Apply KG to enhance Image Description Generation

Le Vinh Thuan[1], Nguyen Minh Khoa[2], Nguyen Vinh Thanh[3], and Nguyen Thi Dinh[4,]

[{1,2,3}]Faculty of Information Technology, University of Science, VNU-HCM, Ho Chi Minh City, Vietnam,
[4]Ho Chi Minh City University of Industry and Trade, Ho Chi Minh City, Vietnam

[1]lvthuan23@apcs.fitus.edu.vn, [2]nmkhoa23@apcs.fitus.edu.vn, [3]23120012@student.hcmus.edu.vn, [4]dinhnt@huit.edu.vn

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

The automatic generation of natural language descriptions from images has emerged as one of the most challenging and important tasks in computer vision and natural language processing [?, ?]. This interdisciplinary field, commonly known as image captioning or image description, requires machines to not only recognize objects and their spatial relationships within images but also to generate coherent, contextually appropriate textual descriptions that capture the semantic essence of visual content [?, ?].

Recent advances in deep learning have significantly improved the performance of image description systems, with approaches ranging from encoder-decoder architectures [?] to attention-based mechanisms [?] and transformer-based models [?]. However, existing methods often struggle with generating rich, contextually aware descriptions that go beyond simple object enumeration and spatial relationships [?, ?].

Knowledge graphs have shown tremendous potential in enhancing various computer vision tasks by providing structured semantic information and enabling reasoning capabilities [?, ?]. The integration of knowledge graphs with image understanding systems allows for more sophisticated semantic reasoning and can bridge the gap between visual perception and high-level conceptual understanding [?, ?]. Furthermore, the incorporation of external knowledge bases enables systems to generate more informative and contextually rich descriptions by leveraging world knowledge beyond what is directly observable in the image [?, ?].

This paper presents a comprehensive four-stage pipeline for generating detailed image descriptions that combines state-of-the-art object detection with

knowledge graph construction and natural language inference. Our approach employs YOLOv11 [?] for robust object detection in the first stage, followed by knowledge graph construction and entity linking. Unlike traditional approaches that rely solely on visual features, our system incorporates enriched contextual data through knowledge graph reasoning to perform sophisticated contextual analysis and inference, ultimately generating more comprehensive and meaningful image descriptions.

The main contributions of this work are: (1) A novel four-stage architecture that seamlessly integrates computer vision, knowledge representation, and natural language processing for image description; (2) An enhanced knowledge graph-based reasoning approach that enriches visual understanding with external knowledge; (3) A streamlined contextual analysis module that focuses on enriched data interpretation; and (4) Comprehensive evaluation demonstrating the effectiveness of our approach in generating high-quality, contextually aware image descriptions.

2 Related Work

Research on semantically-rich Image Description Generation (IDG) intersects multiple key domains, including Computer Vision (CV), Knowledge Representation (specifically Knowledge Graphs, KG), and Natural Language Processing (NLP). Early studies in image description primarily focused on generating fluent but factually limited captions based on simple encoder-decoder architectures. More recent works have shifted towards incorporating explicit semantic structures to enrich the output. At the same time, the field of Knowledge Representation has been exploring effective methods to inject external, common-sense knowledge into visual tasks. This section reviews three main directions relevant to the proposed framework: Object Detection and Feature Extraction for IDG, Knowledge-Enhanced Semantic Grounding, and Contextual Inference and Natural Language Generation.

2.1 Object Detection and Feature Extraction for IDG

High-quality image description relies on robust detection and accurate attribute extraction from the visual input. Object detection frameworks, such as the YOLO family [5] and Faster R-CNN [6], have become the foundation for most modern vision-to-language models. Recent advancements in object detection, including variants that enhance performance on small objects or crowded scenes [2, 8], continuously provide more granular and reliable entity bounding boxes.

While most detection-based IDG models use bounding boxes to apply attention mechanisms [10] or build Scene Graphs [11], they often stop at basic object classification (e.g., "person," "car"). This leaves a gap in research tailored to adapting the raw visual output for explicit knowledge retrieval. Our approach specifically leverages the visual features extracted (e.g., object class, bounding box, detected attributes) as input for a subsequent semantic grounding phase,

using a robust detector like YOLOv11 [1] to ensure high precision in entity localization.

