AUTOMATED CATALOG GENERATION USING DEEP LEARNING

Palak Kakani\textsuperscript{*1}, Shreya Vyas\textsuperscript{*2}

\textsuperscript{1}Department Of Information Technology & Engineering, Shri Govindram Seksaria Institute Of Technology And Science (SGSITS), Indore, M.P., India.

\textsuperscript{2}Department Of Computer Science & Engineering, Institute Of Engineering And Technology (IET DAVV), Indore, M.P., India.

DOI: https://www.doi.org/10.56726/IRJMETS44010

ABSTRACT

Automated catalog generation plays a crucial role in e-commerce, retail, and inventory management systems. Object detection holds immense significance within the realm of computer vision, particularly in grocery item recognition. Deep learning techniques have shown remarkable advancements in various domains, including computer vision. In this research, we propose an automated catalog generation framework utilizing deep learning methods. Specifically, we integrate RoboFlow for image annotation and YOLOv7 and object detection algorithms for dataset training and testing. We present a comprehensive methodology for dataset preparation, annotation, training, and evaluation. The proposed framework demonstrates promising results on diverse datasets, including a novel self-generated benchmark dataset. Our model achieves significant accuracy in object detection and annotation, showcasing its potential for efficient and accurate catalog generation. The scalability and adaptability of our framework open avenues for real-world applications in inventory management and e-commerce platforms. This research focuses on the application of deep learning techniques to automate the generation of product catalogs in supermarkets. The project involves annotating a diverse dataset of grocery item images collected from various sources using Roboflow. The images are then processed and trained using the YOLOv7 and object detection algorithms. The results showcase the successful implementation of Roboflow for automated image annotation, enabling accurate and efficient object detection.

Keywords: Classification, Annotation, Convolutional Neural Network, Deep Learning, Data Set, Yolov7 And Object Detection Algorithms, Roboflow.

I. INTRODUCTION

Automated catalog generation is a vital aspect of modern e-commerce, retail, and inventory management systems. The ability to automatically categorize and annotate images for catalog creation significantly enhances the efficiency and accuracy of these systems. Deep learning techniques, particularly convolutional neural networks (CNNs), have demonstrated exceptional capabilities in image classification, object detection, and semantic segmentation [1]. Leveraging these advancements, we propose a novel framework for automated catalog generation.

The primary focus of this research is to combine RoboFlow for image annotation and the YOLOv7 and object detection algorithms for dataset training and testing. Roboflow provides a robust platform for image annotation, streamlining the process of labeling objects in images. YOLOv7, a state-of-the-art object detection algorithm, is employed to train and test the annotated dataset, enabling accurate object detection and annotation.

Fig 1: Convolutional Neural Network.
In this paper, we present a comprehensive methodology for dataset preparation, annotation using RoboFlow, training and testing using YOLOv7 and object detection algorithms, and performance evaluation. We conduct experiments on diverse datasets, including standard benchmark datasets and a self-generated dataset. The results showcase the effectiveness of our proposed framework in automated catalog generation.

1.1 Motivation

The motivation behind this research stems from the increasing need for efficient and accurate ways to manage supermarket inventory and enhance customer experiences. In our day-to-day life scenario, an object can be categorized into multiple classes. Automated catalog generation using deep learning can streamline the process of identifying and categorizing products, reducing human effort and errors. This research aims to contribute to the development of intelligent systems that can transform the retail industry.

II. LITERATURE REVIEW

Image classification might appear straightforward at first glance, involving the computer’s ability to identify objects within images. Yet, this process is deceptively challenging. Distinguishing between a person and a dog, differentiating human faces from bodies, or discerning cats from people, requires intricate analysis. As the number of categories increases, the complexity of classification escalates. This task is accomplished through a powerful tool in computer science: machine learning.

Machine learning, a discipline enabling algorithms to autonomously learn and enhance, finds application in various domains like web crawling, natural language processing, speech recognition, and image classification. In image classification, we delve into training machine learning algorithms. This process involves instructing an algorithm to recognize specific classes, which is an element of deep learning. Deep learning is a subset of machine learning focused on data representation and usage. Neural networks, the core of deep learning, emulate the human brain with input, hidden, and output layers, composed of interconnected neurons bearing weights and thresholds. The neural network architecture involves data input to the input layer, transmission to hidden layers (the count varies based on the model, affecting accuracy and computational demands), and a nonlinear transformation process in hidden layers. This mapping of data reads it for neural network training. Data then proceeds to the output layer, where linear transformation occurs, shaping it for predictions.

