



Integrating Object Detection With Context-Aware Caption Generation

Manoj Dhanawade, Dhanashree Bhandigare, Rasika Bhor, Prof. Kirti Randhe Department of Artificial Intelligence and Machine Learning
ISB&M College of Engineering, Pune, India

Milind Ankleshwar
Founder and Director, MASS IT Solutions LLP, Pune, India

Abstract: Through this research, we would like to introduce an innovative merger of object detection and context-aware captioning to attain better interpretation of images. While the 2017 COCO dataset is for the YOLO model to recognize the objects accurately, contextually relevant captions are given out for images through the BLIP model. This integration aims to offer richer and more informative descriptions of images by merging object recognition with intelligent contextual captions. Experimental results indicate improvements in both aspects of object detection and caption generation, hence suggesting viable applications in automated image annotation as well as for accessibility purposes.

Keywords: Deep Learning, Object Detection, Caption Generation, Machine Learning, Contextual Captioning.

1 Introduction

Object detection and image captioning being united into one is what most would pin down as the future of image understanding. While object detection attempts to discover and locate objects in images [2], image captioning tries to create a text description [3]. It has been realized that although much development has happened in either of these fields individually, integrating them to produce contextually enriched image descriptions remains a challenge [1], [6].

To address this challenge, this study integrates the YOLO model—an object detector known for its real-time capability [10]—with the BLIP model, recognized for its context-aware caption generation and vision-language understanding [8]. The proposed methodology uses the COCO 2017 dataset, widely adopted in image captioning and object detection tasks [19], to generate improved image descriptions through object detection and context-sensitive captioning.

The present paper outlines our methodology, implementation, and evaluation, thereby demonstrating the future potential of this integrated model in applications such as automated image annotation, accessibility tools, and intelligent visual assistance systems [5].

2 Literature Survey

Object detection and image captioning are two very important aspects in the domain of artificial intelligence. This chapter mainly reviews the very important advances and techniques concerning this arena.

- Object Detection: Recognition of the object has been extraordinarily remarkable improvement and development in the field of deep learning. The YOLO (You Only Look Once) as put forward by Redmon and

his colleagues in changed real-time object detection as a unified regression problem, whereby it was capable of predicting class probabilities and bounding boxes within a single pass. It has become increasingly used owing to its speed and accuracy across a variety of applications [10].

- **Image Captioning:** From its simple template-based beginnings, the journey of image captioning has seen changes through increasingly sophisticated neural network models. The old generation contained an RNN-E-CNN model-using a Recurrent Neural Network to produce captions from the image features captured by convolutional neural networks. Other newer applications, such as BLIP (Bootstrapped Language-Image Pre-Training), are targeting context relevance and fluency of captions generated through transformer architectures [8].
- **Integration Efforts:** Most efforts relate joint work to understand better the context of an object detected and its description. Bottom-Up and Top-Down Attention has therefore enabled the name given to an object from a construction to be guided by an object which has been detected [19]. Other researchers have gone through different approaches in looking at the aligning of detected objects to generated words to encourage the semantic content of their captions.
- **Challenges and Opportunities:** Challenges have not diminished with the progress that has been made. The captions in most models are not reflective of what is happening in complex scenes and really do not capture subtle contextual information. Much research is underway to improve interpretation and generalization of these models, with a great promise to applications in automated image annotation, content-based image retrieval, and accessibility tools [6].

This literature survey creates a snapshot of advances and challenges towards the combination of object detection and image captioning forms the basis of our approach, which aims at leveraging the functions of the YOLO and BLIP models for rich image understanding [13].

3 Problem Statement

Despite the impressive advances that have been made in both object detection and image captioning, the combination of these technologies to render detailed and contextually appropriate image descriptions remains a very challenging task. Many captioning models rely heavily on global image features, leading to the loss of individual object details. Groups of objects are bound together in a single relationship, missing all the intricate relationships present in a scene. Hence, such global approaches are less able to create elaborate and context-aware descriptions that are really needed in some applications, such as automated image annotation programs or accessibility tools for the visually impaired [1, 3, 6].

The main challenge is that, to create the integrated solution, one must build a synergy between the object detection models, such as YOLO, with strong localization of the objects [7, 9] and sophisticated captioning models, such as BLIP, that generate fluent and contextually relevant text [8]. Solving the integration issue will significantly improve the overall credibility of the interpretation provided and, hence, the user experience.

