Improving Efficiency Image Captioning by Using Attention Mechanism Combined with Knowledge Graph

Tam Khoi Tran[1], Nguyen Thi Uyen Nhi[2], Thanh Manh Le[3], and Nguyen Thi Dinh[4(✉)]

[1] Ton Duc Thang University, Ho Chi Minh City, Vietnam
242805006@student.tdtu.edu.vn
[2] University of Economics, The University of Da Nang, Da Nang, Vietnam
nhintu@due.udn.vn
[3] University of Sciences, Hue University, Hue, Vietnam
lmthanh@hueuni.edu.vn
[4] Ho Chi Minh City University of Industry and Trade, Ho Chi Minh City, Vietnam
dinhnt@huit.edu.vn

Abstract. Improving the accuracy of image caption extraction is one of the problems related to computer vision and depends on natural language processing. Therefore, the image captioning problem is performed by many different combination methods. This paper proposes a method to build captions for multi-object images by combining attention mechanisms to focus on the main objects and relationships in the image; thereby building a knowledge graph and combining semantics from ConcepNet. To do this, each object in the input image is recognized and classified using a deep learning network. The objects and relationships between the objects of interest will be added to the knowledge graph. Finally, semantics from the built knowledge graph and ConcepNet are combined to complete the caption for the input image. The method is tested on MS-COCO and Flickr30k image datasets with BLUE-1 and BLUE-4 metrics to evaluate these results. The experimental results are compared with the results of other methods on the same dataset; this proves the feasibility, correctness, and efficiency of the implemented method.

Keywords: Image Captioning · Attention Mechanisms · Knowledge Graph

1 Introduction

Nowadays, multimedia data is increasing rapidly over time, which is both an opportunity and a challenge for image analysis and processing problems. In particular, image captioning is an important problem in the field of computer vision combined with natural language processing applied in many fields such as biomedicine, education, agriculture and transportation [1]. This is a difficult problem because it not only requires extracting the correct information from the image but also combining natural language to describe the image in terms of content and semantics on the image. Currently, there are many

N. T. Nguyen et al. (Eds.): ACIIDS 2025, LNAI 15684, pp. 72–87, 2025.
https://doi.org/10.1007/978-981-96-6005-6_6

techniques used in combination to improve the efficiency of image captioning extraction for images such as using deep learning [2], multi-modal knowledge graph (KG) [3], transformer model [4], using a dual attention mechanism [5], etc. In this paper, the attention mechanism is used to focus on the object of interest compared to the main object, and at the same time combining the knowledge graph to extract image captioning.

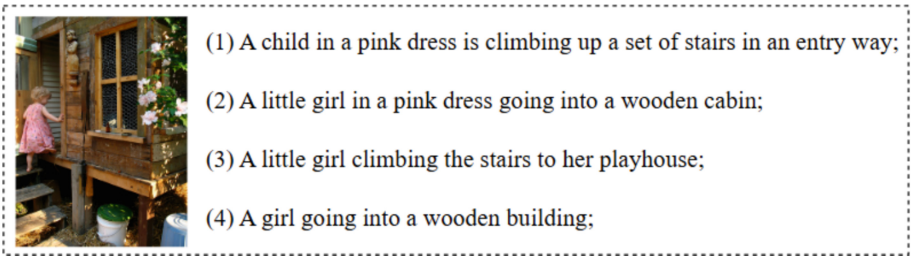

Fig. 1. Illustration of annotations for image 1000268201.jpg (Flickr30k)

Any image can have multiple annotations captions to describe it, the image in Fig. 1 has four annotations as follows: (1) "A child in a pink dress is climbing up a set of stairs in an entryway"; (2) "A little girl in a pink dress going into a wooden cabin"; (3) "A little girl climbing the stairs to her playhouse"; (4) "A girl going into a wooden building". Therefore, image captions that want to be highly accurate must be selective regarding objects and relationships, adding semantics to the most appropriate image context. Thus, paying attention to the object of interest and selecting appropriate keywords will bring high efficiency to image captions.

Nowadays, multi-object images are very rich and diverse, each image not only consists of many objects but also the relationship between objects is very complex and needs to be analyzed to clarify the image semantics. Incorrect image semantic identification can lead to unpredictable consequences; for example, in the medical field, diagnosing diseases based on images and making wrong decisions is a warning. Wrong judgments in agriculture also lead to serious economic losses. Therefore, in order to improve the efficiency of image recognition and semantic analysis, a feasible, suitable, efficient, and accurate implementation method is needed. Therefore, the method of combining attention mechanism and knowledge graph to build image captioning is carried out with the following steps:

(1) Identify objects and determine the main objects in the image using a deep learning network;
(2) Applying attention mechanism to optimize the objects related to the main object and relationships.
(3) Building a knowledge graph based on the results of identifying main objects and relationships using the attention mechanism;
(4) Combining semantics from the knowledge graph and knowledge from ConcepNet to generate captions for input images;
(5) Optimize caption results for images

The main contributions of the paper include: (1) optimizing the identification of objects and relationships between objects using attention mechanism according to the characteristics of separate datasets; (2) constructing a knowledge graph based on the results of objects and object relationships identified by attention mechanism; (3) combining semantics from knowledge graph and ConcepNet to extract captions for images; (4) experimental evaluation of image caption extraction according to $BLEU - n$, $(n = 1, 4)$ measures on two image sets MS-COCO [6] and Flickr30k [7].

