Machine learning algorithms for Computer Vision

Nakul Singh [a]

[a] Bachelors student, Department of Electrical Engineering, Indian Institute of Technology Kanpur, India

Abstract

Computer vision is a well-studied field due to its wide applications in several spheres, from medicine to robotics. The ubiquity of computer vision tasks in almost all domains prompted innovative solutions based on machine learning methods. Consequently, in recent years several developments have been made in the area of Artificial intelligence-aided computer vision. These methods provide a commendable performance improvement over the traditional techniques and work in an end-to-end framework requiring minimal manual intervention. Because of the flexibility offered by machine learning techniques, they are a favorable alternative for tasks like object recognition, activity detection, and image segmentation. This work briefly reviews some popular machine learning algorithms that have gained prominence in computer vision tasks.

Keywords : Machine learning, deep learning, computer vision, face detection, object recognition, CNN, DBN, RBM

1. Introduction

Computer vision is a field of artificial intelligence that deals with how machines can gain valuable insights from visual input like images, videos, etc. Computer vision is a well-studied problem with several methods and techniques proposed to solve machine vision tasks like object recognition, motion detection, Image segmentation, rendering, etc. The tasks associated with computer vision are generally non-trivial as they present themselves as an inverse problem [1], i.e., starting from the effects it solves for the causes of those effects. For instance, consider the inverse problems in Imaging; in inverse imaging problems, we are given a set of noisy measurements (effect) to begin with, and the objective is to recover the original data (cause) [2].

In the past few years, machine learning has become a viable option for several computer vision tasks due to improvements in computational power and the availability of large datasets. This work reviews some prominent machine learning techniques used for computer vision tasks.

2. Background

Earlier methods for computer vision were based on the principles of physics. These techniques model how objects move and animate, how light reflects off their surfaces, is scattered by the atmosphere, refracted through camera lenses (or human eyes), and finally projected onto a flat (or curved) image plane [6]. Stereo matching algorithms can generate a 3D model of an object using photographs on the internet that differ in exposure and lighting conditions [7]. These algorithms use the values in the pixels of the image for 3D reconstruction. For eg. Local stereo matching compares the surrounding of a pixel to a slightly translated positions of the pixel to estimate disparity of that pixel [9]. Such algorithms, which perform feature extraction based on raw pixel values of the images are also used widely in image processing tasks like filtering, corner detection, edge detection etc. Person tracking algorithms which can track the motion and trajectory of a walking person [8]. Structure from motion algorithms can reconstruct a sparse 3D point model of a large complex scene from hundreds of partially overlapping photographs [6].

Artificial intelligence has become a standard go-to nowadays for several complex tasks which were earlier solved using techniques that were non-trivial and difficult to interpret. Machine learning algorithms provide a black box solution where the algorithms learn suitable patterns present in the data without any manual intervention. MRI-based brain tumor image segmentation is a field in which AI-based algorithms have been successfully applied [3]. In [4], the authors used a convolutional neural network (CNN) based architecture as a brain tumor segmentation method on the BRATS 2013 dataset [5]. The performance achieved is only marginally inferior to human inspection, as seen in Table 1.

Method	Whole tumor	Core tumor	Active tumor
Human Inspection	0.88	0.93	0.74
Zikic et al. [4] (CNN based fully-automatic architecture)	0.837	0.736	0.69

3. Applications

In this section some applications of machine learning algorithms for tasks related to computer vision are briefly discussed.

3.1 Object recognition

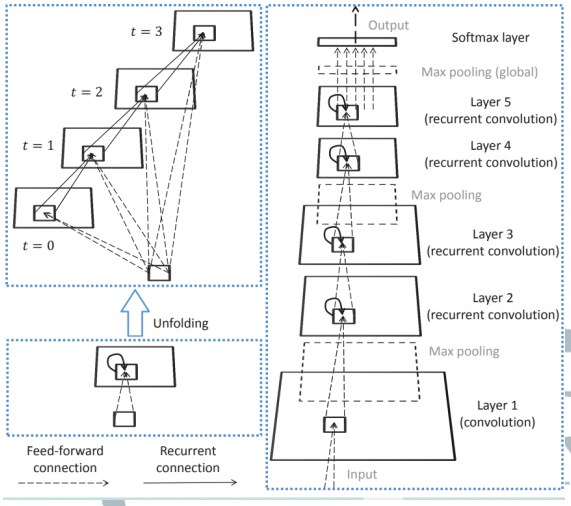
Object recognition is an area of Computer vision that deals with locating and identifying objects in a visual input such as an image or frames of a video. A vast majority of work on object detection using deep learning apply a variation of CNNs. However most CNNs have a feed forward architecture and do not effectively emulate the biological neural networks. In a human brain apart from feed forward synapses there also exists feedback and recurrent synapses that play an important role in context modulation [13].

