Chapter 34
Deep Learning for Robot Vision

Mamilla Keerthikeshwar and S. Anto

Abstract Deep learning comes under a class of machine learning where we use it for extremely high-level output, like recognition of images, etc. It has been used in pattern recognition over a vast area such as handmade crafts to extract the data from learning procedures. At present, it has gained a great significance in robot vision. In this paper, we show how neural networks play a vital role in robot vision. Image segmentation, which is the initial step, is used to preprocess the images and videos. The multilayered artificial neural networks have a lot more applications. It can be applied in drug detection, military bases, and many more. The main objective of this paper is to review how deep learning algorithms and deep nets can be used in various areas of robot vision. There are some predefined deep learning algorithms that are available in the market, which are used here to perform this comparative study. These will help us to have a clear insight while building vision systems using deep learning.

34.1 Introduction

Deep learning is the booming topic in the area of research, and it has gained a lot of attention in the last couple of years. It is involved in machine learning and robotics. A lot of conferences and workshops are being conducted on this [1–3]. The convolutional neural network has a lot of applications in robot vision. Many more algorithms have been developed for robot vision.

The main intention of this paper is it acts as a guide for new developers who are keenly interested in robot vision. In this, convolutional neural network plays an important role, and mostly convolutional neural networks are employed and in

M. Keerthikeshwar (✉) · S. Anto
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India
e-mail: Keerthikeshwarreddy@gmail.com

S. Anto
e-mail: Anto.s@vit.ac.in

A. N. R. Reddy et al. (eds.), *Intelligent Manufacturing and Energy Sustainability*,
Smart Innovation, Systems and Technologies 213,
https://doi.org/10.1007/978-981-33-4443-3_34

some cases for pedestrian detection or to find fast-moving objects using semantic segmentation and how different datasets are used.

34.2 Deep Learning

Deep learning is a subset of machine learning that uses the artificial networks for the preprocessing of the data, and it also creates patterns that are used for decision making. It is a sequential process that takes step by step. Deep learning that includes AI can learn the unstructured data. Deep learning is used to detect fraud.

Deep learning has many applications. It is used in the fields of the research area, military base camps to detect the human movements, and pose estimation, and it is also used in the image processing where it is used to segregate the image by pixel is used to predict the genes.

34.3 Convolutional Neural Networks

The convolutional neural network comes under a class of neural networks. It is used for analyzing visual imagery. It is in the form of layers where one neuron in one layer is connected to all the neurons in the next layer. Image representation is done by convolutional neural networks. Deep neural networks using convolutional temporal architecture and ordered LSTM cells can be used for classifying video. The final layer uses the temporal feature and late pooling for second convolutional and smaller temporal uses for slow pooling [4]. Human pose estimation is one of the applications that uses a convolutional neural network. This pose is estimated which results in deviated pose prediction. This uses novel structure convolutional networks for training deep networks. It has three methods to estimate pose (i) design novel networks, (ii) multi-task network is designed, and (iii) human pose is evaluated [5]. Another convolutional neural network is the residual attention network which incorporates the state-of-the-art bottom-up top-down structure; it has two branches (i) Mask branch, it has four convolutional layers, and it is employed in the image segmentation for robots. (ii) Trunk branch, it is employed for video classification and uses an end-to-end approach to classify the videos, and it classifies video by layer approach [6]. The fully convolutional network is a convolutional network used for the detection of objects regionally. It uses the sensitive position of maps and ROI. The regional proposal came out from RCNN with deep networks [2]. Bayesian convolutional neural network shows six DOF camera pose from a single RDB image [7].

Using convolutional neural networks, one can detect the pedestrians, and LIDAR is used to detect the light and sense the objects. The very popular framework in this is the Caffe framework [8]. Pedestrian detection makes use of support vector machine [SVM], and Kalman filter is used for tracking the pedestrian [9, 10]. Convolutional neural networks have a lot many more applications such as it is used to build 3D scenes from a single image, and this uses machine learning and uses CRF for refining. It also uses a database source, but that does not display 3D, and this type of convolutional network is fully convolutional networks [11]. Different scale-specific detectors are combined to produce strong object detector [12]. The accurate detection of the simple stage is by rolling convolutional. It is a two convolutional neural network used for object detection [1].