The rest of the article includes: Sect. 2 presents a group of related works on object recognition and object classification using deep learning networks, works using attention mechanisms to optimize object recognition and object relationships, and image caption extraction using knowledge graphs. Section 3 presents attention mechanisms to identify objects, relationships in images and presents the proposed model and related algorithms. Section 4 presents experimental results and evaluates and compares the results. Conclusions and development directions are presented in Sect. 5.

2 Related Works

While writing this paper, several related works were surveyed to analyze and evaluate the proposed methods for performing image captioning problems, such as applying knowledge graphs to search for image semantics and using attention mechanisms to extract image captions. Therefore, the group of related works surveyed was divided into three main groups: image captioning, Attention mechanism for extracting image caption, and knowledge graph for image captioning.

2.1 Image Captioning

This section presents some projects that create image captions by using different methods. Songtao Ding, et al. (2019) [8] proposed a model for establishing image captions based on high-level image features. The authors combined low-level information, such as image quality, with high-level features, such as motion classification and face recognition to detect attention regions of an image. This model has brought more efficient than some previous methods by experimenting on MS-COCO and Flickr30k image sets. This is a highly effective work for image caption extraction and is highly appreciated. A method using deep learning for the image captioning problem is mentioned as Bratislav Predic, et al. (2022) [9] proposed a model that combines machine learning algorithms to generate image captions. By changing the parameters on the InceptionV3 ResNet-50 or EfficiencyNet-B1 network, this method gives better results than training the model with MobileNet. Objects in the image are detected and recognized, then a text description with correct syntax is generated to describe each image. The image description results belong to pre-trained convolutional networks. Although this work has achieved quite good results, many issues still need to be improved because some image descriptions are incomplete. Besides, a work that has been evaluated as excellent in extracting image captions was done by the author Madhuri Bhalekar, et al. (2022) [10] proposed an image caption extraction model by combining CNN and LSTM networks to describe

the image's content. In this work, the results are better than the selected works for comparison on the MS-COCO, Flickr8k, and Flickr30k image sets. This paper proposes a system to generate detailed image captions and extract descriptive text for images by combining CNN and LSTM networks. The experiment has brought results similar to some previously proposed solutions that the authors have compared in this work.

2.2 Attention Mechanism for Extracting Image Caption

In recent years, the group of works that extract image captions using mechanisms to pay attention to some main objects on images has been quite rich, specifically: Jin Yuan, et al. [11] demonstrated that the attention mechanism on the image object has contributed significantly to improving the accuracy of image captioning. To do this, the author used the attention mechanism combined with visual concept templates to enhance the ability to predict visual concepts in image captioning. Furthermore, combining diverse visual concept templates from different domains, the proposed model has contributed to narrowing the gap between object domains for image caption extraction, helping to save costs. Experiments were conducted on two image datasets MS-COCO and Flickr30k with positive results compared to previous methods, specifically BLEU-1 and F1 values, which proves that the attention mechanism has contributed significantly to improving the accuracy of image captioning.

Tong Bai et al. (2023) [12] proposed a new method to solve the problem of visual information loss and adjust the input image during decoding to improve the ability to fully recognize information in the image. In addition, DenseNet and Multiple Instance Learning are applied in the encoder, combining nested long-short-term memory to enhance the ability to extract and analyze image information during encoding and decoding. Finally, an attention mechanism to focus on details and build a two-layer decoding structure is applied to increase the detailed description of image semantic information. To demonstrate this proposal, the author trained and experimented on the MS-COCO and Flickr30k datasets; the results showed that the model was improved compared to previous models that did not use the attention mechanism in terms of accuracy evaluation indicators such as BLEU, METEOR, and CIDEr. At the same time, a solution for image captioning systems using CNN networks is also highly mentioned, which is Rashid khan et al. (2023) [13] developed an image captioning system using pre-trained Convolutional Neural Networks to extract features from images, integrate the features with attention mechanisms, and generate captions using Recurrent Neural Networks (RNN). Simultaneously, the system uses multiple pre-trained convolutional neural networks. Then, a language model called GRU was selected as the decoder to construct descriptive sentences for images. Finally, a Bahdanau attention model with GRU to enable focused learning on a specific part of the image to increase efficiency. The MSCOCO dataset was used for experimenting with this system with the results achieving competitive performance compared to other state-of-the-art methods.

