A Mini Project on

Object Detection

by

Team 404
G H Raisoni Nagpur
team404ghrce@gmail.com

TA: Ganesh Chandan sir

International Institute of Information Technology Hyderabad
500 032, India

OCT 2024

Abstract

This paper presents a study on object detection using a custom Convolutional Neural Network (CNN) model trained on the COCO dataset, a widely used benchmark in computer vision. Object detection is a foundational task in various fields, including autonomous driving, security, and augmented reality, and is instrumental in enabling machines to interpret and interact with their environments. The objective of this study is to implement and evaluate a CNN-based approach for detecting multiple object classes within images. Using a subset of 25,000 labeled images from the COCO dataset, the model was designed to classify 80 object categories in a multi-label classification framework. Key preprocessing steps involved filtering missing or invalid annotations and resizing images for consistency, which were essential for maintaining data quality and reducing computational load.

The CNN architecture utilized in this project comprises multiple convolutional layers for feature extraction, interleaved with max-pooling layers for spatial reduction, followed by dense layers to handle classification. The model was trained with the Adam optimizer and binary cross-entropy loss, specifically suited to multi-label classification challenges. Results showed that the model achieved a validation accuracy of 59.50, with clear trends in training and validation metrics indicating the model's learning progress and its limitations due to dataset size and network simplicity. These findings suggest that while CNNs hold promise for robust object detection, future improvements could focus on deeper architectures or integrating transfer learning techniques to achieve higher accuracy and generalization. This work underscores the importance of CNN-based object detection models and opens avenues for further exploration in advanced network designs and data augmentation strategies.

Contents

Chapter 1

Introduction

Object detection is a crucial task in computer vision, with applications spanning numerous fields, including autonomous driving, surveillance, robotics, and augmented reality. It enables machines to identify and locate multiple objects within an image or video frame, facilitating real-time decision-making and interaction with environments. Object detection techniques have evolved significantly with the advent of deep learning, particularly Convolutional Neural Networks (CNNs), which excel in learning hierarchical features and patterns in image data. CNN-based methods have largely replaced traditional detection approaches due to their ability to handle complex visual data, making object detection more accurate, faster, and adaptable across various domains.

A fundamental resource driving progress in object detection research is the COCO (Common Objects in Context) dataset, developed by Microsoft. The COCO dataset is one of the most widely used benchmarks in computer vision, featuring over 200,000 labeled images containing 80 object categories. It is known for its challenging annotation tasks, including multi-label classification, object segmentation, and contextual relationships between objects. The dataset's diversity and complexity make it an ideal benchmark for training and evaluating object detection models. By training on COCO, models are expected to handle complex, real-world environments where multiple objects coexist and interact within varied contexts. Using COCO for this project allows for a robust assessment of the model's performance in identifying a wide range of objects under realistic conditions.

The purpose of this research is to implement and evaluate an object detection model using a CNN trained on the COCO dataset. This project aims to create a model capable of identifying and classifying multiple objects within an image, addressing the challenges of multi-label classification in object detection. Given the complexity of detecting multiple objects within a single image, we have opted for a custom CNN architecture with multi-label classification capabilities, enabling it to handle the distinct characteristics of each category.

Our approach involved designing a custom CNN model with layers optimized for extracting relevant image features, followed by dense layers for classification. We trained this model on a subset of 25,000 COCO images, which required preprocessing to filter out missing or corrupted files and resizing for efficient computation. By evaluating the model's performance, this project seeks to understand the capabilities and limitations of CNN-based approaches in object detection, providing insights into potential

future improvements, such as using deeper networks, data augmentation, or transfer learning to enhance detection accuracy.

1.1 Objective and scope of the proposal

Objectives: The primary objective of this project is to develop a CNN-based object detection model capable of performing multi-label classification on images from the COCO dataset. The model will identify and classify multiple objects from 80 categories, leveraging the strengths of CNNs in feature extraction and classification. Additionally, the project aims to evaluate the model's performance on a subset of 25,000 COCO images, assessing its accuracy and generalization in detecting objects within diverse real-world scenes. The ultimate goal is to identify the strengths and limitations of a simple CNN approach for object detection and explore areas for future improvements.

