INTELLIGENT SHELF MONITORING SYSTEM USING YOLO-NAS

Armaan Haque
Department of Industrial and Systems Engineering
University at Buffalo
Buffalo, NY, USA
armaanha@buffalo.edu

1 INTRODUCTION

In modern retail environments, efficient management of supermarket shelves is critical for ensuring optimal product availability, reducing stockouts, and enhancing the overall shopping experience for customers. Despite its importance, manual shelf monitoring presents several challenges for supermarkets. Firstly, the process is labor-intensive and time-consuming, requiring dedicated staff to periodically check and update shelf inventory. Human error is also a significant concern, leading to inaccuracies in product counts, misplaced items, and discrepancies between shelf displays and inventory records. Additionally, the frequency of manual shelf checks may not be sufficient to keep up with the dynamic nature of retail environments, leading to stockouts, overstocking, and missed sales opportunities. As supermarkets strive to adapt to changing consumer preferences and market trends, the need for more efficient and accurate shelf monitoring solutions becomes increasingly apparent.

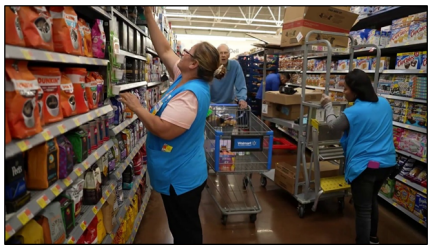

Figure 1: Workers organizing products on retail shelves manually

Automating shelf monitoring processes using advanced technologies such as deep learning offers significant advantages for supermarkets. By leveraging computer vision algorithms and machine learning techniques, supermarkets can enhance their operational efficiency, improve inventory accuracy, and deliver a better shopping experience to customers. Standard **Convolutional Neural Network (CNN)** models are typically utilized for image classification tasks; however, when it comes to detecting empty spaces on shelves, deep-learning models designed for increased speed, efficiency, and real-time detection are far more effective.

2 BACKGROUND

The predetermined organization of products on shelves is known as a **planogram**. It illustrates the placement layout for each product on shelves and specifies the quantity of facings required. Facings represent the number of **stock-keeping units (SKUs)** of the same product that should be visible in the front row of the shelf(1). One of the key factors influencing the effectiveness of a planogram is space utilization. It is essential for employees to consistently monitor the product inventory to quickly identify and replenish out-of-stock items on the shelves. This is especially critical in large retail stores like Walmart, where failing to address stock shortages promptly can significantly hinder

sales and customer satisfaction. By harnessing the power of deep learning algorithms, supermarkets can transform their shelf monitoring operations, improving accuracy, efficiency, and overall store performance. These algorithms can be trained on large datasets of shelf images to detect and recognize products, identify their locations on shelves, and track inventory levels in real-time. ***Deep learning*** belongs to the category of Artificial Neural Networks (ANNs) having numerous processing layers. It represents a machine learning approach dedicated to extracting features from data. ***Object detection*** is the practice of accurately identifying even the minutest elements within an image. This objective is achieved by employing existing classification algorithms, followed by the delineation of a bounding box around the object depicted in the image(2). The big difference between how humans and computers see the world creates problems for computer vision tasks. For instance, computer vision systems often struggle with things like objects from different angles, varying lighting, and being hidden or obscured. These are just some of the many challenges these algorithms face. It's important to note that object detection, and computer vision in general, has come a long way. We've moved from using basic rules defined by hand to the powerful data-driven approaches used today.(3). Image classification aids in categorizing the contents of an image. Image localization pinpoints the position of a single object within an image, while object detection identifies the locations of multiple objects within the image. Lastly, image segmentation generates a pixel-wise mask for each object depicted in the images(4).

Figure 2: Image Classification Vs Object Detection Vs Image Segmentation

Single-pass object detection involves analyzing an input image in one go to predict the existence and positions of objects within it. This method processes the entire image in a single step, leading to computational efficiency. ***YOLO (You Only Look Once)*** exemplifies a single-pass detector that employs a fully convolutional neural network (CNN) to analyze an image(5).

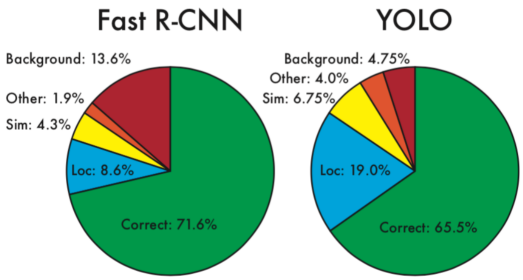

Figure 3: **Error Analysis: Fast RCNN vs YOLO** The pie chart shows the percentage of localization and background errors in the top N detection of various categories(6)

3 Related Work

Numerous deep-learning models have been developed for ensuring planogram compliance, predominantly employing Regional Convolutional Neural Networks (R-CNN), as depicted in the accompanying figure. However, there has been relatively little focus on establishing a distinct class for empty shelf spaces. Typically, classification has revolved around identifying the various products placed on

the shelves. One of the earliest successful attempts to address the object detection challenge using

Figure 4: R-CNN model detecting items of different types

deep learning was the R-CNN (Regions with CNN features) model. This strategy integrated region proposal algorithms with ***convolutional neural networks (CNNs)*** to recognize and precisely locate objects within images.