4 Methodology

4.1 Data Collection

The COCO 2017 dataset was chosen for its vast array of photos, capturing several object instances and containing extensive annotations, ideal for object detection and captioning tasks [11, 19].

4.2 Pre-Processing of Data

- Images were finally downscaled to a size that corresponded to the YOLO and BLIP models.
- Annotations for input requirement represented as:
 - For YOLO, we took coordinate information for the boxed classification [7].
 - For BLIP, aligned image caption pairs were trained to maintain consistency between visual and textual data [8].

4.3 Model Selection and Training

- Object Detection:
 - YOLO was chosen for its efficiency and accuracy in real-time object detection [9].
 - Fine-tuning was done for the COCO 2017 dataset, with adjustments in hyperparameters to improve detection over varied object class performance.
- Caption Generation:
 - BLIP was used for its ability to generate meaningful captions [8].
 - Fine-tuning on the COCO 2017 dataset was done to ensure that it can generate coherent and context-aware captions.

4.4 Model Evaluation

That is, evaluate the integrated model through commonly used metrics:

- Mean Average Precision (mAP) for object detection, detaining localization and classification accuracies, and BLEU, METEOR, CIDEr
- For caption generation, enhancing quality, relevance, and fluency of generated captions.

4.5 Optimization

Special efforts towards improving the harmony between YOLO and BLIP are duly attested to:

- Tuning of both hyper-parameters and model ensemble options
- Attention mechanisms were further fine-tuned to improve the overall captioning performance.
- The iterative testing and validation phase(s) will serve to enhance the ability of the model in generating highly complex and accurate image descriptions.

5 Algorithm

1. Initialization:
 - Start with the loading of the 2017 COCO data set. - Then, checks for object detection using the YOLO algorithm.
 - Finally, this is synced with the generating captions BLIP model.
2. Data Preparation:
 - First, prep the images per the YOLO model's input requirements.
 - Data structures have to be established to hold the detected objects with their attributes.
3. Object Detection:
 - As each image walks into the dataset-set:
 - The YOLO object-detection model is performed.
 - Each detected object has stored its bounding box, class label, and confidence level.
4. Contextual Caption Generation:
 - Then, one by one, for each of the detected objects:
 - The object is contextually captioned about the image through the BLIP model.
 - Naturalize some scenes and interaction context.
5. Integrated Result:
 - A combined caption for all the detected objects for a single description of the image.
 - Preferably, the flowing captions will be both contextually fitting and logical.
6. Evaluation and Improvement:
 - Evaluation of the system using captioning metrics, namely, BLEU and METEOR, and for object detection, using mean Average Precision.
 - Improvement of the models and integration based on feedback from evaluation.
7. Final Output:
 - The output is the final one, showing detected objects and context-based captions.

