 Non-linearity layer

Nonlinear functions are very significant and have degree higher than one, and when they are plotted, they used to have a curvature. The main objective of this layer is to transform the input signal to the output signal, and that signal will be utilized as an input in the next layer. Some popular types of Non-linearity layers are: sigmoid or logistic, Tanh, ReLU, PReLU, ELU, etc.

4.3 Pooling layer

CNN can have local or global subsampling layers that add the outputs of a neuron at one layer into an individual neuron in the following layer. Its main task is to scale down the spatial size of the representation to diminish the no. of parameters and calculations in the model. It not only speeds up the calculations but also averts the problem of overfitting. The most common form of pooling layer is the max pooling.

4.4 Fully connected layer

FC layers are standard deep NN, which seeks to build the predictions from the activations, to be used for classification or regression. It has a similar principal as the conventional multi-layer perceptron neural system (MLP). This layer acquires the full connections to every each activation in the antecedent layer, and the activations can be calculated by using matrix multiplication followed by a bias offset.

4.5 Loss/classification layer

The loss layer determines how the training prohibits the deviation among the true and anticipated labels, meaning that it is mainly used to guide the training process of NN. Different loss functions suitable for various errands might be utilized in DCNN, like softmax, cross-entropy, etc. Softmax loss is utilized for foreseeing a solitary class of K mutually exclusive classes, and the softmax layer yields a likelihood appropriation, i.e., the values of the output total to 1. Moreover, this layer is a delicate form of the max-output layer so it is differentiable and furthermore versatile to anomalies. Sigmoid cross-entropy loss is utilized for foreseeing K-free probability values. The sigmoid capacity yields negligible probabilities, and along with these probabilities, lines can be utilized for numerous classes grouping. An issue with sigmoid is that after reaching the saturation, the gradient became vanished. Euclidean loss is utilized for regression to truly valued labels.

Here is the overview of systems in terms of neural network model, database, result and applications.

5 CNN models

Deep CNN made a noteworthy contribution in several domains like image classification and recognition; therefore, they become widely known standards. This section describes various classical and modern architectures of Deep CNN, which are currently utilized as a building block of several segmentation architectures. Major CNN models are as follows (Tables 4, 5, 6, 7):

1. LeNet.
2. AlexNet.
3. ZFNet.
4. GoogleNet.
5. VGGNet.
6. ResNet.
7. Inception model.
8. ResNeXt.
9. SENet.
10. MobileNet V1/V2.
11. DenseNet.
12. Xception model.
13. NASNet/PNASNet/ENASNet.
14. EfficientNet.

5.1 LeNet model

This traditional NN architecture was successfully applied on MNIST handwritten digit recognizer patterns. LeNet

Table 1 Noticeable papers of 2014

System (year)	NN model	Dataset	Accuracy	Application
Ross Girshick [50] (2014)	CNN, RNN	PASCAL VOC dataset	Achieving a mAP of 53.3%	Object detection and segmentation
Volodymyr Mnih [55] (2014)	CNN, RNN	MNIST	Reaching roughly 1.9% error	Image classification and detection

Table 2 Noticeable papers of 2015

System (year)	NN model	Dataset	Accuracy	Application
Oriol Vinyals [47] 2015	RNN, CNN LSTM	Flickr30k, SBU, COCO dataset	BLEU-1 score is 59	Computer vision and NLP
Anran Wang [56] 2015	Deep CNN	2D3D dataset	mAP of 88.4%	RGB-D object detection
Andrej Karpathy [57] 2015	CNNs, R-CNN, multimodal RNN	Flickr8K, Flickr30K, MSCOCO dataset	Achieves maximum BELU of 66.0	Generating image descriptions
Jeff Donahue [58] 2015	RNN, CNN	UCF-101 dataset	BELU score of 28.8	Visual recognition and description
Yang Hua. [59] 2015	CNN, SVM	Online object tracking benchmark	Precision score of 0.798	Online tracking
Kaiming He [60] 2015	Deep CNN	Pascal VOC, ImageNet	9.14% top-5 error rate	Visual recognition
Li Yao [61] 2015	RNN, 3-D CNN	Youtube2 text dataset	Achieves BELU of 0.4192	Video description

Table 3 Noticeable papers of 2016

System (year)	NN Model	Dataset	Accuracy	Application
Mohammad Havaei [62] 2016	DNN, CNN	2013 BRATS test	Specificity of 0.88 Sensitivity 0.84	Brain tumor diagnosis
Milyaev [23] 2016	CNN VOC 2007	Pascal dataset	MAP of 56.6% is 26.5	Image denoising, detection in noisy images

Table 4 Noticeable papers of 2017

System (year)	NN model	Dataset	Accuracy	Application
Xinyi Zhou [24] 2017	SPPNet, R-CNN	ImageNet PASCAL VOC, COCO	max max mAP of 0.8377 for Soccer Goal	Object detection, image classification, identification of face
Wang Zhiqiang [45] 2017	CNN, R-CNN Yolo	ImageNet, PASCAL VOC, MSCOCO	mAP of 73.2% on VOC 2007	Image processing, object detection
Bo Zhao [46] 2017	CNNs, RNN	CUB 200 dataset, ImageNet	80.3%	Segmentation
Yuanyuan Ding [63] 2017	CNN, ReLUs dropout	CASIA-WebFace and LFW dataset	Highest averaged recognition rate (93.78%)	Face recognition under noise
Ke Shan [64] 2017	Deep CNN	JAFFE and extended Cohn–Kanade	mAP of 76.75 and 80.303 on test datasets	Facial expression recognition
Jian Gang Wang [65] 2017	Deep affordance l earning, SSD, Faster R-CNN,	VOC07/12 dataset, ILSVRC and MSCOCO	max mAP of 0.9894 on single cup class	Object detection
Bing Tian [66] 2017	CNN	Six different classes from traceable video	mAP of 99.99%	Traceability application, video-based object detection
Jifeng Dai [67] 2017	CNNs, STN	Pascal VOC	Increased mAP to 37.5%	Object detection, semantic segmentation
Gao Huang [68] 2017	CNN DenseNet	CIFAR-100, SVHN ImageNet and 17.18 on	3.46% error rate on C10+	Visual object

[76] acquires an input image of $32 \times 32 \times 1$ (grayscale image), which goes to the convolution layers and then to the subsampling layer. After that, there is other succession of convolution layers followed by a pooling layer. Finally, there are 3 FC layers, involving the output layer at the end. The main objective of architecture was to acknowledge the

Table 5 Noticeable papers of 2018

System (year)	NN model	Dataset	Accuracy	Application
Junwei Han [69] 2018	SOD (CNN based), COCO	PASCAL VOC, MS COD	Obtains a mAP of 78.6%	Salient and class-specific object detection
Duc Tung [70] 2018	CNN, Deep learning	CDNet Wallflower	Highest FM of 0.95	Video-based application
Shuai [71] 2018	DNN, Faster R-CNN	VOC	Increases 25.8% processing time	Object recognition

Table 6 Noticeable papers of 2019

System (year)	NN model	Dataset	Accuracy	Application
Zifeng Wu [72] 2019	SOD (CNN based), COD	PASCAL VOC MSCOCO	mAP of 78.6%	Salient class-specific object detection
Rui Zhao [39] 2019	CNN, deep learning	Holiday and UKBench datasets	Map of 0.837%	Feature fusion
Ghulam Muhammad [73] 2019	DNN, Faster R-CNN	VOC	Increases 25.8% processing time	Object recognition
Rajeev Ranjan [74] 2019	DNN, Faster R-CNN	VOC	Increases processing time	Object recognition
Zhong-Qiu Zhao [39] 2019	DNN, Faster R-CNN	VOC	Increases processing time	Object recognition
Shuai Zhang [75] 2019	DNN, Faster R-CNN	VOC	Increases processing time	Object recognition

Table 7 Summary of the models. Adopted from Matcon-vNet

Model	Year	Top-1 Error	Top-5 Error	Images (Size)
resnet-50-dag	2015	24.6	7.7	396.3
googlenet-dag	2014	34.2	12.9	770.6
matconvnet-vgg-f	2013	41.4	19.1	2482.7
vgg-f	2013	41.4	18.8	1118.9
caffe-alex	2012	42.6	19.6	1379.8
ResNeXt50	2017	22.260	6.190	25.1 M
InceptionV3	2015	22.102	6.280	23.9 M
Xception	2017	20.994	5.548	22.9 M
DenseNet121	2018	25.028	7.742	8.1 M
NASNetLarge	2018	17.502	3.996	93.5 M

handwritten digit patterns and the zip code recognition in post offices. It utilizes a 5×5 filter along with stride of 1.

The author in [76] proposed a method with convenient network architecture and gradient-based learning algorithms that were utilized to classify the large dimensional sequences like: handwritten words, with minimal preprocessing. On automatically classification machine, 82% words were correctly identified, 1% errors and 17% rejects (Table 8).

5.2 AlexNet model

AlexNet [77] was the pioneering deep architecture, with top-5 test accuracy of 84.6% on ImageNet data. It utilized those data augmentation methods that are comprised of image translation, patch extractions and horizontal reflection. This CNN model implements dropout layers with a particular end goal to battle the issue of overfitting to the training data. It is trained by batch stochastic gradient descent, along with particular values for weight decay and momentum. It was trained on two GTX 580 GPUs for 5–6 days and contains five convolutional layers, one max pooling, ReLU as non-linearities, three FC layers and dropout. Howard [78] proposed a method known as MobileNets, which uses CNN for mobile vision approaches. Reduced MobileNet is 4 times better than AlexNet, was smaller in size, i.e., 45×45, and requires less computation time than AlexNet.

Huang [79] propounded a training method for stochastic depth that permits the contrary setup to train small networks and utilize the deep networks at the testing time. By using stochastic depth, the depth of residual networks can be increased over 1200 layers and attains advancements in test error (4.91% on CIFAR-10) (Figs. 1, 2).

Table 8 Publications relevant to LeNet

CNN model (year)	The work	Error rate	Dataset	Importance
LeNet [76] 1998	Yann Lecun	82% accurately identified checks, 1% errors, 17% rejects	Modified NIST Set	Document recognition

Remarks on AlexNet model

- This network is trained on ImageNet dataset that has images more than 15 million.
- For nonlinear functions, ReLU is used.
- Dropout layers are implemented to encounter the issue of overfitting.
- Trained on two GTX 580 GPUs for 5–6 days (Table 9).

5.3 ZFNet model

This model [85] was a fine-tuning to the past AlexNet model. This model utilized the filters of size 7×7 and diminished stride esteem. The idea beyond this change is that a small filter size in the primary convolution layer holds a considerable measure of unique pixel data in the input volume. This model was trained on the GTX 580 GPU for 12 days and evolved a visualization technique named deconvolutional network. Zeiler [85] proposed visualization framework which provides the insights into the function of intermediary feature layers and the working of the classifier. It was trained on ImageNet 2012, Caltech-101 and Caltech-256 databases and gains the error rate of 0.1%.

- It is significantly improved compared to AlexNet.
- It reduced the filter size of the first layer from 11×11 to 7×7.
- Its top-5 validation error rate is 16.5
- It was trained on a GTX 580 GPU for 12 days.