Scope: This project focuses on implementing a custom CNN architecture for object detection, avoiding more complex models like Faster R-CNN or YOLO. The dataset used will be a subset of 25,000 images from the COCO dataset, which limits the scale of training but still provides a diverse range of object categories and scenes. The scope of this research is to build a foundational object detection model, with the potential for future enhancements through deeper architectures, data augmentation, or transfer learning techniques.

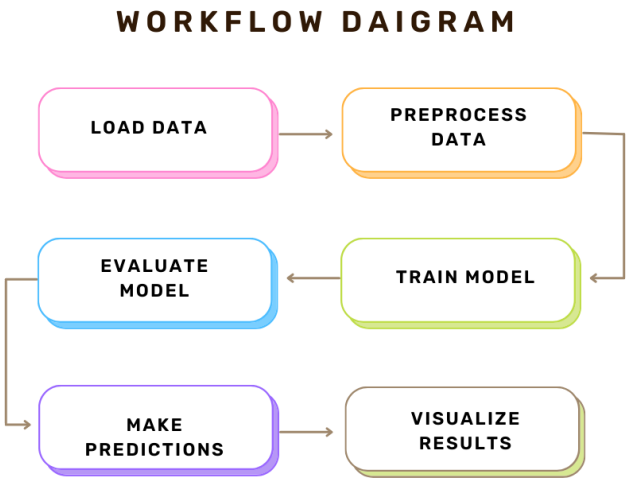

Fig. Algorithm flow diagram

1.2 Related Work

Object detection has evolved from traditional methods like sliding windows and handcrafted features (e.g., SIFT, HOG) to CNN-based models. A. B. Amjoud noted that" previous research has categorized object detection methods into anchor-based, anchor-free, and transformer-based approaches, with future directions focusing on overcoming current model limitations."

Prominent CNN models like YOLO and Faster R-CNN have set benchmarks. Tembhurne highlighted that" previous research categorizes object detectors into two-stage and single-stage models, with YOLO excelling in speed and Faster R-CNN in accuracy." Wang, Y further noted that" research on object detection networks emphasizes advancements from region-based methods like R-CNN to single-shot networks such as YOLO and SSD."

The COCO dataset has driven significant advances. R. Girshick introduced" Region Proposal Networks (RPNs) that share convolutional features with detection networks, enhancing efficiency and accuracy in models like Faster R-CNN, which set new standards in object detection."

Our approach, using a simpler CNN architecture for multi-label classification, contrasts with YOLOand Faster R-CNN. E. P. Dadios observed that" research in vision systems for mobile robotics highlights the use of CNNs for object detection, comparing SSD with MobileNetV1 for real-time efficiency and Faster-RCNN with InceptionV2 for higher accuracy." Sonam Srivastava proposed" a real-time object detection model using CNNs, employing SSD and Faster R-CNN with multi-scale feature maps and aspect ratio filters to optimize accuracy and performance."

