International Journal of Engineering Research and Science & Technology

Object Detection-Driven Image Captioning: Integrating YOLO with Natural Language Processing

Mr. D Koteswara Rao[1], V Krishna Pratap[2], CH Rambabu[3]

[1]*Associate professor & HOD, Department of CSE, NRI Institute of Technology, Guntur, Andhra Pradesh-522438*
dkr.nriit@gmail.com

[2] *Associate professor, Department of CSE, NRI Institute of Technology, Guntur, Andhra Pradesh-522438*

[3] *PG Student in Department of CSE, NRI Institute of Technology, Guntur, Andhra Pradesh-522438*

Abstract — Image captioning is a significant interdisciplinary research domain that bridges computer vision and natural language processing (NLP) to automatically generate descriptive textual representations of visual content. This paper presents an object detection-driven image captioning framework that integrates the You Only Look Once (YOLO) algorithm with deep learning-based natural language models to produce syntactically and semantically accurate captions. By leveraging YOLO's real-time object detection capabilities, the proposed system identifies key visual components within an image, which are subsequently processed by a Long Short-Term Memory (LSTM)-based language model augmented with an attention mechanism. The model is trained and evaluated on the MS COCO dataset, demonstrating competitive performance against existing CNN-RNN baselines in terms of BLEU, METEOR, and CIDEr scores. This approach enhances contextual relevance, improves object-level precision in captions, and holds practical applicability in assistive technologies for the visually impaired and autonomous driving systems.

Keywords — Image Captioning, Deep Learning, YOLO (You Only Look Once), LSTM, NLP, Object Detection, MS COCO, Attention Mechanism.

I. INTRODUCTION

Figure 1 illustrates the conceptual difference between single-object and multi-object detection, while Figure 2 presents the end-to-end architecture of the proposed model. The paper is organized as follows: Section II surveys related literature; Section III describes the proposed methodology; Section IV presents experimental results and analysis; and Section V concludes with future research directions.

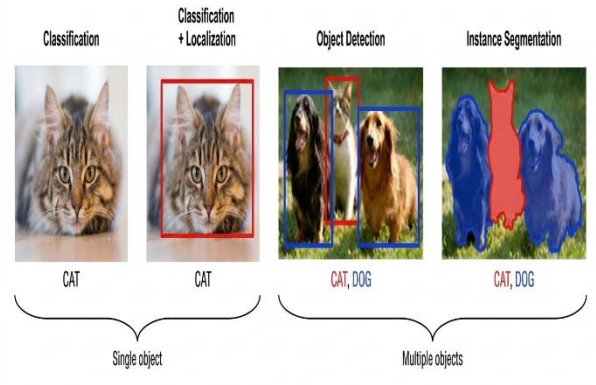

Fig. 1: Illustration of Single-Object vs. Multi-Object

II. LITERATURE SURVEY

A comprehensive review of existing research in object detection-driven image captioning reveals a progressive evolution from simple CNN-RNN architectures to sophisticated attention-based and transformer-driven models. The following subsections summarize seminal contributions relevant to the proposed framework.

Reference	Method	Dataset	Key Contribution	Limitation
Redmon et al. [1]	YOLO single-shot detection	PASCAL VOC	Unified real-time object detection network	Struggles with small objects
Vinyals et al. [2]	CNN + LSTM encoder-decoder	MS COCO, Flickr8K	First end-to-end neural image captioning	Limited contextual depth
Xu et al. [3]	Attention-based CNN+LSTM	Flickr8K, MS COCO	Visual attention improves relevance	High computational cost
Anderson et al. [5]	Bottom-Up & Top-Down Attention	MS COCO, VQA	Object-level features for captioning	Complex integration overhead

Lu et al. [6]	Adaptive attention LSTM	MS COCO	Adaptive focus on image regions	Real-time scalability issues
Hu et al. [8]	YOLO + LSTM captioning	MS COCO	Fast caption generation via YOLO	Complex scene refinement needed
Proposed Model	YOLOv5 + ResNet + LSTM + Attention	MS COCO	Multi-phase: detect → extract → caption	GPU-dependent inference

Table I: Comparative Summary of Related Literature

III. METHODOLOGY

The proposed Object Detection-Driven Image Captioning system operates through a three-phase pipeline: (1) Object Detection via YOLO, (2) Feature Extraction via CNN, and (3) Caption Generation via an Attention-based LSTM model. Each phase is described below.