2.3 Knowledge Graph for Image Captioning

Knowledge graph is a tool built quite specifically for the image captioning problem. In the scope of this paper, some typical works using knowledge graphs for image captioning extraction as Wentian Zhao et al. (2021) [3] addressed the complexity of extracting detailed relationships between entities to generate informative descriptions for images. To solve this problem, the authors proposed a novel method to construct a multimodal knowledge graph to associate visual objects with named entities and identify relationships between entities simultaneously with the help of external knowledge collected from the web. Specifically, a sub-knowledge graph is constructed by extracting named entities and their relationships. On this basis, a multimodal knowledge graph is constructed using Wikipedia to combine corresponding concepts. Finally, the multimodal knowledge graph is integrated into the annotation model through the graph attention mechanism. Extensive experiments on both the GoodNews and NYTimes800k datasets have demonstrated the effectiveness of the proposed method. Mohammad Saif Wajid et al. (2023) [14] conducted a survey to evaluate the image caption extraction problem with related factors such as natural language processing in terms of grammar or semantic gap between low-level features and high-level semantics of images as factors affecting the image caption results. This work analyzes related methods such as deep learning methods, knowledge graph methods to combine the advantages of each method to improve the accuracy in image caption extraction. At the same time, the authors also study effective experimental datasets commonly used for image captioning problems such as Open-I, MIMIC-CXR, Flickr30k, and MS-COCO. Besides, Xin Wang et al. (2023) [15] evaluated that multimodal knowledge graphs play an important role in integrating text, images, videos, and audio to serve the image captioning problem. However, existing multimodal knowledge graphs do not cover all four elements, so there are still certain limitations. Therefore, in this work, the authors proposed TIVA-KG, a multimodal knowledge graph that integrates Text, Images, Videos, and Audio to apply to the image captioning problem.

From the works that have surveyed the combination methods to perform the image captioning problem, it has been done quite richly to obtain different results. Therefore, a solution is important: how to improve the accuracy of extracting image captions for images. After conducting a survey of related methods, this paper proposes a solution to combine knowledge graphs using attention mechanisms with ConceptNet to improve the efficiency of the process of building image captions for input images.

3 Applying Attention Mechanism to Extract Objects in Image

3.1 Attention Mechanism

Attention mechanism plays a very important role in machine translation, especially deep learning models, and natural language processing, and is quite effective for the image caption problem. For the image captioning problem, combining deep learning to extract objects on the image with the attention mechanism will focus on the objects of interest and the main relationships between these objects. This is the basis for solving the problem of low performance when extracting image captions for images because the

process of building captions for images depends on many factors such as context, objects, relationships, and language. In addition, the attention mechanism also greatly affects the problem of natural language processing, this mechanism helps machine learning algorithms focus on words and phrases that are considered to have high weight to select and build descriptive sentences with the closest accuracy to reality. Meanwhile, the image caption problem is a combination of content description, image semantic description, and natural language processing, so applying an attention mechanism to this problem contributes to reducing noise and distraction factors and improving the accuracy of image description [16].

The Deep learning models used for object recognition and extraction in images mostly recognize all objects appearing in the image based on trained objects. Meanwhile, the problem we are building focuses on a few main objects. This leads to each image using YOLO or RCNN deep learning network to recognize objects, several objects in each image are extracted a lot, which affects the subsequent results of the problems being implemented. Therefore, this paper uses the attention mechanism to select the objects of interest and the main relationships to contribute to the image caption construction process.

In addition, the attention mechanism is also applied in the stage of selecting the words that are of interest to extract captions for images. In this stage, each object and relationship after being encoded into feature vectors will be decoded to form a text sentence describing the content and semantics of the image [17]. To illustrate this process, Algorithm 1 presents the process of using the attention mechanism to select the objects and relationships of interest.

Algorithm 1: Extracting attention objects and keywords using Attention Mechanism

Input: Image Dataset ID and captions from the dataset CID;
Output: Attention Objects and Words AOW;
Function ATOW (ID, CID, AOW)
Begin
    K = 3;
Obs = Extracting all objects on input image by using YOLO model (ID);
$Feature(Obs)$ = Extracting features of objects from detected (Obs);
    $AObs =$ Attention Mechanism(Obs) *with* K Objects;
$AKWs =$ Attention mechanism(CID) with words in captions;
    $AOW = Visual\ Embedding(AObs, AKWs)$;
Return (AOW);
End.

In this paper, each image identifies a main object and three attention objects that need attention according to the priority of frequency of appearance. At the same time, between the main object and related objects as well as between each pair of related objects, between the main objects on the images, semantic relationships are determined to serve the process of building a knowledge graph. The process of building and enriching semantics into the knowledge graph is presented in Sect. 3.2.

3.2 A Knowledge Graphs for Image Semantic Description

In the works [18, 19] the process of building knowledge graphs for images still has some limitations such as the number of objects on each image after extraction and the corresponding relationships are used entirely for building the knowledge graph. Therefore, some objects that appear with a relatively low frequency are also included in the construction of knowledge graph and there are redundant relationships as well as the number of duplicate objects. This leads to a rather complicated graph when having to store quite a lot of unnecessary objects, the time to query and extract information from knowledge graph is quite large. To reduce this limitation, an attention mechanism is used in this work to limit the duplication of objects as well as the relationships between objects. At that time, the constructed knowledge graph significantly reduces the query time, improving the accuracy when searching and extracting image semantics from the knowledge graph.