To train a machine learning algorithm, problem definition is paramount. In this instance, the goal is to teach an algorithm to recognize human faces in images. The task involves defining and extracting features that characterize the object class. For example, facial features might encompass eyes, nose, lips, and hair. Algorithm training starts with features extracted from images, which describe various aspects such as color, brightness, contrast, texture, and shape. These features, translated into numerical values, are input into a classification algorithm. Classification algorithms make predictions based on these feature vectors. Often, the predicted label is the most prevalent one in the training data, derived from a dataset of feature vectors and corresponding predicted labels. Different classification algorithms exist, each with its strengths and limitations. A widely used choice is the Convolutional Neural Network (CNN), a deep learning model characterized by layers, including the Pooling layer for dimension reduction.

![Convolutional Neural Network using Deep Learning](image)
Convolutional neural networks are a type of neural network that have become popular in recent years. They have been applied to a variety of problems, including: speech recognition, image classification, and object detection. CNN is a deep learning algorithm that can classify images and videos. It has been trained to detect objects in images. It does this using a convolutional neural network.

The research builds upon previous work in multi-label image annotation, object detection, and cross-modal retrieval. Techniques like YOLOv7 and object detection algorithms have shown success in object detection tasks, while attention-based methods have improved cross-modal retrieval. The use of Roboflow for image annotation streamlines the data preprocessing pipeline.

RELATED WORK

In the realm of Computer Vision, multi-label image annotation has long been recognized as a fundamental task, surpassing the challenges posed by single-label image classification due to the intricate nature of the output space's combinatorial structure. Over the past two decades, researchers have devoted considerable effort to addressing this open problem. Earlier approaches predominantly relied on non-deep learning methodologies, employing simplistic nearest neighbor techniques for label transfer. Many of these methods attempted to create independent binary classifiers for each label, utilizing both non-deep learning and deep learning models. However, these strategies often neglected to account for the inherent co-occurrence patterns among labels in real-world scenarios. Chen et al. made substantial progress in mitigating label dependency concerns, thereby surpassing the performance of prior techniques. By introducing approaches that consider label co-occurrence, they demonstrated the potential to enhance multi-label image annotation. Recent innovations have sought to reframe the multi-label image annotation task as a sequence prediction challenge, leveraging Recurrent Neural Networks (RNNs). Despite the intrinsic significance of label order in training recurrent models for sequence prediction, these techniques adeptly managed the label order-agnostic nature of the annotation task, yielding promising outcomes. Cross-modal retrieval, a key facet of Computer Vision, necessitates effective matching between images and captions. Previous endeavors predominantly focused on visual-semantic embedding to gauge image-text similarity in a joint embedding space. These methodologies entailed representing images via Convolutional Neural Networks (CNNs) and captions using basic Recurrent Neural Network (RNN) models. Training these models involved ranking loss, though Faghri et al. proposed an enhancement by introducing hardest negative pairs for loss calculation. Notably, attention mechanisms have emerged as a recent paradigm shift in cross-modal retrieval, emphasizing local features instead of global representations as previously pursued. This transition to attention-based methodologies has led to the introduction of fine-grained cross-modal interactions, particularly to address instances where specific textual words correlate with distinct image regions. Consequently, attention-based techniques currently stand as the state-of-the-art solutions for cross-modal retrieval tasks.