Shared convolutional neural networks is another neural network used for object discovery, where the speed is about 100 frames per second on GPU [13]. ADE20K is a dataset used for perspective adaptive convolutions. For this, an efficient algorithm for online training of network trajectories is used, and this analyzes problems and uses XNOR and SqueezeNet as detectors [14]. Image is segmented pixel-wise by encoding and decoding the image at SegNet. It uses the VGG16 network which contains 13 convolutional layers, and it is specially designed for the robotic vision. It takes input picture and differentiate the image into several colors, for the image which is of the same type, and then, they have the same color and the images which are different have other colors [15]. The 3D scene uses an NYU depth V2 dataset and SUNRGBD dataset [11]. Segmentation using convolution networks uses PASCAL [16]. VOC, NYUDv2, and SIFT as its datasets [10, 17]. CNN can be also used for playing games such as soccer using back propagation through time (BPTT) and real-time recurrent learning (RTRL) [18].

34.3.1 Fast RCNN

We mainly focus on the fast RCNN for the object detection. In this, convolutional feature maps are generated by taking in the input images. These images are then extracted, and then these are again reshaped into fixed size.

Algorithm:

Step 1: Image is taken as the input, and these are processed.

Step 2: ConvNet generates the regions which we wanted to extract. The image is sent to ConvNet for extraction.

Step 3: For resizing the extracted image, RoI pooling layer is used for the images generated from the ConvNet, and later, this is passed to fully connected network.

Step 4: After getting the result from the fully connected layer, these are further passed on to the Softmax layer and linear regression layer.

34.4 Generative Adversarial Networks

Generative adversarial networks are used for semi-supervised learning. It helps the robot to interact with objects. It is a minimax game between a generator (G) and a discriminator (D) by mapping noise z from p(z), where p(z) is noise distribution. The function of the discriminator is that it differentiates between real date and generator sample. For robots like self-driving or warehouse robots, it uses depth perception that also makes use of 3D object recognition and path planning. So to use this, the robot should also know the depth of the ground, to know the depth of the ground it makes use of some highly expensive hardware tools, and to overcome this problem, YGAN is used, that uses data from cameras from all the sides to estimate the depth [19, 20]. GAN has a lot of applications, where image resolution and classification is one of them. To classify GAN's accuracy, we take datasets consisting of repeated results approximately 7000 series with 1000 series in each material. It got the result of 7000 by sixfold cross-checking, with each fold being 1000 samples [21].

34.5 Restricted Boltzmann Machine

Restricted Boltzmann machine is an unsupervised model that produces never seen data from the original data. It is of the form of layers with one visible layer and several hidden layers. But Boltzmann machine is different from restricted Boltzmann machine, in restricted Boltzmann machine the visible node and the hidden nodes are not linked to one another, whereas in Boltzmann machine, the nodes are linked to each other. A deep belief network is a process where multiple RBMs are stacked together can be fine-tuned by process and back propagation. In the restricted Boltzmann machine, all the neurons behave individually [22].

Restricted Boltzmann machine has a lot many applications, and automatic hand sign language is one among them. Input is taken by RGB and depth. These inputs are sent to the RBM. The output RBM is simplified to another RBM, and this model is trained by datasets [23].

34.6 Recurrent Neural Networks

Recurrent neural networks are used to join images that are drawn. Kazuma Sasaki stated that they have conducted two experiments. First experiment tells that model can learn 15 drawing shapes by the bottom-up process. In the second experiment, four images are trained with four deformed variations per each type, and the images are segregated based on their type using drawings and image classification [24].

