A Novel Approach for Object Detection using NLP

*** Prabhati Bharti**
**** Asst. Prof. Ayush Kumar**

ABSTRACT

In this paper, we tackle the problem of natural language object retrieval, where the goal is to locate a target object within an image based on a natural language description. Unlike text-based image retrieval, natural language object retrieval requires understanding the spatial relationships between objects in the scene and the overall context of the image. To address this, we introduce a novel Context Recurrent ObjNet (CRO) model that serves as a scoring function for candidate bounding boxes, integrating both spatial configurations and scene-level contextual information. Our model processes query text, local image features, spatial configurations through a recurrent network, producing a probability score for each candidate box based on the query. Additionally, the model leverages visual-linguistic knowledge from the image captioning domain to enhance retrieval accuracy. Experimental results show that our method effectively incorporates both local and global context, significantly outperforming previous baselines across various datasets and scenarios and demonstrates the ability to utilize large-scale vision and language datasets for knowledge transfer.

Keywords:- *Part of Speech, Hindi, Emission probabilities, Transition probability***.**

*Prabhati Bharti, Research Scholar, Department of Computer Science & Engineering, Radharaman Engineering College, Bhopal , itsprabhati16@gmail.com
**Ayush Kumar, Asst. Professor, Department of Computer Science & Engineering, Radharaman Engineering College, Bhopal, India

I. INTRODUCTION

Object detection and retrieval have been crucial tasks within the field of computer vision, traditionally approached through text-based image retrieval or object recognition systems that rely on predefined categories. However, as vision and language-based AI applications expand, the need for more nuanced approaches to object retrieval has become apparent. Specifically, in scenarios where users specify target objects using natural language rather than predefined labels, there is an increasing demand for systems capable of understanding complex, free-form descriptions. In this paper, we tackle the problem of natural language object retrieval, where the objective is to locate a specified object within an image based on a descriptive query provided in natural language [1]. This task represents a significant advancement over traditional object detection by requiring the model to capture subtle relationships between objects in the scene and utilize both spatial and contextual cues to achieve accurate localization show in figure 1.

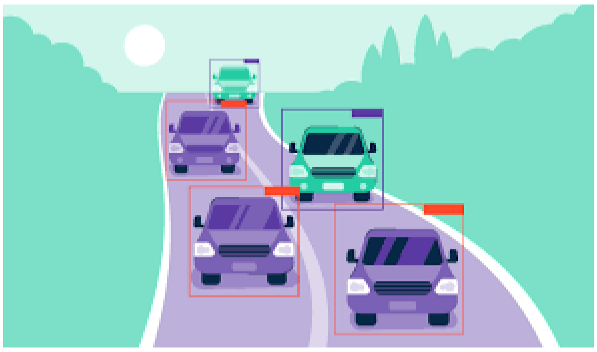

Figure 1: Object detection by using traditional model

Unlike conventional image retrieval, which typically focuses on retrieving entire images based on keyword or feature-based matches, natural language object retrieval necessitates the understanding of object interactions, spatial configurations, and overall scene context within a single image. A text-based image retrieval system might be capable of retrieving an image that contains "a cat on a sofa," but natural language object retrieval goes a step further: it must not only recognize that both objects (cat and sofa) are present but also discern their spatial relationship to each other. Additionally, the retrieval process must consider the entire context of the image, differentiating between multiple objects and their relative positions based on the nuances provided by the query text. For example, in a scenario where there are multiple objects, our system should be able to distinguish between “the book on the table next to the lamp” and “the book on the table by the window,” accurately localizing the intended target [2].

To address these challenges, we propose a novel Context Recurrent ObjNet (CRO) model as a scoring mechanism for candidate bounding boxes. This model integrates both spatial configurations and scene-level contextual information, allowing it to assess the relevance of each candidate region in relation to the natural language query. The CRO model leverages a recurrent neural network to effectively process query text and analyze spatial configurations, enabling it to assign a probability score to each candidate box. The probability score reflects the likelihood that the candidate box aligns with the query, offering a highly interpretable scoring method that facilitates robust object retrieval. Our CRO model’s design allows it to incorporate multiple sources of information, such as local image features, spatial configurations, and global context, yielding a more comprehensive understanding of the scene and enhancing retrieval accuracy.

