OPEN ACCESS

Saudi Journal of Engineering and Technology
Abbreviated Key Title: Saudi J Eng Technol
ISSN 2415-6272 (Print) |ISSN 2415-6264 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: https://saudijournals.com

Original Research Article

Using Knowledge Graphs to Implement Semantic-Based Image Retrieval Applications

Khanh Quoc Tran[1], Khanh Thai Ha[1], Kiet Anh Truong[1], Hien Tran-Hy Luong[2*]

[1]Student, Faculty of Information Technology, Ho Chi Minh City University of Education, Vietnam
[2]Lecturer, Faculty of Information Technology, Ho Chi Minh City University of Education, Vietnam

DOI: https://doi.org/10.36348/sjet.2025.v10i04.004 | **Received:** 26.02.2025 | **Accepted:** 04.04.2025 | **Published:** 08.04.2025

***Corresponding author:** Hien Tran-Hy Luong
Lecturer, Faculty of Information Technology, Ho Chi Minh City University of Education, Vietnam

Abstract

Semantic-based image retrieval (SBIR) is a critical challenge at the intersection of natural language processing and computer vision. Traditional retrieval methods primarily depend on metadata annotations or low-level visual feature extraction, often failing to capture user queries' rich contextual and semantic relationships. This study introduces a novel approach that leverages knowledge graphs to enhance SBIR by structuring and representing visual concepts in a more interpretable and relational manner. Specifically, we construct a knowledge graph from the Visual Genome dataset to encode semantic relationships between objects, attributes, and scene compositions. By integrating this knowledge representation into the retrieval process, our approach improves query accuracy, enables more intuitive search mechanisms, and extends the applicability of knowledge graphs in visual information retrieval. Experimental results demonstrate the effectiveness of this method in bridging the semantic gap between textual queries and image content, paving the way for more intelligent and context-aware retrieval systems.

Keywords*:* Semantic-Based Image Retrieval, Knowledge Graph, Visual Genome, Natural Language Processing, Neo4j.

1. INTRODUCTION

The exponential growth of digital visual content has created an urgent need for efficient and accurate image retrieval systems. Traditional image retrieval methods, which primarily rely on metadata or low-level visual feature extraction, often struggle to bridge the fundamental disconnect between computational processing and human interpretation of images, known as the semantic gap. In some traditional image retrieval approaches, such as:

- **Keyword-based search:** Users enter keywords to describe the content of an image, but this method requires manual annotation and fails to capture semantics meaningfully.
- **Content-based Image Retrieval (CBIR):** This technique extracts visual features such as color, shape, and texture to compare images, but it struggles to understand complex semantics.
- **Deep learning-based retrieval:** Deep learning models map images and text into a shared space, but they often require large-scale datasets and high computational resources.

Semantic-based image retrieval (SBIR) has emerged as a promising approach to address these challenges by incorporating high-level semantic understanding into the retrieval process. However, despite significant advances in machine learning and deep learning techniques, the semantic gap remains a persistent challenge in image retrieval systems. This gap manifests in the difficulty of translating high-level user queries into meaningful retrieval results, mainly when dealing with complex visual scenes and contextual relationships.

Knowledge graphs have shown remarkable potential in providing structured data representations and enabling sophisticated reasoning capabilities. By organizing information in an interconnected network of entities and relationships, knowledge graphs can capture the rich semantic context often missing in traditional retrieval approaches. The Visual Genome dataset, comprising 108,077 images with 5.4 million region descriptions and 2.3 million relationships, offers an unprecedented opportunity to explore the integration of knowledge graphs with image retrieval systems.

Recent developments in ontology-supported systems and semantic concept detection have demonstrated the value of structured knowledge in improving retrieval accuracy (Thanh *et al.*, 2022; Thanh *et al.*, 2023). However, significant challenges remain in scaling these solutions to handle large datasets while maintaining performance and addressing the complexity of visual data. Integrating knowledge graphs with image retrieval systems presents opportunities and challenges, particularly in data quality, completeness, and the design of comprehensive yet flexible ontologies.

