isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 6, Issue: 3,Special Issue: 3 ,Apr,2024 ,ISSN_NO: 2321-3337

DEEP LEARNING FOR VISUAL TRACKING: A COMPREHENSIVE SURVEY

Appunraj C[1], Dineshkumar D[2], Sathya V[3], Kalaivani S[4*]

[123]UG Scholar-Dept. CSE, GRT Institute of Engineering and Technology, Tiruttani, India.

[4]Assistant Professor- Dept. CSE, GRT Institute of Engineering and Technology, Tiruttani, India.

appunr473@gmail.com, dinesh9655djs@gmail.com, sathyavtrt@gmail.com,

[*]**Corresponding Author:**kalaishan@gmail.com

Abstract

This project introduces a cutting-edge approach that combines advanced computer vision techniques with speech synthesis capabilities to achieve groundbreaking results. The impact of this innovation extends beyond technical advancements, finding relevance in diverse real-time applications. The fusion of speech synthesis with object detection adds a layer of accessibility and convenience to multiple scenarios. This system can revolutionize the way individuals with visual impairments interact with their surroundings, offering an auditory understanding of their environment. Furthermore, the model's ability to accurately measure distances between objects and the camera has far-reaching implications, spanning from enhanced object recognition in autonomous vehicles to optimized industrial processes and security monitoring. This project focus on combining speech and vision modalities to yield accurate object detection and distance calculation outcomes. This innovative endeavor sets the stage for intelligent systems that not only visualize the world but also communicate findings audibly, opening doors to novel applications and possibilities across various sectors.

Keywords*: Object detection, Web Camera, Image Preprocessing, DNN Algorithm.*

1. Introduction

Over the past few years, we have witnessed the success of deep neural network (DNN) for visual object detection. To represent objects of various appearance, aspect ratios and poses with limited convolutional features, many DNN-based detectors leverage anchor boxes as reference points for object localization. By assigning each object to a single anchor or multiple anchors at proper scales and aspect ratios, convolutional features are determined and two fundamental detection procedures, classification and localization, are carried out. Anchor-based detectors leverage spatial alignment, i.e., Intersection over Union (IoU) between objects and anchors, as the criterion for anchor assignment. Each assigned anchor independently supervises network learning for object prediction, based upon the assumption that the anchors spatially aligned with objects are always appropriate for classification and localization. In what follows however, we argue that such an assumption is implausible and the spatial alignment should not be the sole criterion for anchor assignment. On the one hand, for objects of acentric features, e.g., slender objects, the most representative features are not close to their geometric centers. A spatially aligned anchor might correspond to less representative features, which deteriorate classification and localization performance. On the other hand, it is implausible to match objects with proper anchors/features using the IoU criterion when multiple objects come together. These issues arise from pre-defining single anchors for specific objects which then independently supervise network learning for object predictions. The open problem is how to flexibly match anchors/features with objects, which is the focus of this study.

2. Related Work

Recurrent Neural Networks (RNNs) are commonly used in computer vision tasks like figuring out what's in a picture, especially when there are multiple things to recognize (multi-label classification). RNNs produce outputs one after another, so we need to decide the order of the labels. [1]

In recent years, researchers have been focusing a lot on two types of learning: multi-label learning and zero-shot learning. Multi-label learning is about predicting multiple labels for one thing, like identifying both a cat and a dog in a picture. Zero-shot learning is when a model learns to recognize new things it hasn't seen before by transferring knowledge from things it has seen.[2]

This article introduces a straightforward solution for classifying images with multiple labels, achieving competitive results on popular benchmarks like COCO and PASCAL VOC. The main idea is to mimic how humans recognize objects in images: by understanding both what objects are present and how they relate to each other. They use a standard Convolutional Neural Network (ConvNet) to recognize the objects in the image, and then they go further by considering how these objects are related to each other. [3]

Multi-label image classification involves predicting multiple labels associated with an image. While many existing methods focus on unified image representations, we propose a framework that extracts label-specific features and exploits label correlations to improve classification accuracy. For input space learning, we introduce a method called label-specific feature pooling, which refines convolutional features to capture features specific to each label. For output space learning, we present a Two-Stream Graph Convolutional Network (TSGCN). This network learns multi-label classifiers by considering both spatial object relationships and semantic label correlations.[4]