Based on the inheritance and improvement of knowledge graph from works [18–20], in this paper, a knowledge graph is illustrated in Fig. 2. In which objects are added to the knowledge graph only once after extracting from the original image. Therefore, in this knowledge graph, the enrichment of relationships and objects requires checking whether their existence in the knowledge graph already exists. If the objects and relationships do not exist in the knowledge graph, they continue to enrich the knowledge graph. Therefore, this knowledge graph simplifies the duplicate relationships and objects, which will streamline the knowledge graph and increase the accuracy of image captions.

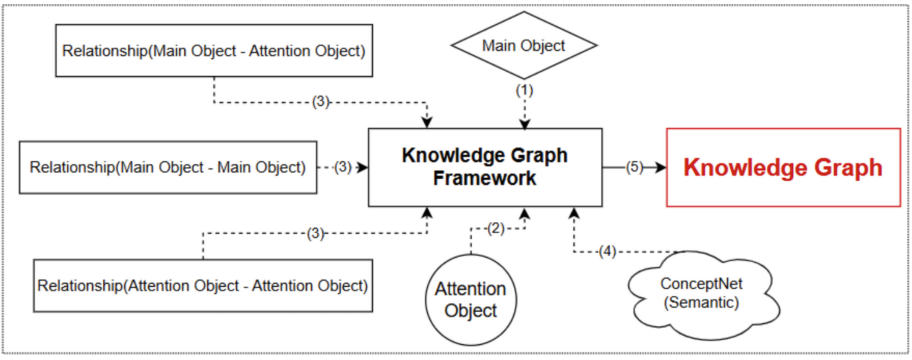

Fig. 2. Illustration of the process of building and adding semantics to the KG

(1) Some of the contents that need to be focused on in the process of building a knowledge graph are as follows:
(2) Improve the knowledge graph in which only one main object with the largest area in the image is interested in using the attention mechanism. Each central image only cares about 3 objects with high-frequency relationships.
(3) Reduce the complexity of the knowledge graph by reducing unnecessary relationships between many pairs of objects, which helps the search process faster because it only focuses on the really necessary objects.

(4) The number of relationships (edges) on the knowledge graph will be reduced, which will reduce the complexity of searching for the semantics of the image on the knowledge graph.
(5) The knowledge graph only focuses on the main semantics of the image that needs to be built, so extracting captions also contributes to increasing the accuracy of the semantic description of the input image.

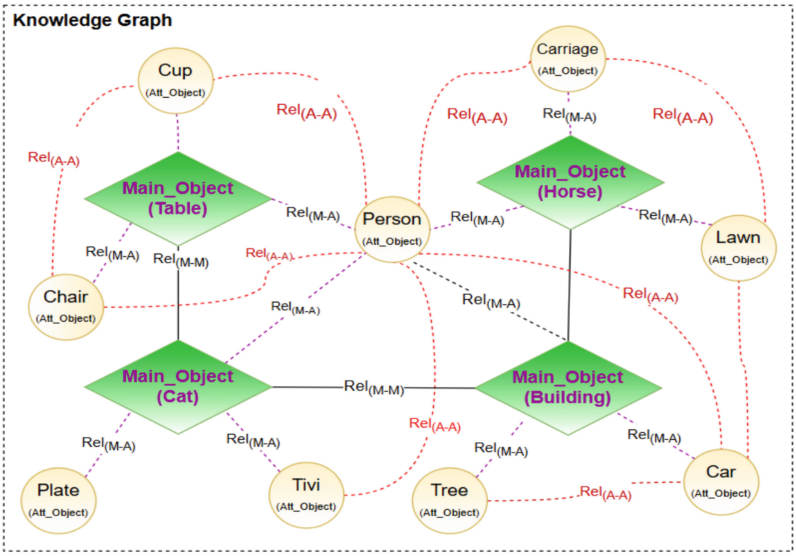

Fig. 3. The visual illustration structure of knowledge graph for image semantic

In the Fig. 3, the components of the Knowledge Graph are described as follows:

(1) *Main_Object* describes the main object of each image in the experimental image dataset.
(2) *Att_Object* (Cup, Chair,…) describes each noticed object that is related to the main object in each image, in this knowledge graph each main object determines 3 related objects according to algorithm 1 with $K = 3$.
(3) $Rel_{(M-M)}$is the relationship between two main objects. That is the relationship between two images.
(4) $Rel_{(M-A)}$ is the relationship between the main object and the related object in each image.
(5) $Rel_{(A-A)}$ is the relationship between two related objects, also known as two noticed objects in the same image or between any two images containing related objects.
(6) The relationships $Rel_{(M-M)}$, $Rel_{(M-A)}$, $Rel_{(A-A)}$ can be *IsA*, *PartOf* , *UsedFor* or extended relationships extracted from other knowledge systems.

Building the knowledge graph, Algorithm 2 performs the following steps:

Step 1: Identifying the main object of the image along with related objects ($K = 3$) using the attention mechanism.

Step 2: Describe the relationship between each pair of objects identified using ConceptNet.

Step 3: Adding more relationships between pairs of objects identified using ConceptNet and other knowledge models.

Step 4: Using the attention mechanism to extract keywords describing the relationship from the original image caption;

Step 5: Add Objects types ($Main\_Object$, Att_Object), relationship types ($Rel_{(M-M)}$, $Rel_{(M-A)}$, $Rel_{(A-A)}$) and selected keywords to knowledge graph.