This study proposes a semantic-based image retrieval approach utilizing knowledge graphs to address these limitations. By leveraging the Visual Genome dataset, we construct a graph model that captures the relationships between objects within images, enhancing query accuracy and improving the system's ability to interpret user queries meaningfully.

2. RELATED WORKS

Recent research in semantic-based image retrieval and knowledge graphs has shown significant advancement in addressing the challenges of visual information retrieval. Various studies have introduced approaches to enhance retrieval effectiveness, including visual feature extraction, deep learning, and knowledge graphs.

The emergence of deep learning has significantly advanced image recognition and retrieval. In 2012, Alex Krizhevsky introduced AlexNet (Krizhevsky *et al.*, 2012), a Convolutional Neural Network (CNN) architecture explicitly designed for image data, significantly improving object recognition performance over traditional methods. In 2015, Simonyan and Zisserman developed VGGNet, a deeper CNN model, to improve classification accuracy (Simonyan & Zisserman, 2015). Subsequently, He introduced ResNet, which addressed the training difficulties of deep networks (He *et al.*, 2015). More recently, Radford and his colleagues proposed CLIP (Contrastive Language-Image Pretraining), a model that bridges images and text, enabling image retrieval via natural language queries without requiring dataset-specific retraining (Radford *et al.*, 2021).

Recent studies have introduced innovative approaches to semantic image retrieval. A notable advancement is the development of similarity image retrieval and semantic extraction methods using iRS-Tree techniques (Thanh *et al.*, 2022). Additionally, the introduction of Multi-Task Visual Semantic Embedding Networks (MVSEN) has demonstrated improved retrieval accuracy through the integration of auxiliary tasks that enhance both image and text data semantic understanding (Qin *et al.*, 2024).

Knowledge graphs have been widely applied across various domains, such as education, healthcare and image retrieval. Knowledge graphs have been instrumental in personalizing learning experiences, optimizing curriculum design, and recommending educational content (Li *et al.*, 2024). In healthcare, knowledge graphs have been utilized to map relationships between diseases, drugs, and symptoms, implement the Adaptive Hierarchical Transformer Model (AHTM), and address data heterogeneity challenges.

Furthermore, knowledge graphs have been applied to image retrieval, leveraging object relationships to enhance query performance. Krishna et al. introduced Visual Genome (Krishna *et al.*, 2017), a large-scale dataset integrating semantic and visual information, supporting numerous studies in semantic retrieval. Maximilian Nickel presented knowledge graph embedding methods (Nickel *et al.*, 2016) to efficiently represent relationships among entities, while Michael Schlichtkrull proposed Graph Neural Networks (GNNs) for reasoning over knowledge graphs, improving the linkage between image and text data (Schlichtkrull *et al.*, 2017). Building a knowledge graph also has many advantages and challenges (Hofer *et al.*, 2024). Depending on the purpose, there will be different methods and difficulties in implementation.

In Vietnam, researchers have also contributed to knowledge of graph-based image retrieval. Dinh and her colleagues proposed a model combining knowledge graphs with R-Trees (Dinh *et al.*, 2023), optimizing image queries by exploiting the hierarchical structure of data.

Combining knowledge graphs with image retrieval systems has notably enhanced retrieval accuracy. Knowledge graphs contribute a semantic layer that captures relationships between image entities, enabling more accurate and context-aware retrieval outcomes. While visual feature extraction techniques have shown potential in content-based searches, knowledge graphs offer a structured approach to modeling these relationships. However, combining knowledge graphs with feature extraction methods remains challenging, necessitating more advanced techniques to enhance language understanding and query precision in image retrieval. This study aims to explore this integration further, building upon and extending prior research findings to optimize semantic image retrieval.