Multi-label image classification involves predicting a set of labels corresponding to objects, attributes, or other entities present in an image. In this study, we propose the Classification Transformer (C-Tran), a versatile framework for multi-label image classification that harnesses Transformers to capture complex dependencies among visual features and labels. Our approach utilizes a Transformer encoder trained to predict a set of target labels using an input set of masked labels and visual features from a convolutional neural network. A crucial aspect of our method is a label mask training objective that employs a ternary encoding scheme to represent label states as positive, negative, or unknown during training.[5]

3. Objective

The primary objective of the project is to develop an innovative system that combines object detection using MobileNet architecture and OpenCV, accurate distance calculation through focal length principles, and speech synthesis for real-time communication of detected objects. This project aims to create a versatile solution with applications in accessibility, autonomous driving, industrial automation, and security contexts, revolutionizing how objects are recognized, understood, and interacted with across various domains.

4. Proposed System

This project seamlessly integrates several intricate components. Beginning with input frames sourced from a web camera, the system employs the powerful MobileNet architecture for robust object detection, producing precise bounding boxes and object classifications based on the extensive COCO dataset. The subsequent stage leverages fundamental concepts of focal length and known object dimensions, enabling meticulous distance computation, thereby enhancing spatial awareness. A pioneering feature is the integration of speech synthesis, transforming the detected objects into coherent auditory

feedback, fostering multi-modal engagement. The proposed system operates at the nexus of various technical disciplines, amalgamating real-time object detection, deep neural network (DNN)-based classification, principles of focal length for accurate distance measurement. This fusion of advanced computer vision and natural language processing paradigms culminates in a sophisticated platform with a myriad of applications. The system's versatile applications span from revolutionizing the navigation capabilities of autonomous vehicles to optimizing industrial automation processes and enhancing accessibility tools for differently-abled individuals. Notably, the project underscores the innovative convergence of cutting-edge technologies to shape the future landscape of intelligent systems.

5. Architecture diagram

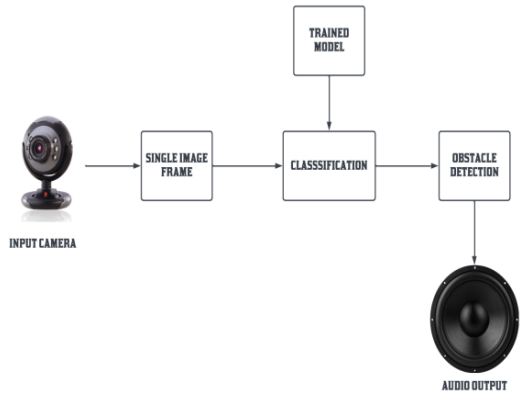

Fig. 5.1 Architecture Diagram

6. Algorithm

Deep Neural Network (DNN):

A deep neural network (DNN) is an ANN with multiple hidden layers between the input and output layers. Similar to shallow ANNs, DNNs can model complex non-linear relationships. The main purpose of a neural network is to receive a set of inputs, perform progressively complex calculations on them, and give output to solve real world problems like classification. We restrict ourselves to feed forward neural networks. We have an input, an output, and a flow of sequential data in a deep network. Neural networks are widely used in supervised learning and reinforcement learning problems. These networks are based on a set of layers connected to each other. In deep learning, the number of hidden layers, mostly non-linear, can be large; say about 1000 layers. DL models produce much better results than normal ML networks.

7. Implementation

7.1 Input Web Camera

The input camera is using an obstacle detection algorithm to analyze the input camera is identify obstacle within it. These algorithms use machine learning techniques like deep learning to recognize patterns and features in the obstacle. First, we gather the camera to classify the obstacle detection. The camera is trained using the DNN Detection model. The live camera is streaming the camera portal. The Collected data are clearly and neatly to find the exact accuracy to the solution. The streaming are categories on camera to image.