This concept of context was not explored intensively in earlier approaches whose architectures were inspired from a feed forward CNN. Contextual information is important for object recognition since the processing of visual signals is strongly modulated by the underlying context. One approach to make use of contextual information is by using a feedback connections which is used in convolutional deep belief networks [14]. Another approach to handle context information is using recurrent CNN. Recurrent CNN architecture offers lateral connections that can allow context modulation, thus boosting the performance of deep learning model [12]. In a recurrent CNN, the convolutional layers provide an additional recurrent input along with the feed forward input to update the current state of the layer.

$$z_{ijk}(t) = (w^{f}_{k})^{T}(u^{(i,j)}(t)) + (w^{r}_{k})^{T}(x^{(i,j)}(t-1)) + b_{k}$$

The above equation represents the update equation for the $(i,j)^{th}$ element in the k^{th} feaure map of the recurrent convolutional layer. The first term represents the standard CNN feed forward update, the second term is induced by the recurrent connections and the last term is a bias. Fig.1 provides the architectural design of the RCNN proposed in [12]. Selective search is another popular object recognition model that uses segmentation and full object search in a combined framework [15]. The usage of the textural and spatial features of the image, as well as the quick method of selecting regions in the image, led to the widespread practical application of the selective search.

Figure 1 RCNN architecture [12]



3.2 Face detection

Several machine learning based approaches to face detection have been studied in the past [10], [11]. In [11], the authors approach the problem of face detection as a binary classification problem with two classes (defined as face and not face). To boost the classification performance they use an Adaboost inspired algorithm called Floatboost(FB) classifier. For face detection they apply the Floatboost boosted classifier over subwindows obtained by scanning the input image. Over each window the classifier marks the sub window as having a face (or not a face). Fig.2 provides an example of a face detection task using the FB classifier.

Figure 2 Face detection using FB classifier [11]

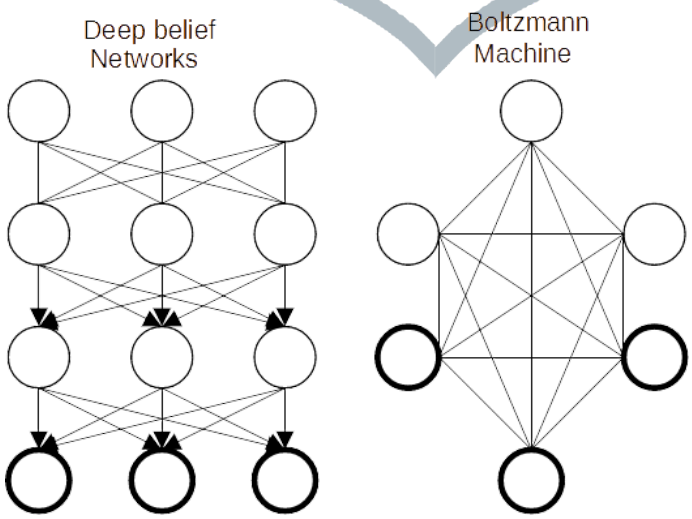
Some traditional methods for face detection include algorithms based on cascaded models like the Viola-Jones face detector which uses AdaBoost cascade scheme using Haar-like features [16]. Traditional methods like the Viola-Jones

face detector [16], ACF-Multiscale [17] use hand crafted features which cannot handle non-trivial image conditions like occlusion, contrast, pose, expression etc.

On the other hand, deep learning-based methods for face detection perform automatic feature extraction from the input image. Thus deep learning methods perform better than the traditional methods in complicated face variances. Cascade CNN, YOLO, and Faster R-CNN are some popular deep learning methods used for face detection. Cascade CNN was first proposed to address the problem of high computational cost and high variances of face detection. The intuition of cascade structure is to reject simple negative samples at the early stages and refine the results later [18]. YOLO(you only look once) is a deep learning framework developed for object recognition and was later applied successfully to face detection [19]. YOLO offers faster computational speed due to simpler architecture and simultaneous classification and bounding box regression than other CNN architectures like R-CNN. However, the advantage of speed comes at the expense of lower localization accuracy. In R-CNN (region-based CNN), the pipeline comprises two stages. First, the selective search algorithm generates a set of class agnostic object proposals. In the second stage, the image region within each submission is mapped to a feature vector. This feature vector is then provided as input to a classifier and a regressor that refines the detection position [20]. However, R-CNN is computationally inefficient due to the forward pass required for each object proposal. Two variants were proposed after its inception, Fast R-CNN and Faster R-CNN [21], to improve the speed of R-CNN.