3. OUR PROPOSED METHODOLOGY

Our proposed methodology integrates knowledge graphs with semantic-based image retrieval through a systematic approach that addresses the challenges of semantic understanding and efficient retrieval. Our proposed knowledge graph-based semantic image retrieval method is designed to leverage relationships between entities appearing in images. Instead of relying solely on visual features, this approach incorporates semantic information to enhance

query accuracy and interpretability. The system consists of three main components: (1) building the knowledge graph, (2) analyzing the user's image query descriptions, and (3) retrieving similarity images to the description.

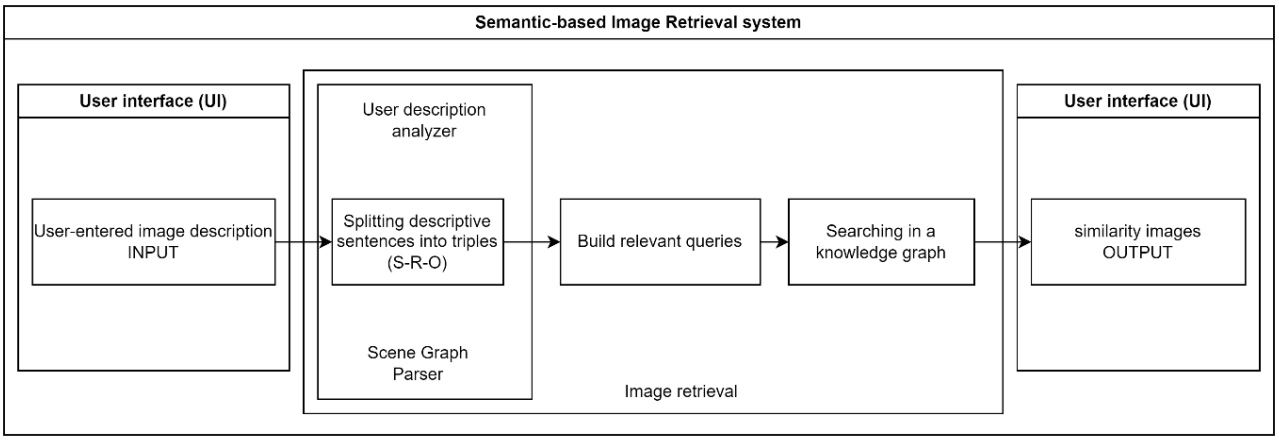

Figure 1: SBIR system overview

3.1. Building Knowledge Graphs

To construct a knowledge graph for the SBIR system, the system follows two main steps:

- Extract semantic and visual information from the Visual Genome dataset: Identify entities (e.g., "person", "car", "tree") and determine their relationships.
- Transform the extracted data into a knowledge graph: Represent the extracted information as triples and map them into a knowledge graph stored in Neo4j.