In a research paper titled "You Only Look Once," Joseph Redmon introduced a new object detection system called YOLO. Unlike other methods, YOLO uses a single neural network to analyze an entire image at once. This approach prioritizes speed (up to 45 frames per second) over accuracy, which can be lower due to errors in pinpointing objects. However, there are even faster variations reaching 155 frames per second.(7) The YOLO model works by dividing the image into a grid. Each cell in the grid is tasked with predicting a bounding box, as long as the object's center falls within that cell. For each cell, the model predicts a bounding box with its location (x and y coordinates) and size (width and height). It also assigns a confidence score, indicating how certain the model is about the box containing an object. Additionally, each cell predicts the class of the object in the box(7). The new YOLO-NAS delivers state-of-the-art performance with the unparalleled accuracy-speed performance, outperforming other models.

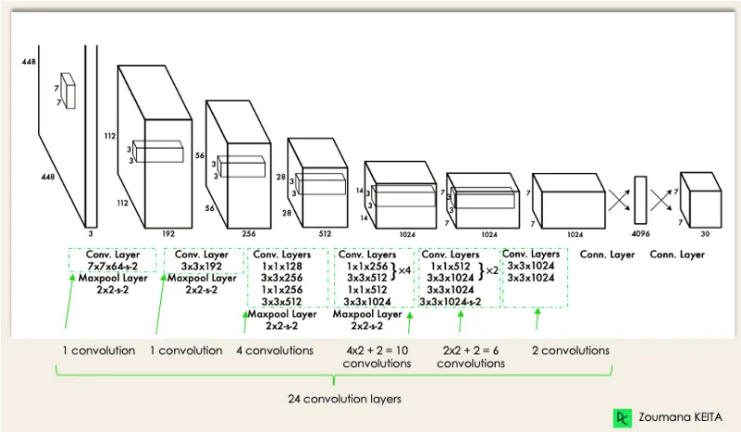

Figure 5: Architecture of YOLO NAS (8)

Instead of relying on multiple steps, YOLO utilizes a powerful deep convolutional neural network (CNN) to directly detect objects in an image.

4 METHODOLOGY

The core of this deep learning model is the integration of YOLO(You Only Look Once)-NAS. YOLO-NAS developed by Deci-AI revolutionizes object detection with fast and accurate real-time detection capabilities suitable for production(9). YOLO-NAS's multi-phase training process involves pre-training on Object365, COCO Pseudo-Labeled data, Roboflow100, Knowledge Distillation (KD), and Distribution Focal Loss (DFL). For this project, COCO dataset will be used.

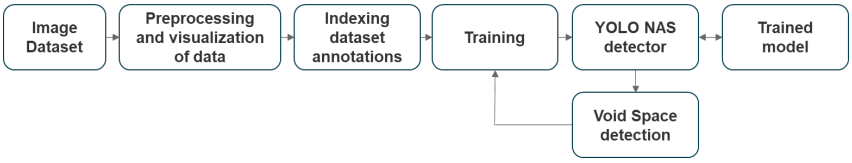

Figure 6: Modular representation of the training process

The deep learning model is built by Pytorch and to fine-tune the pre-trained YOLO NAS model super-gradients is used. The dataset is downloaded from Github, it contains images of supermarket shelves and annotations (labels) outling the empty spaces on the shelves, For this paper, YOLO NAS is used as it provides a good balance between accuracy and efficiency, making it suitable for real-time object detection tasks. The dataset has 3 directories namely train, test and valid folders. The train contains 1530 images along with annotations in YOLO format, valid contains 174 and test contains 54. The annotations ared indexed with the training images and passed through the YOLO NAS detector to detect the void spaces on the rack.