7.2 Image preprocessing

An image classification task determines the category of a given input image in the clear dataset. It is a basic task in high –level image understanding and can be divided into binary and multi classification tasks. An image is classified in the output layer following the requirements. Activation function of the output layer is the only difference between binary and multi classification tasks. An image classification task for visual image analysis easily identified and then necessary actions can be taken to prevent visual tracking is a high performance in natural image classification, including DNN Detection model can be used in JPG/PNG image classification.

isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 6, Issue: 3,Special Issue: 3 ,Apr,2024 ,ISSN_NO: 2321-3337

7.3 Feature Extraction

In machine learning, pattern recognition, and image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be redundant, then it can be transformed into a reduced set of features (also named a feature vector). Determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

7.4 Obstacle Detection

Obstacle Detection is a very prominent feature in the image since obstacle in detected. The visual obstacle needs to be found when it's running. Since the obstacle is detected, edges can be used for the same. Canny edge detection is found to give very good results once the thresholds are tuned properly. Image can be filtered before edge detection to remove noise. Edge detection results in a cluster of number of lines. We need to extract the obstacle out of it. The Obstacle Detection can be detected something alarm detected automatic send message to User.

8. Experimental Results

We have implemented two different outputs one for single object detection with distance information (Fig 8.1), and another for multiple objects detection with distance and output as voice (Fig 8.2). It working on a system that can detect objects in an environment and provide information about their distance, and possibly even verbally announce the detections in the case of multiple objects.

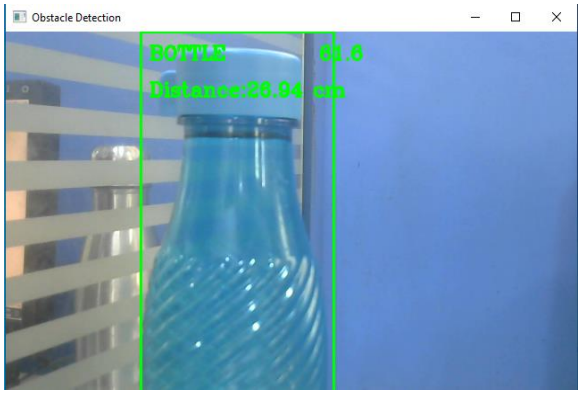

Fig 8.1 Single Object detection with distance

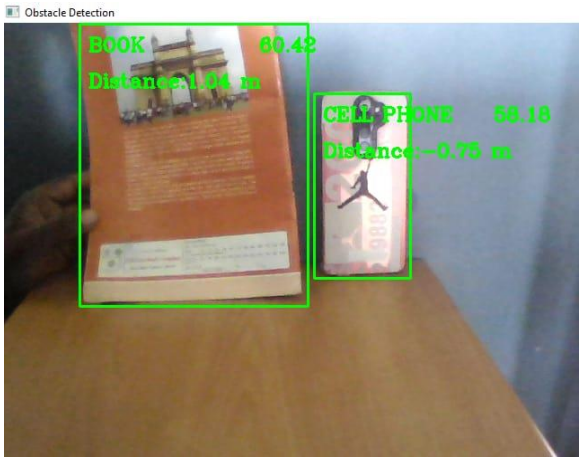

Fig 8.2 Multiple Objects detection with distance

9. Conclusion & Future Work

In conclusion, the envisioned project presents a remarkable integration of intricate components, seamlessly blending advanced computer vision and natural language processing paradigms. The utilization of the MobileNet architecture for robust object detection, coupled with precise distance computation through focal length principles, exemplifies a comprehensive approach to enhance spatial awareness. The innovative inclusion of speech synthesis adds a pioneering dimension, fostering multi-modal engagement and accessibility. This convergence of real-time object detection, DNN-based classification, and multi-modal interaction not only signifies the project's

technical prowess but also its potential to redefine industries. From autonomous vehicles' navigation advancements to revolutionizing industrial automation and facilitating accessibility for differently-abled individuals, the system's versatility is evident. In essence, this project stands as a testament to the transformative synergy achieved through the amalgamation of cutting-edge technologies, poised to shape the future landscape of intelligent systems with its profound applications.