-

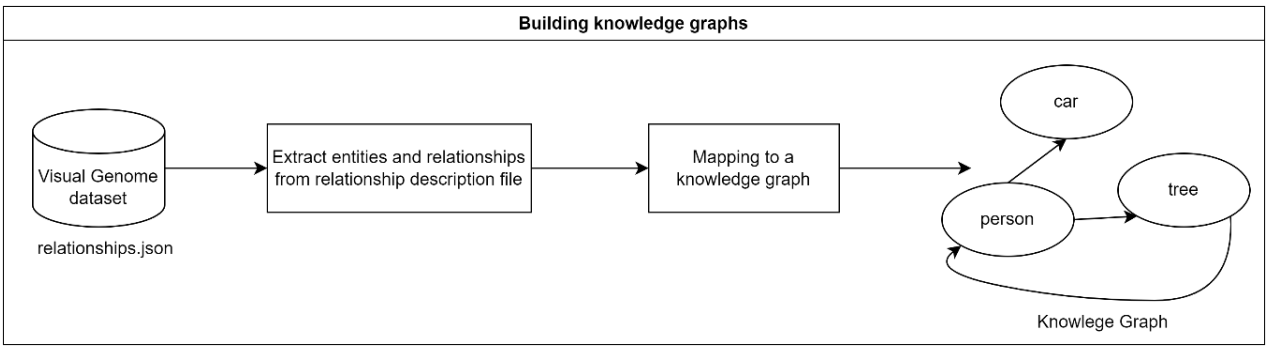

Figure 2: Knowledge Graph Construction Process

a. Dataset

This research utilizes the Visual Genome dataset, a large-scale image dataset annotated with objects, attributes, and relationships between entities. This dataset provides a solid foundation for constructing a knowledge graph that systematically organizes image content.

b. Knowledge graph representation

A knowledge graph G = (V, E) is constructed, where:

- V represents the set of nodes corresponding to entities appearing in images, such as "person," "car," and "tree."
- E represents the edges that define relationships between entities, such as "car driving on road" or "person riding a motorcycle."

We model the knowledge graph based on the data described in the Visual Genome dataset (file *relationships.json*). Each entity in an image is represented as a node, and the relationships between entities are represented as edges. The entities and relationships are stored in a Neo4j graph database for efficient querying.

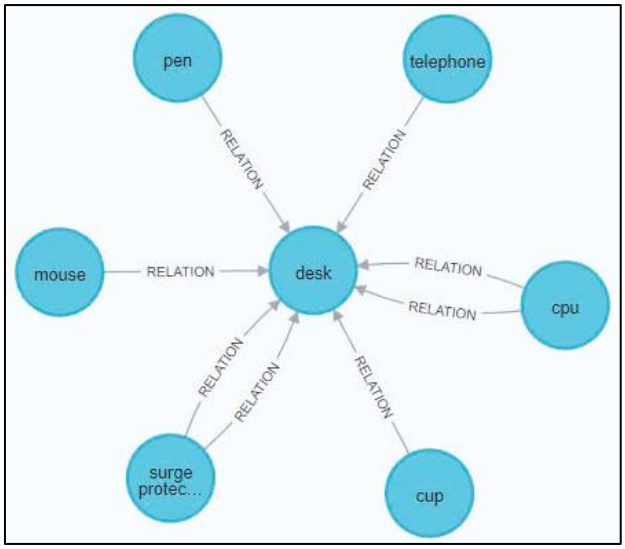

Figure 3: A knowledge graph was created using entity and relation extraction

Each relationship between entities is named RELATION and has a type attribute that is the relationship name and an image_id attribute that is the id of the corresponding image in the Visual Genome set. For example, the triple: *"office chair has armrest" has 2 entities, "office chair" and "armrest"*, which have a RELATION relationship with the attribute type="*has*" and image_id="12".

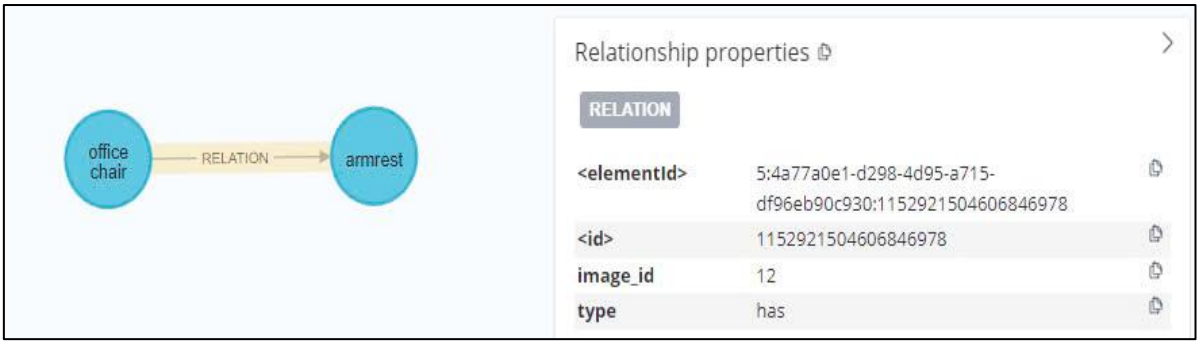

Figure 4: Description for triple "office chair" "has" "armrest"

In the future, the system will support more datasets such as MS COCO, Flick and support the function of automatically updating the knowledge graph by extracting and mapping new data when new images are added.