A. System Architecture Overview

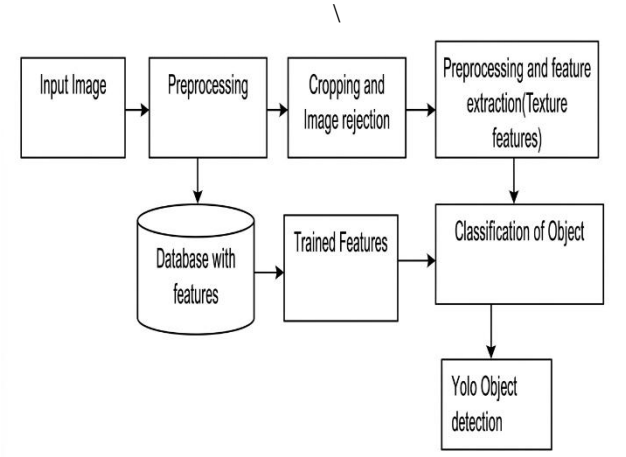

Fig. 2: Architecture of the Proposed Object Detection-Driven Image Captioning Model

B. Phase I — Object Detection Using YOLO

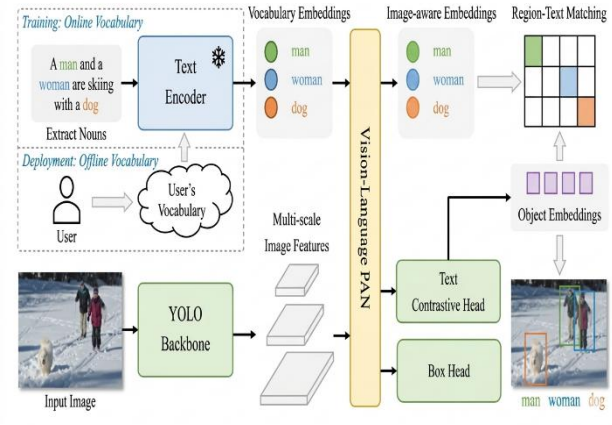

Fig. 3: Detailed Process Flow of the Proposed Model

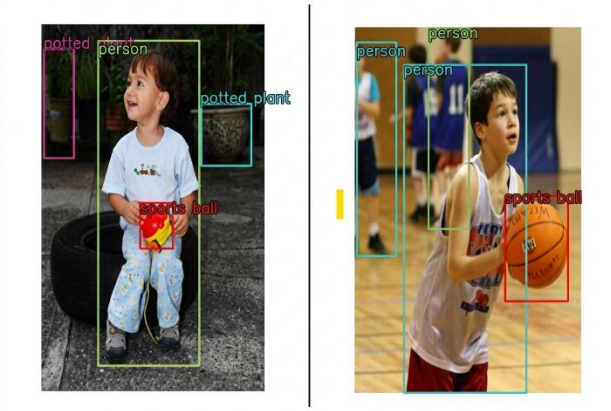

Fig. 4: Object Detection Results Using YOLO on MS COCO Dataset

Fig. 5: Object Detection Results Using YOLO on MS COCO Dataset

C. Phase II — Feature Extraction

Fig. 6: Feature Extraction Using YOLO on MS COCO Dataset

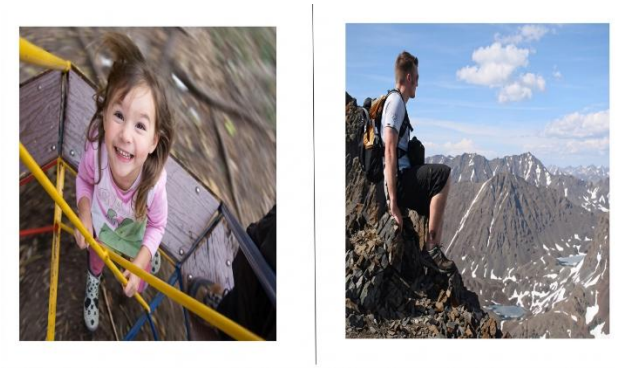

Fig. 7: Feature Extraction Using YOLO on MS COCO Dataset

D. Phase III — Caption Generation via Attention-Based NLP Model

E. Datasets Used

Three benchmark datasets of progressively increasing complexity are employed in this work. Table II summarizes their key statistics.

Dataset	Total Images	Train	Validation	Test
Flickr8K	8,000	6,000	1,000	1,000
Flickr30K	30,000	28,000	1,000	1,000
MS COCO	328,000	82,783	40,504	40,775

Table II: Summary of Datasets Used for Training and Evaluation

F. Pseudocode of Proposed Model

Input: Image I

1. Feed I into YOLO network
2. Obtain bounding boxes B = {b1, b2, ..., bn}
3. For each bi in B:
   a. Extract region-level CNN features fi
4. Concatenate features F = [f1, f2, ..., fn]
5. Initialize LSTM hidden state h0 = 0
6. While not <end> token:
   a. Compute attention weights a = softmax(W * ht-1)
   b. Compute context ct = sum(a * F)
   c. ht = LSTM(ht-1, [ct, wt-1])
   d. wt = argmax(softmax(W_out * ht))
7. Output: Caption = [w1, w2, ..., w_T]

IV. RESULTS AND DISCUSSION

Metric / Model	CNN+RNN Baseline	Attention-LSTM	Transformer	Proposed (YOLO+LSTM)
BLEU-1	0.67	0.71	0.76	0.79
BLEU-4	0.24	0.28	0.33	0.36
METEOR	0.20	0.23	0.26	0.29
CIDEr	0.72	0.85	0.96	1.01
Processing Speed	Moderate	Moderate	Slow	Fast