5 EXPERIMENT

There is only one class for this object-detection model i.e [Void-Detected] which will detect the empty spaces on shelves. The maximum epochs is 100.
Learning rate parameters specified for different stages of the training:

- Warmup Initial Learning Rate: is set to 1e-6
- Learning Rate After Warmup: is set to 5e-4
- Learning Rate Schedule: is set to cosine, that means a cosine annealing schedule will be used to adjust the learning rate during training.
- Final Learning Rate Ratio: is set to 0.1, that means the LR at the end of the training will be 10

The model will focus on maximizing the mean Average Precision at an IoU threshold of 0.50. To assess the effectiveness of this entire system, we'll employ ***Mean Average Precision (mAP)*** for a quantitative evaluation. This metric is widely used to measure the performance of object detection and segmentation systems (10).
The mean Average Precision (mAP) is determined by computing the Average Precision (AP) for each class and subsequently averaging across the total number of classes. (10)

$$mAP = \frac{1}{N} \sum_{i=1}^{N} APi \tag{1}$$

mAP formula is based on the following sub metrics:

- **Confusion Matrix:** In machine learning, a confusion matrix is a helpful tool to assess how well a classification model performs.
- **Intersection over Union(IoU):**It measures how much a predicted bounding box overlaps with the actual location of an object (ground truth It box)
- **Precision and Recall:** Precision measures how well you can find true positives out of all positive predictions whereas Recall measures how well you can find true positives out of all predictions.

Figure 7: Intersection of Union

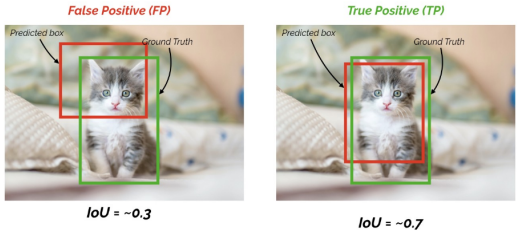

Figure 8: Calculating IoU threshold

The model will focus on maximizing the mean Average Precision at an IoU threshold of 0.50.

6 RESULTS

Upon completing the training and evaluation phases, the model demonstrated its ability to accurately identify the void spaces within a test image, as depicted in the figure 9.

Figure 9: YOLO-NAS model successfully detecting the void spaces on a test image

The detection of void spaces can be served as an alert for staff, prompting immediate action to replenish the inventory. The mAP for the test image was 0.72 and with other images it values ranges from 0.70 to 0.85. A mAP of above 70% is usually quite effective for real-world applications.

7 DISCUSSION

The diversity, volume, and quality of annotated training data are critical. More comprehensive and representative data can lead to better learning and generalization, pushing mAP values higher. A model with high mAP should also be robust to changes in object scale, lighting conditions, and background clutter. Further testing might be needed to ensure the model performs well across different scenarios.
Overall, an mAP range of 0.70 to 0.85 for the model is indicative of robust detection capabilities, but continuous improvements and optimizations might be required to ensure the model performs optimally across all expected conditions and use cases.

8 LIMITATIONS

Detecting void spaces poses unique challenges. These spaces are often small and located in low-light environments, making them difficult for object detection models to identify accurately. While YOLO NAS is a powerful tool, these specific conditions can still lead to some inaccurate predictions. The model can be further optimized by incorporating data specifically focused on void spaces in low-light settings. Additionally, future upgrades to YOLO versions might offer even higher efficiency for this specific task.

9 FUTURE WORK

Enhancing this model's capability to detect misplaced items on shelves could significantly improve its functionality. To achieve this, additional classes should be defined for each item type. A monitoring system with misplaced and low-stock alerts can streamline product management, freeing up staff time for other tasks.

REFERENCES

[1] Mehwish Saqlain, Saddaf Rubab, Malik M Khan, Nouman Ali, and Shahzeb Ali. Hybrid approach for shelf monitoring and planogram compliance (hyb-smpc) in retails using deep learning and computer vision. *Mathematical Problems in Engineering*, 2022:1–18, 2022.

[2] Chamarty Anusha and PS Avadhani. Object detection using deep learning. *International Journal of Computer Applications*, 182(32):18–22, 2018.

[3] Fei-Fei Li, Andrej Karpathy, and Justin Johnson. Convolutional neural networks for visual recognition. *Available in http://cs231n. github. io/convolutional-networks*, 2015.

[4] Pulkit Sharma. Image classification vs. object detection vs. image segmentation. *https://medium.com/analytics-vidhya/image-classification-vs-object-detection-vs-image-segmentation-f36db85fe81*, 2019.

[5] Rohit Kundu. Yolo: Algorithm for object detection explained [+examples]. *https://www.v7labs.com/blog/yolo-object-detection*, 2023.

[6] Jiacheng Li. Deep learning for object detection: Fundamentals. *http://home.ustc.edu.cn/ jclee/posts/2018/03/object-detection/*, 2018.

[7] John Ajala. Object detection and recognition using yolo: Detect and recognize url (s) in an image scene. 2021.

[8] Luís Fernando Torres. Introducing yolo-nas: One of the most efficient object detection algorithms. *https://medium.com/latinxinai/introducing-yolo-nas-one-of-the-most-efficient-object-detection-algorithm-d24303de542*, 2023.

[9] Rohini Vaidya. Deci's yolo-nas: Next-generation model for object detection. *Rohini Vaidya*, 2023.

[10] Deval Shah. Mean average precision (map) explained: Everything you need to know. *https://www.v7labs.com/blog/mean-average-precision*, 2022.