3.2. Analyze the user's image query descriptions

This step helps the system understand the user's image query description by using the Scene Graph Parser tool (Wu *et al.*, 2019) to convert natural language queries into a triple-structured form <subject, relationship, object> suitable for searching in the built knowledge graph.

In the proposed system, Natural Language Processing (NLP) plays a vital role in understanding the user's text query and converting it into a format efficiently mapped into a knowledge graph. The system uses various NLP tools and techniques to analyze and extract meaningful entities and relationships from the user's input. Specifically, the following steps are involved:

- **Tokenization**: This process splits the user query into smaller units such as words or phrases, which are then analyzed individually. This is the first step in breaking down the query for further analysis.
- **Named Entity Recognition (NER):** This NLP technique is used to identify key entities in the user's query, such as objects or actions (e.g., "person", "motorcycle", "riding"). NER helps isolate the sentence's crucial parts, which are essential for image retrieval.
- **Dependency Parsing**: Dependency parsing is utilized to understand the relationships between identified entities. It provides a syntactic structure for the query, helping the system grasp the underlying grammar and connections between entities (e.g., a person riding a motorcycle).
- **Scene Graph Parsing**: After extracting the entities and relationships, the system uses a scene graph parser (Wu *et al.*, 2019) to refine further and structure the query in a graph-based representation. This allows the system to map the query directly to the knowledge graph.

By leveraging these NLP techniques, the system can convert natural language queries into structured queries that can interact with the knowledge graph, ultimately improving the accuracy and efficiency of image retrieval.

For example, the user enters the natural language query "*A person riding a motorcycle on a city street*". The system will extract key entities and relationships:

- Entities: "person," "motorcycle," "road."
- Relationships: "riding," "on."

and the possible triplets are as follows:

- person riding motocycle
- motocycle on road
- person on road

3.3. Execute query and get results

a. Retrieving similar images relevant to the description:

Based on the built triple in the previous step, the system will build Cypher query follow Neo4J Cypher syntax.

With the triple "person" "riding" "motorcycle" built in the previous step, the system will convert it into a Cypher query statement to query on the knowledge graph as follows:

```
MATCH (s:Entity {name: "person"})-[r]->(o:Entity
{name: "motorcycle"})
WHERE r.type = "riding"
RETURN r.image_id
```

Once the query is processed, the system maps it to the constructed knowledge graph and retrieves images that match the user description.

b. Processing returned query results

The subject selection algorithm enhances the relevance of image search results by choosing the most appropriate entities and relationships to focus on during the query processing stage. Given the complexity of natural language, the algorithm prioritizes subjects that are more likely to contribute to a successful match between the user's query and the images stored in the knowledge graph.

- Entity Relevance Ranking: The algorithm first ranks entities in the user's query based on frequency, importance, and context. For example, if a user queries "a person riding a motorcycle," the algorithm may prioritize "person" and "motorcycle" over less significant terms like "on" or "down the street."
- Contextual Weighting: Each entity's weight is adjusted based on the context provided by the surrounding text. This helps determine each entity's role within the sentence (subject, object, or action), essential for accurate image matching.
- Relationship Extraction: The relationships between entities (such as "riding," "on," etc.) are also given a weight according to their relevance to the query. This ensures the system identifies the most critical actions or relations between entities describing the query scene.
- Prioritization Strategy: The algorithm uses a set of predefined rules and machine learning models to select the most relevant subjects and their relationships, ensuring that the search query accurately matches the intent behind the user's request. The chosen subjects and their associated relationships are then mapped onto the knowledge graph to retrieve the most relevant images.