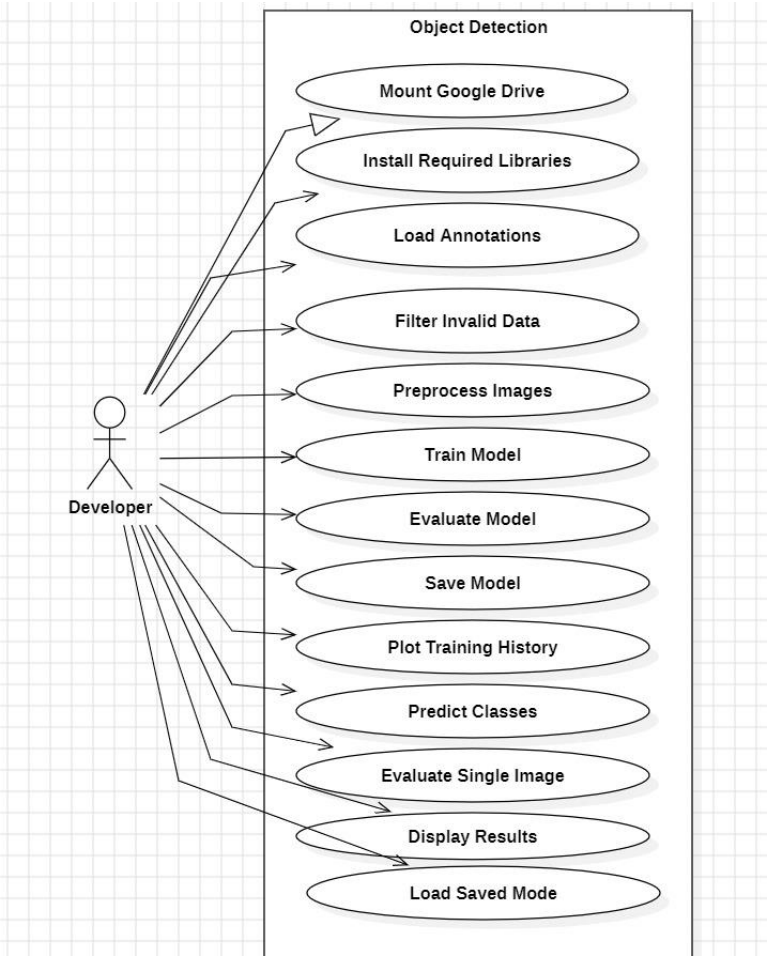

Fig. Use Case diagram

1.3 Methodology

Methodology

The methodology for this object detection project consists of several key steps, including data preparation, model architecture design, training, and evaluation. The primary goal was to create a Convolutional Neural Network (CNN)-based model for multi-label classification on the COCO dataset, enabling the detection of multiple objects within an image. Data Preparation involved preparing the dataset for training. The COCO dataset contains over 200,000 images, but for this project, a subset of 25,000 images was selected to fit computational constraints. The images were filtered to exclude those associated with missing or invalid category IDs. A list of known missing IDs was created, and any images corresponding to these IDs were discarded. This step ensured that the dataset used for training and evaluation was clean and consistent.

Model CNN architecture used in this project is a relatively simple yet effective design tailored for multi-label classification. The network starts with three convolutional layers, each followed by a max-pooling layer. These layers are designed to extract hierarchical features from the images, starting from basic edge and texture detection in the earlier layers to more complex object features in the deeper layers. The convolutional layers use ReLU activation functions to introduce non-linearity and allow the model to learn complex patterns. The pooling layers reduce the spatial dimensions of the image, helping to minimize computation and avoid overfitting by making the model more invariant to small translations of objects.

Training Process of the model used a batch size of 256, which is a balance between training speed and memory usage. The Adam optimizer was chosen due to its ability to adaptively adjust learning rates, making it suitable for complex models like CNNs. The binary cross-entropy loss function was selected because the problem at hand is multi-label classification, where each label is independent, and the task is to classify each object as present or not in the image. The model was trained for 15 epochs, with the training data being fed into the model in batches, and the validation data was used to monitor overfitting and generalization.