Datasets: In recent years, the surge in machine learning performance has driven the creation of diverse datasets within the computer vision and robotics communities. These datasets serve as benchmarking tools for object recognition tasks, encompassing a wide spectrum of object categories and natural images. While aiding in the comparison of machine learning approaches, these datasets also facilitate the development of domain-specific perceptions and robotic systems. The domestic environment domain, wherein autonomous robots must recognize everyday objects like groceries, presents a formidable challenge necessitating access to real-world training data that extends beyond readily available online product images. One notable dataset is the FREIBURG DATASET, specifically designed to address the recognition of a wide variety of everyday objects in domestic environments. This dataset provides a valuable resource for evaluating object recognition methodologies in challenging real-world settings. In a similar vein, the SKU-110K dataset targets the detection of densely packed scenes containing numerous objects positioned closely together. Scenes of this nature, frequently encountered in retail shelf displays, urban landscapes, and traffic scenarios, pose unique challenges for object detection due to the abundance of similar or identical objects in close proximity. Despite the prevalence of such scenarios, existing object detection benchmarks often lack representation of densely packed scenes, thereby challenging state-of-the-art object detection models.
Data Preprocessing with Roboflow:
Roboflow offers a suite of data preprocessing capabilities aimed at enhancing datasets for training computer vision models. These capabilities encompass diverse transformations, including resizing, cropping, rotation, and application of image filters, all of which are vital for preparing data for effective training. Moreover, augmentation techniques such as random flipping, rotation, and brightness adjustments contribute to enhancing dataset diversity and robustness, crucial for model performance. Roboflow's toolset empowers researchers to optimize data quality, a pivotal step in the success of subsequent model training endeavors.

Annotations in Computer Vision:
Annotations are foundational to training effective computer vision models, offering labeled data that enable algorithms to learn and make accurate predictions based on visual information. Various annotation types serve distinct purposes, aiding diverse computer vision tasks:

Bounding Box Annotation:
Rectangular boxes drawn around objects enable object detection and localization. Bounding box annotation involves drawing a rectangular box around an object of interest within an image. The box is typically defined by its coordinates, which represent the top-left and bottom-right corners of the box. Bounding box annotation is widely used for object detection and localization tasks. It provides information about the position and size of the object within the image, enabling algorithms to recognize and locate objects accurately.

Polygon Annotation:
More precise delineation of irregularly shaped objects is achieved by outlining contours with connected points, beneficial for semantic segmentation tasks. Semantic Segmentation: Assigning labels to individual pixels creates detailed object boundaries, pertinent for image segmentation and object recognition.

Landmark Annotation:
Marking points of interest, or landmarks, is crucial for tasks like facial recognition, pose estimation, and object tracking. Landmark annotation involves marking specific points of interest within an image. These points, also known as landmarks or key points, are used to identify and locate specific features or structures. Landmark annotation is commonly used in facial recognition, facial landmark detection, pose estimation, and object tracking tasks. Examples of landmarks could include facial features like eyes, nose, and mouth, or key joints in the human body for pose estimation.

Image Classification:
Assigning labels to entire images facilitates categorization and recognition tasks. Image classification annotation involves assigning one or more labels to an entire image. It is used to categorize images into different classes or categories. In this type of annotation, the entire image is labeled with a specific class, allowing algorithms to learn to recognize and classify images based on their content. Image classification is widely used in various applications, including content-based image retrieval, object recognition, and quality control in manufacturing.
Line Annotation:
Annotating lines, curves, or shapes aids tasks such as lane detection and architectural design digitization. Line annotation is used to annotate lines, curves, or shapes within an image. It helps in tasks like lane detection, road marking recognition, and digitization of architectural designs. Lines can represent road boundaries, paths, or any other linear features within an image. Line annotation provides crucial information for algorithms to understand and analyze the geometrical properties of objects or structures.

Text Annotation:
Incorporating textual information into images assists optical character recognition and document analysis tasks. Text annotation involves adding textual information to images, such as captions or descriptions. It aids in tasks like optical character recognition (OCR) and document analysis. Text annotation is valuable for indexing and retrieving information from images, enabling algorithms to understand and extract text-based information for further processing.

These different types of annotations play a vital role in training computer vision models by providing labeled data that allows algorithms to learn and make accurate predictions based on visual information. These annotations collectively contribute to the training and success of computer vision models across diverse applications. Incorporating these insights from related work into our research paper enhances our understanding of the landscape surrounding multi-label image annotation, cross-modal retrieval, dataset creation, and annotation methodologies, thus providing valuable context and establishing a foundation for our proposed methodology and experiments.
III. PROPOSED METHODOLOGY

Our proposed framework for automated catalog generation integrates RoboFlow for image annotation and YOLOv7 and object detection algorithms for dataset training and testing. The methodology encompasses dataset preparation, annotation, model training, testing, and performance evaluation.

Data Collection: In this step, a wide range of grocery item images are collected from various sources such as supermarkets, online product listings, and other publicly available sources. To ensure diversity, different cameras are utilized to capture images under varying lighting conditions, angles, and backgrounds. The collected images may include different types of grocery items, packaging variations, and real-world scenarios to make the dataset representative.