5.4 GoogleNet model

GoogleNet gives the state-of-the-art performance on ImageNet ILSVRC14 detection and classification challenge, with top-5 test accuracy of 93.3%. The primary indication of this model is the enhanced use of computer resources into the model. This is accomplished through a building block known as "inception module" which takes into account the enlarged depth and width of the model. This was one of the first CNN architectures which deviate from the common

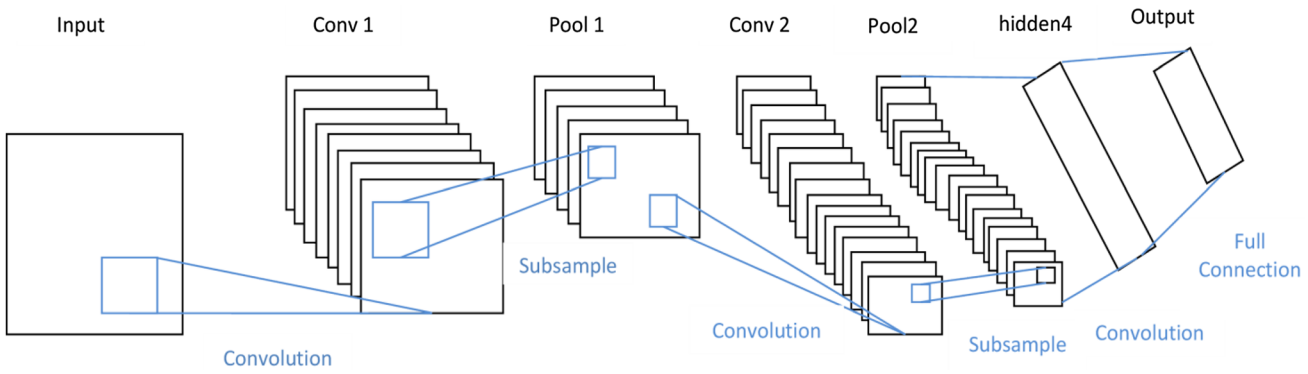

Fig. 1 Architecture of LeNet [76]

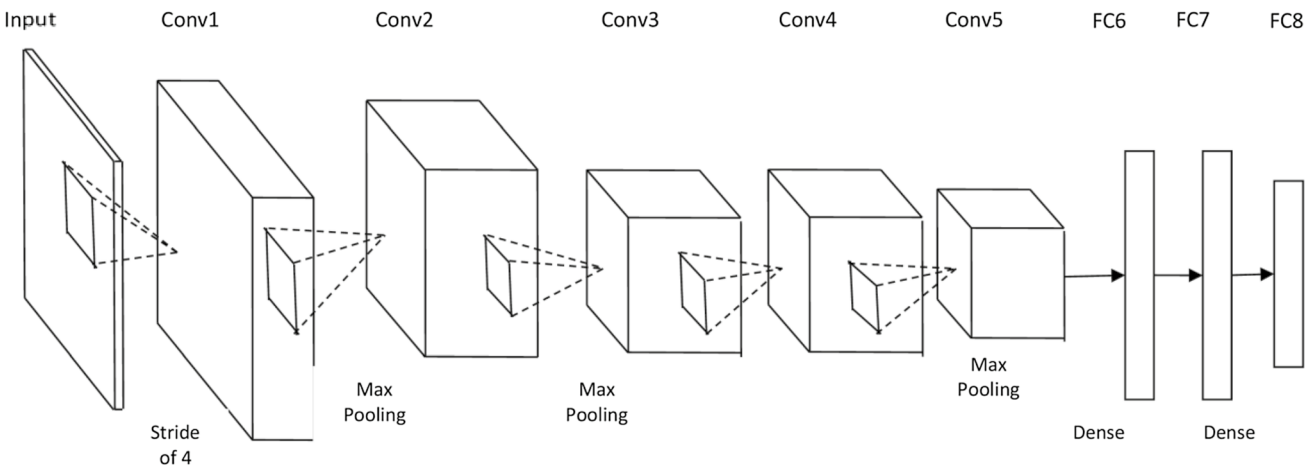

Fig. 2 Architecture of AlexNet [77]

Table 9 Publications relevant to AlexNet

CNN model (year)	The work	Error rate	Dataset	Importance
AlexNet 2012	Krizhevsky [77]	15.3%	ILSVRC	Data augmentation
AlexNet 2017	Howard [78]	mAP of 83%	ILSVRC 2012, YFCC100M	Mobile vision application, fine-grained classification
AlexNet 2016	Gao Huang [79]	4.91% error rate	CIFAR-100 ImageNet	Computer vision
AlexNet 2017	Sang-Il Oh [80]	77.72% mean average precision	KITTI benchmark database	Object detection and classification
AlexNet 2018	Haoyu Xu [81]	Accuracy is enhanced by 39.6%	FOD dataset	Foreign object debris material recognition
AlexNet 2016	Hieu Minh [82] Bui	mAP of 89.7%	W-RGB-D dataset	Object recognition
AlexNet 2016	Xiaoheng Jiang [83]	MR increases by 1.04%	INRIA dataset	Pedestrian detection
AlexNet 2016	Tom [84]	Miss rate is 0.199	NVIDIA	Pedestrian detection

approach of simply stacking convolution and pooling layers on top of each other in a sequential structure (Fig. 3).

Szegedy [87] propounded a deep CNN model that was accountable for creating the modern state of the art for detection and classification on the ImageNet dataset. The depth of this network is 22 layers, and the superiority is evaluated in relation to detection and classification (Table 10).

Silver [88] propounded a framework for computer Go which utilizes the value networks for evaluating the board locations and policy networks for selecting the moves. By utilizing this model, AlphaGo gained a 99.8% victorious rate across alternative Go programs (Fig. 4).

Remarks on GoogleNet model

- It contains 1×1 convolution at the middle of the network.
- Global average pooling is used at the end of the network.
- It utilizes $12\times$ fewer parameters than AlexNet.
- It is trained on a few high-end GPUs within a week.
- It achieved a top-5 error rate of 6.67 (Table 11).

5.5 VGGNet model

Visual geometry group (VGG) [90] is a CNN model introduced by VGG from University of Oxford. It expands the intensity of the neural network that not just accomplishes the state-of-the-art precision over ILSVRC [91] datasets, yet in addition is relevant to another image recognition databases. The VGG-16 contains 13 convolutional layers with 3 FC layers; on the other hand, the VGG-19 contains 3 extra convolutional layers. They utilize channels with the little receptive field: 3×3, and every hidden layer is outfitted with the rectification non-linearity function. Straightforwardness and depth: this is the meaning of this model and best used with the error rate of 7.3%.

Ren [49] propounded a region proposal network which uses the convolutional features of full image by utilizing the detection network. In ILSVRC and COCO 2015, Faster R-CNN and RPN come to the first place in various records.

He [92] proposed residual learning methods which achieve great accuracy from noticeably expanded depth. This method reformulates the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. They gain a 3.57% error on ImageNet dataset.

Remarks on VGGNet model

- It utilizes 3×3 filters in the initial layer.
- It performs well on the classification of images and their localizations.
- This model is built along with Caffe toolbox.
- It utilize the scale jittering as the data augmentation method at the time of training.
- ReLU layer is used after every convolutional layer.
- Trained on 4 Nvidia Titan Black GPUs for 2–3 weeks (Table 12).

5.6 ResNet model

ResNet [92] is another 152 layer model that makes advanced traces in categorization, localization and detection with one magnificent model. Residual block: It addresses the issue of training a deep model by commencing identity skip connections so that layers can copy their inputs to the next layer. The idea behind this approach is the next layer should learn something new and divergent from what the input is already given (Fig. 5).

He [92] proposed a deep residual learning technique for recognition of images. They perform their observations on ImageNet database which contains 1000 classes. ResNet decreases top-1 error by 3.5.

Szegedy [98] proposed an inception architecture that performs really good and has a relatively low computational cost. With a group of three residual and one Inception-v4,

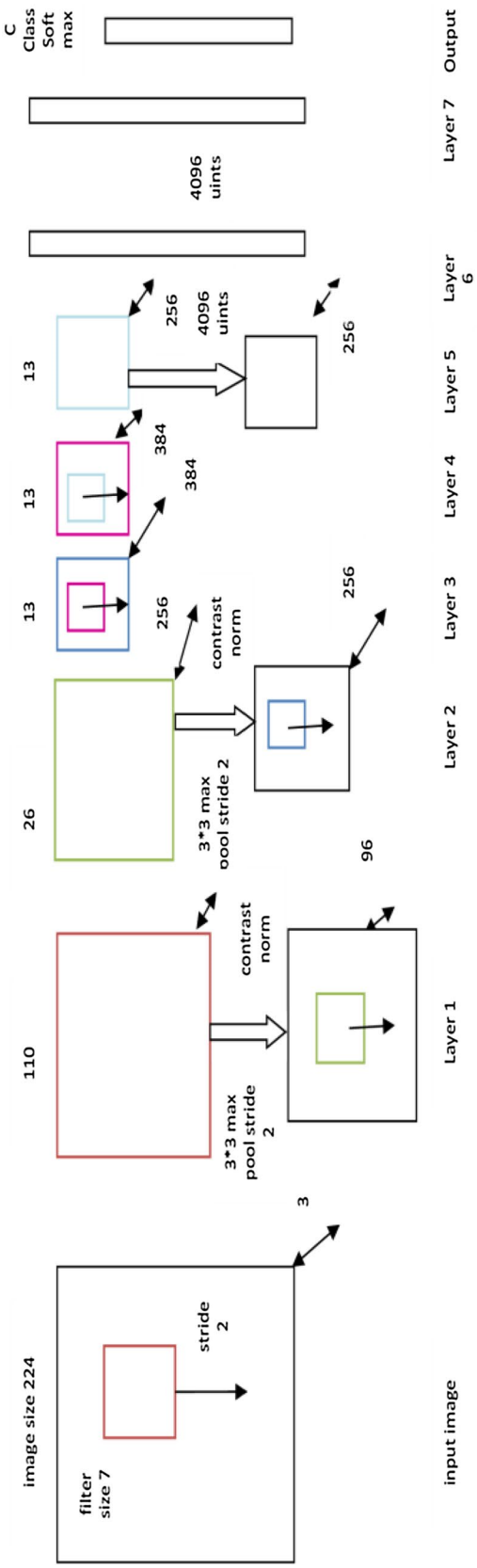

Fig. 3 Architecture of ZFNet [85]

they accomplished a 3.08 top-5 error on the test database of the ImageNet.

Remarks on ResNet model

- Extremely deep model, with 152 layers.
- The spatial size will be reduced after the initial 2 layers.
- By trying 1202 layers, the accuracy rate became lower.
- Trained on 8 GPU for 2–3 weeks (Table 13).

5.7 Inception model

The inception network was an important milestone in the development of CNN classifiers. The main idea of the Inception model is in view of discovering how an ideal local sparse design in a convolutional vision framework can be relative and secured by promptly available dense components. A layer-by-layer construction is done here, where one ought to dissect the relationship statistics of the end layer and bunch them into gatherings of units with the high connection. These bunches shape the units of the following layer and are associated with the units in the past layers. It is accepted that every unit from the prior layer is compared with some locale of the input image and these units are assembled into filter banks. In the below layers, associated units would focus in nearby locales regions.

Christian Szegedy [98] proposed an Inception model for computer vision. They get 21.2 top-1 and 5.6 top-5 error on single frame calculation. Xia [101] propounded an Inception-v3 model for flower classification. The tests are done on the Oxford-17 and the Oxford-102 flower dataset. They get mAP of 95 on OXFORD-17 and 94 on OXFORD-102 flower database (Figs. 6, 7).

Remarks on inception model

- For reduction in dimension and rectified linear activation, 1×1 convolution uses 128 filters.
- It utilizes $12\times$ fewer parameters than other competitors.
- It takes advantage of multi-level feature extraction.
- It comes in the first place in ILSVRC2014, with 6.67% top-5 error rate (Table 14).