Algorithm 2: Building Knowledge Graph

Input: Visual Embedding of Objects (Obs) and Key Words (KW);
Output: Knowledge Graph KG;
Function BKG(Obs, KW, KG)
Begin
  Initialize KG = Main object of an image on the Dataset;
  $ReObs$ = Describe the relationship between Objects using ConceptNet;
  $Rels$ = Extract the relationship($IsA, PartOf, UsedFor$) from ConceptNet;
  ARO = Attention mechanism to determine the relationship ($Rels$);
  AKW = Attention mechanism to determine Key word from the Dataset;
  $KG = KG \cup (ARO, AKW)$;
Return KG;
End.

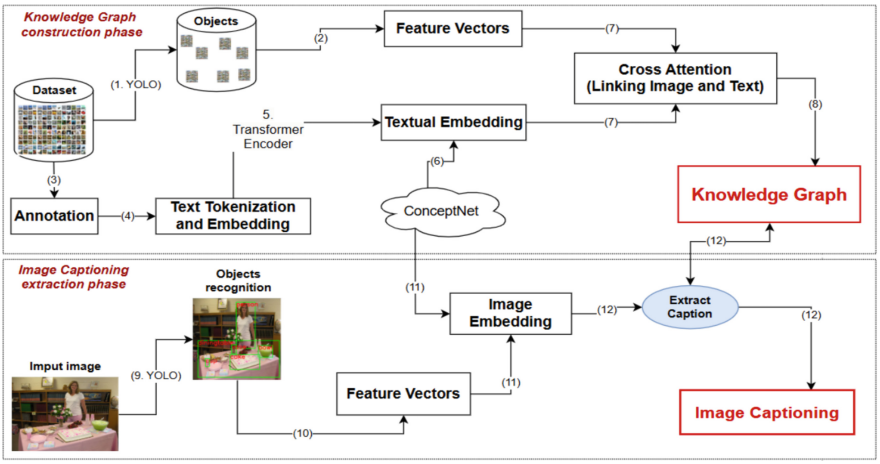

Fig. 4. A proposed model of extracting image captioning

The Fig. 4 describes the process of building a knowledge graph and extracting image captioning for the input image with the following steps:

(1) Identifying objects in images using the YOLO network;
(2) Extracting feature vectors for segmented image sets;
(3)–(4) Using annotations of experimental image sets and separating them into keyworks;
(5)–(6) Using the Transformer Encoder model combined with ConceptNet to create Textual Embedding.
(7)–(8) Combining Textual Embedding and segmented image features to build a knowledge graph;
(9) Segmenting objects using the YOLO model for each input image;
(10) Extracting feature vectors for segmented images belonging to input images;
(11) Combining ConceptNet and input image features to create Image Embedding;
(12) Combining knowledge graph and Image Embedding to build captions for input images.

Algorithm 3 performs the semantic description of the image from the knowledge graph based on the steps of the model in Fig. 4.

Algorithm 3: Semantic description for images using Knowledge Graph

Input: Knowledge Graph KG, input Image iI;
Output: Semantic of image SeI;
Function SoI(KG, TE, SeI)
Begin
    KR = Associates Key word and Relationship from ConceptNet and other Knowledge models;
    Rel = = Apply Attention to identify relationships between objects in the image;
    $Ann =$ Apply Attention to identify keywords in annotations from CID;
    Inf(iI) = Integrate information (Rel) and (Ann) from KG;
        $SeI =$ Extract semantics for images from KG;
    Return (SeI);
End.

The process of retrieving captions for images involves computer vision and natural language processing. Therefore, algorithm 4 identifies objects, relationships, and image semantics from the knowledge graph combined with knowledge from ConceptNet to build captions for an input image. In the algorithm 4, the loss function is optimized to reduce the computational cost and increase the efficiency of creating image captioning for the input image (iI) through the parameter IC = Optimizing the caption generation process using loss functions using Cross-Entropy Loss ∪ SoI(iI);

Algorithm 4: Combining Knowledge Graph and ConceptNet to extract image caption

Input: Input image iI, Textual Embedding and ConceptNet from Knowledge Graph TCK,
Output: Image Caption IC;
Function SoI(iI, TCK, IC)
Begin
    Semantic Enrichment (iI) = ConceptNet on Textual Embedding ∪ Rel;
    Multimodal Fusion = Using Cross-Attention to combine Visual Embeddings with Textual Embeddings;
    Using Transformer Decoder to generate Image Caption combined with embeddings;
    IC = Optimizing the caption generation process using loss functions using Cross-Entropy Loss $\cup SoI(iI)$;
    Return (IC);
End.

4 Experimental Image Caption Extraction

4.1 Environment and Experimental Data

The experiment is built on the Python language platform. The graphs are built based on the Python Matplotlib library. The process of the experiment is built on the Lenovo IdeaPad Gaming 3 (2020) computer environment, which includes an Intel(R) Core(TM) i7-10750H processor, 16GB RAM, GTX 1650 GPU, and Windows 10 Home Single Language operating system. The experimental data sets are MS-COCO and Flickr30k multi-object image sets, which are described in Table 1.