Image Annotation: To train an object detection model like YOLOv7 and object detection algorithms, accurate annotation of the dataset is crucial. Roboflow, a popular annotation tool, is employed for this purpose. Bounding box annotations are applied to each grocery item in the images. These bounding boxes define the exact location and dimensions of each item within the image, enabling the model to learn how to recognize and localize objects accurately.

Data Preprocessing: The collected dataset undergoes preprocessing to ensure consistent quality and remove any irrelevant or duplicate images. Common preprocessing steps include resizing images to a uniform resolution, normalizing pixel values, and removing any noise or artifacts. Data augmentation techniques may also be applied to artificially increase the dataset's diversity, such as rotation, cropping, and flipping of images.

Training: The preprocessed dataset is then used to train the YOLOv7 and object detection algorithms object detection algorithm. YOLOv7 and object detection algorithms (You Only Look Once version 7) is a state-of-the-art deep learning architecture designed for real-time object detection. During training, the model learns to identify different grocery items by analyzing the annotated bounding boxes. The training process involves optimizing the model's parameters using a loss function that penalizes discrepancies between predicted bounding boxes and ground-truth annotations.

Testing and Validation: Once the YOLOv7 and object detection algorithms model is trained, it is tested on a separate set of images that were not used during training. This evaluation process helps assess the model's accuracy, generalization capability, and efficiency in detecting grocery items. Various evaluation metrics, such as precision, recall, and mean Average Precision (mAP), are calculated to quantify the model's performance.

Throughout this methodology, it's essential to follow best practices in machine learning, such as splitting the dataset into training, validation, and test sets, tuning hyperparameters, and optimizing the model's architecture. Regular monitoring and fine-tuning of the model based on validation results can lead to improved performance over time.

IV. RESULT AND DISCUSSION

5.1 Dataset

The dataset consists of grocery item images collected from various sources, capturing the diversity of products in supermarkets. Image annotation using Roboflow involves precise bounding box annotations around individual items. Performance is evaluated using metrics such as precision, recall, and mean average precision (MAP). The model is capable of multi-label classification to categorize different grocery products.

Fig - 6: Object Annotation
5.2 Classification and Annotation

The application of Convolutional Neural Networks (CNNs) extends to classification tasks, and a comprehensive understanding of their operational flow can be gleaned from Figure 5. The input data, generated autonomously, undergoes successive transformations through distinct CNN layers, including the convolutional layer, pooling layer, and fully connected layer. Each of these layers plays a unique role in capturing specific features from the input image. The primary objective of this study involves the precise delineation of semantic elements within both the image's background and foreground.

![A typical Convolutional Neural Network Architecture](image)

**Fig -7:** CNN Layers for input image.

5.3 Work Comparison

Compared to traditional manual catalog generation methods, the automated approach using deep learning offers increased accuracy, efficiency, and scalability. The use of Roboflow for image annotation accelerates the data preparation process, and the YOLOv7 and object detection algorithms outperform previous methods in object detection tasks.

V. CONCLUSION

This research demonstrates the successful implementation of automated catalog generation using deep learning techniques. The combination of Roboflow for image annotation and YOLOv7 and object detection algorithms for object detection results in an efficient and accurate system for identifying and categorizing grocery items. This work contributes to the advancement of intelligent systems for the retail industry, with potential applications in inventory management and customer experiences. In this research, we presented an automated catalog generation framework that combines RoboFlow for image annotation and YOLOv7 and object detection algorithms for object detection. Our proposed methodology demonstrated impressive results on both benchmark datasets and a self-generated dataset. The integration of these technologies streamlines the catalog generation process, offering accurate and efficient solutions for e-commerce, retail, and inventory management systems.

The promising outcomes of our research underscore the potential for real-world applications of the proposed framework. Future work could focus on optimizing hyperparameters, exploring additional deep learning architectures, and extending the framework to handle multi-modal data.

VI. REFERENCES


[4] Pin Wang, En Fan, Peng Wang, Comparative Analysis of Image Classification Algorithms Based on
Traditional Machine Learning and Deep Learning, Pattern Recognition Letters (2020).


[6] Rajnish K. Ranjan and Anupam Agrawal, "Video Summary Based on F-Sift, Tamura Textural and Middle
level Semantic Feature," Procedia Computer Science, 12th International Conference on Image and