2.2 Knowledge-Enhanced Semantic Grounding

To move beyond superficial descriptions, several studies have focused on incorporating external knowledge. Knowledge Graphs (KGs) are the most widely adopted form of structured external knowledge, proving vital in providing commonsense or domain-specific facts [9].

The core challenge is Semantic Grounding: effectively mapping coarse visual detections to specific KG entities. Techniques like Entity Linking [3] are used to resolve ambiguities (e.g., distinguishing "apple" as a fruit vs. a company). Some models apply KG features during the decoding phase, primarily to aid vocabulary selection or fact-checking [13]. However, this late integration often fails to fundamentally change the input structure of the generation model. Our methodology differentiates by placing the KG at the center of the pipeline, using it to systematically enrich all raw visual entities and attributes with specific semantic details before the generation phase, thereby transforming a simple object list into a dense semantic structure.

2.3 Contextual Inference and Natural Language Generation (NLG)

The final stage of IDG involves converting the enriched semantic data into coherent, natural sentences. Conventional methods often rely on powerful sequence-to-sequence models to implicitly learn the mapping from features to text [7]. However, when dealing with highly structured, fact-dense input (like the output of a KG), implicit mapping can lead to factual omissions or illogical sentence structure.

Early Template-based models [4] offered logical structure but lacked fluency. More sophisticated methods employ explicit Relational Inference to discover implicit actions or states between objects [12]. Our work emphasizes a dedicated inference layer that operates on the structured, KG-enriched data. This layer's role is to perform Contextual Analysis and Logical Structuring—essentially prioritizing the semantic facts and organizing them into a logical flow. This is crucial for satisfying the goal of generating a detailed "description" (which requires a narrative flow) rather than a simple "caption" (a single descriptive sentence). The structured output then feeds into a robust NLG module (potentially template-based or an advanced decoder) to ensure both high fluency and factual accuracy.

2.4 Research Gap and Contribution

Previous research highlights the complementary strengths of advanced object detectors and external knowledge sources. However, most methods remain limited either by their reliance on implicit reasoning within the decoder or by a superficial integration of knowledge (e.g., only using it for fact-checking). A unified

framework that systematically leverages structured knowledge for entity enrichment and uses this enriched data for explicit, structured inference before NLG remains a significant challenge. Addressing this gap requires a unified framework that tightly couples robust visual analysis with a structured KG and a dedicated inference mechanism to deliver accurate, scalable, and semantically deep Image Description Generation.

3 Theoretical Basis

The development of our Knowledge Graph-enhanced Image Description Generation (IDG) system necessitates a solid foundation built upon advanced Computer Vision, structured knowledge representation, and natural language processing techniques. Core methods include deep convolutional networks for object detection, graph-based mechanisms for semantic grounding, and language models for coherent text synthesis. Our framework utilizes these techniques in sequence: raw visual features are extracted, enriched with external knowledge, and finally converted into descriptive language.

3.1 Core Concepts and Definitions

Image Description Generation (IDG): A task that involves generating a textual output that is not only visually accurate but also semantically rich, often incorporating contextual or world knowledge. This differs from standard Image Captioning by prioritizing factual detail and narrative depth.

Knowledge Graph (KG): A structured representation of knowledge composed of entities, attributes, and their relationships (often expressed as subject-predicate-object triples). KGs serve as the external memory for our system, providing the necessary factual context to enrich object detections.

Semantic Grounding / Entity Linking: The process of resolving ambiguity and connecting detected entities from the visual domain to specific, canonical entities within the Knowledge Graph. This mechanism transforms raw detection labels into Enriched Entities with specific attributes.

Semantic Inference: The intermediate process between KG enrichment and NLG. It involves reasoning over the enriched entities and their inferred relationships to construct a structured, logical sequence of facts (triples) that will guide the text generation.