A core feature of the CRO model is its ability to transfer knowledge from visual-linguistic tasks, such as image captioning, to the natural language object retrieval domain. By leveraging learned representations from image captioning datasets, the model gains access to a wealth of cross-modal knowledge that aids in interpreting diverse language expressions and object features. This knowledge transfer enables our model to better generalize across different datasets and retrieval scenarios, a significant advantage given the vast variability in natural language descriptions. For example, if the query refers to “the blue book,” the model can utilize its prior knowledge of visual-linguistic associations to more effectively identify objects with similar attributes, regardless of their specific spatial configurations.

Our approach builds on recent advancements in computer vision and natural language processing, particularly in the integration of convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) for sequential data processing. By combining these architectures, the CRO model can process multiple types of input data in parallel. The CNN component is responsible for extracting local image features from candidate bounding boxes, providing detailed information about the visual characteristics of each region. Meanwhile, the RNN component interprets the query text and analyzes the spatial configurations of candidate boxes, generating probability scores that reflect the degree to which each candidate matches the query’s description. This multimodal approach allows the CRO model to maintain high levels of accuracy across a range of retrieval tasks.

Our experimental results demonstrate that the CRO model effectively incorporates both local and global contextual information, achieving significant improvements over baseline methods. We evaluated our model across multiple datasets, including standard benchmarks and specialized datasets designed for natural language object retrieval, to validate its generalizability and robustness. In each case, the CRO model outperformed prior approaches, achieving higher precision in locating target objects based on natural language descriptions. Notably, our method shows particular strength in scenarios involving complex queries with multiple spatial and contextual cues, underscoring the importance of a context-aware approach in this domain.

In addition to achieving state-of-the-art performance, the CRO model demonstrates the scalability needed for real-world applications. By training on large-scale vision and language datasets, our approach capitalizes on the wealth of available visual-linguistic data, further enhancing its ability to handle diverse queries. This scalability is essential in practical applications, such as robotics and human-computer interaction, where the system must interpret and respond to natural language instructions with high accuracy. For example, a robotic assistant could use our model to respond to commands like “pick up the red coffee mug next to the laptop” by accurately identifying and locating the specified object in its environment.

II. LITERATURE REVIEW

Kaur and Singh (2023) provide a thorough review of object detection using deep learning techniques [3]. Their work highlights the major advancements deep learning has brought to the field, focusing on different architectures and models that have significantly enhanced both detection accuracy and speed. By examining the impact of these techniques, their review underscores the transformative role deep learning plays in improving object detection systems. This comprehensive analysis offers valuable insights into how deep learning innovations have shaped the current landscape and advanced the effectiveness of object detection technologies.

Kaur and Singh (2023) provide an in-depth review of object detection advancements driven by deep learning [4]. Their analysis highlights the substantial impact of deep learning techniques on enhancing detection accuracy and speed. The review covers a range of architectures and models that have been pivotal in advancing the field, illustrating how these innovations have contributed to more effective and efficient object detection. By examining key developments and breakthroughs, their work underscores the transformative role of deep learning in improving performance and capabilities within the object detection domain.

Alzahrani and Al-Baity (2023) propose a novel object recognition system aimed at aiding the visually impaired [5]. Their approach leverages deep learning techniques combined with Arabic annotations to improve accessibility and usability specifically for Arabic-speaking users. By incorporating language-specific annotations, the system enhances the effectiveness of object recognition and provides more relevant feedback to users. This development marks a significant advancement in creating inclusive technology solutions, addressing the unique needs of Arabic-speaking visually impaired individuals and demonstrating progress towards more universally accessible assistive technologies.

Wase et al. (2023) investigate the integration of object detection models with large language models (LLMs) to boost safety and security applications [6]. Their research showcases how combining these advanced technologies can enhance the robustness and accuracy of object detection systems in critical situations. By fusing object detection capabilities with the linguistic understanding of LLMs, the study highlights significant improvements in system performance, making it more reliable and effective in complex and high-stakes environments. This approach represents a promising advancement in developing sophisticated safety and security solutions through the synergy of object detection and language processing technologies.

Chiu et al. (2024) present a study on integrating object detection and natural language processing (NLP) models to develop a personalized attraction recommendation agent [7]. Their research, published in Advanced Engineering Informatics, emphasizes the use of combined object detection and NLP technologies within a smart product service system. This integration aims to enhance user experience by providing personalized recommendations based on visual and textual data. The study illustrates how advancements in these technologies can be leveraged to create more sophisticated and user-centric applications in recommendation systems.

Ma (2024) explores the application of artificial intelligence (AI) technologies in both NLP and image processing, focusing on their ethical and social impacts [8]. Published in the International Journal of Computer Science and Information Technology, this paper reviews various AI technologies and their implications. The author provides a comprehensive overview of how AI-driven advancements in NLP and image processing are reshaping industries and the associated ethical considerations, highlighting the broader societal impact of these technologies.