Recurrent neural network can also be a part to develop a 3D scene layout. A robotic camera is installed, and it captures images, and these images are filtered using RGB, depth, and foreground, and later after that, the image is converted into a 3D tensor. To convert it into a 3D scene layout, it is passed through recurrent neural network [25]. Recurrent neural networks can be also used planners for the bio-inspired robotic motion, it uses long-term memory networks of sequential data, and it also makes use of the simulated fish trajectories. Using this, it can be implemented in the robots without even knowing their position. This work is related to animal behavior and then used to operate the robots [26]. Recurrent neural networks works effectively for path planning and also in object avoidance which is generally represented in the form of neurons [27].

34.7 CNN Architectures

In the last couple of years, we have witnessed a numerous increase in the CNN architectures. These architectures are used by giving input datasets. In this paper, we have taken the most used CNN architectures and described them how are they useful in the robot vision.

34.7.1 AlexNet

AlexNet is one of the convolutional neural networks. It has over eight layers out of which five are convolutional layers, and the remaining three layers are fully connected layers. Some networks use tanh function, but AlexNet uses rectified linear units. It also has multiple GPU training where the time is reduced and makes it run faster. AlexNet checks that it is overfitting [4]. To print a real-time 3D scene fully, convolutional neural network uses AlexNet instead of VGG [11]. Image segmentation by a convolutional network uses AlexNet [10]. This has the same architecture like that of the LeNet. AlexNet uses rectified linear units instead of tanh functions since this accelerates the speed six times at the same accuracy and uses dropout as it overcomes the overfitting but doubles the time for 0.5.

34.7.2 GoogLeNet

GoogLeNet is another convolutional neural network that is pretrained. Similar to AlexNet, GoogLeNet also has 22 layers. The image in this can be trained using an image net or place S65 dataset. It allows only one unique video to be processed by multiple image processing [4]. Semantic segmentation uses GoogLeNet as one of the datasets [10]. The performance of GoogLeNet is similar to that of human-level performance, and it requires human training to beat the GoogLeNet. This is the combination of a group of small convolutions which is to reduce the parameters. AlexNet has over 60 million parameters, but GoogLeNet has over 4 million parameters, which is quite small. GoogLeNet is more accurate.

34.7.3 RGB-D

RGB-D is the mixture of depth with the RGB image. In this, depth image is an image in which distance is calculated from plane to RGB image. It is used for object detection. The architecture for RGB-D has three steps: (i) input image is processed, (ii) train network, (iii) classifying depth images [28, 29]. RGB-D dataset is the largest dataset that holds household objects. The object recorded in a video sequence and the objects is rotated in a circular plane, so that the object is captured from all the sides. The video is recorded using a Kinect style 3D camera. This even has other indoor and outdoor environments such as garden, kitchen, living room, and it can capture these scenes from long distances even though they are partially included or fully involved in the frame.

34.7.4 KITTI

KITTI is the dataset used for the detection of moving objects. As pedestrians move from one place to another there, KITTI dataset is used to detect them [8, 9]. This KITTI dataset also uses Velodyne LIDAR. Fast object detection is done using KITTI and Caltech dataset [15]. Single-stage detectors RRC using novel recurrent rolling has achieved a benchmark in KITTI. Scale-dependent, pooling, and cascaded classifiers make use of the PASCAL object detection challenge and KITTI. It has two colored cameras and a gray-scale camera installed which are used to detect the objects [16].

34.7.5 ImageNet

ImageNet is a database that contains hundreds of images per a single node. The performance of residual attention is done by the ImageNet [5]. ImageNet was created for educators and researchers for those who use a lot of images for training. To make it easy, a large database of images is created, and this database is termed as ImageNet. It does not own any of the copyrights for images, but instead, it only holds URLs and images all together.

34.7.6 CIFAR-100

CIFAR-100 is the same as CIFAR-10 but with 100 classes and over 500 images in each class, and these 100 classes are divided into 20 superclasses. It contains thousands of images per class. It has five training batches and one test batch. They may contain images from any class. It is used in residual attention network evaluation [6].