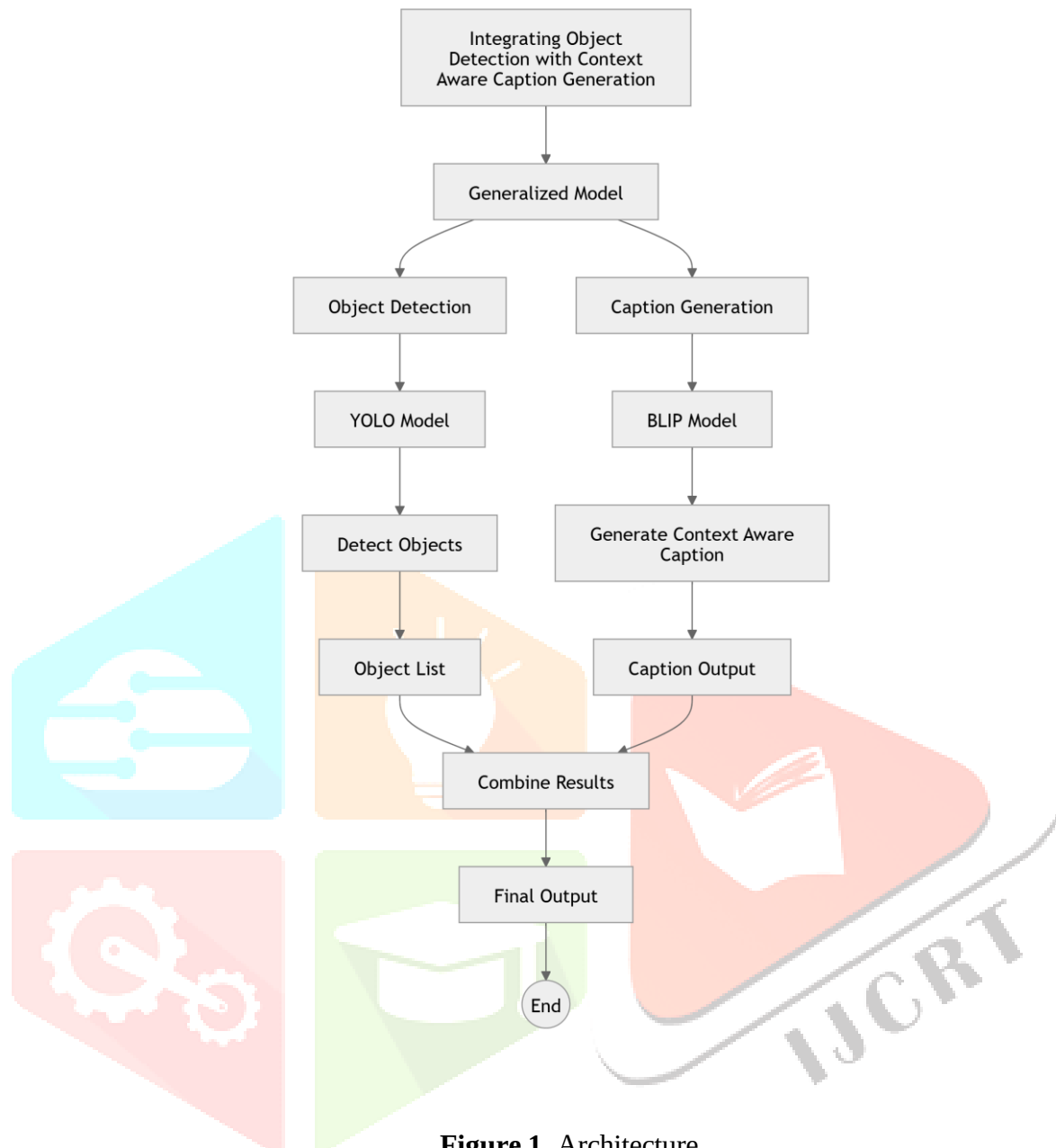


Figure 1. Architecture

- **System Overview:**

The specification related to the concept is either combining or integrating object detection and caption generation for the sole purpose of jointly describing the image [1, 6].

- **Generalized Model:**

The generalized model would trigger both very primary works of object detection and caption generation [5, 8].

- **Object Detection Path:**

- **YOLO Model:** This model is responsible for detecting as well as locating the objects into images [9].

- **Detect Objects:** YOLO Model measures the image to locate objects, gives bounding box coordinates, and labels.

- **Object List:** A list of all the detected objects with their necessary parameters.

- **Caption Generation Path:**

- **BLIP Model:** In this captioning model works [8].

- **Generate Context-Aware Caption:** Generate the caption according to the detected objects and scene contextualization.

- **Caption Output:** Now the captions are ready to be combined

- So here, the system is combining the generated lists of objects together with their captions to come up with what will be probably among the fullest descriptions such an image receives, one giving the vision while the other makes text out of it [6].

- The final output will be that fused one to really give a rich description of the image with its context.

especially in the case of frequent occurrence. In other words, it indicates which words the model tends to confuse with one another that may need to be improved by the means of processing the language to bring it up to the mark.

7.2 Precision-Recall Curves:

- Precision-Recall Curves of Object Detection Model

This curve shows the general performance of the model in detecting and classifying objects with an

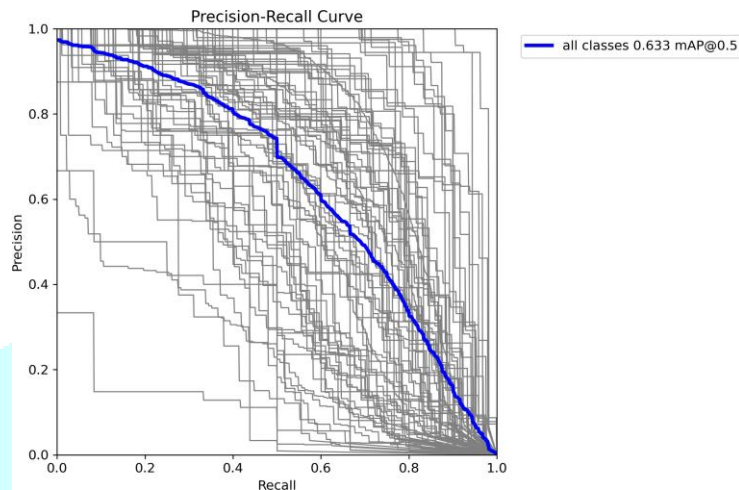


Figure 4. Precision-Recall Curves

average precision (mAP) score of 0.633. A high mAP value indicates good accuracy.

