Image and its Coordinates Detection in Convolution Neural Network Using YOLO Framework

Abhilasha Gupta, Krishna Joshi, Umesh Diwedi

Babu Banarasi Das Northen Indian Institute of Technology, Lucknow

Corresponding author: Abhilasha Gupta, Email: guptaabhilasha0@bbdniit.ac.in

Computer vision is a field that deals with high level under- standing from digital Images. It provides new directions for making machines attentive and responsive to man. In this paper we are propos- ing an approach that is used for image detection. Face detection is a major problem in this area and many dataset are created for the pur- pose of computer vision. We have various deep learning models like convolutional neural network, recurrent neural network etc. But among all, deep convolutional neural networks are the best model for finding patterns from images. In the same direction we designed a model which receives its input from three fully connected layer that predict the coor- dinates and probability, and fully connected layer gets its input from Pooling layer that would scale back number of parameters when pic- tures are over large. To detect the image features and its coordinates we are using bounding boxes. Image detection is performing with the help of convolution neural network that is getting its input from three fully connected layers and fully connected layer is getting its input from Pooling layer. We used Relu layer to provide input to pooling layer. In this way images features are classified and the image is predicted on the basis of this model. So our output layer predicts both class proba- bility and coordinates of bounding boxes. The work is tested with the help of two experiment setup on two different machines. Results gave remarkable achievement and future directions.

Keywords: Image detection, Convolution Layer, Relu Layer.

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 905–909. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-955020-5-9-86

1. Introduction

1.1 Computer vision

Computer vision is a combining scientific field that deals with how computers can take high-level understanding from digital images or videos. From the perspective of engineering, it understand andautomate tasks that human can perform. It extends many new dormant for making machines attentive and responsive to man. Face detection is a major problem in this era. So many dataset are created for computer vision. In this era, research was run out on the various handcrafted images extraction methods, which were used in training the traditional ML algorithms for detection and recognition. It guides to an increase in the computation power and time for extracting features and gives less accurate results. To overcome the computation time, power and accuracy, the same was implemented using the models of neural networks and thereafter deep neural networks. [2]

1.2 Face Recognition

Face analytics is an active area of re- search. It involves extracting informationlike expressions, pose, gender, age, etc. It has several applications like law enforcement, face biometrics for payments, self-driving vehicles etc. Face identification and verification systems typically have three different types of modules. First, a face detector is needed for detecting and localizing faces in an image. Desirable properties of a face detector are robustness to variations in pose, illumination, and scale. Also, a good face detector should have consistent output and well localized bounding boxes. The second module is to localize the facial identity such as eye centers, tip of the nose, corners of the mouth, tips of earlobes, etc. These identities are used toalign faces, which mitigates the effects ofin-plane rotation and scaling. Third, a feature extractor encodes the identity information in a high-dimension descriptor. These descriptors are then used to compute a similarity score be- tween two faces. An effective feature extractor needs to be robust to errors introduced by previous steps in the pipe- line: face detection, landmark localization, and face alignment [1] There are various deep learning [1] models like convolutional neural net- work, recurrent neural network etc. But among all, deep convolutional neural networks (CNNs) [2] are the best model for finding patterns from images. CNN also has the capability to classify, detect and label the object with high accuracy. Region-based CNN (R-CNN) [3], Fast R- CNN [4], Faster R-CNN [5], and YOLO

[6] are popular object detection networks in recent years. Face detection has a plethora of applications. It plays a crucial role in face recognition algorithms. Face recognition has several applications such as person identification in surveillance and authentication for a security system. It also help for emotion recognition and based on detected emotion, further analysis can be used for emotion-based applications. Hence, it is considered to be a way to deliver rich information like age, emotion, gender and many more about an individual. Other applications of face detection are to automatically focus on human faces in camera, to give tag and to identify different parts of faces. Automated face detection has gained attention in computer vision and pattern recognition. Earlier face detection systems could handle only simple cases but now it has outperformed in various situations using deep learning algorithms. Due to large variation caused by occlusions, illumination and viewpoints, face detection remains a challenging problem in the area of computer vision. So accuracy, training time and processing time in real- time videos for detecting faces are still research issues. In this paper, section two presents related work of face detection algorithms. Section three describes the working of YOLO framework for detecting objects. Proposed work is explained in section four. Experimental setup and datasetinformation are discussed in section five. Results are analyzed in section six. Final-ly, conclusion and future work are de- scribed in section seven.

2. Related Work

Face detection is one of the most daring problems to the detection of face in pattern. Early in 1994 Vaillant et al. [6] had applied the algorithm named neural networks for detecting the faces. They had proposed a model, which could detect the absence or presence of the face in an image by training a neural network. In this method, the entire image was scanned with the network **at** all possible locations**.** Face can be detected in different area (front, back or semi frontal) through the network .In the year 1998, [8] rotation invariant face detection method was used wherein a "router" network estimated the orientation of the face and proper detector network was applied. For detecting the semi-frontal face from a complex image, a neural network was developed by Gracia in the year 2002 [9]. Convolutional neural network for pose estimation and detection of the face was proposed by Osadchy [10]. Wilson et al. presented har cascading for facial feature detection [11]. But limitation arises for [10, 11] when the face is exposed to various illuminations, poses and expressions. In this era, face detection is defined with using deep learning models. One of the most popular models for DL is CNN (convolutional neural network) [12]. Faster R-CNN is also accomplishing mark able results for object detection. This paper proposes architecture of a convolutional neural network to detect the face using the YOLO framework. Our architecture is tried rely on the handcrafted features. Faces are detected based on the CNN, which extract features, by itself. Training and testing of a model are carried out on two GPU and it detects the faces at a faster rate in real time YOLO framework is detect the real life objects. Other detectors are not good as compared to YOLO. It's an deep learning framework for detection of moving objects. Objects detection should be accurate and faster. YOLO prepare datasets for object detection.