Table 1. Detailed description for the MS-COCO and Flickr30k image sets

Image Datasets	Images in Training set	Images in Testing set	Images in Validation set	Concept in the Visual Concept set
MS-COCO [6]	118,287	40,670	5,000	80
Flickr30k [7]	29,000	1,783	1,000	79

4.2 Evaluation of Experimental Results

The experimental results of building captions for images on the MS-COCO and Flickr30k datasets are shown in Figs. 5 and 6. Figure 5 shows the results of building captions for images 000000162252.jpg and 000000223569.jpg (MS-COCO). Figure 6 shows the results of building captions for images 1028205764.jpg and 1053116826.jpg (Flickr30k). The average accuracy results of these two image sets are presented in Table 2 with BLEU-1 and BLEU-4 metrics.

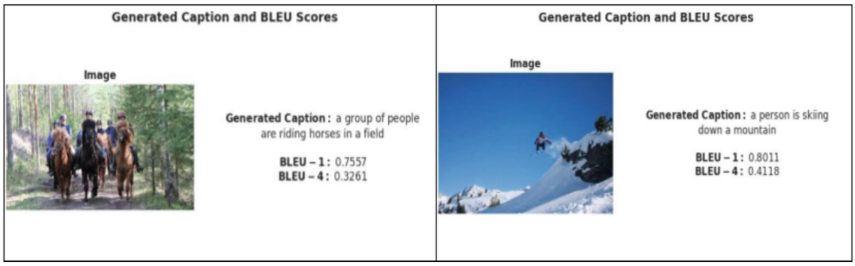

Fig. 5. Image captioning results 000000162252.jpg, 000000223569.jpg (MS-COCO)

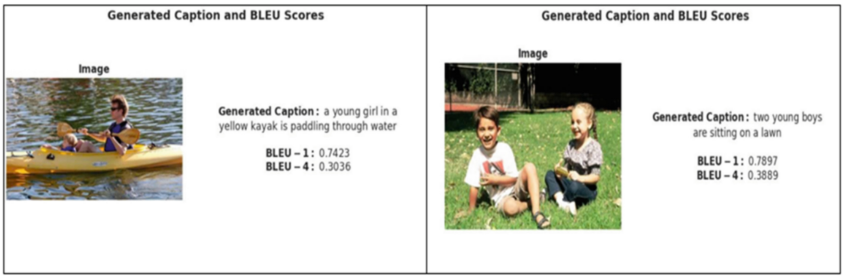

Fig. 6. Image captioning results for 1028205764.jpg, 1053116826.jpg (Flickr30k)

Figure 7 is the attention heatmap graph on images and vocabulary generated experimentally on MS-COCO and Flickr30k image sets.

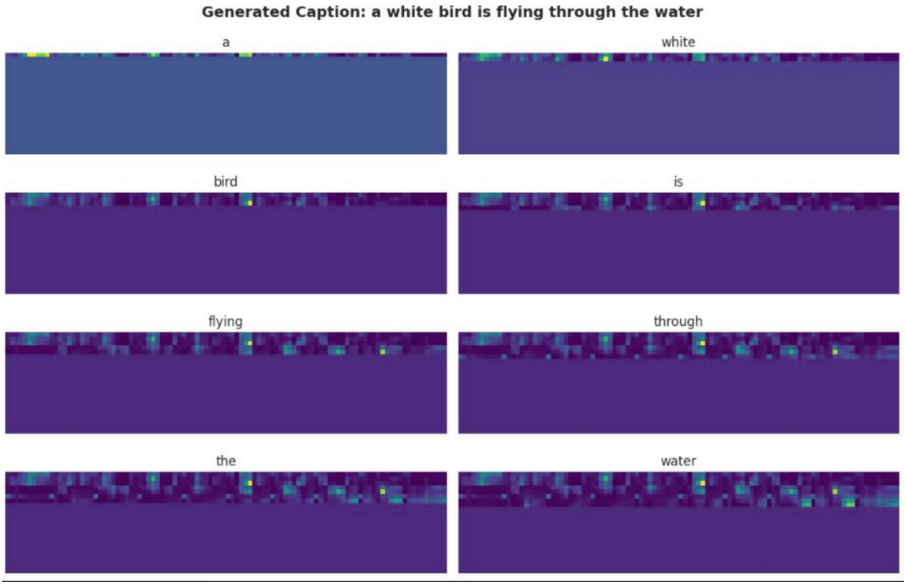

Fig. 7. The attention heatmap graph on images and vocabulary

Figure 8 is the graph showing the length of sentences generated from the model, compared with ground-truth.