Future Enhancement: Enhancing the project to include voice feedback, where the detected objects are verbally communicated to the user, ensures accessibility and a seamless interaction. Continual updates to the COCO dataset, coupled with periodic model retraining, will enable the project to stay abreast of evolving object categories and maintain high accuracy. These future enhancements collectively aim to transform the project into a versatile and user-friendly object detection system, fostering accessibility, real-time responsiveness, and adaptability to emerging technological trends.

Reference

[1] V.O. Yazici, A. Gonzalez-Garcia, A. Ramisa, B. Twardowski, and J. van de Weijer, “Orderless recurrent models for multi-label classification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 13440–13449.

[2]. Z. Ji et al., “Deep ranking for image zero-shot multi-label classification,” IEEE Trans. Image Process., vol. 29, pp. 6549–6560, 2020.

[3]. S. Wen et al., “Multilabel image classification via feature/label coprojection,” IEEE Trans. Syst., Man, Cybern. Syst., vol. 51, no. 11, pp. 7250–7259, Nov. 2021.

[4]. J. Xu, H. Tian, Z. Wang, Y. Wang, W. Kang, and F. Chen, “Joint input and output space learning for multi-label image classification,” IEEE Trans. Multimedia, vol. 23, pp. 1696–1707, 2021.

[5]. J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, “General multilabel image classification with transformers,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 16478–16488.

[6]. J R. You, Z. Guo, L. Cui, X. Long, Y. Bao, and S. Wen, “Cross-modality attention with semantic graph embedding for multi-label classification,” in Proc. AAAI Conf. Artif. Intell., vol. 34, Apr. 2020, pp. 12709–12716.

[7]. Y. Liu, W. Chen, H. Qu, S. M. H. Mahmud, and K. Miao, “Weakly supervised image classification and pointwise localization with graph convolutional networks,” Pattern Recognit., vol. 109, Jan. 2021, Art. no. 107596.

[8] P. Huang, J. Han, N. Liu, J. Ren, and D. Zhang, “Scribble-supervised video object segmentation,” IEEE/CAA J. Autom. Sinica, vol. 9, no. 2, pp. 339–353, Feb. 2022.

[9]. H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, “Semantic segmentation with context encoding and multi-path decoding,” IEEE Trans. Image Process., vol. 29, pp. 3520–3533, 2020.

[10]. Z. Tang, X. Liu, and B. Yang, “PENet: Object detection using points estimation in high definition aerial images,” in Proc. 19th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2020, pp. 392–398.

[11] Choi, J. Kwon, and K. M. Lee, “Real-time visual tracking by deepreinforced decision making,” Comput. Vis. Image Und., vol. 171, pp.10–19, 2018.

[12] X. Wang, C. Li, B. Luo, and J. Tang, “SINT++: Robust visual tracking via adversarial positive instance generation,” in Proc. IEEE CVPR, 2018, pp. 4864–4873.

[13] E. Park and A. C. Berg, "Meta-tracker: Fast and robust online adaptation for visual object trackers," in Proc. ECCV, 2018.

[14] L. Zhang, A. Gonzalez-Garcia, J. v. d. Weijer, M. Danelljan, and F.S. Khan, "Learning the model update for Siamese trackers," in Proc. IEEE ICCV, 2019.

[15] F. Du, P. Liu, W. Zhao, and X. Tang, "Correlation-guided attention for corner detection based visual tracking," in Proc. IEEE CVPR, 2020.

[16] B. Yan, D. Wang, H. Lu, and X. Yang, "Cooling-shrinking attack: Blinding the tracker with imperceptible noises," in Proc. IEEE CVPR, 2020.

[17] X. Chen, X. Yan, F. Zheng, Y. Jiang, S.-T. Xia, Y. Zhao, and R. Ji, "One-shot adversarial attacks on visual tracking with dual attention," in Proc. IEEE CVPR, 2020.

[18] M. Danelljan, L. V. Gool, and R. Timofte, "Probabilistic regression for visual tracking," in Proc. IEEE CVPR, 2020.

[19] J. Gao, W. Hu, and Y. Lu, "Recursive least-squares estimator-aided online learning for visual tracking," in Proc. IEEE CVPR, 2020.

[20] T. Yang, P. Xu, R. Hu, H. Chai, and A. B. Chan, "Roam: Recurrently optimizing tracking model," in Proc. IEEE CVPR, 2020.