5.8 ResNeXt model

This is the manageable architecture that acquires the ResNet/VGGs approach of replicated layers and utilizes the split–transform–merge procedure. This model contains the stack of residual blocks, which use similar topology and rules, which are influenced by the ResNet/VGGs. These rules are as follows: (a) the blocks will split the hyper parameters if the produced spatial maps are of the same size and (b) every time the spatial mapped is pooled by 2 factors;

Table 10 Publications relevant to ZFNet

CNN model (year)	The work	Error rate	Dataset	Importance
ZFNet 2013	Fergus [85]	14.8%	ILSVRC 2013	Visualization approaches
ZFNet 2017	Lisha Xiao [86]	mAP of 84.08%	ImageNet	Scene classification

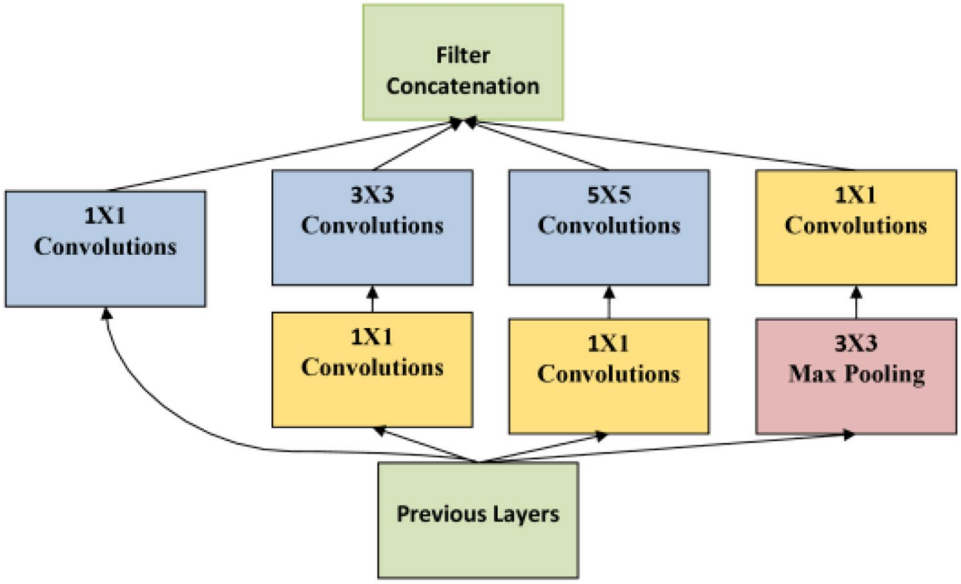

Fig. 4 Architecture of GoogleNet [87]

Table 11 Publications relevant to GoogleNet

CNN model (year)	The work	Error rate	Dataset	Importance
GoogleNet 2014	Liu [87]	6.67%	ILSVRC14	Very close to human-level performance
GoogleNet 2014	David Silver [88]	Top-5 error of 6.67%	ILSVRC14	NN for computer vision
GoogleNet 2017	Yuhang Zhang [89]	FAR is 16.32	Dataset collected by QuickBird	Aircraft detection

Table 12 Publications relevant to VGGNet

CNN model (year)	The work	Error rate	Dataset	Importance
VGGNet 2014	Andrew [90]	Top-5 test error of 6.8%	ILSVRC	Features extraction
VGGNet 2015	Jifeng Dai [93]	mAP of 44.3%	PASCAL VOC, MSCOCO	Semantic segmentation
VGGNet 2015	Shaoqing Ren et al. [49]	Detection mAP of	PASCAL VOC 2012	Object detection
VGGNet 2015	Kaiming He [92]	detection mAP of 48.4%	ImageNet, ILSVRC 2015,, COCO	Image recognition
VGGNet 2017	Guangxing Han [94]	mAP of 73.2	PASCAL VOC 2007	Object detection
VGGNet 2017	Hao Wu [95]	mAP of 0.74%	112,902 images from Internet and personal database	Object recognition
VGGNet 2017	Shervin Minaee [96]	Best accuracy rate of 99.4%	Iris databases	Iris recognition
VGGNet 2017	Quanquan Li [97]	Highest precision is 97.26%	Caltech, Pascal VOC	Object detection

block width will be multiplied by 2 factor. This model utilized the Faster R-CNN technique, and for making it simple, they do not contribute the features between Faster R-CNN and RPN.

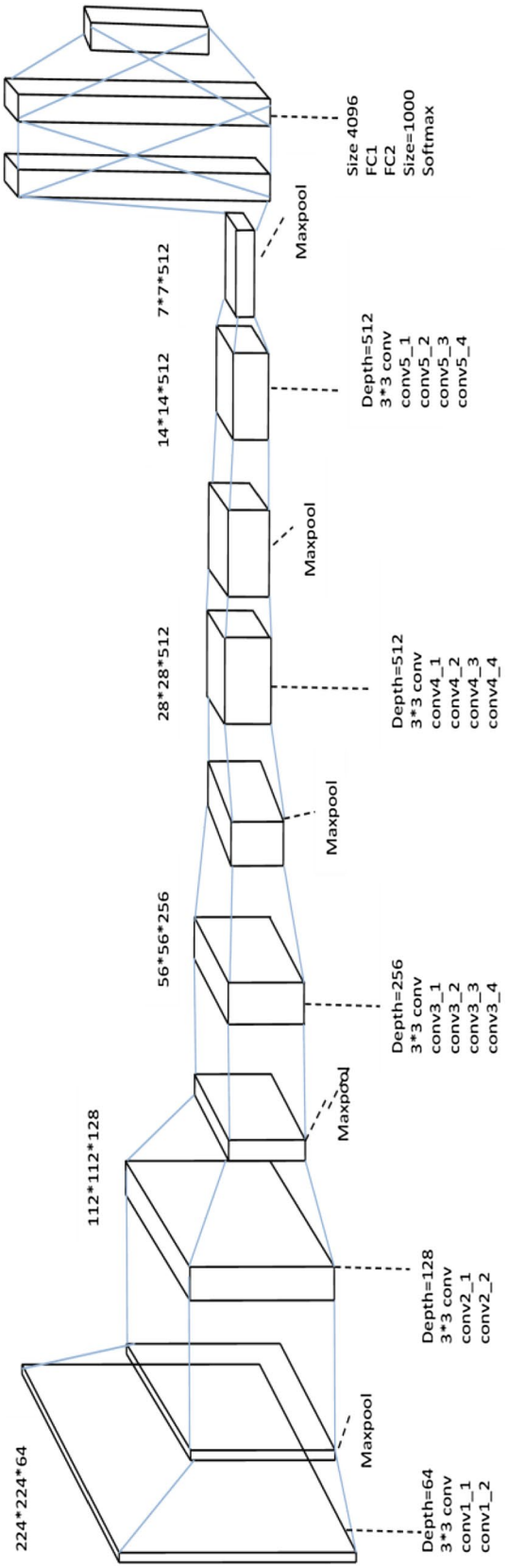

Fig. 5 Architecture of VGG19 [90]

Xie [102] proposed a method for DNN, which is known as aggregated residual transformations. This model is built by replicating a building block which collects the group of transformations along with the similar topologies. This model is the winner of the ILSVRC 2017 classification challenge. The experiments are carried out on ImageNet-5K set, and the COCO detection set and show better results than ResNet model. ResNeXt improves AP@0.5 by 2.1% and AP by 1.0%, without enhancing the complexity (Fig. 8).

Remarks on ResNeXt model

- It is multi-branch model which contains few hyperparameters.
- It contains new dimension, known as cardinality.
- Better results than ResNet (Table 15).

5.9 SENet model

This model is propounded to enhance the representational power of the network, and this is carried out by empowering it to achieve the dynamic channel-wise feature recalibration. This model mainly concentrates on the relationship between the channels and then propounded an innovative architectural system termed as: squeeze and excitation (SE) block (Fig. 9).

SENets assembled groundwork on ILSVRC 2017 categorization submission that won the first category and automatically diminishes the top-5 error to 2.251%. Hu [103] propounded a model known as squeeze and excitation. It enhances the interdependencies of the channel with no computational cost, and global average pooling is implemented on GPU. SENets come on the first position on ILSVRC 2017 classification submission and diminishes the top-5 error to 2.251% (Table 16).

Remarks on SENet model

- Use global average downsampling to achieve channel-wise statistics.
- Features are first moved across the squeeze action.
- Sample-specific activations are held in excitation operation.

5.10 MobileNet V1/V2

MobileNets V1 In this model [78], the normal convolution is replaced by the depthwise convolution followed by pointwise convolution which is called as depthwise separable convolution. It performs a solitary convolution on every color channel as opposed to joining every one of the three and smoothing it, and it can be done by utilizing the depthwise separable convolutions. This architecture was propounded by

Table 13 Publications relevant to ResNet

CNN model (year)	The work	Error rate	Dataset	Importance
ResNet 2015	Kaiming He [92]	3.57% error rate	ILSVRC 2015	Residual learning
ResNet 2016	Christian Szegedy [98]	3.08% top-5 error	ILSVRC 2012	Image recognition
ResNet 2017	Youngwan Lee [99]	1.33% and 2.83% lower error rates	KITTI dataset	Real-time object detection
ResNet 2018	Chang Liu [100]	Test accuracy 81.5%	UEC-256, Food-101	Dietary assessment
ResNet 2016	Jifeng Dai [93]	83.6% mAP on the test set	PASCAL VOC datasets	Object detection

Google. Overall architecture of the MobileNet is described as: having 30 layers with (1) convolutional layer with stride 2, (2) depthwise layer, (3) pointwise layer that doubles the number of channels, (4) depthwise layer with stride 2, (5) pointwise layer that doubles the number of channels, etc. (Table 17).

MobileNetV2 Basic idea of this model [104] is to utilize the depthwise separable convolution as proficient building block. It acquaints two new highlights with the design: (1) linear bottlenecks among the layers and (2) shortcut connections among the bottlenecks. It utilizes depthwise separable convolutions and contains three convolutional layers in the block. MobileNetV2 architecture is faster for a similar accuracy over the whole latency spectrum. Specifically, this model accomplishes higher accuracy by utilizing $2\times$ less operations and 30% less parameters, and it is around 30–40% quicker on a Google Pixel phone than MobileNetV1 architecture (Fig. 10).

Remarks on MobileNet architecture

- Utilizes depthwise separable convolutions.
- Two global hyperparameters are introduced foe efficient trade off among accuracy and latency.
- Smaller model size and smaller complexity.
- Outperforms GoogleNet and VGGNet.

5.11 DenseNet

This architecture is developed by [68] where every layer is coupled in feed forward manner. This model with L layer has $L(L+1)/2$ direct connections. It concatenate the output feature maps along with incoming feature maps; therefore, each layer acquires collective knowledge from all previous layers. This work is useful in terms of reducing the vanishing gradient problem, minimizing the number of parameters and concept of feature reuse. They attain state-of-the-art performance on several databases including CIFAR-100, ImageNet and SVHN with the only drawback that it utilizes a lot of extra memory, because of tensors concatenation.

Zhu [106] proposed a method named sparsely aggregated convolutional networks for the utilization of deeper layers. It shows the enhancements in efficiency over ResNet and DenseNet, on CIFAR and ImageNet datasets. Zhou [107] propounded a method named STDN, for detecting the multi-scale objects. It utilized DenseNet as the base model and achieves great results in terms of accuracy and speed (Table 18).

Remarks on DenseNet model

- Quite similar to ResNet, but has some fundamental differences.
- For reducing complexity and size, BN-ReLU—1×1 Conv is done before BN-ReLU—3×3 Conv.
- Diminishes the vanishing gradient problem.
- One of the lowest error rates on CIFAR/SVHN datasets.

5.12 Xception model

Xception architecture [108] from is the extended version of inception model, which is completely based on depthwise distinguishable convolutions, followed by pointwise convolution. They used the hypothesis that: spatial correlations and cross-channel correlations can be adequately decoupled. This architecture gives better results on ImageNet than Inception V3, ResNet-50, ResNet-101, ResNet 152, VGGNet (Figs. 11, 12).

Remarks on Xception model

- Uses depthwise separable convolutions.
- The number of connections are fewer, and the model is lighter.
- Shows better results than Inception V3 (Table 19).

5.13 NAS\PNAS\ENAS

Neural architecture search (NASNet) [109] is an algorithm for exploring the best model on neural network. It has attained significant results on semantic segmentation and image classification. For optimizing the configurations of a model, it makes the use of reinforcement learning search method. In progressive neural architecture search (PNAS) [109], sequential model-based optimization (SMBO)