- Precision-Recall Curves of Caption Generation Model

This curve shows the trade-off between precision and recall when generating captions. The opposite

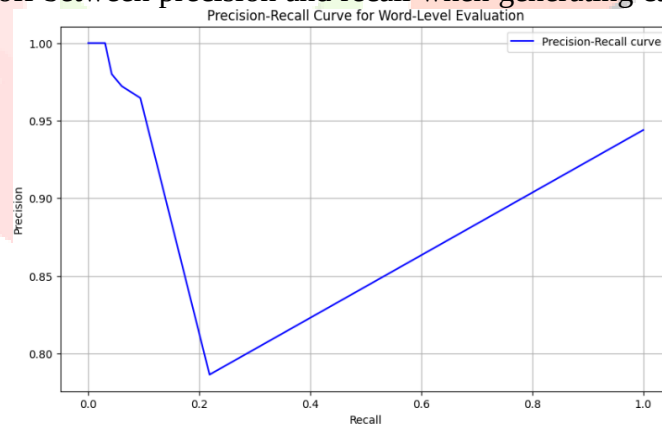


Figure 5. Precision-Recall Curves

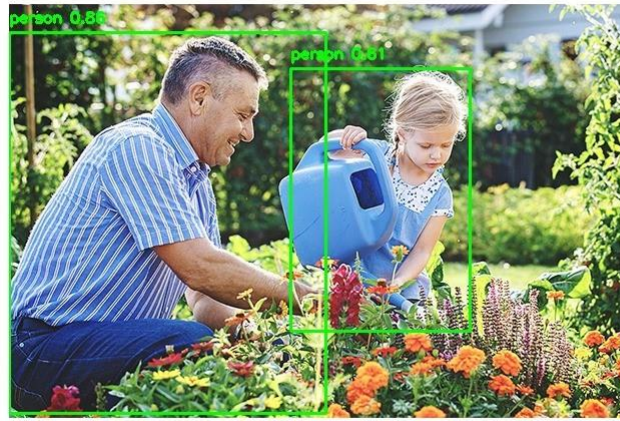
case of high precision and low recall may indicate that the model is over-adjusted to be not losing accuracy at the expense of coverage.

7.3 Qualitative Results:

- Qualitative Result1

In this result, people detection is given a very high confidence score. The caption "A man and a little girl

0: 448x640 2 persons, 69.6ms
 Speed: 2.5ms preprocess, 69.6ms inference, 1.5ms postprocess per image at shape (1, 3, 448, 640)
 Processed image saved at: processed_image.jpg



Generated Caption:
 a man and a little girl watering the flowers

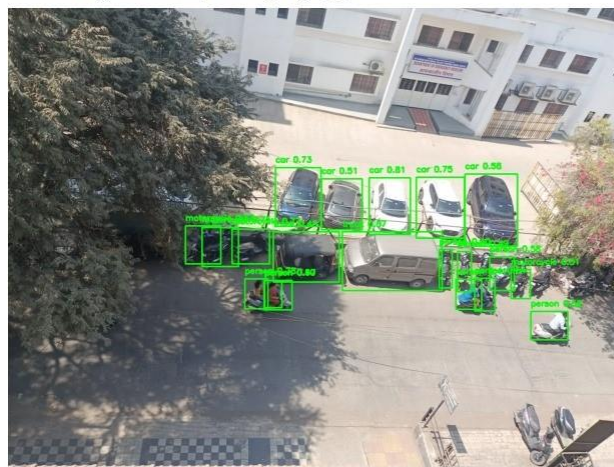
Figure 6. Qualitative Result1

watering the flowers” describes the event quite well and shows proper cooperation between detection and captioning.

- **Qualitative Result2:**

This image captures a busy scene with cars and people. The caption ”An area with vehicles parked on the street” effectively conveys the context, showcasing the model’s ability to describe complex scenes.

0: 480x640 5 cars, 4 motorcycles, 9 persons, 2 trucks, 62.1ms
 Speed: 3.4ms preprocess, 62.1ms inference, 2.3ms postprocess per image at shape (1, 3, 480, 640)
 Processed image saved at: processed_image.jpg



Generated Caption:
 an area with vehicles parked and parked on the street.