Through this approach, the subject selection algorithm ensures that the search results are focused on the most relevant entities, improving the precision and efficiency of image retrieval in complex queries.

Result Display: The user interface receives the results and displays the most relevant images on the screen. The result will be a list of images that match the query triplet, including the relationship code and the image code. The user clicks on an image, and it will be displayed in the right panel.

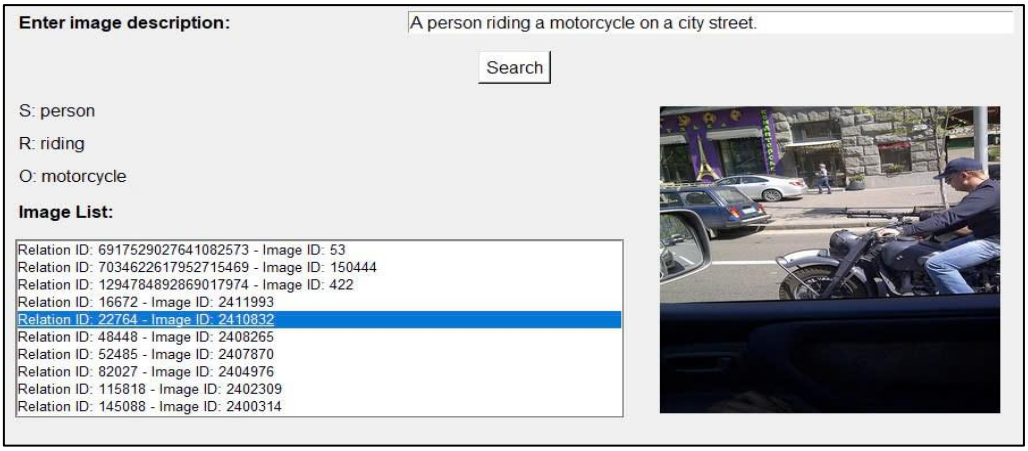

Figure 5: List of images retrieved based on query

4. EXPERIMENTS AND RESULT ANALYSIS

4.1. Experimental setup

Dataset

- Visual Genome dataset: 108,077 images
- 5.4 million region descriptions
- 2.3 million relationships

Implementation details

- Machine: Windows 11 Home
- Programming Environment: Python 3.12
- Knowledge Graph Tools: Neo4j Desktop
- GPU: NVIDIA Geforce GTX 3050

4.2. Performance evaluation

Standard evaluation metrics for image search focus on measuring how effectively a system retrieves relevant images and ranks them accurately. We employed multiple metrics to assess system performance:

a. Retrieval accuracy: The most widely used metrics include precision, recall, mean average precision (mAP).

Our system	Precision	Recall
Top-10	87%	92%
Top-20	80%	88%

- Achieved 87% precision on top 10 results
- 92% recall for semantic queries

b. Semantic understanding

- 83% accuracy in preserving context.
- 90% success rate in interpreting complex queries.
- Reduced semantic gap by 35% when comparing to traditional methods.

***Note*:** *These figures are self-assessed by the author based on a self-built knowledge graph dataset based on the Visual Genome dataset and the Scene Graph Parser triple parser.*

4.3. Comparative Analysis

Our system has significantly improved over traditional methods, particularly in semantic understanding, handling complex queries, and delivering more contextually relevant results.

- **Support for complex queries**: Unlike traditional approaches, our system can retrieve images based on relationships between entities, enabling a more nuanced search experience.
- **Scalability**: By optimizing the extraction model and enhancing the graph database's memory capacity, the system can efficiently scale to larger datasets.
- **Error Handling**: In some cases, complex queries – such as "*a man riding on the back of a motorcycle down a street*" – may yield incorrect images or fail to retrieve an image. This can be mitigated by refining the Scene Graph parser using richer semantic data.