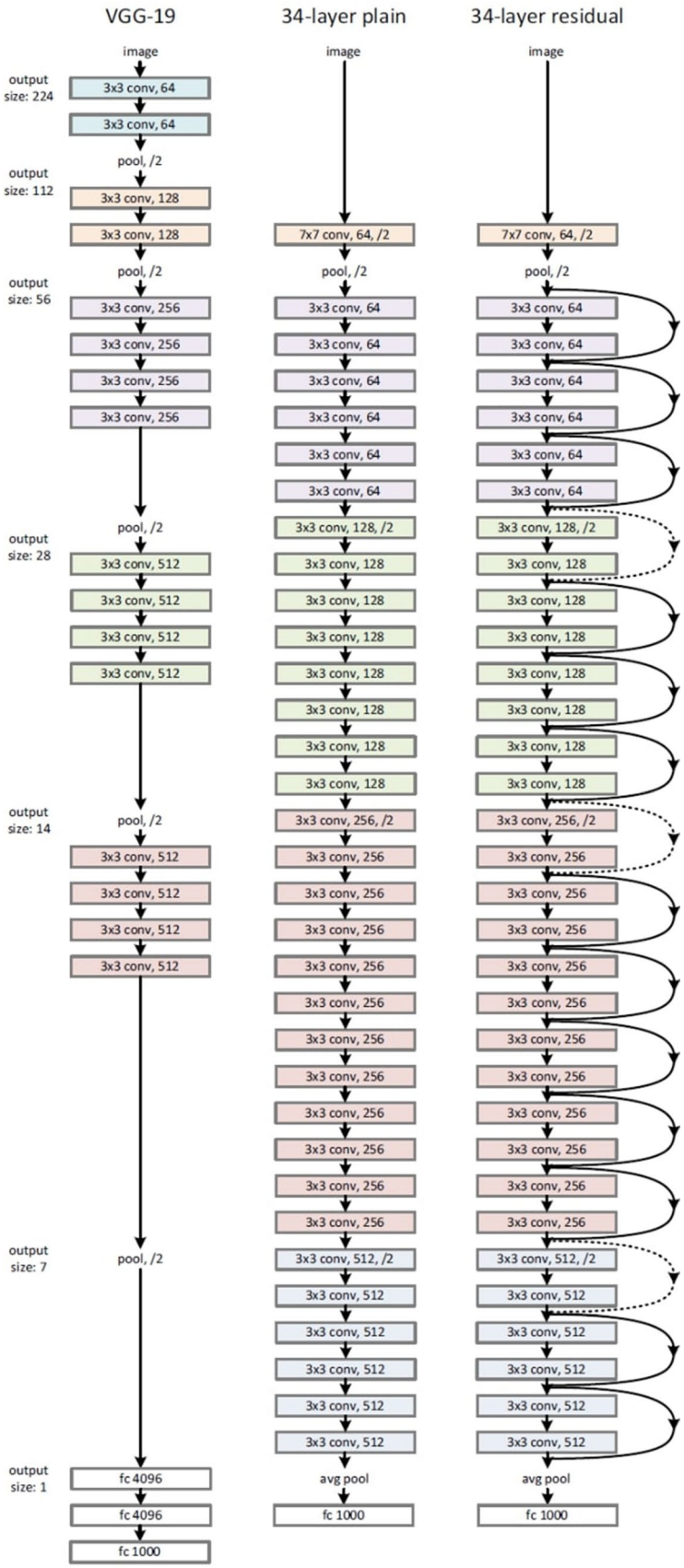

Fig. 6 Architecture of ResNet [92]

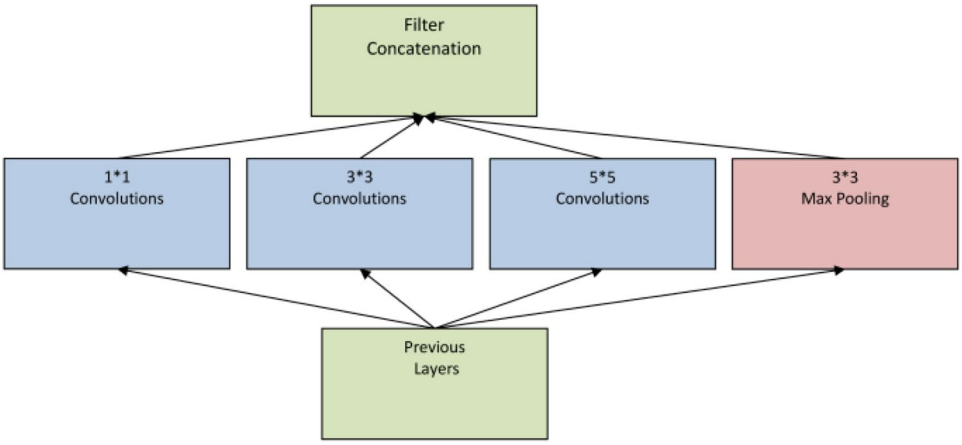

Fig. 7 Architecture of Inception model [87]

Table 14 Publications relevant to inception model

CNN model (year)	The work	Error rate	Dataset	Importance
Inception Model 2015	Christian Szegedy [98]	Top-5 error rate is 3.5%; top-1 is 17.3%	ILSVRC 2012	High-performance vision networks
Inception Model 2017	Xiaoling Xia [101]	mAP of 95%	Oxford-17 flower dataset	Flower classification

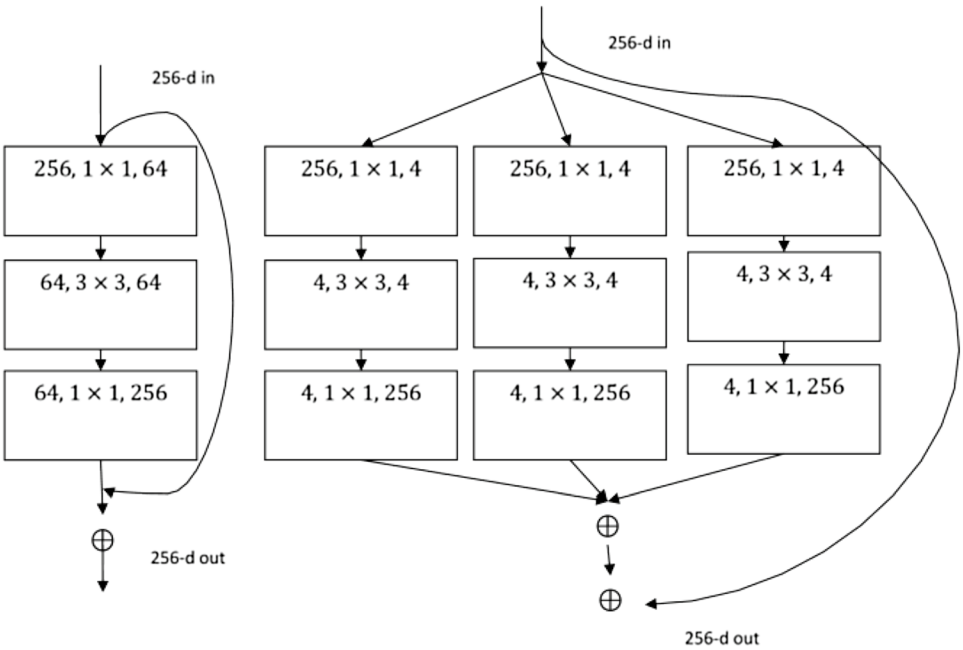

Fig. 8 Architecture of ResNeXt [102]

Table 15 Publications relevant to ResNeXt model

CNN model (year)	The work	Error rate	Dataset	Importance
ResNeXt Model 2017	Saining Xie [102]	Improves AP @ 0.5 by 2.1% AP by 1.0%	ImageNet-5K, COCO detection set	Object detection and instance segmentation

technique is utilized instead of reinforcement learning. The major objective of this graph is to convey that better CNN architecture search can be stimulated by utilizing a progressive approach, where learned prediction functions are

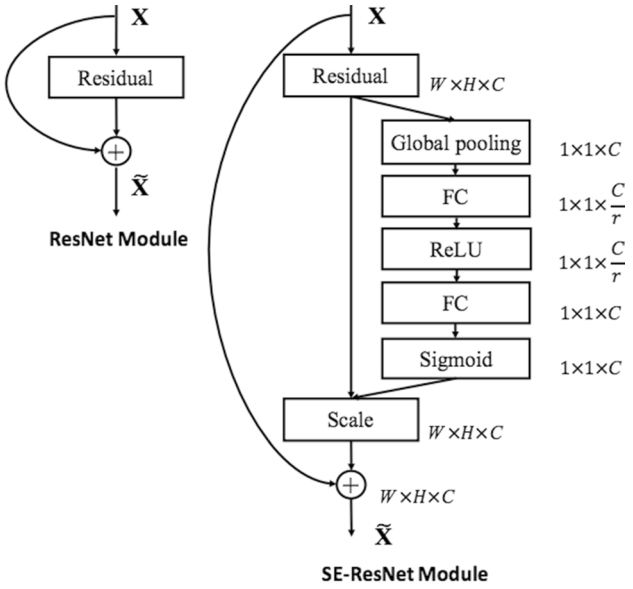

Fig. 9 ResNet and SE-ResNet module

merged with the space of progressively complex graph. This architecture has various merits over other models, i.e., it trains the basic structure very fast, to anticipate the quality of structures they ask for surrogate and the search space is partitioned into product of small scale search spaces (Table 20).

Efficient neural architecture search (ENAS) [110] makes use of macro- and microsearch techniques to generate the two types of neural network, namely the controller and the child model. The controller is a predefined RNN, where child model is the required CNN for classification of images. It decreases the computational time by 1000* via sharing of parameters within the model (Figs. 13, 14).

Remarks on PNAS model

- It utilizes sequential model-based optimization (SMBO) strategy.
- Structures are searched in increasing order complexity.
- Attains higher performance than SENet, NASNet5, etc.

Remarks on ENAS model

- NN architectures are discovered by searching for optimal subgraph.
- Utilizes much fewer GPU's hours.
- 1000 times less expensive than standard NAS.
- Attains 2.89% test error.

5.14 EfficientNet

It is a scaling technique [113] whose core id's are to utilize the effective compound coefficient to scale up CNNs in a progressively organized way. It consistently scales every dimension like depth, width and resolution, with a fixed set of scaling coefficients. Other than ImageNet, EfficientNets likewise transfer well and accomplish state-of-the-art accuracy exactness on 5 out of 8 broadly utilized datasets, while decreasing parameters by up to 21 × than existing ConvNets (Figs. 15, 16).

Remarks on EfficientNet architecture

- It event scales every dimension with a defined set of scaling coefficients.
- Huge reduction in parameters and computations.
- 5–10 times more efficient than present CNN's.
- 8.4 × smaller and 6.1 × faster on CPU inference (Table 21).

6 Transfer learning

This is one of the methods of machine learning where the model, which is created for an errand, is reutilized as the beginning stage for a model on a second errand. Transfer learning [115] can be used for the prophetic modeling issues. Two basis perspectives are described as:

1. Develop model approach.
2. Pre-trained model approach.

Table 16 Publications relevant to SENet Model

CNN model (year)	The work	Error rate	Dataset	Importance
SENet Model 2018	Jie Hu [103]	Top-5 error rate is 2.251%	COCO dataset	Dynamic channel recalibration

Table 17 Publications relevant to DenseNet Model

CNN model (year)	The work	Error rate	Dataset	Importance
MobileNet 2017	Howard [78]	70.6% accuracy	ImageNet	Mobile vision applications
MobileNet 2019	Mark Sandler [104]	75.32% mIOU	ImageNet, COCO	Object detection

Develop model approach The first step is to select the Source errand, and for that, the relevant predictive modeling issue should be chosen with an affluence of data and there should be a relation between input, output and the concepts, which are learned at the time of mapping from input to output data. Then, the source model is developed. In the further step, the skillful model should be developed for this first errand. To assure that this model is more appropriate than the original model, few feature learning should be executed.

Reuse model The model trained on the source errand would then be able to be utilized, as the beginning stage for a model on the second errand of interest. This may include utilizing all or parts of the model, contingent upon the modeling strategy utilized. Alternatively, tune model should be adjusted or refined on the input–output data accessible for the errand of interest.

Pre-trained model approach Select the source model. A pre-trained source technique is browsed from the accessible models. Numerous research foundations release the models on expansive and testing datasets that might be incorporated into the aspirant models.