Figure 7. Qualitative Result2

8 Conclusion and Future Work

8.1 Conclusion

This project applies object detection for context-aware caption generation using various models, such as YOLO and BLIP. In addition, using the COCO 2017 dataset, we demonstrated a system that exactly identifies objects and then generates a relevant caption capturing the context of the scene.

The results, in terms of confusion matrices and precision-recall curves, show the ability of the model in both object detection and language generation. Qualitative examples further show the system’s ability to describe even some of the most complicated scenes accurately.

While the system shows strong performance, there is potential for enhancing caption clarity and reducing repetition.

8.2 Future Work

The future line of efforts could be improving the language model to allow more diversity and coherence of captioning. It would also be good to include a variety of datasets to make the system more generalizable to different contexts. Also, such advanced techniques as attention mechanisms could increase the model's capability of detailing parts of images.

References

- [1] Liu, Y., Zhang, Z., Wang, X., & Lin, H., "An image caption method based on object detection," *Multi-media Tools and Applications*, Springer, 2019. Available: link.springer.com
- [2] Abdelrahman, D., & Abdelrahman, O., "Application of Deep Learning for Object Detection," *Procedia Computer Science*, Elsevier, 2018. Available: sciencedirect.com
- [3] Zhao, C., Lin, X., et al., "Image captioning model using attention and object features to mimic human image understanding," *Journal of Big Data*, SpringerOpen, 2022. Available: journalofbig-data.springeropen.com
- [4] Liu, B., Yao, T., Pan, Y., Li, Y., & Mei, T., "Region-Object Relation-Aware Dense Captioning via Transformer," *IEEE Transactions*, 2022. Available: ieeexplore.ieee.org
- [5] Shukla, S., Singh, H., & Singh, R. K., "Image Captioning Using Improved YOLO V5 Model and Xception V3 Model," *Research Square*, 2023. Available: researchsquare.com
- [6] Gu, X., Huang, Z., Chen, X., et al., "Beyond a Pre-Trained Object Detector: Cross-Modal Textual and Visual Context for Image Captioning," *arXiv preprint arXiv:2205.04363*, 2022. Available: arxiv.org
- [7] Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M., "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv preprint arXiv:2004.10934*, 2020. Available: arxiv.org
- [8] Li, J., Li, D., Xiong, C., & Hoi, S. C. H., "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *arXiv preprint arXiv:2201.12086*, 2022. Available: arxiv.org
- [9] Jocher, G., et al., "YOLOv8: Cutting-edge object detection with Ultralytics," *Ultralytics Documentation*, 2023. Available: docs.ultralytics.com
- [10] Redmon, J., & Farhadi, A., "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, 2018. Available: arxiv.org
- [11] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P., "Focal Loss for Dense Object Detection," *ICCV*, 2017. Available: arxiv.org
- [12] Huang, T., Bai, S., Liu, Z., et al., "YOLO-World: Real-Time Open-Vocabulary Object Detection," *CVPR*, 2023. Available: arxiv.org
- [13] Li, J., Zhang, P., Li, C., et al., "BLIP-2: Bootstrapped Language Image Pretraining with Querying Transformer," *arXiv preprint arXiv:2301.12597*, 2023. Available: arxiv.org
- [14] Zhang, H., Li, X., Hu, X., et al., "VinVL: Making Visual Representations Matter in Vision-Language Tasks," *CVPR*, 2021. Available: arxiv.org
- [15] Wang, W., Li, X., Ma, Z., et al., "SimVLM: Simple Visual Language Model Pretraining with Weak Supervision," *ICLR*, 2022. Available: arxiv.org
- [16] Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R., "Meshed-Memory Transformer for Image Captioning," *CVPR*, 2020. Available: arxiv.org
- [17] Zhou, L., Palangi, H., Zhang, L., et al., "Unified Vision-Language Pre-Training for Image Captioning and VQA," *AAAI*, 2020. Available: arxiv.org
- [18] Chen, Y. C., Li, L., Yu, L., et al., "UNITER: UNiversal Image-TEXT Representation Learning," *ECCV*, 2020. Available: arxiv.org
- [19] Anderson, P., He, X., Buehler, C., et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," *CVPR*, 2018. Available: arxiv.org
- [20] Xu, K., Ba, J., Kiros, R., et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *ICML*, 2015. Available: arxiv.org