34.8 Discussion

As mentioned earlier, deep neural networks are suitable for robotics applications to deal with the limited camera resolution and human pose. For the biomedical image to be segmented, we provide elastic deformation with fully based implementation [30]. As said, convolutional neural networks are used to estimate human pose, and PoseNet is used to discriminate the original pose and fake pose [5]. Image classification is done by residual attention network which can capture attention and can be extended to convolutional network [6]. ResNet is used for image classification in region-based fully convolutional network [2]. Trajectory-centric RL algorithm will able to learn different type of skills, and these are used to autoencoders [31]. For pedestrian detection, we employ LIDAR and fusion methods, where fusion performs well [8]. A convolutional neural network is also employed in the real-time 3D scene and uses CRF to refine the boundary and remove extra groups [11]. For scene parsing, we employ perspective adaptive convolutions in parallel GPU and this improves accuracy [32]. Signet is another architecture that is smaller, faster, and more efficient [15]. Noise-aware training is accurate, and it also improves recognition accuracy [28]. Shared convolutional neural networks employed in object detection and have better performance than a single model [13]. Gradient-based algorithm for online uses XNOR, SqueezeNet [14]. GTX.TITAN X GPU is used for testing grasp detection [13]. Image description in the wild (IDW) is used to improve segmentation accuracy using weak supervisions [33].

34.9 Conclusion

This paper has addressed the use of convolutional networks and image segmentation in the area of robot vision. It mainly focused on object detection, pedestrian detection, and showed how different a networks have been used in robot vision development [9]. This paper will help and provide a valuable guide for developers and researchers who are working in the robot vision since it gives them the basic idea of all the algorithms used and different datasets that have been used in it. Training with more sets in RGB-D object detection does not show any better results [28].

It is expected that the robot vision using deep learning will increase in the next few years and thanks for better thinking and adapting the DNN for robot vision. We all know that robots should interact with the environment and human beings. It should be adapted to the surroundings, and to do this, it should be trained properly. Geometry-based and deep-based methods will be a part of state-of-the-art vision systems which leads to the increase of robotic autonomy and training of DNNs. Different CNN architectures are made use of for robot vision. There are many other architectures of CNN and other networks. GoogLeNet and AlexNet are the most used CNN architectures. When a single model architecture for original images is considered, ResNet has top accuracy than other architectures and followed by VGG architecture. AlexNet and GoogLeNet have the least accuracy. When the error rate is considered, GoogLeNet has more error rate than other architectures. When the preprocessed images are taken into the consideration, AlexNet tops the list.

References

1. J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y.-W. Tai, L. Xu, Accurate single stage detector using recurrent rolling convolution, in *CVPR* (2017)
2. J. Dai, Y. Li, K. He, J. Sun, R-FCN: object detection via region-based fully convolutional networks. arXiv:1605.06409
3. G. Wang, P. Luo, L. Lin, X. Wang, Learning object interactions and descriptions for semantic image segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 5859–5867
4. Y.H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: deep networks for video classification, in *IEEE Conference on Computer Vision and Pattern Recognition[CVPR]* (2015), pp. 4694–4702
5. Y. Chen, C. Shen, X.-S. Wei, L. Liu, J. Yang, Adversarial PoseNet: a structure-aware convolutional network for human pose estimation. arXiv:1705.00389
6. F. Wang, M. Jiang, C. Qian, et al. Residual attention network for image classification (2017). arXiv preprint arXiv:1704.06904
7. A. Kendall, R. Cipolla, Modelling uncertainty in deep learning for camera relocalization, in *IEEE International Conference on Robotics and Automation [ICRA]* (May 2016)
8. J. Schlosser, C.K. Chow, Z. Kira. Fusing LIDAR and images for pedestrian detection using convolutional neural networks, in *IEEE International Conference on Robotics and Automation [ICRA]* (May 2016)
9. M. Szarvas, A. Yoshizawa, M. Yamamoto, J. Ogata, Pedestrian detection with convolutional neural networks. IEEE Proc. Intel. Veh. Sympos. **2005**, 224–229 (2005). https://doi.org/10.1109/IVS.2005.1505106
10. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in *CVPR 2015* [best paper honorable mention]
11. S. Yang, D. Maturana, S. Scherer, Real-time 3D scene layout from a single image using convolutional neural networks, in *IEEE International Conference on Robotics and Automation [ICRA]* (2016)
12. Z. Cai, Q. Fan, R.S. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, in *European Conference on Computer Vision* (Springer International Publishing, 2016), pp. 354–370
13. D. Guo, T. Kong, F. Sun, H. Liu, Object discovery and grasp detection with a shared convolutional neural network, in *IEEE International Conference on Robotics and Automation [ICRA]* (2016)