Reuse model The pre-trained model would then be able to be utilized as the beginning stage, for a model on the second

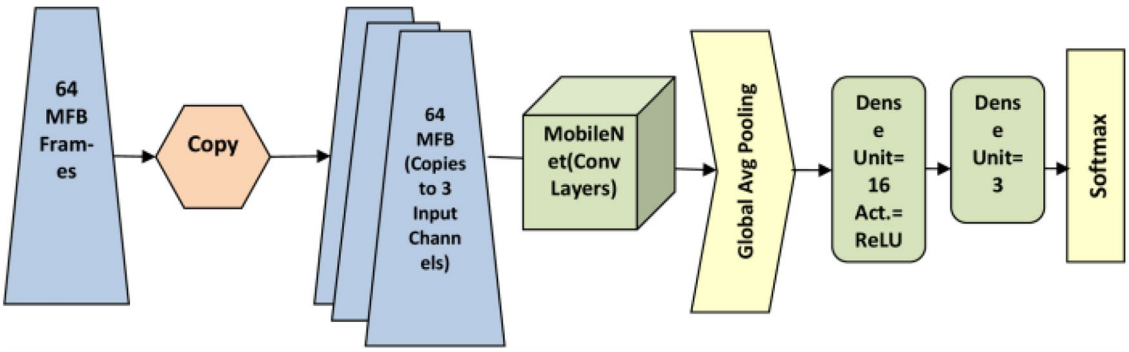

Fig. 10 Architecture of MobileNet [105]

Table 18 Publications relevant to DenseNet Model

CNN model (year)	The work	Error rate	Dataset	Importance
DenseNet 2018	Gao Huang [68]	Reduce error rate by approx. 2%	CIFAR100, ImageNet	Visual object recognition
DenseNet 2018	Peng Zhou [107]	Achieves 80.9% mAP on VOC 2007	PASCAL VOC2007	Object detection
DenseNet 2018	Ligeng Zhu [106]	Error rate of 18.22%	CIFAR 100, ImageNet	Visual recognition

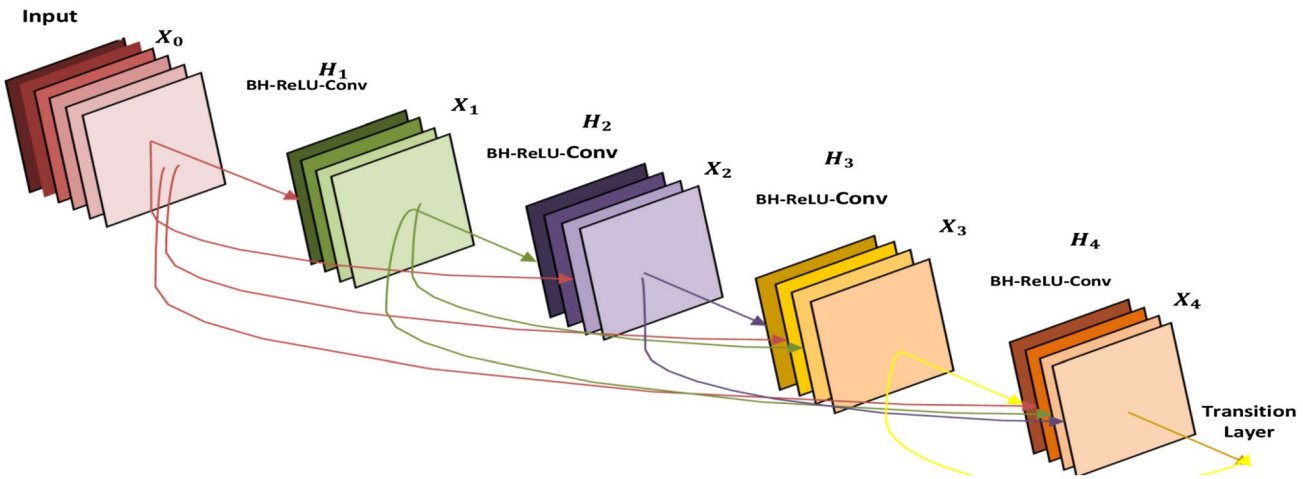

Fig. 11 Architecture of DenseNet model [68]

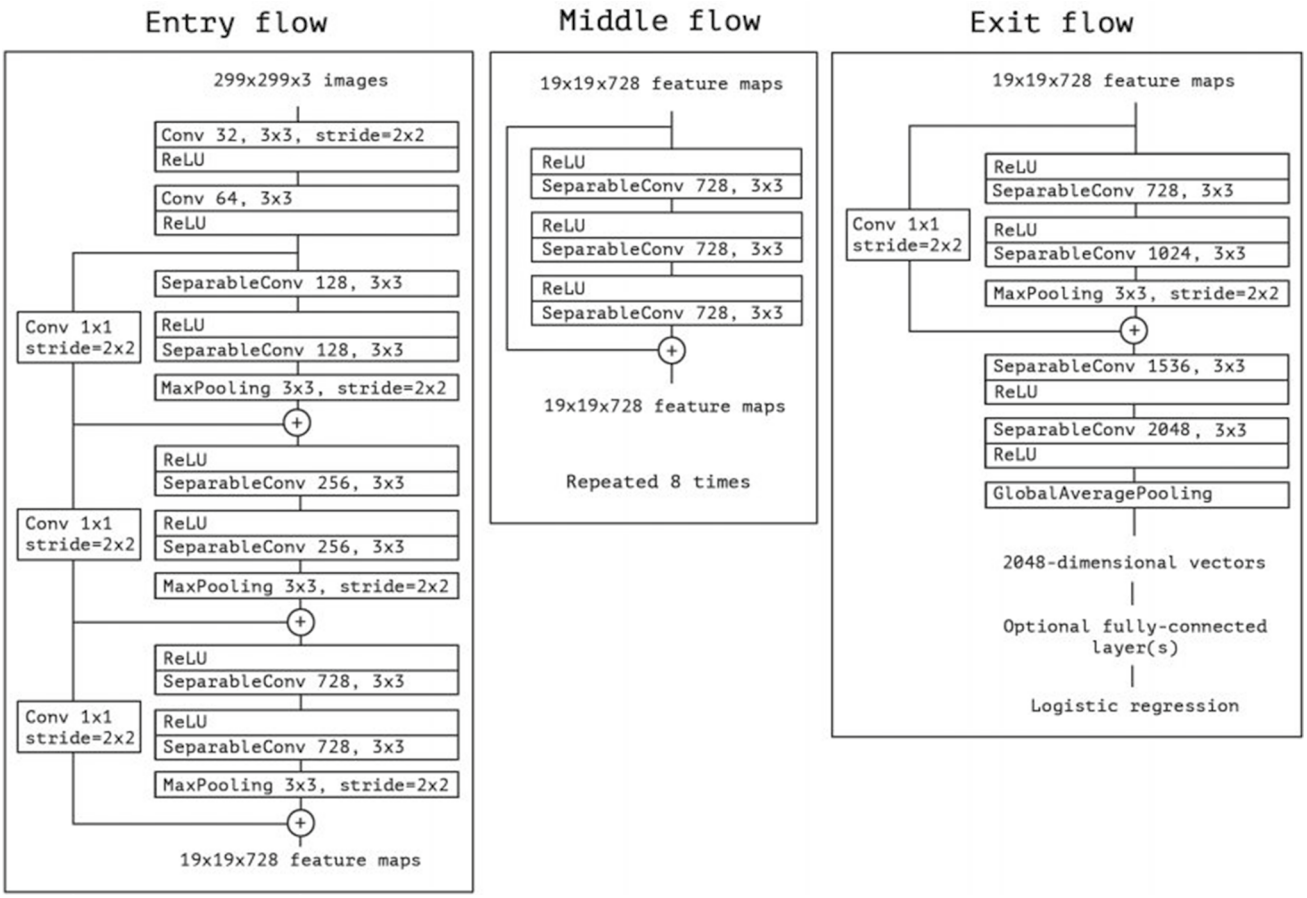

Fig. 12 Architecture of Xception model [108]

errand of interest. This may include utilizing all or parts of the model, contingent upon the modeling method utilized. Alternatively, tune model should be adjusted or refined on the information yield match, accessible or concentrated for the errand of interest.

All in all, there are two sorts of transfer learning with regards to deep learning:

1. Transfer learning via feature extraction.
2. Transfer learning via fine-tuning.

When accomplishing feature extraction, the pre-trained network is treated as a discretionary feature extractor, which enables the input image to transmit forward, halting at pre-indicated layer and taking the outputs of that layer as the features.

Fine-tuning [116] delicately adjusts the weights of the pre-trained model. It necessitates that the model architecture should be updated itself by evacuating the previous full connected layer heads, giving new, naturally introduced ones, and afterward training the latest FC layers to anticipate the information classes.

There are several ways to **Fine tune** the model:

- Linear SVM on top of bottleneck features: SVM can be trained on the top of the convolutional layers just before the FC Layers.
- Just replace and train the last layer: Train only last few layers.
- Freeze, pre-train and fine tune (FPT) weights of the initial layers of the network can be frizzed, while training can be done only on the higher layers.
- Train all the layers.

Table 19 Publications relevant to Xception Model

CNN model (year)	The work	Error rate	Dataset	Importance
Xception 2017	Francois Chollet [108]	Top-5 accuracy rate is 0.945	ImageNet	Image classification

Table 20 Publications relevant to PNASNet\ENASNet

CNN model (year)	The work	Error rate	Dataset	Importance
PNAS 2018	Chenxi Liu [41]	Top-5 accuracy of 96.2%	CIFAR10, ImageNet CIFAR10	Performance prediction
ENAS 2018	Hieu Pham [110]	2.89% error rate	CIFAR10	Automatic model design
DetNAS 2019	Yukang Chen [111]	Achieves 40.0 mAP on COCO	COCO	Image classification

Deep transfer learning Deep transfer learning [117] is the new approach which incorporates the deep learning methods with the transfer learning methods and examines about how to use the information from different fields by deep neural model. It is defined as: Given a transfer learning task as: Ds, Ts, Dt, Tt, fT, it is a deep transfer learning task where fT is a nonlinear function, which reflects the deep neural network. It is categorized into four classifications: instance-based deep exchange, mapping-based deep transfer learning, network-based deep transfer learning and adversarial-based deep transfer learning.

Applications of transfer learning

- Transfer learning for NLP.
- Transfer learning for audio/speech.
- Transfer learning for computer vision.

Transfer learning advantages

- Helps solve complex real-world problems with several constraints.
- Tackle problems like having little or almost no labeled data availability.
- Ease of transferring knowledge from one model to another based on domains and tasks.
- Provides a path toward achieving artificial general intelligence some day in the future!

Transfer learning challenges

- Negative transfer: Negative transfer alludes to situations where the transfer of information from the source to the objective does not prompt any improvement, but instead

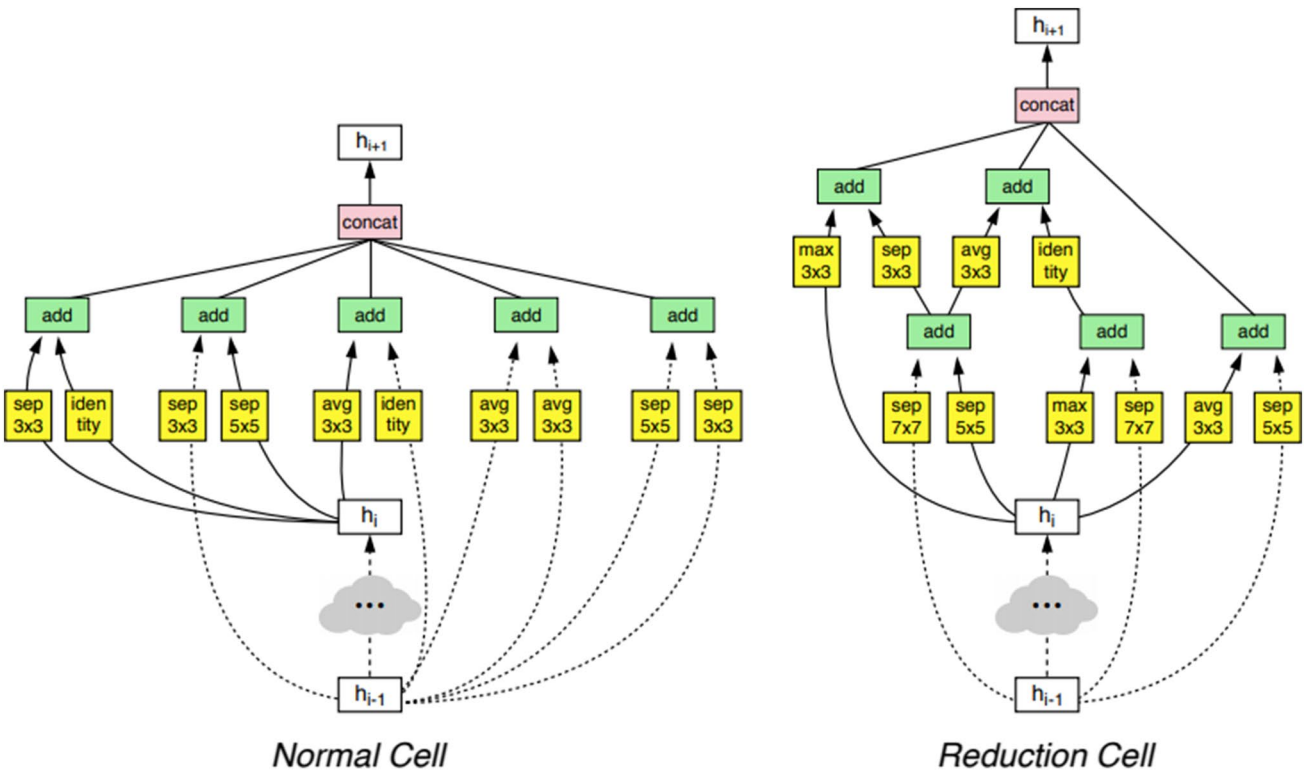

Fig. 13 Architecture of NAS architecture [112]