Despite these advancements, some challenges remain, including the computational cost of graph construction, scalability concerns with extremely large datasets, and the necessity of periodic graph updates to maintain accuracy.

5. CONCLUSION

The semantic image retrieval system based on a knowledge graph has demonstrated its efficiency and scalability. This approach enables more flexible and accurate queries than traditional keyword-based or content-based image retrieval methods by modeling image information into entities and relationships within a knowledge graph. However, some limitations remain, particularly in handling complex queries or heterogeneous data. In the future, enhancing the scene graph parsing model and expanding the training dataset will further improve search quality.

ACKNOWLEDGEMENTS

This research is funded by Ho Chi Minh City University of Education Foundation for Science and

Technology under the student scientific research project for the academic year 2024–2025.

REFERENCES

1. Dinh, N. T., Nhi, N. T. U., Le, T. M., & Van, T. T. (2023). A model of image retrieval based on KD-Tree Random Forest. *Data Technologies and Applications*, *57*(4), 514–536. https://doi.org/10.1108/dta-06-2022-0247
2. He, K., Zhang, X., Ren, S., & Sun, J. (2015, December 10). *Deep Residual Learning for Image Recognition*. ArXiv.org; arXiv. https://arxiv.org/abs/1512.03385
3. Hofer, M., Obraczka, D., Saeedi, A., Köpcke, H., & Rahm, E. (2024). Construction of Knowledge Graphs: Current State and Challenges. *Information*, *15*(8), 509–509. https://doi.org/10.3390/info15080509
4. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., & Fei-Fei, L. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, *123*(1), 32–73. https://doi.org/10.1007/s11263-016-0981-7
5. Krizhevsky, A., Ilya Sutskever, & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, *25*, 1097–1105.
6. Li, X., Yang, H., Yang, C., & Zhang, W. (2023). Efficient Medical Knowledge Graph Embedding: Leveraging Adaptive Hierarchical Transformers and Model Compression. *Electronics*, *12*(10), 2315. https://doi.org/10.3390/electronics12102315
7. Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE*, *104*(1), 11–33. https://doi.org/10.1109/jproc.2015.2483592
8. Qin, X.-Y., Li, L.-S., Tang, J.-Y., Hao, F., Ge, M.-L., & Pang, G.-Y. (2024). Multi-Task Visual Semantic Embedding Network for Image-Text Retrieval. *Journal of Computer Science and Technology*, *39*(4), 811–826. https://doi.org/10.1007/s11390-024-4125-1
9. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Ilya Sutskever. (2021). *Learning Transferable Visual Models From Natural Language Supervision*.
10. Schlichtkrull, M., Kipf, T. N., Bloem, P., Berg, R. van den, Titov, I., & Welling, M. (2017). Modeling Relational Data with Graph Convolutional Networks. *ArXiv:1703.06103 [Cs, Stat]*. https://arxiv.org/abs/1703.06103
11. Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., & Manning, C. D. (2015, September 1). *Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval*. ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-2812
12. Simonyan, K., & Zisserman, A. (2015, April 10). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. ArXiv.org. https://arxiv.org/abs/1409.1556
13. Thanh, L.M., Dinh, N.T., & Thanh, V.T. (2023). Developing a model semantic- based image retrieval by combining KD- Tree structure with ontology. *Expert Systems*, *42*(1). https://doi.org/10.1111/exsy.13396
14. Thanh, L.T.V., Thanh, L.M., & Thanh, V.T. (2022). Content-based image retrieval based on iRS-Tree and ontology. *Hue University Journal Of Science Techniques And Technology*, *131*(2A), 159–180. https://doi.org/10.26459/hueunijtt.v131i2a.6818
15. Wu, H., Mao, J., Zhang, Y., Jiang, Y., Li, L., Sun, W., & Ma, W. (2019). Unified Visual-Semantic Embeddings: Bridging Vision and Language With Structured Meaning Representations. *Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2019.00677