Robust Brand Logo Detection Under Adversarial Conditions

Katherine Xu Jason Sun
Stanford University
Email: {kwx04, jasonsun}@stanford.edu

Abstract

Object detection models like YOLOv8 achieve high accuracy under clean conditions but degrade significantly when exposed to adversarial perturbations and real-world distortions. This project initially explored YOLOv8 for detecting Coca-Cola logos in waste images but found that its robustness to various corruptions (blur, occlusion, noise) was insufficient for real-world applications. As an alternative, we developed a custom Convolutional Neural Network (CNN) to better understand and control feature extraction, robustness, and classification strategies. Our work evaluates the trade-offs between state-of-the-art object detection models and simpler yet adaptable architectures, demonstrating the impact of adversarial training and data augmentation on detection accuracy.

1. Introduction

Brand logo detection plays a crucial role in various applications, including waste sorting, counterfeit detection, and retail monitoring. Automated waste management, in particular, relies on logo detection to classify and recycle branded packaging. However, real-world conditions introduce significant challenges due to image corruptions such as occlusions, noise, and motion blur. Object detection models like YOLOv8, despite their high accuracy on clean datasets, suffer from considerable performance degradation in these scenarios.

Deep learning models for object detection have primarily been trained on clean, well-annotated datasets that do not account for real-world variations. As a result, they often fail to generalize to distorted or adversarially perturbed images. Unlike human vision, which is highly resilient to noise and occlusion, deep learning models exhibit vulnerabilities that can lead to unreliable predictions. Addressing these robustness challenges is critical for deploying object detection models in industrial and commercial settings where reliability is paramount.

2. Related Works

The vulnerability of deep learning models to adversarial attacks and real-world corruptions has been well-documented. Early research by Szegedy et al. (2013) demonstrated that small, imperceptible perturbations could significantly degrade neural network performance. This was extended to object detection models, where Zhang et al. (2019) showed that adversarial attacks targeting classification and localization losses could cause models like Faster R-CNN to misdetect or entirely miss objects. To mitigate this, adversarial training has been widely explored—methods like Det-AdvProp (Chen et al., 2021) introduce adversarial perturbations as a form of augmentation, improving robustness without significantly sacrificing clean-image accuracy. Similarly, RobustDet (Dong et al., 2022) integrates adversarial-aware convolutions to disentangle adversarial and clean loss gradients, improving resilience under strong perturbations. However, adversarial training increases computational overhead and does not generalize equally well across all corruption types, requiring additional strategies such as preprocessing and synthetic data augmentation.

Beyond adversarial training, dataset augmentation techniques have been leveraged to improve robustness to real-world distortions. Studies on common corruption benchmarks (COCO-C, Pascal-C) by Michaelis et al. (2019) found that standard detectors suffered up to 60 percent accuracy drops under noise, blur, and occlusions. Augmentations such as Gaussian noise, synthetic occlusions (cutout, random erasing), and style transfer techniques have been shown to improve generalization to unseen distortions. In the domain of logo detection, robustness has been particularly challenging due to the high variability in logo appearances under different lighting conditions and backgrounds. One-stage object detectors like YOLO, which process images holistically, have been found to exhibit strong generalization capabilities due to built-in augmentations like mosaic augmentation and self-adversarial training. However, both one-stage and two-stage detectors remain vulnerable to structured adversarial noise that exploits biases in feature extraction layers.

Our approach builds on prior robustness research by leveraging targeted augmentations and adversarial training in a domain-specific manner. Instead of fine-tuning large-scale object detection models like YOLOv8, we opted for a lightweight custom CNN trained on a dataset with synthetic distortions tailored to the challenges of logo detection. Inspired by Det-AdvProp, we applied adversarial noise in a data-centric way—simulating realistic edge noise, geometric distortions, and synthetic occlusions—to enhance generalization without excessive model complexity. Unlike previous work that primarily evaluates robustness on broad object detection benchmarks, our study focuses on a single-category classification task, demonstrating that a well-curated training pipeline can improve robustness with minimal computational overhead. The +13 percent accuracy gain on adversarially perturbed test images reinforces findings from prior adversarial training studies while highlighting a practical, small-scale implementation of robust detection techniques.

3. Methodology

Our project initially explored YOLOv8, a widely used object detection model, for the task of detecting Coca-Cola logos in waste images. YOLOv8 was chosen due to its strong performance in real-time object detection tasks, making it a natural candidate for brand logo identification in cluttered environments. However, after initial testing, it became evident that YOLOv8 struggled significantly when faced with real-world distortions such as occlusions, noise, and perspective shifts. The model frequently produced false negatives in cases where the logo was partially visible or blended into a complex background. Additionally, YOLOv8's reliance on bounding box regression made it less effective in scenarios where the primary task was **classification rather than localization**. Given these challenges, we opted to transition to a **custom Convolutional Neural Network (CNN)** designed specifically for binary classification (Coca-Cola vs. Non-Coca-Cola), allowing for more targeted feature extraction and adversarial robustness training.