Manakitsa et al. (2024) offer a review of machine learning and deep learning techniques applied to object detection, semantic segmentation, and human action recognition within machine and robotic vision systems [9]. Their article in Technologies provides an extensive examination of recent developments in these areas, focusing on how these technologies enhance the capabilities of machine vision systems. The review covers both theoretical and practical aspects, offering insights into the current state and future directions of research in these fields.

Mangalika (2024) provides a comprehensive study on the progression from object recognition to content-based image retrieval within computer vision [10]. This review examines key developments and applications in the field, highlighting advancements in technologies that enable more effective image recognition and retrieval. The paper discusses how object recognition has evolved to support content-based image retrieval systems, enhancing their ability to search and retrieve images based on visual content rather than metadata alone. Mangalika’s study underscores the significance of these advancements in improving image processing capabilities and offers insights into the ongoing evolution and future directions of computer vision technologies.

Mahesh Babu et al. (2024) investigate the integration of object recognition into conversational chatbots using deep learning and machine learning techniques [11]. Their research, published in Conversational Artificial Intelligence, explores how object recognition can enhance chatbot interactions by enabling them to understand and respond to visual inputs. The study demonstrates the potential of combining NLP and computer vision to create more interactive and responsive conversational agents.

Buettner and Kovashka (2024) examine the role of attribute context in vision-language models for object recognition and detection [12]. Presented at the IEEE/CVF Winter Conference on Applications of Computer Vision, their research delves into how contextual information influences the performance of vision-language models. The paper highlights the significance of incorporating attribute context to improve the accuracy and efficiency of object recognition and detection tasks.

Chen et al. (2024) propose Taskclip, a method for extending large vision-language models to enhance task-oriented object detection [13]. Their preprint on arXiv introduces a novel approach that integrates vision-language models with task-specific objectives to improve object detection performance. The study demonstrates the potential of leveraging large-scale vision-language models for more precise and context-aware detection tasks.

III. PROPOSED WORK

In this section, we present our Context Recurrent ObjNet (CRO) model for natural language object retrieval. During testing, the model is provided with an image, a natural language object query, and a set of candidate bounding boxes (e.g., generated by object proposal methods such as EdgeBox [14]). The system's task is to identify and select the subset of bounding boxes that best match the query text.

3.1 Context Recurrent ObjNet (CRO)

The model comprises three Long Short-Term Memory (LSTM) units—$LSTM_{language}$, $LSTM_{local}$, and $LSTM_{global}$ as well as a local and a global Convolutional Neural Network (CNN), a word embedding layer and a word prediction layer. During testing, given an image I, a query text sequence S, and a set of candidate bounding boxes $\{b_i\}$ in I, the network produces a score s_i for each candidate box b_i. This score is based on the local image descriptors x_{box} on b_i, the spatial configuration $x_{spatial}$ of the box in relation to the scene, and the global contextual feature $x_{context}$.

In this work, the local descriptor x_{box} is extracted by CNN_{local} from local region I_{box} on b_i and we use feature extracted by another network CNN_{global} on the whole image I_{im} as scene-level contextual feature $x_{context}$. The spatial configuration of b_i is an 8-dimensional representation.

$$x_{spatial} = [x_{min}, y_{min}, x_{max}, y_{max}, x_{center}, y_{center}, w_{box}, h_{box}] \quad (1)$$

where w_{box} and h_{box} are the width and height of b_i. We normalize image height and width to be 2 and place the origin at the image center, so that coordinates range from 1 to 1.

At test time, given an input image I, a query text S and a set of candidate bounding boxes $\{b_i\}$, the query text S is scored on i-th candidate box using the likelihood of S conditioned on the local image region, the whole image and the spatial configuration of the box, computed as

$$s = p(S|I_{box}, I_{im}, x_{spatial})$$

$$= \prod_{\omega_t \in s} p(\omega_t|\omega_{t-1}, \ldots, \omega_1, I_{box}, I_{im}, x_{spatial}) \quad (7)$$

and the highest scoring candidate boxes are retrieved.

IV. EXPERIMENTS

We evaluate our method across multiple datasets, ranging from small to relatively large scales. Each candidate box is assigned one of these classes, and a bag of words is created from the predicted class's ImageNET [15] synset, including the category name and synonyms. This word bag is then projected into a vector space and matched with the projected query text using cosine similarity to generate a score. The sentence embedding in [16] is predefined, with the 7K object classifier being the only component trained. [16] Also proposes an instance-matching model that uses online APIs during testing; however, since our work assumes a self-contained system without external API reliance, we only compare against the CAFFE-7K category model in [16].