14. N. Cruz, K. Lobos-Tsunekawa, J. Ruiz-del-Solar, Using convolutional neural networks in robots with limited computational resources: detecting NAO robots while playing soccer (2017). arXiv:1706.06702
15. V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for image segmentation (2015). arXiv:1511.00561
16. F. Yang, W. Choi, Y. Lin, Exploit all the layers: fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers, in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition* (2016)
17. J. Fu, J. Liu, Y. Wang, H. Lu, Stacked deconvolutional network for semantic segmentation. arXiv preprint arXiv:1708.04943 [2017]
18. R.J. Williams, J. Peng, An efficient gradient-based algorithm for on-line training of recurrent network trajectories. Neural Comput. **2**(4), 490–501 (1990)
19. M. Alonso Jr, Y-GAN: a generative adversarial network for depthmap estimation from multi-camera stereo images, 3 Jun 2019. arXiv preprint arXiv:1906.00932
20. A. Pronobis, R.P. Rao, Learning deep generative spatial models for mobile robots, in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 24 Sep 2017. IEEE, pp. 755–762
21. Z. Erickson, S. Chernova, C.C. Kemp, Semi-supervised haptic material recognition for robots using generative adversarial networks (2017). arXiv:1707.02796v2
22. A. Al, M. Zain Amin, A briefly explanation of restricted boltzmann machine with practical implementation in pytorch
23. R. Rastgoo, K. Kiani, S. Escalera, Multi-modal deep hand sign language recognition in still images using restricted boltzmann machine, in *Entropy* 23 Oct 2018
24. K. Sasaki, K. Noda, T. Ogata, Visual motor integration of robot's drawing behavior using recurrent neural network. Rob. Auton. Syst. **86**, 184–195 (2016)
25. R. Cheng, Z. Wang, K. Fragkiadaki, Geometry-aware recurrent neural networks for active visual recognition, in *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada
26. A. Khan, F. Zhang, Using recurrent neural networks (RNNs) as planners for bio-inspired robotic motion, in *2017 IEEE Conference on Control Technology and Applications (CCTA)*, 27 Aug 2017. IEEE, pp. 1025–1030
27. N. Bin, C. Xiong, Z. Liming, X. Wendong, Recurrent neural network for robot path planning, in *International Conference on Parallel and Distributed Computing: Applications and Technologies*, 8 Dec 2004 (Springer, Berlin, Heidelberg, 2004), pp. 188–191
28. A. Eitel, J.T. Springenberg, L. Spinello, M. Riedmiller, W. Burgard, Multimodal deep learning for robust RGB-D object recognition, in *IEEE/RSJ International Conference on Intelligent Robots and Systems [IROS]*, Hamburg, Germany (2015)
29. M. Schwarz, H. Schulz, S. Behnke, RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features, in *ICRA* (2015)
30. O. Ronneberger, P. Fischer, T. Brox. U-net: convolutional networks for biomedical image segmentation (2015). arXiv preprint arXiv:1505.04597
31. C. Finn, X.Y. Tan, Y. Duan, T. Darrell, S. Levine, P. Abbeel, Deep spatial autoencoders for visuomotor learning, in *IEEE International Conference on Robotics and Automation [ICRA]* (2016)
32. R. Zhang, S. Tang, Y. Zhang, J. Li, S. Yan, Perspective-adaptive convolutions for scene parsing, in *IEEE Transaction on Pattern Analysis and Machine Intelligence [Early Access]*
33. P. Luo, G. Wang, L. Lin, X. Wang, Deep dual learning for semantic image segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 2718–2726