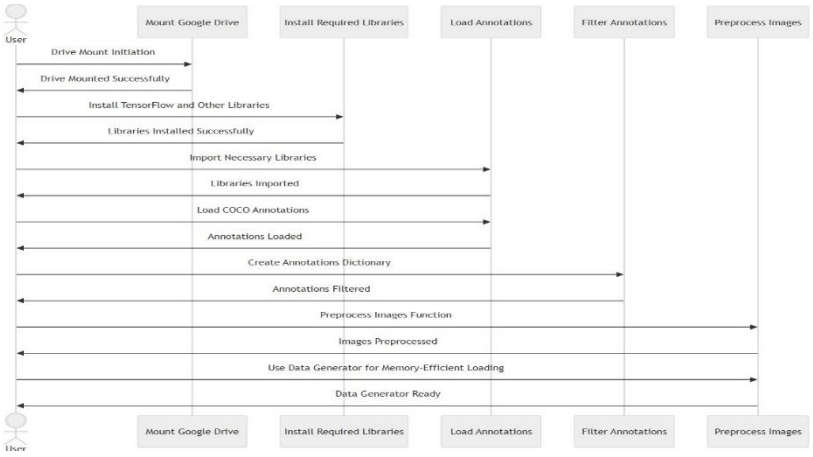

Fig. Sequence diagram

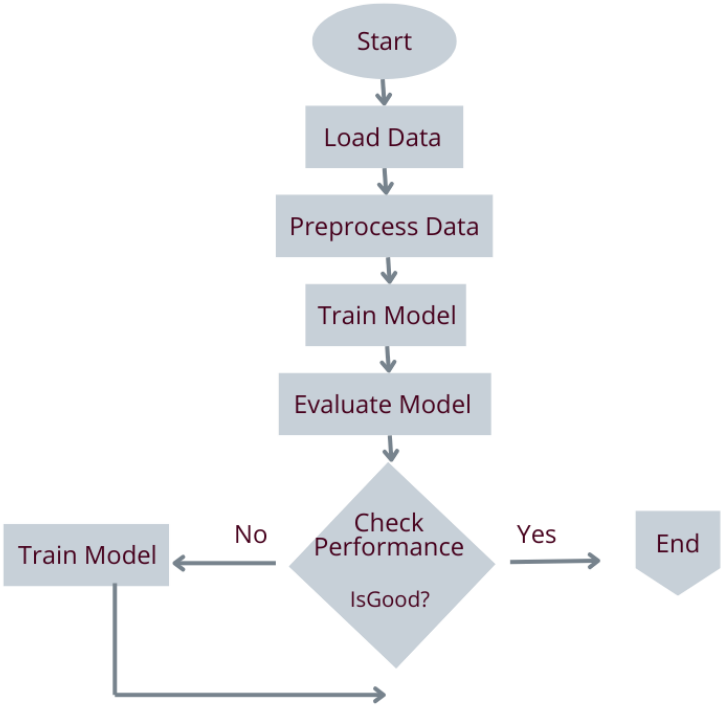

Fig. work flow diagram

Chapter 2

Title of the chapter goes here...

The CNN-based object detection model achieved an accuracy of 59.50 on the validation dataset. This performance was observed after training the model on a subset of 25,000 images from the COCO dataset, with 80 categories for multi-label classification. The model's accuracy, while not exceptionally high, reflects the challenges associated with training deep learning models for complex object detection tasks with relatively simple architectures.

2.1 Training History

The training process was monitored using accuracy and loss graphs across the epochs. The accuracy graph shows a steady increase in the model's performance as it learned to identify objects within the images. However, there was a noticeable plateau in accuracy around the middle epochs, indicating that the model struggled to make further improvements without more advanced techniques like data augmentation or a deeper network. On the other hand, the loss graph showed a consistent decline, suggesting that the model was successfully minimizing the error between its predictions and the ground truth.