Our recurrent architecture is also inspired by LRCN [17], which has proven effective for image captioning and retrieval. We adapt the LRCN model, trained on MSCOCO [18] for captioning, as an object retriever by applying it to candidate bounding boxes. Given an image I, a set of candidate boxes and a query text S_{query}, we compute $p(S_{query}|I_{box})$, the probability of the query text S_{query} given the local image region I_{box} as output by LRCN—to score each candidate box, retrieving the top-scoring matches.

V. RESULTS

Table 1 displays the top-1 precision, the percentage of cases where the highest-scoring region is correct, in the first scenario, where the candidate set includes all annotated boxes in the image.

Notably, the CAFFE-7K model struggles to provide informative results when none of the query words match its category names, resulting in an empty word bag and the same score across all regions. In contrast, our CRO model consistently provides deterministic results, as it can represent unknown words with an "<unk>" token. Following [16], we evaluate using "P@1-NR," which measures non-random top-1 precision for informative cases, and "P@1," which represents top-1 precision across all cases, including non-informative ones where a random guess is used.

Table 1: Top-1 precision of our method compared with baselines on annotated bounding boxes in ReferIt dataset.

Method	**P@1**
CAFFE-7K	27.73%
LRCN(17)	38.38%
CRO	72.74%

Results show that our full CRO model achieves the highest top-1 precision. Table 1 also reveals that pretraining on image captioning, incorporating spatial configuration, and adding scene-level context each contribute to improved performance, with spatial configuration $x_{spatial}$) yielding the most substantial boost. This improvement is expected, as spatial configuration enhances retrieval not only in queries with explicit spatial relationships (e.g., "the man on the left") but also by allowing the model to learn prior distributions of object locations, such as "sky" usually appearing at the top of the scene and "ground" at the bottom. Figure 2 show the Probability Comparison Chart of previous and proposed model.

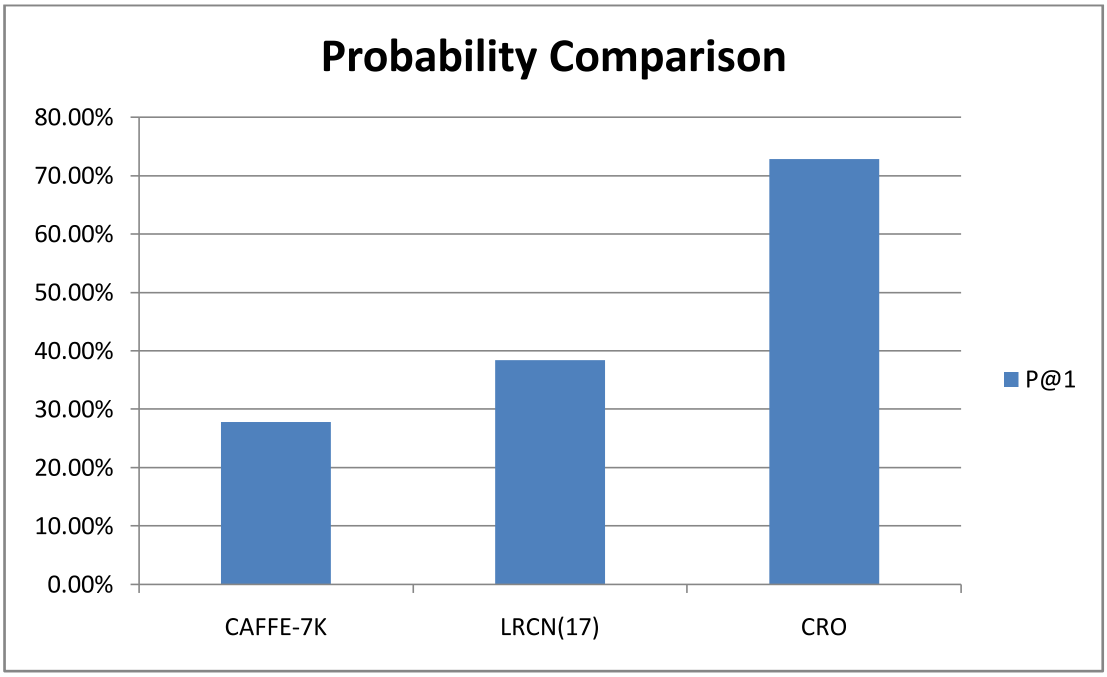

Figure 2: Probability Comparison Chart

VI. CONCLUSION

In this paper, we tackle the task of natural language object retrieval with our Context Recurrent ObjNet (CRO), a recurrent neural network model that scores candidate bounding boxes based on local image descriptors and spatial configurations scene context. Our findings demonstrate that integrating spatial configurations significantly enhances retrieval performance. The recurrent structure of our model provides an end-to-end trainable scoring function, resulting in notable improvements over baseline methods.