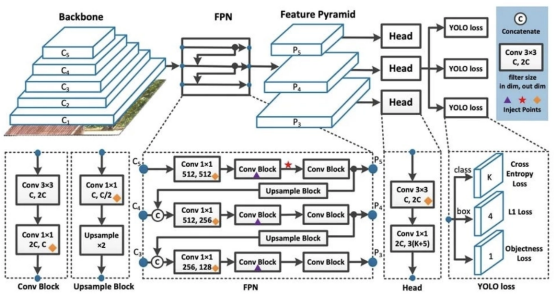

Figure 1. YOLO v8 Model Architecture

3.1. Custom CNN Architecture

The CNN model was designed as a hierarchical feature extractor, progressively capturing meaningful visual patterns through a **three-layer convolutional network**.

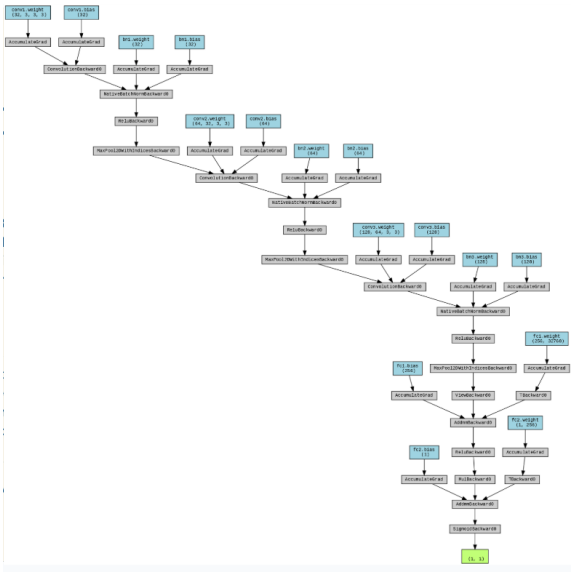

Figure 2. CNN Model Architecture

The architecture consisted of:

- Three Convolutional Layers: These layers were responsible for extracting spatial features at increasing levels of abstraction. The first layer captured edge and contour-based features, while deeper layers identified textures, shapes, and high-level patterns relevant for distinguishing Coca-Cola logos from other background elements.

- Batch Normalization: Applied after each convolutional layer to stabilize learning by normalizing feature activations, improving gradient flow, and reducing internal covariate shift.

- Max Pooling Layers: Downsampled feature maps to reduce spatial dimensions while retaining essential information, enhancing computational efficiency and translation invariance.

- Fully Connected Layers: The final feature maps were flattened and passed through dense layers, transforming extracted features into a high-dimensional representation for classification. The last layer used a sigmoid activation function, outputting a probability between 0 and 1, where values above 0.5 were classified as containing a Coca-Cola logo.

- Dropout Regularization: A dropout layer was included to reduce overfitting by randomly deactivating a fraction of neurons during training, ensuring the model learned more generalized patterns.

This architecture allowed the model to focus solely on classification rather than bounding box prediction, leading to improved performance under distorted and noisy conditions.

3.2. Adversarial Training and Data Augmentation

To enhance robustness, we incorporated adversarial training by introducing synthetic distortions and dataset augmentation techniques. These methods aimed to expose the model to challenging real-world variations, forcing it to generalize beyond clean training examples. We retrieved the coca-cola images through a Roboflow dataset consisting of the Coca-cola brand logo in different manifestations. We retrieved the non-Coca-cola images through web-scraping search queries of "random image" and so on. In total our dataset consisted of 1324 images, half being coke and half being non-coke.

Figure 3. Coca-Cola Image Example

Figure 4. Non-Coca-Cola Image Example

In terms of the corrupted/perturbed dataset we did the following perturbations, drawing upon course material to implement.

- Edge Noise Injection: This technique simulated wear and tear by applying Canny edge detection to highlight high-contrast regions, followed by the injection of random pixel noise along detected edges. This method ensured the model was not overly reliant on clean, well-defined logos.
- Perspective Transformations: Using geometric distortions, we applied random warping to images, replicating variations in viewpoint and camera tilt effects often encountered in real-world scenarios.
- Synthetic Logo Placement: Coca-Cola logos were artificially placed onto random background images using OpenCV's alpha blending technique. This ensured the model learned to differentiate true logos from cluttered environments, rather than memorizing dataset-specific backgrounds.

Figure 5. Synthetic Logo Placement Example

- General Data Augmentation: Additional transformations, including random rotations, brightness/contrast variations, Gaussian blurring, and flipping, were used to further enhance variability in training data.

To ensure a well-balanced dataset, we maintained a 50/50 distribution of Coca-Cola vs. Non-Coca-Cola images. The non-Coca-Cola category included background clutter, other beverage containers, and generic waste materials to prevent the model from developing biases based on contextual elements rather than the presence of a logo.