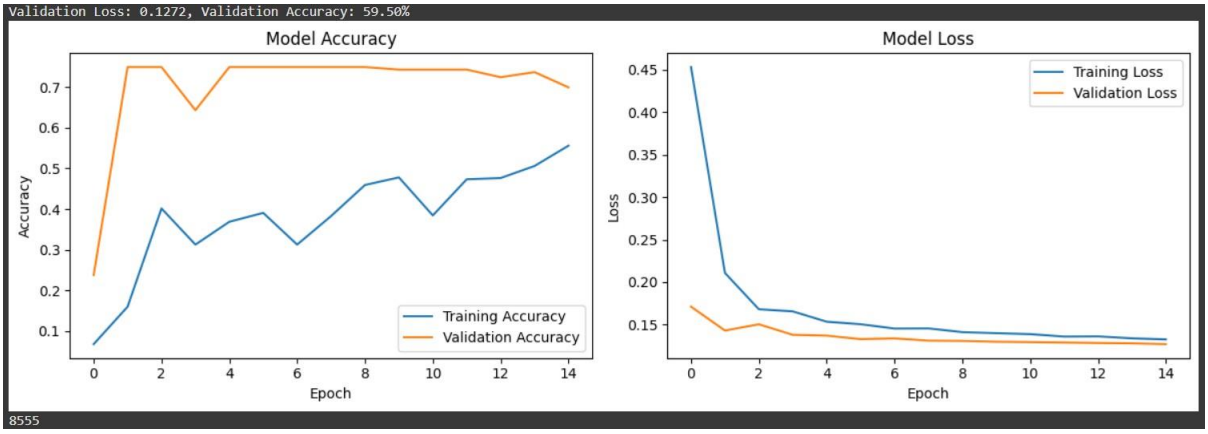

Figure 2.1 Result / Output

Fig. Sample result

2.2 Discussion

The model's 59.50 accuracy is a reasonable result for a baseline CNN architecture on a complex multi-label classification task. Several factors could explain the accuracy level. First, the architecture of the model, while effective for basic object detection tasks, is relatively shallow and may not capture the complex relationships between objects in an image as effectively as more advanced models like YOLO or Faster R-CNN. Additionally, the dataset's diversity, which includes a wide variety of object categories and scene conditions, presents a significant challenge for a model with limited capacity.

Another limitation is the lack of data augmentation and regularization techniques, which could improve the model's generalization. Additionally, training with only 25,000 images rather than the full COCO dataset may have limited the model's ability to generalize to unseen data. More epochs, larger batches, and deeper architectures could improve performance.

Despite these limitations, the model demonstrated reasonable performance in identifying objects and is a promising starting point for future exploration and improvement. Further work could involve optimizing the model with more advanced architectures and techniques to handle the challenges of multi-label object detection more effectively.

Chapter 3

Conclusion

In this study, a CNN-based object detection model was trained on a subset of the COCO dataset, achieving a validation accuracy of 59.50. This performance was obtained by training on 25,000 images, which represented a fraction of the full dataset. Despite the relatively modest accuracy, the results demonstrate the potential of CNNs for object detection tasks in multi-label classification. However, the model's performance was constrained by several limitations. The primary limitation was the relatively simple architecture, which may not have been complex enough to capture the intricate features of the dataset. Additionally, due to resource constraints, the entire COCO dataset could not be processed, limiting the model's training and generalization capabilities.

For future work, scaling the model with a deeper architecture, such as utilizing pre-trained models or transfer learning, could significantly improve performance. Increasing the dataset size and incorporating data augmentation techniques could also help address overfitting and enhance the model's ability to generalize. The model could also benefit from more advanced object detection techniques, such as YOLO or Faster R-CNN, which are known for their superior accuracy and speed. With improved resources and more training data, the model's performance can be scaled further, paving the way for more robust object detection systems.

Chapter 4

Refrences

1. B. Amjoud and M. Amrouch, ”Object Detection Using Deep Learning, CNNs and Vision Transformers: A Review,” in IEEE Access, vol. 11, pp. 35479-35516, 2023, doi: 10.1109/ACCESS.2023.3266093.

2. T. Diwan, G. Anirudh, and J. V. Tembhurne, ”Object detection using YOLO: challenges, architectural successors, datasets and applications,” *Multimedia Tools and Applications*, vol. 82, pp. 9243–9275, 2023, doi: 10.1007/s11042-022-13644-y.

3. J. Ren and Y. Wang, ”Overview of Object Detection Algorithms Using Convolutional Neural Networks,” *Journal of Computer and Communications*, vol. 10, pp. 115-132, 2022, doi: 10.4236/jcc.2022.101006.

4. S. Ren, K. He, R. Girshick, and J. Sun, ”Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, June 2017, doi: 10.1109/TPAMI.2016.2577031.

5. M. Tan, R. Pang, and Q. V. Le, ”EfficientDet: Scalable and Efficient Object Detection,” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 10778-10787, doi: 10.1109/CVPR42600.2020.01079.

6. R. L. Galvez, A. A. Bandala, E. P. Dadios, R. R. P. Vicerra, and J. M. Z. Maningo, ”Object Detection Using Convolutional Neural Networks,” *TENCON 2018 - 2018 IEEE Region 10 Conference*, Jeju, Korea (South), 2018, pp. 2023-2027, doi: 10.1109/TENCON.2018.8650517.

7. A. Kumar and S. Srivastava, ”Object Detection System Based on Convolution Neural Networks Using Single Shot Multi-Box Detector,” *Procedia Computer Science*, vol. 171, pp. 2610-2617, 2020, ISSN 1877-0509, doi: 10.1016/j.procs.2020.04.283.