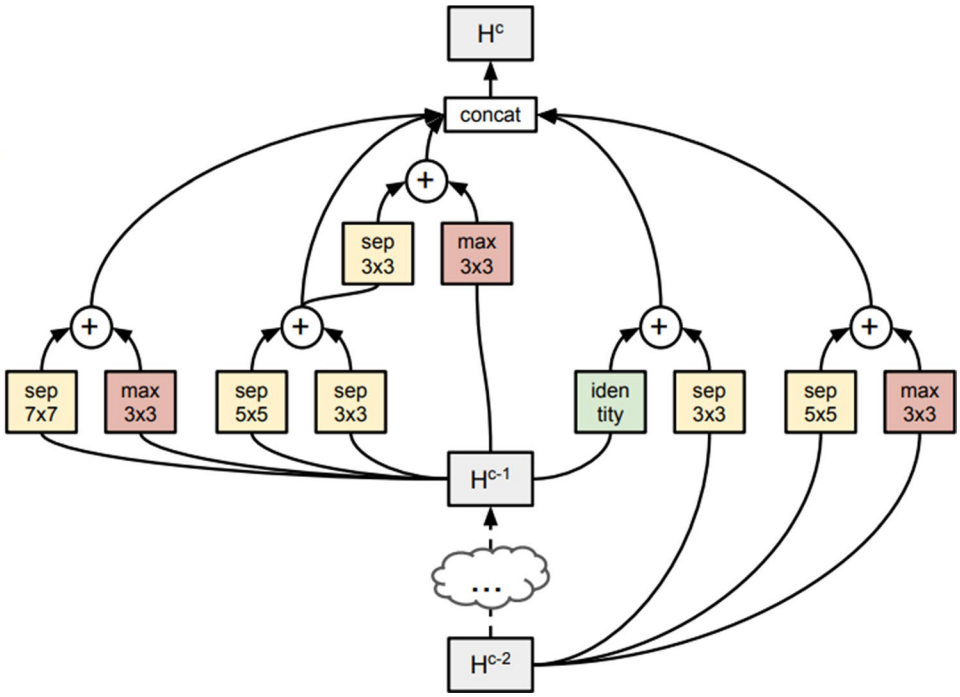

Fig. 14 Architecture of PNAS architecture [41]

purposes a drop in the general execution of the target errand.

- Transfer bounds: Evaluating the transfer in transfer learning is additionally significant that influences the nature of the transfer and its viability. To check the amount for the transfer, utilized Kolmogorov complexity is utilized which demonstrates certain hypothetical bounds to break down the transfer learning and measure relatedness between tasks.

Highlights of transfer learning

- This learning works in deep learning if the features of models which are learned from the principal errand are general.
- Here, firstly we train a basic framework in light of a base dataset and errand, and after that, we remodel the learned features, or exchange them, to a second objective system to be trained on an objective dataset and errand.
- This method will learn to work if the features are generic, which means appropriate to both base and target errands, rather than particular to the base errand.

7 Applications of CNN for object detection

In this section, we review the primary applications of CNN in the field of object detection. A remarkable work toward the applications of deep learning has grown steadily in the last few years. Detection of the object is a demanding issue because of the major amount of variables or factors that must be taken care of, like an assortment of available objects

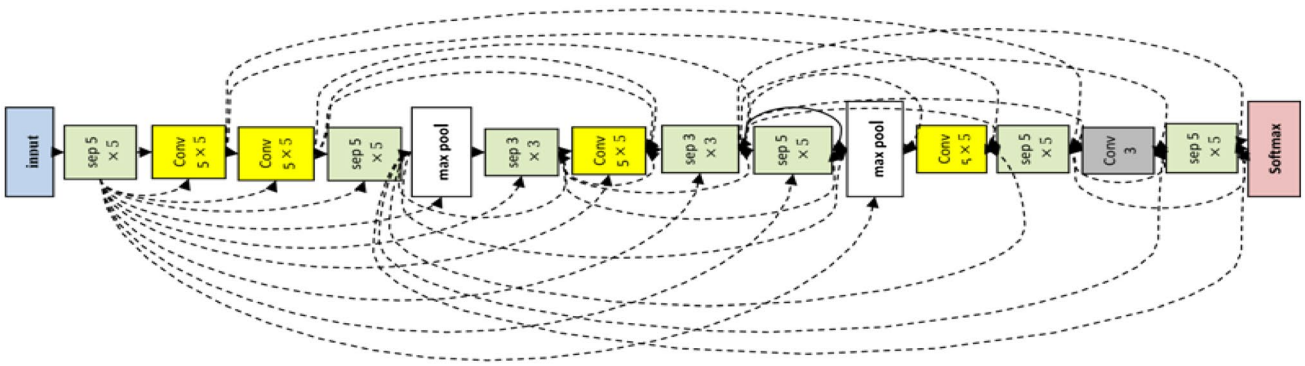

Fig. 15 Architecture of ENAS architecture [110]

Fig. 16 Architecture of EffcientNet architecture [114]

shapes and colors, lighting conditions, occlusion and so forth. In this section, we primarily focus on: wild animal detection, arms detection, human being detection and miscellaneous object detection.

7.1 Wild animal detection

Animal detection-based researches are helpful for those applications which are used in real life and these techniques are useful for those researches which are associated with the locomotive behavioral of the intended animal and furthermore to counteract the dangerous animal interruption in the residentiary zone.

Nguyen [34] propounded a technique for an automated animal recognition in the wildlife. The technique is based on automated wildlife monitoring procedure. They have achieved the accuracy of 96.6 on the animal images. Nguyen [34] propounded a technique for automatically identifying, computing and recognizing the wild animals on camera-trap snapshot Serengeti dataset, and they have got the accuracy of 93.8. Guignard [118] demonstrated the method for animal detection by utilizing a remote camera dataset. They train the multi-layer perceptron network by considering R-CNN, and finally, bounding boxes are applied. They got the accuracy of 92 on RGB images. Alexander [119] propound an automatic wild animal monitoring system by using DCNN. They achieved the accuracy of 88.9% in top-1 and 98.1% accuracy in the top-5 evaluation set, with the help of the residual network. Okafor [120] examined a relative study of DCNN and bag of visual words (BOW), which are the alternative for detecting the wild animals. On wild animal dataset, they achieved the topmost accuracy of 99.38%.

Fang [121] described the complication of detecting the animals in natural habitat from aerial videos. They propounded a technique which combines the neural network with the robust classifier and then discriminates the animals from background subclasses. Yu [122] demonstrated an automated species recognition technique to capture the wildlife images. They have integrated the local feature classifiers, namely SIFT and cLB and SVMs, to categorize the global feature. They achieved the classification accuracy of 82.

7.2 Human being detection

Human activity recognition is a method to perceive different activities of the human through the outer sensors, for example, inertial or video sensors. As of late, HAR has invoked critical enthusiasm among analysts in the regions of human care and life care services, since it permits automated checking of patient activities.

Abousaleh [53] proposed a method known as comparative region convolutional neural network (CR-CNN), for facial

Table 21 Publications relevant to EfficientNet

CNN model (year)	The work	Error rate	Dataset	Importance
EfficientNet 2019	Mingxing Tan [113]	84.4% top-1, 97.1%	ImageNet top-5 accuracy	Scaling method

age estimation. Moreover, they propounded to include the technique of auxiliary coordinates (MAC) for training which diminishes the ill-conditioning issue of deep neural network and manages a proficient and appropriated enhancement. Their future work includes further improvement in the baseline election because achieving an adequate baseline is vital in their provisional advent.

Ranjan [22] propounded a deep learning technique for understanding the Faces. The paper presents deep convolutional neural network (DCNN)-based techniques for detection of faces, and it can be partitioned into two categories: sliding window approach and the region-based approach. But here the main issues to be addressed are as follows: minimizing the reliance on huge training database, controlling data bias and mortification in training data and minimizing the time of training when the network is wider and deeper (Table 22).

Zhang [123] propounded a technique to survey the conceivable connections between the images of face and personality characteristics. With the altogether of three residual and one Inception-v4, they accomplish 3.08 error on the ImageNet dataset.

Shan [64] propounded a technique for automatic recognition of facial expressions on the basis of Deep CNN. The proposed framework is made up of the input module, the pre-preparing module, the recognition module and the output module. They have achieved the following accuracy on JAFFE and CK+ datasets: 76.7442 and 80.303.

Zhao [124] proposed a method for the recognition of facial expressions by joining multi-layer perceptron and DBNs. Accuracy was tested on two datasets, i.e., JAFFE and the Cohn–Kanade. The recorded accuracy is: 88.57 for 16×16 images, 89.05 for 32×32 images and 90.95 for 64×64 images.

7.3 Small arm (gun) detection

The vast majority of the criminal actions are occurred by utilization of handheld arms especially gun, revolver, and knife. A few overviews uncover that handheld firearm is the preeminent weapon utilized for differing violations like thievery, assault, etc. Automated strategies for identification of the weapon have begun to develop lately, for the most part for the reason to stop criminal exercises. Amid late years, an expansion in the no. of incidents with the utilization of hazardous devices in common places can be noticed (Table 23).

Grega [127] proposed a method for the detection of firearms and knives. Their methods are based on fuzzy classifier and SVM classification. In this paper, they created an independent dataset for firearms and knives and get the maximum accuracy of 91%.

Olmos [37] proposed a self-activating method for the detection of guns by applying Faster R-CNN. In this paper, they have used two perspectives: sliding window and region approaches and get the precision of 84.21. Lai [128] proposed a method for detection of handheld weapons in real-time environment, by using convolutional neural networks. Here the implementation of overfeat network is done on tensorflow, and they get training accuracy of 93 and test accuracy of 89. Muhamad [129] propounded an automatic system of the gun turret, by using deep multi-layer CNN. First of all, the object is manually detected, and then, those patches are extricated which are enclosed by the next frames. Their proposed technique will track the target without misplacing it. Glowacz [130] propounded a knife detection method by using Harris interest point detector. They got the classification accuracy of 92.5 and false positive rate of 0. Yin [36] propounded a deep learning technique for intrusion detection. The propounded method is applied to NSL-KDD dataset and gets 83.28 accuracy rate on testing data. Farahnakian [131] propounded a deep autoencoder method for the detection of intrusion. The method is applied on the KDD-CUP 99 dataset and achieved the accuracy of 94.71%. Recently, some new works are done in the area of gun detection; for example, Castillo [132] proposed a detection method for cold steel weapons. The basic idea of this paper is to evolve an automatic cold weapon detection method in video inspection and to propound preprocessing methods known as darkening and contrast at learning and test stages (DaCoLT). They obtained an F1 measure of 93%. Olmos [133] proposed a binocular image fusion method for processing a vast number of false positives in the RGB surveillance videos. By obtaining the videos from the symmetric cameras, a disparity map is calculated. Then, by removing the background objects, preselection of areas is carried out. The final detection is carried out by applying the obtained mask to one of the authentic frames. They minimize the false positives by 49.47% (Table 24).

7.4 Miscellaneous object detection

Detection of an object is one of the critical applications in the area of computer vision, and it becomes the focal

Table 22 Summary of previous work on wild animal detection

System (year)	NN	Dataset	Accuracy	Application
Hung Nguyen [34] 2017	Deep learning, CNN	Single-labeled database from Wildlife Spotter project	mAP of 96.6% mAP of 96.6	Monitoring of wild animals
Anh Nguyen [35] 2017	DL, CNN	Snapshot serengeti dataset	mAP of 99.3%	Wild animals identification
Lewis Guignard [118] 2016	NN	Snapshot serengeti dataset	mAP of 92%	Animal identification
Alexander Gomez [119] 2016	Deep CNN	Snapshot serengeti dataset	Top-1 accuracy of 88.9% and 98.1% in top-5	Identification of animal species
Emmanuel Okafor [120] 2016	CNN	Wild animal dataset	Top-1 loss rate of 0.07	Wild animal recognition
Yunfei Fang [121] 2016	Neural network	Aerial video acquired from natural environment	True positive rate 83.2%	Animal categorization

Table 23 Summary of previous work on human being

System (year)	NN (year)	Dataset	Accuracy	Application
Abousaleh [53] 2016	R-CNN	FG-NET aging database	Improvement of 23.20% (on MORPH), 13.24% (on FGNET) and 4.74% (IoG)	Facial age estimation
Rajeev Ranjan [22] 2018	CNN	VGGFace, MegaFace, CASIA-WebFace	TPE of 98.8%	Face detection
Yuanyuan Ding [63] 2017	CNN	CASIA-WebFace, LFW dataset	93.78% highest averaged recognition rate	Face recognition under noise
Xiaoming Zhao [124] 2015	DBN, Deep learning	JAFFE and Cohn–Kanade database	Highest recognition accuracy is 90.95%	Facial expression recognition
Yaniv Taigman [125] 2014	DNN	Faces in the wild (LFW) dataset, SFC dataset, YTF dataset	mAP of 97.35%	3D face modeling
ByungIn Yoo [126] 2018	CNN	MORPH-II, FG-Net dataset	Performance measure of 99.28%	Age estimation, gender recognition

point of research, and convolution neural network has gained extraordinary ground in object detection. Xu [51] propounded a technique for the detection of RGB-D detection. This method mainly contains two phases: estimation of object and recognition of an object in a region-wise manner. The training and testing are performed on two datasets, i.e., SUN RGB-D and NYU Depth v2, and their method gets the mAP of 52.9. Girshick [52] and Ren [49] propounded Fast R-CNN and Faster R-CNN, respectively. They get the much better quality of detection (mAP) than R-CNN, SPPNet, and for feature caching, there is no requirement for storage of disk. Fang [54] propounded a deep learning technique for facial and emotion recognition. They get the facial recognition accuracy of 97.55 and emotion recognition accuracy of 90.97%.