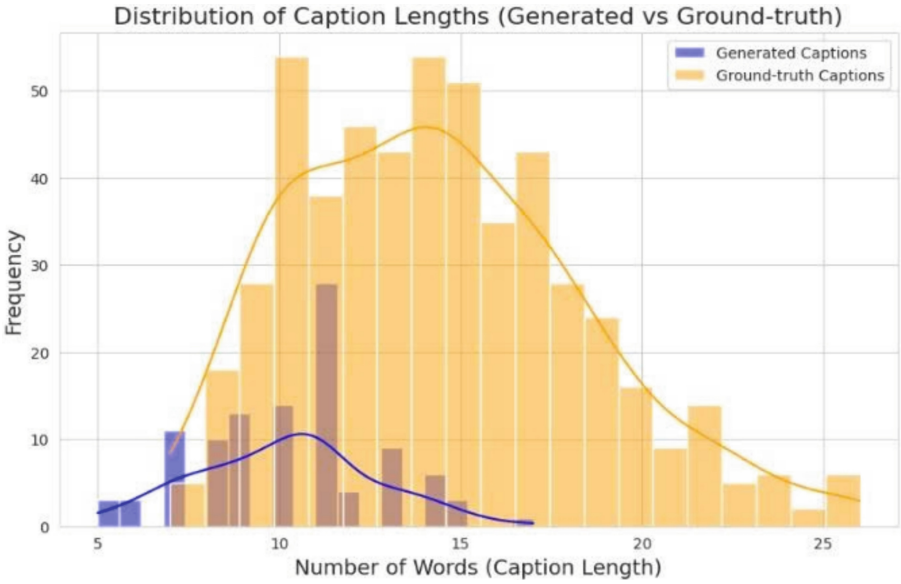

Fig. 8. The graph showing the length of sentences generated from the model

The experimental results for the Loss by Epochs graph are shown in Fig. 9.

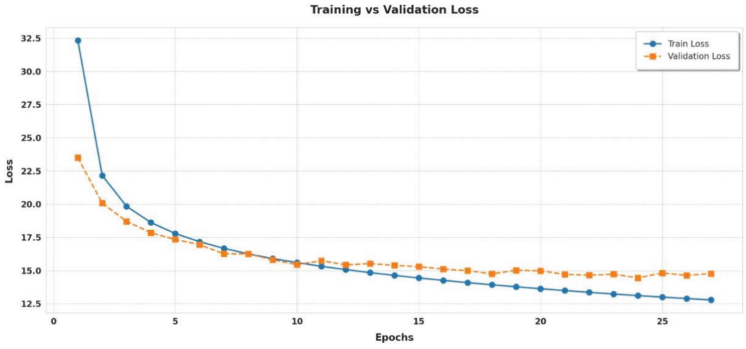

Fig. 9. The experimental results for the Loss by epochs

Table 2. Experimental results image caption by using BLEU-1 and BLEU-4

Datasets	Number of image validation	BLEU-1	BLEU-4
MS-COCO [6]	5,000	0.7679	0.3932
Flickr30k [7]	1,000	0.7586	0.3578

In the experiment with extracting captions for images on the MS-COCO and Flickr30k image sets, Tables 3 and 4 compare the accuracy with BLEU-1 and BLEU-4 measures to demonstrate the superiority of our proposed model when combining the attention mechanism into building the knowledge graph, integrating semantics for the process of extracting captions for input images.

Table 3. Comparing image caption accuracy with some other methods on MS-COCO

Method	BLEU-1	BLEU-4
Trans [D2GPO + MLE], 2021 [21]	0.7639	0.3438
ResNet101 Features, 2020 [11]	0.7610	0.3350
DRL and attention mechanism, 2023 [12]	0.7520	0.3440
Context-aware attention, 2020 [22]	0.7600	0.3600
Attention Mechanism and Knowledge Graph	**0.7789**	**0.3932**

Table 4. Comparing image caption accuracy with some other methods on Flickr30k

Method	BLEU-1	BLEU-4
Trans[D2GPO + MLE] + KG, 2021 [21]	0.6836	0.2655
ResNet101 Features, 2020 [11]	0.6860	0.2640
DRL and Attention Mechanism, 2023 [12]	0.7380	0.3350
Context-aware attention, 2020 [22]	0.6980	0.2770
Attention Mechanism and Knowledge Graph	**0.7586**	**0.3578**

The experimental results of image caption accuracy are presented in Table 2, and the comparison results with some other methods presented in Tables 3 and 4 show that our proposed method outperforms these methods. This result is obtained due to the combination of factors: (1) the attention mechanism only focuses on objects and relationships of interest with a high frequency of occurrence in contexts; (2) the integration of the semantics of Knowledge Graph and ConceptNet contributes to increasing the accuracy of semantic descriptions for images; (3) the combination of the attention mechanism in selecting keywords in image annotation also contributes to improving the accuracy of input image captions.

The main contribution of this paper is to use the attention mechanism to focus on key objects and relationships in images, thereby avoiding duplication of objects and relationships when adding to the knowledge graph. The semantic enrichment process for Knowledge Graph has contributed to improving the accuracy of image caption extraction.