Machine learning models like deep belief networks and Boltzmann machines are also used for solving vision tasks. A deep belief network is a graphical generative model that estimates the joint distribution of the observed data and the labels. It is made of stacked layers of a Restricted Boltzmann machine(RBM) and uses a greedy approach for training[22]. Using greedy learning algorithms makes the training process unsupervised, thus eliminating the need for labels [23]. Boltzmann machine is a generative model that does not have output nodes in the network, and all the neurons are connected. These connections allow them to interact with each other and learn patterns in the data. The neurons in a Boltzmann machine are divided into hidden and visible nodes. The input is fed through the visible nodes, and all the observations are made through them. RBMs are a particular case of Boltzmann machines that requires the nodes in the network to form a bipartite graph between the hidden and the visible neurons. This means there should be no connections between two neurons if they belong to the same group. Due to the lesser number of links present in RBM, they are faster than the Boltzmann machine [24]. Fig.3 provides a graphical representation of the DBN and Boltzmann machine. The figure shows that the DBN is a bipartite-directed graphical generative model, whereas the Boltzmann machine represents an undirected graphical generative model. In both representations, the darker nodes represent the visible neurons through which input is fed, and the lighter nodes represent the hidden neurons.

Figure 3 DBN and boltzmann machine



4. Conclusion

Over the last years, most of the development in computer vision has been because of advancements in machine learning-based algorithms. CNN-based architectures have wide-ranging applications in image classification and object recognition tasks. In some cases, CNN-based classification methods have surpassed human inspection in performance. However, CNN requires the existence of a labeled dataset due to its supervised learning framework for training. In tasks that predominantly have unlabelled datasets, Boltzmann machine-based models like Deep belief nets(DBN) and Deep Boltzmann machine(DBM) are used as they use an unsupervised learning framework for training. This work briefly overviews the recent developments in the domain of AI inspired computer vision.

References

[1] Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

[2] McCann, Michael T., Kyong Hwan Jin, and Michael Unser. "Convolutional neural networks for inverse problems in imaging: A review." *IEEE Signal Processing Magazine* 34.6 (2017): 85-95.

[3] Işın, Ali, Cem Direkoğlu, and Melike Şah. "Review of MRI-based brain tumor image segmentation using deep learning methods." *Procedia Computer Science* 102 (2016): 317-324.

[4] Zikic D. et al. Segmentation of brain tumor tissues with convolutional neural networks. MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS) 2014:36–39

[5] https://www.smir.ch/BRATS/Start2013

[6] Szeliski, Richard. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[7] Goesele, Michael, et al. "Multi-view stereo for community photo collections." *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007.

[8] Sidenbladh, Hedvig, Michael J. Black, and David J. Fleet. "Stochastic tracking of 3D human figures using 2D image motion." *European conference on computer vision*. Springer, Berlin, Heidelberg, 2000.

[9] Stankiewicz, Olgierd, Gauthier Lafruit, and Marek Domański. "Multiview video: Acquisition, processing, compression, and virtual view rendering." *Academic Press Library in Signal Processing, Volume 6*. Academic Press, 2018. 3-74.

[10] Sivic, Josef, C. Lawrence Zitnick, and Richard Szeliski. "Finding People in Repeated Shots of the Same Scene." *BMVC*. Vol. 2. 2006.

[11] Li, Stan Z., and ZhenQiu Zhang. "Floatboost learning and statistical face detection." *IEEE Transactions on pattern analysis and machine intelligence* 26.9 (2004): 1112-1123.

[12] Liang, Ming, and Xiaolin Hu. "Recurrent convolutional neural network for object recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[13] P. Dayan and L. F. Abbott. Theoretical neuroscience. Cambridge, MA: MIT Press, 2001.

[14] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML), pages 609–616, 2009

[15] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective Search for Object Recognition," International Journal of Computer Vision

[16] P. Viola and M. J. Jones, "Robust real-time face detection," IJCV, vol. 57, no. 2, pp. 137–154, 2004.

[17] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," in IJCB, pp. 1–8, IEEE, 2014

[18] Zhou, Yuqian, Ding Liu, and Thomas Huang. "Survey of face detection on low-quality images." *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018.

[19] Garg, Dweepna, et al. "A deep learning approach for face detection using YOLO." *2018 IEEE Punecon*. IEEE, 2018.

[20] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, pages 580–587, 2014

[21] Jiang, Huaizu, and Erik Learned-Miller. "Face detection with the faster R-CNN." *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 2017.

[22] I. Arel, D. C. Rose, and T. P. Karnowski, "Deep machine learning—a new frontier in artifcial intelligence research," IEEE Computational Intelligence Magazine, vol. 5, no. 4, pp. 13–18, 2010.

[23] Voulodimos, Athanasios, et al. "Deep learning for computer vision: A brief review." *Computational intelligence and neuroscience* 2018 (2018).

[24] Zhang, Nan, et al. "An overview on restricted Boltzmann machines." *Neurocomputing* 275 (2018): 1186-1199.