Milyaev [23] propounded a technique for objects detection in noisy images. Training and testing are carried out on two benchmarks, namely Pascal Visual Object Classes and KAIST multispectral pedestrian detection benchmarks. Their work outperforms other denoising methods by adding significant distortions by synthetic and real noise. Ning [134] propounded a deep learning method for recognition, detection, and segmentation of white background photographs and got an accuracy of 96 in recognition and 94 in detection. Chin [135] propounded an object detection technique, and they research the exchange off, among the accuracy and speed with domain particular approximations. Cao [136] propounded a deep neural network technique for detection of objects in real-time videos. The data are trained on the CIFAR-10 DATASET. The propounded strategy can accelerate the system by up to 16 times while keeping up the performance of object detection (Table 25).

8 Conclusion

In this article, we have reviewed deep learning-based object detection methods. First of all, various generic CNN models are proposed along with the LeNet, AlexNet, VGG net, GoogleNet, ResNet, ResNeXt, SENet, DenseNet, etc. A brief summary of transfer learning approaches was also given. Then, the applications of deep learning in the areas

Table 24 Summary of previous work on arms detection

System (year)	NN (year)	Dataset	Accuracy	Application
Roberto Olmos [37] 2017	DL, CNN	ImageNet	Zero false positives, 100% recall and precision 84.21%	Pistol detection
Justin Lai [128] 2016	CNN	IMFDB	MAP of 93% on training data and 89% on testing data	Gun detection
Anwar [129] 2017	DL, CNN	V-REP1 robotic simulation	False positive rate is 0 system	Automatic targeting system of gun turret
Chuanlong Yin [36] 2017	RNN-IDS	NSL-KDD dataset	Detection rate 97.09% on testing dataset	Intrusion detection
Fahimeh Farahnakian [131] 2018	DNN stacked autoencoders	KDD-CUP 99 dataset	Achieved detection rate of 94.71%	Intrusion detection

Table 25 Summary of previous work on object detection

System (year)	NN	Dataset	Accuracy	Application
Xiangyang Xu [51] 2017	Faster R-CNN, Caffe	NYU Depth v2 and SUN RGB-D	53.0% mAP on test dataset	RGB-D object detection
Ross Girshick [52] 2015	Deep VGG16, R-CNN	PASCAL VOC 2012	mAP of 65.7%	Object detection
Shaoqing Ren [49] 2015	SPPNet, fast R-CNN VGG-16	PASCAL VOC 2007, PASCAL 2012	73.2% mAP on testing dataset	Object detection
Xing Fang [54] 2017	CNN	Radboud faces database (RaFD)	Precision for facial recognition is 97.55% and for emotion recognition 90.97%	Identify activated pixels from feature maps
Milyaev [23] 2016	Standard deformable parts model	Pascal VOC 2007	mAP of 56.6%, and for test data PSNR results are 26.5%	Low-cost methods for image denoising
Xiaofeng Ning [134] 2017	Faster R-CNN, RPN (region proposal network)	VOC 2012 dataset	mAP of 96%	White background photographs, recognition segmentation
Cong Tang [32] 2017	DNN, R-CNN	ILSVRC 2012	mAP is 74.3% on test dataset	Object detection
Yan Liu [137] 2016	DL model	MNIST, CIFAR-10, BioID face dataset	Max. Avg. recognition accuracy is 98.92%	Image recognition with incomplete data
Honglak Lee [16] 2009	DBN, RBM	Kyoto natural image dataset, Caltech-101	Obtained 0.8% test error MNIST	Learns useful high-level visual features
Kaiming He [60] 2015	Deep residual nets, VGG	ImageNet, ILSVRC 2015	Top-5 error rate of 3.57%	Image recognition
Wei Liu [41] 2016	R-CNN	PASCAL VOC, COCO, ILSVRC	74.3% mAP	Detecting objects
Enver Sangineto [138] 2018	Faster R-CNN	ILSVRC, Pascal VOC 2007, 2010	Attains mAP of 12.13%	Object detection
Shabab Bazrafkan [139] 2018	DL	CASIA Thousand, Bath 800 datasets	Segmentation results are 98.55%	Detection of handheld devices
Hongyu Xu [140] 2018	DCNN, deep regionlets COCO dataset	PASCAL VOC and Microsoft	mAP of 82.0%	Object detection

of object detection, such as animal detection, handheld arms detection, human detection and some miscellaneous object detection, are introduced. Deep learning-based object detection methods have achieved astounding progress in recent years, so it will remain an active research area. In future directions, we can design a CNN model, by considering one or more existing models or by modification of the current models to optimize the object detection problems. We can also evaluate the system for reducing the false positive rate or by preprocessing the images and using videos. Further analysis and study of major CNN models will be carried out.

References

1. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)
2. Hong, Z.: A preliminary study on artificial neural network. In: 2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference, vol. 2, pp. 336–338 (2011)
3. Wang, X.J., Zhao, L.L., Wang, S.: A novel SVM video object extraction technology. In: 2012 8th International Conference on Natural Computation, pp. 44–48. IEEE (2012)
4. Rish, I.: An empirical study of the naive Bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3, no. 22, pp. 41–46 (2001)
5. Islam, N., Zeeshan I., Nazia N.: A survey on optical character recognition system. arXiv preprint arXiv:1710.05703 (2017)
6. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial networks. arXiv:1406.2661 (2014)
7. Besbinar, B., Alatan, A.A.: Visual object tracking with autoencoder representations. In: 2016 24th Signal Processing and Communication Application Conference (SIU), pp. 2041–2044 (2016)
8. Ma, X., Geng, J., Wang, H.: Hyperspectral image classification via contextual deep learning. EURASIP J. Image Video Process. **2015**(1), 20 (2015)
9. Hinton, G.: A practical guide to training restricted Boltzmann machines. Momentum **9**(1), 926 (2010)
10. Shin, H., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)
11. Li, W., Fu, H., Yu, L., Gong, P., Feng, D., Li, C., Clinton, N.: Stacked Autoencoder-based deep learning for remote-sensing image classification: a case study of African land-cover mapping. Int. J. Remote Sens. **37**, 5632–5646 (2016)
12. Vincent, P.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**, 3371–3408 (2010)
13. Feng, F., Wang, X., Li, R.: Correspondence autoencoders for cross-modal retrieval. ACM Trans. Multimed. Comput. Commun. Appl. **12**(1), 1–22 (2015)
14. Hutchison, D.: LNCS 8588—Intelligent Computing Theory. Springer, Berlin (2014)
15. Koushik, J.: Understanding convolutional neural networks. arXiv preprint arXiv:1605.09081 (2016)
16. Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 609–616. ACM (2009)
17. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. **36**, 193–202 (1980)
18. Papakostas, M., Giannakopoulos, T., Makedon, F., Karkaletsis, V.: Short-term recognition of human activities using convolutional neural networks. In: 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 302–307. IEEE (2016)
19. Yudistira, N., Kurita, T.: Gated spatio and temporal convolutional neural network for activity recognition: towards gated multimodal deep learning. EURASIP J. Image Video Process. **2017**, 85 (2017)
20. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2011)
21. Zhou, X., Gong, W., Fu, W., Du, F.: Application of deep learning in object detection. In: 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), pp. 631–634. IEEE (2017)
22. Ranjan, R., Sankaranarayanan, S., Bansal, A., Bodla, N., Chen, J.-C., Patel, V.M., Castillo, C.D., Chellappa, R.: Deep learning for understanding faces: machines may be just as good, or better, than humans. IEEE Signal Process. Mag. **35**(1), 66–83 (2018)
23. Milyaev, S., Laptev, I.: Towards reliable object detection in noisy images. Pattern Recognit. Image Anal. **27**(4), 713–722 (2017)
24. Zhou, X., Gong, W., Fu, W., Du, F.: Application of deep learning in object detection, pp. 631–634 (2017)
25. Druzhkov, P.N., Kustikova, V.D.: A survey of deep learning methods and software tools for image classification and object detection. Pattern Recognit. Image Anal. **26**(1), 9–15 (2016)
26. Sze, V., Chen, Y.-H., Yang, T.-J., Emer, J.S.: Efficient processing of deep neural networks: atutorial and survey. Proc. IEEE **105**, 2295–2329 (2017)
27. Park, S.U., Park, J.H., Al-masni, M.A., Al-antari, M.A., Uddin, Z., Kim, T.: A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services. Procedia Comput. Sci. **100**, 78–84 (2016)
28. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Sequential deep learning for human action recognition. In: International workshop on human behavior understanding, pp. 29–39. Springer, Berlin, Heidelberg (2011)
29. Zhao, X., Shi, X., Zhang, S.: Facial expression recognition via deep learning. IETE Tech. Rev. **32**(5), 347–355 (2015)
30. Xie, S., Yang, T., Wang, X., Lin, Y.: Hyper-class augmented and regularized deep learning for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2645–2654 (2015)
31. Floyd, M.W., Turner, J.T., Aha, D.W.: Using deep learning to automate feature modeling in learning by observation: a preliminary study. In: 2017 AAAI Spring Symposium Series
32. Tang, C., Feng, Y., Yang, X., Zheng, C., Zhou, Y.: The object detection based on deep learning. In: 2017 4th International Conference on Information Science and Control Engineering (ICISCE), pp. 723–728 (2017)
33. Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Hasan, M., Van Esesn, B.C., Awwal, A.A.S., Asari, V.K.: The history began from AlexNet: a comprehensive survey on deep learning approaches. arXiv:1803.01164 (2018)
34. Nguyen, H., Maclagan, S.J., Nguyen, T.D., Nguyen, T., Flemons, P., Andrews, K., Ritchie, E.G., Phung, D.: Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 40–49. IEEE (2017)
35. Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J.: Automatically identifying, counting, and describing wild animals incamera-trap images with deep learning. Proc. Nat. Acad. Sci. **115**(25), E5716–E5725 (2018)
36. Yin, C., Zhu, Y., Fei, J., He, X.: A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access **5**, 21954–21961 (2017)
37. Olmos, R., Tabik, S., Herrera, F.: Automatic handgun detection alarm in videosusing deep learning. Neurocomputing **275**, 66–72 (2018)
38. Lee, J., Bang, J., Yang, S.I.: Object detection with sliding window in images including multiple similar objects. In: 2017 International Conference on Information and Communication Technology Convergence (ICTC), pp. 803–806 (2017)

39. Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., Gao, R.X.: Deep learning and its applications to machine health monitoring. Mech. Syst. Signal Process. **115**, 213–237 (2019)
40. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2015)
41. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision, pp. 21–37. Springer, Cham (2016)
42. Li, Y., Ren, F.: Light-Weight RetinaNet for Object Detection. arXiv preprint arXiv:1905.10011 (2019)
43. Lin, T.-Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007 (2017)
44. Lin, T.-Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. CoRR. arXiv:1612.03144 (2016)
45. Zhiqiang, W., Jun, L.: A review of object detection based on convolutional neural network. In: 2017 36th Chinese Control Conference (CCC), pp. 11104–11109 (2017)
46. Zhao, B.: A survey on deep learning-based fine-grained object classification and semantic segmentation. Int. J. Autom. Comput. **14**, 119–135 (2017)
47. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
48. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3150–3158 (2015)
49. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing System, pp. 91–99 (2015)
50. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
51. Xu, X., Li, Y., Wu, G., Luo, J.: Multi-modal deep feature learning for RGB-D object detection. Pattern Recognit. **72**, 300–313 (2017)
52. Girshick, R., Donahue, J., Darrell, T., Malik, J.: 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 (2014)
53. Abousaleh, F.S., Lim, T., Cheng, W.H., Yu, N.H., Anwar Hossain, M., Alhamid, M.F.: A novel comparative deep learning framework for facial age estimation. EURASIP J. Image Video Process. **2016**(1), 47 (2016)
54. Fang, X.: Understanding deep learning via back-tracking and deconvolution. J. Big Data **4**, 40 (2017)
55. Mnih, V., Heess, N., Graves, A.: Recurrent models of visual attention. In: Advances in Neural Information Processing Systems, pp. 2204–2212 (2014)
56. Wang, A., Lu, J., Cai, J., Cham, T., Wang, G.: Large-margin multi-modal deep learning for RGB-D object recognition. IEEE Trans. Multimed. **17**(11), 1887–1898 (2015)
57. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
58. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634 (2015)
59. Hua, Y., Alahari, K., Schmid, C.: Online object tracking with proposal selection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3092–3100 (2015)
60. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
61. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4507–4515 (2015)
62. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Med. Image Anal. **35**, 18–31 (2017)
63. Ding, Y., Cheng, Y., Cheng, X., Li, B., You, X., Yuan, X.: Noise-resistant network: a deep-learning method for face recognition under noise. EURASIP J. Image Video Process. **2017**(1), 43 (2017)
64. Shan, K., Guo, J., You, W., Lu, D., Bie, R.: Automatic facial expression recognition based on a deep convolutional-neural-network structure. In: 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), pp. 123–128 (2017)
65. Wang, J.G., Mahendran, P.S., Teoh, E.K.: Deep affordance learning for single- and multiple-instance object detection. In: TENCON 2017-2017 IEEE Region 10 Conference, pp. 321–326 (2017)
66. Tian, B., Li, L., Qu, Y., Yan, L.: Video object detection for tractability with deeplearning method. In: 2017 Fifth International Conference on Advanced Cloud and Big Data (CBD), pp. 397–401 (2017)
67. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 764–773 (2017)
68. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
69. Han, J., Zhang, D., Cheng, G., Liu, N., Xu, D.: Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Process. Mag. **35**(1), 84–100 (2018)
70. Babaee, M., Tung, D., Rigoll, G.: A deep convolutional neural network for video sequence background subtraction. Pattern Recogn. **76**, 635–649 (2018)
71. Li, S., Luo, Y., Sun, K., Choi, K.: Heterogeneous system implementation of deep learning neural network for object detection in OpenCL framework. In: 2018 International Conference on Electronics, Information, and Communication (ICEIC), pp. 1–4 (2018)
72. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: revisiting the ResNet model for visual recognition. Pattern Recogn. **90**, 119–133 (2019)
73. Hossain, M.S., Muhammad, G.: Emotion recognition using deep learning approach from audio and visual emotional big data. Inf. Fusion **49**, 69–78 (2019)
74. Ranjan, R., Patel, V.M., Chellappa, R.: HyperFace: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans. Pattern Anal. Mach. Intell. **41**(1), 121–135 (2019)
75. Zhang, S., Yao, L., Sun, A., Tay, Y.I.: Deep learning based recommender system: a survey. ACM Comput. Surv. **52**(1), 5 (2019)
76. Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)

77. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
78. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
79. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: European Conference on Computer Vision, pp. 646–661 (2016)
80. Oh, S.I., Kang, H.B.: Object detection and classification by decision-level fusion for intelligent vehicle systems. Sensors **17**(1), 207 (2017)
81. Xu, H., Han, Z., Feng, S., Zhou, H., Fang, Y.: Foreign object debris material recognition based on convolutional neural networks. EURASIP J. Image Video Process. **2018**, 21 (2018)
82. Bui, H.M., Lech, M., Cheng, E.V.A., Neville, K., Burnett, I.S.: Object recognition using deep convolutional features transformed by a recursive network structure. IEEE Access **4**, 10059–10066 (2017)
83. Jiang, X., Pang, Y., Li, X., Pan, J.: Neurocomputing speed up deep neural network based pedestrian detection by sharing features across multi-scale models. Neurocomputing **185**, 163–170 (2016)
84. Tomè, D., Monti, F., Barof, L., Bondi, L., Tagliasacchi, M., Tubaro, S.: Deep convolutional neural networks for pedestrian detection. Signal Process. Image Commun. **47**, 482–489 (2016)
85. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision, pp. 818–833. Springer, Cham (2014)
86. Xiao, L., Yan, Q., Deng, S.: Scene classification with improved AlexNet model. In: 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), pp. 1–6. IEEE
87. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
88. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K.: Mastering the game of Go with deep neural networks and tree search. Nature **529**(7585), 484–489 (2016)
89. Zhang, Q., Yang, L.T., Chen, Z., Li, P.: A survey on deep learning for big data. Inf. Fusion **42**, 146–157 (2018)
90. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv :1409.1556 (2014)
91. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a largescale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
92. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
93. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3150–3158 (2016)
94. Han, G., Zhang, X., Li, C.: Revisiting faster r-cnn: a deeper look at region proposal network. In: International Conference on Neural Information Processing, pp. 14–24 (2017)
95. Wu, C.H., Huang, Q., Li, S., Kuo, C.C.J.: A Taught-Obesrve-Ask (TOA) Method for Object Detection with Critical Supervision. arXiv preprint arXiv:1711.01043
96. Minaee, S., Abdolrashidiy, A., Wang, Y.: An experimental study of deep convolutional features for iris recognition. In: 2016 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1–6 (2016)
97. Li, Q., Jin, S., Yan, J.: Mimicking very efficient network for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6356–6364 (2017)
98. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
99. Lee, Y., Kim, H., Park, E., Cui, X., Kim, H.: Wide-residual-inception networks for real-time object detection. In: 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 758–764 (2017)
100. Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., Ma, Y.: Deepfood: deep learning-based food image recognition for computer-aided dietary assessment. In: International Conference on Smart Homes and Health Telematics, pp. 37–48. Springer, Cham (2016)
101. Xia, X., Xu, C., Nan, B.: Inception-v3 for flower classification. In: 2017 2nd International Conference on Image, Vision and Computing (ICIVC), pp. 783–787. IEEE (2017)
102. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
103. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
104. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
105. Hussain, M., Haque, M.A.: Swishnet: a fast convolutional neural network for speech, music and noise classification and segmentation. arXiv preprint arXiv:1812.00149 (2018)
106. Zhu, L., Deng, R., Maire, M., Deng, Z., Mori, G., Tan, P.: Sparsely aggregated convolutional networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 186–201 (2018)
107. Zhou, P., Ni, B., Geng, C., Hu, J., Xu, Y.: Scale-transferrable object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 528–537 (2018)
108. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
109. Adam, G., Lorraine, J.: Understanding Neural Architecture Search Techniques. arXiv preprint arXiv:1904.00438 (2019)
110. Pham, H., Guan, M.Y., Zoph, B., Le, Q.V., Dean, J.: Efficient neural architecturesearch via parameter sharing. arXiv preprint arXiv:1802.03268 (2018)
111. Chen, Y., Yang, T., Zhang, X., Meng, G., Pan, C., Sun, J.: Detnas: Neural Architecture Search on Object Detection. arXiv preprint arXiv:1903.10979 (2019)
112. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8697–8710 (2018)
113. Tan, M., Le, Q.V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv preprint arXiv :1905.11946 (2019)
114. Google AI Blog: EfficientNet: Improving Accuracy and Efficiency through AutoML and Model Scaling. https://ai.googleblog.com/2019/05/efficientnet-improvingaccuracy-and.html. Accessed 8 June 2019
115. Torrey, L., Shavlik, J.: Transfer learning. In: Handbook of Research on Machine Learning Applications and Trends:

Algorithms, Methods, and Techniques, pp. 242–264. IGI Global (2010)
116. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks?. In: Advances in Neural Information Processing Systems, pp. 3320–3328 (2014)
117. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. In: International Conference on Artificial Neural Networks, pp. 270–279. Springer, Cham (2018)
118. Guignard, L., Weinberger, N.: Animal identification from remote camera images (2016)
119. Villa, A.G., Salazar, A., Vargas, F.: Towards automatic wild animal monitoring: identification of animal species in camera-trap images using very deep convolutional neural networks. Ecol. Inform. **41**, 24–32 (2017)
120. Okafor, E., Pawara, P., Karaaba, F., Surinta, O., Codreanu, V., Schomaker, L., Wiering, M.: Comparative study between deep learning and bag of visual words for wild-animal recognition. In: 2016 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–8. IEEE (2016)
121. Fang, Y., Du, S., Abdoola, R., Djouani, K.: Background categorization for automatic animal detection in aerial videos using neural networks. ANNPR **2016**, 220–232 (2016)
122. Yu, X., Wang, J., Kays, R., Jansen, P.A., Wang, T., Huang, T.: Automated identification of animal species in camera trap images. EURASIP J. Image Video Process. **2013**(1), 52 (2013)
123. Zhang, T., Xu, H., Hu, Z.: Physiognomy: personality traits prediction by learning. Int. J. Autom. Comput. **14**, 386–395 (2017)
124. Zhao, X., Shi, X., Zhang, S., Zhao, X., Shi, X., Zhang, S.: Facial expression recognition via deep learning facial expression recognition via deep learning. IETE Tech. Rev. **32**(5), 347–355 (2015)
125. Taigman, Y., Yang, M., Ranzato, M.A., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
126. Yoo, B., Kwak, Y., Kim, Y., Choi, C., Kim, J.: Multitask learning with weak label expansion. IEEE Signal Process. Lett. **25**(6), 808–812 (2018)
127. Grega, M., Matiolański, A., Guzik, P., Leszczuk, M.: Automated detection of firearms and knives in a CCTV image. Sensors **16**(1), 47 (2016)
128. Lai, J., Maples, S.: Developing a Real-Time Gun Detection Classifier (2017)
129. Anwar, M.K., Risnumawan, A., Darmawan, A., Tamara, M.N., Purnomo, D.S.: Deep multilayer network for automatic targeting system of gun turret. In: 2017 International Electronics Symposium on Engineering Technology and Applications (IES-ETA), pp. 134–139 (2017)
130. Glowacz, A., Kmieć, M., Dziech, A.: Visual detection of knives in security applications using active appearance models. Multimedia Tools Appl. **74**(12), 4253–4267 (2015)
131. Farahnakian, F., Heikkonen, J.: A deep auto-encoder based approach for intrusion detection system. In: 2018 20th International Conference on Advanced Communication Technology (ICACT), pp. 178–183 (2018)
132. Ning, X., Zhu, W., Chen, S.: Recognition, object detection and segmentation of white background photos based on deep learning. In: 2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC), pp. 182–187 (2018)
133. Olmos, R., Tabik, S., Lamas, A., Pérez-Hernández, F., Herrera, F.: A binocular image fusion approach for minimizing false positives in handgun detection with deep learning. Inf. Fusion **49**, 271–280 (2019)
134. Ning, X., Zhu, W., Chen, S.: Recognition, object detection and segmentation of white background photos based on deep learning, pp. 182–187 (2017)
135. Chin, T.-W., Halpern, M.: Domain-specific approximation for object detection. IEEE Micro **38**, 31–40 (2018)
136. Cao, W., Yuan, J., He, Z.: Fast deep neural networks with knowledge guided training and predicted regions of interests for real-time video object detection. IEEE Access **6**, 8990–8999 (2018)
137. Liu, Y., Hua, K.A.: Field effect deep networks for image recognition. ACM Trans. Multimed. Comput. Commun. Appl. **12**(4), 1–22 (2016)
138. Sangineto, E., Nabi, M., Culibrk, D., Sebe, N.: Self paced deep learning for weakly supervised object detection. IEEE Trans. Pattern Anal. Mach. Intell. **14**(8), 712–725 (2015)
139. Bazrafkan, S., Corcoran, P.: Enhancing iris authentication on handheld devices using deep learning derived segmentation techniques. In: 2018 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–2 (2018)
140. Xu, H., Lv, X., Wang, X., Ren, Z., Bodla, N., Chellappa, R.: Deep regionlets for object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 798–814 (2018)