5 Conclusion and Development

In this paper, an image caption extraction system based on the semantics of knowledge graphs combined with ConceptNet is developed. The knowledge graph is constructed from the objects in the image after using the attention mechanism to enhance the relationship and interest in related objects from the main objects identified in the image. The accuracy scores of the new system for the MS-COCO and Flickr30k image sets are 0.7679 and 0.7586 (BLEU-1), 0.3932 and 0.3575 (BLEU-4) respectively. The process of constructing the knowledge graph using the attention mechanism to retrieve image semantics is a major contribution to this work. At the same time, combining the semantics from ConceptNet has contributed to improving the accuracy of image content description and semantics. However, there is still a problem with the improvement of knowledge graph development by combining transfer learning models to improve the accuracy of image caption retrieval. Besides, some issues surrounding knowledge graphs need to be implemented and our research team is also aiming to implement such as developing the visual question answering problem.

Acknowledgments. The authors would like to thank the Faculty of Information Technology, University of Sciences, Hue University for their professional advice for this study. We would also like to thank HCMC University of Industry and Trade, University of Economics, The University of Da Nang, Ton Duc Thang University which are sponsors of this research. We also thank anonymous reviewers for their helpful comments on this paper.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Xu, L., Tang, Q., Lv, J., Zheng, B., Zeng, X., Li, W.: Deep image captioning: a review of methods, trends and future challenges. Neurocomputing **546**, 126287 (2023)
2. Chun, P.J., Yamane, T., Maemura, Y.: A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage. Comput.-Aided Civil Infrastruct. Eng. **37**(11), 1387–1401 (2022)
3. Zhao, W., Wu, X.: Boosting entity-aware image captioning with multi-modal knowledge graph. IEEE Trans. Multimedia **26**, 2659–2670 (2023)
4. Elbedwehy, S., Medhat, T., Hamza, T., Alrahmawy, M. F., Efficient image captioning based on vision transformer models. Comput. Mater. Continua **73**(1) (2022)
5. Liu, M., Li, L., Hu, H., Guan, W., Tian, J.: Image caption generation with dual attention mechanism. Inf. Process. Manage. **57**(2), 102178 (2020)
6. MS-COCO Image Dataset. https://cocodataset.org/#download. Accessed 10 Oct 2024

7. Flickr Image Dataset. https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset. Accessed 10 Oct 2024
8. Ding, S., et al.: Image caption generation with high-level image features. Pattern Recogn. Lett. **123**, 89–95 (2019)
9. Predić, B., et al., Automatic image caption generation based on some machine learning algorithms. Math. Prob. Eng. (2022)
10. Bhalekar, M., Bedekar, M.: D-CNN: A new model for generating image captions with text extraction using deep learning for visually challenged individuals. Eng. Technol. Appl. Sci. Res. **12**(2), 8366–8373 (2022)
11. Yuan, J., Zhang, L., Guo, S., Xiao, Y., Li, Z., Image captioning with a joint attention mechanism by visual concept samples. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **16**(3), 1–22 (2020)
12. Bai, T., Zhou, S., Pang, Y., Luo, J., Wang, H., Du, Y.: An image caption model based on attention mechanism and deep reinforcement learning. Front. Neurosci. **17**, 1270850 (2023)
13. Khan, R., Islam, M.S., Kanwal, K., Iqbal, M., Hossain, M.I., Ye, Z.: A deep neural framework for image caption generation using gru-based attention mechanism. arXiv preprint arXiv: 2203.01594 (2022)
14. Wajid, M.S., Terashima-Marin, H., Najafirad, P., Wajid, M.A.: Deep learning and knowledge graph for image/video captioning: a review of datasets, evaluation metrics, and methods. Eng. Rep. **6**(1), e12785 (2024)
15. Wang, X., Meng, B., Chen, H., Meng, Y., Lv, K., Zhu, W.: TIVA-KG: a multimodal knowledge graph with text, image, video and audio. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 2391–2399 (2023)
16. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4634–4643 (2019)
17. Li, W., Liu, K., Zhang, L., Cheng, F.: Object detection based on an adaptive attention mechanism. Sci. Rep. **10**(1), 11307 (2020)
18. Dinh, N.T., Linh, T.T., Van, T.T., Le, T.M.: A technique of knowledge graph construction applied to the image retrieval. In: Proceedings of the National Conference on Basic Research and Application of Information Technology (FAIR 2024) (2024). https://doi.org/10.15625/vap.2024.0244
19. Khoi, T.T., Phuoc, T.T., Van, T.T.: Images retrieval based on deep learning and knowledge graph. In: Proceedings of the National Conference on Basic Research and Application of Information Technology (FAIR2024) (2024). https://doi.org/10.15625/vap.2024.0278
20. Dinh, N.T., Le, T.M., Van, T.T.: Using knowledge graph and KD-tree random forest for image retrieval. In: Rocha, Á., Adeli, H., Dzemyda, G., Moreira, F., Poniszewska-Marańda, A. (eds.) Good Practices and New Perspectives in Information Systems and Technologies: WorldCIST 2024, Volume 5, pp. 13–25. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-60227-6_2
21. Zhang, Y., Shi, X., Mi, S., Yang, X.: Image captioning with transformer and knowledge graph. Pattern Recogn. Lett. **143**, 43–49 (2021)
22. Wang, J., Wang, W., Wang, L., Wang, Z., Feng, D.D., Tan, T.: Learning visual relationship and context-aware attention for image captioning. Pattern Recogn. **98**, 107075 (2020)