3.3. Experimental Setup

We conducted a two-stage training process:

1. CNN Model: Trained on clean, unaltered Coca-Cola images to establish a performance benchmark.
2. Adversarially Trained CNN Model: Trained on the augmented dataset incorporating distortions and adversarial corruptions, testing whether exposure to challenging conditions improved robustness.

The models were trained using the Adam optimizer with a learning rate of 5×10^{-4}, a batch size of 16, and Binary Cross-Entropy (BCE) loss. Early stopping was implemented, halting training if validation loss failed to improve for three consecutive epochs, preventing overfitting to noisy data.

4. Results and Discussion

Both models were evaluated on a held-out test set containing a mix of clean and adversarial images. Performance was measured using accuracy, precision, and recall. The adversarially trained CNN demonstrated a +13.63% improvement in test accuracy over the baseline model, highlighting the effectiveness of robustness-focused training. The model's improved recall indicated it was better at identifying Coca-Cola logos under challenging conditions, reducing false negatives.

The baseline model achieved a test accuracy of 50 percent, suggesting overfitting to clean images. The adversarially trained model achieved 63.63 percent accuracy, indicating improved generalization. Precision and recall values are summarized in Table 1.

This methodological framework demonstrates the effectiveness of custom feature extraction and adversarial training in enhancing logo classification robustness, making it a more viable alternative to traditional object detection approaches in adversarially distorted environments.

Figure 6. Example output, can see areas of improvement

Model	Precision	Recall
Baseline CNN	50.0%	100.0%
Adversarially Trained CNN	65.0%	70.0%

Table 1. Performance comparison between baseline and adversarially trained models.

5. Conclusion

Our study demonstrates the challenges of brand logo detection in real-world environments, where adversarial corruptions such as noise, occlusion, and geometric distortions significantly impact model performance. Initially, YOLOv8 was selected due to its efficiency in object detection, but its reliance on bounding box regression proved limiting for classification-focused tasks. The model exhibited poor generalization to distorted images, leading us to develop a custom CNN tailored for binary classification. Through adversarial training and targeted augmentations, we were able to improve test accuracy by 13.63 percent, reinforcing the importance of robust training methodologies in deep learning-based classification models.

The findings highlight the effectiveness of synthetic distortions and adversarial augmentations in training models that can generalize beyond clean, well-annotated datasets. By exposing our CNN to edge noise, perspective transformations, and synthetic logo placements, the model learned to extract invariant features that were less susceptible to minor perturbations. Additionally, the use of dropout, batch normalization, and carefully tuned hyperparameters contributed to the model's ability to reduce overfitting and maintain performance across diverse test conditions. While our approach yielded promising results, there remains significant room for improvement, particularly in refining feature extraction techniques and balancing computational efficiency with robustness.

Future work could explore the integration of transfer learning to improve model generalization while reducing the need for extensive training from scratch. Pretrained architectures such as ResNet, EfficientNet, or Vision Transformers (ViTs), which have been trained on large-scale datasets like ImageNet, could provide more robust feature representations for logo detection, particularly under varying environmental conditions. Beyond deep learning approaches, incorporating traditional feature extraction methods like SIFT (Scale-Invariant Feature Transform) or ORB (Oriented FAST and Rotated BRIEF) could improve detection under extreme occlusion or low-resolution conditions where deep networks may struggle. Hybrid models that combine deep learning with classical keypoint detection and matching techniques could enhance robustness, especially in scenarios where logos appear partially visible or deformed. Furthermore, domain adaptation techniques could be employed to improve performance on unseen datasets by minimizing domain shift, ensuring that models trained on one dataset generalize effectively to another.

6. Contributions

Katherine: Initial YOLO v8 Implementation, CNN Model Implementation, Data Augmentation, Data Collection,

Jason: Initial Idea, Dataset scraping, Data Augmentation, Presentation Creation

Other Resources: ChatGPT used for debugging purposes, Google Scholar and Perplexity for Prior research/related works, Roboflow for dataset, Google Images

for scraping

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, et al. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[2] C. Zhang, C. Xie, Y. Wang, et al. Towards adversarially robust object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[3] X. Chen, C. Liu, D. Song, et al. Det-AdvProp: Detecting adversarial perturbations with adversarially trained object detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[4] Y. Dong, T. Pang, H. Su, and J. Zhu. RobustDet: Learning robust representations for object detection under adversarial attacks. In *NeurIPS*, 2022.

[5] C. Michaelis, I. Ustyuzhaninov, M. Bethge, and W. Brendel. Benchmarking robustness in object detection: Common corruptions and perturbations. *arXiv preprint arXiv:1906.03963*, 2019.

[6] R. Geirhos, P. Rubisch, C. Michaelis, et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

[7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[8] J. Redmon and A. Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[9] D. Hendrycks and T. G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.