We also show that natural language object retrieval benefits from knowledge transfer through pretraining on image captioning tasks, which helps bridge the gap created by limited datasets with object-level annotations. Since object-level annotations are scarce and often challenging to obtain, we illustrate that leveraging datasets with image-level annotations offers a feasible alternative, as these are generally easier to collect. Building on this work, subsequent results show promising advances through approaches that directly encode phrases for retrieval and methods that predict image segmentations instead of bounding boxes.

REFERENCES

1. Torfi, Amirsina, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A. Fox. "Natural language processing advancements by deep learning: A survey." arXiv preprint arXiv:2003.01200 (2020).
2. Li, Yanfen, Hanxiang Wang, L. Minh Dang, Tan N. Nguyen, Dongil Han, Ahyun Lee, Insung Jang, and Hyeonjoon Moon. "A deep learning-based hybrid framework for object detection and recognition in autonomous driving." IEEE Access 8 (2020): 194228-194239.
3. Kaur, Ravpreet, and Sarbjeet Singh. "A comprehensive review of object detection with deep learning." Digital Signal Processing 132 (2023): 103812.
4. Sharada, K., Wajdi Alghamdi, K. Karthika, Ahmed Hussein Alawadi, Gulomova Nozima, and V. Vijayan. "Deep Learning Techniques for Image Recognition and Object Detection." In E3S Web of Conferences, vol. 399, p. 04032. EDP Sciences, 2023.
5. Alzahrani, Nada, and Heyam H. Al-Baity. "Object recognition system for the visually impaired: a deep learning approach using Arabic annotation." Electronics 12, no. 3 (2023): 541.
6. Wase, Zeba Mohsin, Vijay K. Madisetti, and Arshdeep Bahga. "Object Detection Meets LLMs: Model Fusion for Safety and Security." Journal of Software Engineering and Applications 16, no. 12 (2023): 672-684.
7. Chiu, Ming-Chuan, Cheng-Zhou Tsai, and Yu-Chen Huang. "Integrating object detection and natural language processing models to build a personalized attraction recommendation agent in a smart product service system." Advanced Engineering Informatics 61 (2024): 102484.
8. Ma, Wenxiang. "The Application of Artificial Intelligence Technology in Natural Language Processing and Image Processing and Its Ethical and Social Impact." International Journal of Computer Science and Information Technology 2, no. 2 (2024): 124-128.
9. Manakitsa, Nikoleta, George S. Maraslidis, Lazaros Moysis, and George F. Fragulis. "A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision." Technologies 12, no. 2 (2024): 15.
10. Mangalika, Udula. "Object Recognition to Content Based Image Retrieval: A Study of the Developments and Applications of Computer Vision." (2024).
11. Mahesh Babu, A., Malik Jawarneh, José Luis Arias- Gonzáles, Meenakshi, Kishori Kasat, and K. P. Yuvaraj. "Conversational Chatbot With Object Recognition Using Deep Learning and Machine Learning." Conversational Artificial Intelligence (2024): 335-352.
12. Buettner, Kyle, and Adriana Kovashka. "Investigating the Role of Attribute Context in Vision-Language Models for Object Recognition and Detection." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5474-5484. 2024.
13. Chen, Hanning, Wenjun Huang, Yang Ni, Sanggeon Yun, Fei Wen, Hugo Latapie, and Mohsen Imani. "Taskclip: Extend large vision-language model for task oriented object detection." arXiv preprint arXiv:2403.08108 (2024).
14. C. L. Zitnick and P. Doll´ar. Edge boxes: Locating object proposals from edges. In Proceedings of the European Conference on Computer Vision (ECCV), pages 391–405. Springer, 2014.
15. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 248–255. IEEE, 2009.
16. S. Guadarrama, E. Rodner, K. Saenko, N. Zhang, R. Farrell, J. Donahue, and T. Darrell. Open-vocabulary object retrieval. In Robotics: Science and Systems, 2014.
17. J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2625–2634, 2015.

18. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014, pages 740–755. Springer, 2014.