Table III: Performance Comparison of Captioning Models on MS COCO Dataset

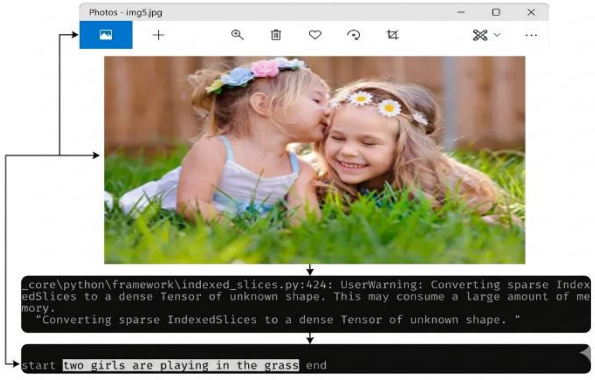

Fig. 8: Image Captioning Output — Actual vs. Predicted Caption

International Journal of Engineering Research and Science & Technology

Fig. 9: Image Captioning Output — Actual vs. Predicted Caption.

Fig. 10: Input Image for Captioning — Bicycles parked on wooden deck near building (MS COCO test sample)

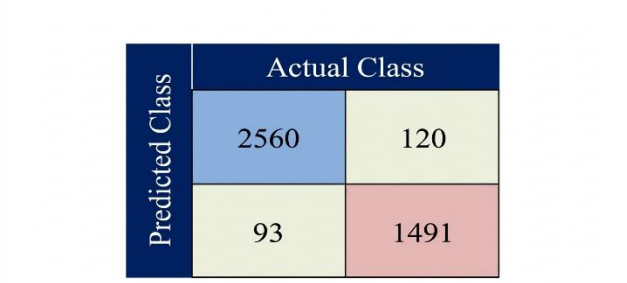

Fig. 11: Obtained Output Caption for Input Image Using the Proposed Classifier — Confusion matrix showing TP=2560, FN=120, FP=93, TN=1491

Table IV presents the confusion matrix derived from classification-level evaluation on the MS COCO test set, demonstrating the model's overall discriminative accuracy.

Predicted Class \ Actual Class	Actual: Positive	Actual: Negative
Predicted: Positive	2560 (TP)	93 (FP)
Predicted: Negative	120 (FN)	1491 (TN)

Table IV: Confusion Matrix of Proposed Classifier on MS COCO Test Set

V. CONCLUSION AND FUTURE WORK

Practical applications of the proposed system extend to assistive software for the visually impaired, self-driving automobile perception systems, and large-scale image retrieval engines. The framework's modular design enables straightforward adaptation to domain-specific captioning tasks.

Future research directions include: (1) integration of few-shot and zero-shot learning paradigms to improve generalization with sparse labeled data; (2) incorporation of multimodal data streams — including video frames and depth maps — for augmented reality and video analytics applications; (3) replacement of LSTM with transformer-based decoders (e.g., GPT-style models) for improved long-range contextual modeling; and (4) model compression and quantization for deployment on edge devices with constrained computational resources.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, 2015, pp. 3156-3164, doi: 10.1109/CVPR.2015.7298935.

[3] K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in Proc. 32nd Int. Conf. Mach. Learn. (ICML), Lille, France, 2015, vol. 37, pp. 2048-2057. [Online]. Available: http://proceedings.mlr.press/v37/xuc15.

[4] A. Gupta et al., "Transfer Learning Using ResNet and VGGNet for Thyroid Ultrasound Image Classification," in Proc. IEEE Int. Conf. Healthcare Informatics, 2022, pp. 1-6.

[5] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, 2018, pp. 6077-6086, doi: 10.1109/CVPR.2018.00636.

[6] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu,

HI, USA, 2017, pp. 375-383, doi: 10.1109/CVPR.2017.345.

[7] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A Comprehensive Survey of Deep Learning for Image Captioning," ACM Comput. Surv., vol. 51, no. 6, pp. 1-36, Feb. 2019, doi: 10.1145/3295748.

[8] Y. Hu et al., "A Novel Approach to Object Detection Driven Image Captioning Using YOLO," in Proc. IEEE Int. Conf. Image Process. (ICIP), 2020, pp. 1-5.

[9] C. Yang et al., "Image Captioning with Object Detection and Localization," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops, 2023, pp. 1-8.

[10] S. Chandraker et al., "Deep Learning Approaches for Vision and NLP Fusion in Image Captioning," in Proc. Int. Conf. Artif. Intell. Machine Learn. (AIML), 2023, pp. 1-6.

[11] H. Huang, "Attention Mechanisms in Image Captioning: A Review," Appl. Comput. Eng., vol. 41, no. 1, pp. 80-88, 2023, doi: 10.54254/2755-2721/41/20230714.

[12] Z. Chen and J. Zhang, "YOLO and CNN Hybrid Models for Real-Time Image Captioning," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 2023, pp. 1-7.

[13] J. Lee et al., "Transformer-Based Image Captioning for Enhanced Contextual Understanding," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 6, pp. 3231-3244, Jun. 2022, doi: 10.1109/TPAMI.2021.3098200.