3.2 Visual Feature Extraction and Object Detection

The initial step in our pipeline is to robustly identify and localize entities within the input image. This relies on state-of-the-art Object Detection models, such as those in the YOLO family [5]. YOLO (You Only Look Once) is preferred for its real-time capability and high accuracy in simultaneously predicting bounding boxes and class probabilities.

The output of this phase is the set of Raw Visual Features, which includes:

- **Raw Entities:** Class labels and bounding box coordinates for each detected object.
- **Visual Attributes:** Low-level visual properties extracted or inferred, such as color, relative size, and spatial location.

These features are the initial, ambiguous input that the system must semantically enrich.

3.3 Knowledge Graph and Entity Semantic Grounding

The Knowledge Graph serves as the domain-specific ontology, built upon a rich dataset like Visual Genome. It is used to overcome the limited scope of raw visual detection.

Semantic Grounding (Entity Linking): Given a raw entity e_{raw} and its visual attributes A_v, the system seeks to find the best matching canonical entity e_{KG} in the Knowledge Graph. This process typically involves a scoring mechanism that combines the confidence score of the raw detection and the degree of match between A_v and the known attributes of e_{KG} within the KG.

$$\text{Score}(e_{\text{KG}}|e_{\text{raw}}, A_v) = f(P(e_{\text{raw}}) \cdot \text{Similarity}(A_v, \text{Attributes}(e_{\text{KG}})))$$

Where $P(e_{\text{raw}})$ is the confidence from the object detector, and Similarity measures the alignment between the visual features and the KG's structural knowledge. An entity is considered Enriched once it is successfully linked, gaining access to all associated facts and relationships in the KG.

3.4 Semantic Inference and Text Generation

The final component ensures the conversion of structured, factual knowledge into coherent, high-quality descriptive text.

Semantic Inference: This layer utilizes the enriched entities and their explicit relationships from the KG to perform Contextual Analysis. It generates a set of ordered Semantic Triples $\mathcal{T} = \{(S_i, P_i, O_i)\}$ that best describe the scene. The triples are prioritized based on their semantic relevance and visual saliency. This structured output ensures that the final description maintains factual accuracy and a logical flow.

Natural Language Generation (NLG) via Template-Based Method: Given the need for factual accuracy and a controlled narrative flow in IDG, a Template-Based NLG approach can be utilized [4]. This method maps the ordered set of semantic triples $\mathcal{T}$ onto a pre-defined syntactic structure (template). This allows for a balance between descriptive richness (provided by $\mathcal{T}$) and linguistic fluency/correctness (provided by the template), offering high control over the output style, which is often a requirement for specialized description systems.

References

1. Cao, C., Song, Y., He, J.: An overview of yolo series: From yolov1 to yolov8. Journal of Physics: Conference Series **2685**(1), 012002 (2023)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV). pp. 213–229 (2020)
3. Chen, Y., Ma, L., Huang, Y., Zhang, S., Qian, Y.: A survey of knowledge graph-based entity linking from text to knowledge graph. IEEE Access **9**, 13000–13020 (2021)
4. Kulkarni, M., Kulkarni, G.S., Krishna, V.S.R., Singh, S.K.: Generating image descriptions using template and semantic composition. In: European Conference on Computer Vision (ECCV). pp. 126–140 (2014)
5. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016)
6. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **39**(6), 1137–1149 (2017)
7. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **39**(10), 1974–1985 (2017)
8. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies for real-time object detection. arXiv preprint arXiv:2207.02696 (2022)
9. Wang, P., Wu, Q., Shen, C., van den Hengel, A.: Fvqa: Fact-based visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **40**(10), 2413–2427 (2018)
10. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning (ICML). pp. 2048–2057 (2015)
11. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph generation with global context. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5831–5840 (2018)
12. Zhao, S., Sun, Z., Li, X., Gong, M., Tao, D., Wei, W., Zhang, X., Li, X.: Attentional relational reasoning for image captioning. In: AAAI Conference on Artificial Intelligence (AAAI). pp. 9463–9470 (2019)
13. Zhu, Y., Yang, Z., Salakhutdinov, R., Xing, E.P.: Incorporating commonsense knowledge into image captioning via graph convolutional networks. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9008–9017 (2019)