Data sets are limited but if we use classification and tagging techniques so that object detection of the images with their tags becomes easy. Assigning a label for classification is cheaper then assigning a tag for an object. YOLO stands for you only look once because on the entire image a single neural network is applied. Only image detection is not our purpose but also its coordinates with its size are required. For this purpose we use bounding boxes, which provides the object's center, coordinates (x,y) and its size that is width w and height h andconfidence that is calculated by class probabilities incorporating the individual box confidence prediction. These values are multiplied to get the overall confidence of the object. YOLO first version having the prob- lem of localization errors and having low recall in comparison to the other region based detection methods. So this model was trained at 224*224 network model and detected at 448*448 resolution. But YOLOv2 is trained at 448*448 resolution.

3. Proposed Architecture

In our proposed architecture output layer is getting its input from three fully cone layer that predict the coordinates and probability, and fully connected layer gets its input from Pooling layer that would scale back number of parameters when pictures are over large. Pooling is a down sampling operation that reduces the dimensionality of the feature map. It has to reduce the computation and make the feature detector moreuniform to its position in the input.

ReLu Layer represents for Rectified Linear Unit for a non-linear operation. The output is f(x)=max(0,x) i.e. it replaces all negative values with zero. The usage of Relu helps to prevent the exponential growth in the computation required to operate the neural network. Convolutional layer that finds the simple features to complex features from the images. Output layer predicts both class probability and coordinates of bounding boxes.

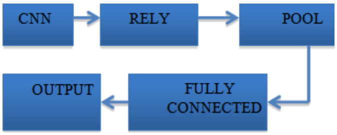

Fig.1: Output Layer

4. Experimental Setup and Data set Information

In our experimental setup we perform it on two machines. First experiment is performed on core i7 processor 16 GB ram and 4GB NVIDIA GTX 1050 Ti GPU and second machine perform on core i5 processor, 8 GB ram, and 2 GB GeForce 820M GPU. Convolutional neural network is tested and trained for FDDB(FACE DETECTION DATASET AND BENCHMARK). Our dataset consist of 5171 faces in a set of 2845 images.Where 70 percent dataset is used for training and 30 percent is used for testing.

5. Conclusion And Future Work

In this paper Image detection is performing with the help of convolution neural network. We used Relu layer to provide input to pooling layer. In this way images features are classified and the image is predicted on the basis of this model. The work is tested with the help of two experiment setup on two different ma- chines. Results are remarkable.

There are some areas where CNN is not performing well such in the case of rotating objects because there lighting condition changes continuously. Similarly transaction invariance and pooling layer problem is also faced by CNN. It pro- vides future scope to handle these problems and to reach at some point where we can get the solution of these problems.

References

[1] Rajeev Ranjan, Ankan Bansal, Jingxiao Zheng, Hongyu Xu, Joshua Gleason, BoyuLu“A Fast and Accurate System for Face Detection, Identification, and Verification “JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, AUGUST 2015

[2] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Image Net classification with deep convolutional neural networks,” Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. Cur- ran Associates Inc., pp. 1097– 1105, 2012

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[5] R. Girshick, “Fast R-CNN,” Proc. IEEE International Conference on Computer Vision, ICCV 2015, pp. 1440–1448, 2015.

[6] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real- time object detection with region proposal networks,” Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. MIT Press, pp. 91–99, 2015.

[7] R. Vaillant, C. Monrocq, and Y. Lecun, “Original approach for the localisation of objects in images,” IEEE Proceedings on Vision, Image, and Signal Processing, vol. 4, 1994.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object De- tection,” 2015.

[9] H.A. Rowley, S. Baluja, T. Kanade, “Neural network-based face detec tion”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 1, pp. 23–38, 1998.

[10] C. Garcia and M. Delakis, "A neural architecture for fast and robust face detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 11, pp. 1408–1423, 2004

[11] M. Osadchy, Y. Le Cun, and M. L. Miller, “Synergistic Face Detection and Pose Estimation with Energy- Based Models,” Journal of Machine Learning Research, vol. 8, pp. 1971,215, 2007.

[12] F. J. Phillip Ian, “Facial feature detection using Haar classifiers,” J. Comput. Sci. Coll., vol. 21, no. 4, pp. 127–133, 2002.

[13] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, “A Convolutional Neural Network Cascade for Face Detec- tion.”, IEEE International Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 5325- 5334, 2015.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisser- man, “The PASCAL Visual Object Classes (VOC) Challenge”, International Journal of Computer Vision, vol. 88, no. 2, pp. 303-338, 2010.

[15] T.-Y. Lin et al., “Microsoft COCO: Common Objects in Context,” European Conference on Computer Vision, ECCV 2014, Lecture Notes in Computer Science, vol 8693. Springer Cham, pp. 740-755.

[16] C. Szegedy, A. Toshev, and D. Erhan, “Deep neural networks for object detection,” Proceedings of the 26th International Conference on Neural In- formation Processing Systems - Volume 2. Curran Associates Inc., pp. 2553–2561, 2013.

[17] V. Jain and E. Learned-Miller, “FDDB: A Benchmark for Face Detection in Unconstrained Settings.”, Technical Report UM-CS- 2010-009, Dept. of Computer Science, University of Massachusetts, Amherst. 2010.