Vietnam Journal of Science and Technology ***64****(1) (2026) 123–138*

DOI: https://doi.org/10.15625/2525-2518/22381

RGTranCNet: Effective image captioning model using cross-attention and semantic knowledge

Nguyen Van Thinh[1,2]**, Tran Van Lang**[3,*]**, Van The Thanh**[2]

[1]*Graduate University of Science and Technology, Vietnam Academy of Science and Technology (VAST), 18 Hoang Quoc Viet Street, Nghia Do Ward, Ha Noi, Viet Nam*
[2]*Faculty of Information Technology, Ho Chi Minh City University of Education (HCMUE), 280 An Duong Vuong Street, Cho Quan Ward, Ho Chi Minh City, Viet Nam*
[3]*Journal Editorial Department, Ho Chi Minh City University of Foreign Languages and Information Technology (HUFLIT), 828 Su Van Hanh Street, Hoa Hung Ward, Ho Chi Minh City, Viet Nam*

[*]Email: *langtv@huflit.edu.vn*

Received: 06 February 2025; Accepted for publication: 14 July 2025

Abstract. Generating captions for images is a key endeavour that connects visual processing and linguistic analysis. However, techniques relying on long short-term memory (LSTM) units and conventional attention systems face restrictions in managing intricate interconnections and supporting effective parallel processing. Additionally, precisely depicting elements absent from the training data presents a significant challenge. To overcome these obstacles, the present research introduces an innovative framework for image description, employing a Transformer architecture augmented by cross-attention processes and semantic insights sourced from ConceptNet. This setup follows an encoder-decoder paradigm, where the encoder derives features from object areas and assembles a graph of associations to depict the visual scene. At the same time, the decoder merges visual and semantic aspects through cross-attention to produce captions that are both accurate and varied. The inclusion of ConceptNet-derived knowledge enhances precision, particularly when handling items not encountered during training. Tests conducted on the standard MS COCO dataset reveal that this approach outperforms recent state-of-the-art approaches. Moreover, the semantic integration strategy outlined here can be readily adapted to alternative image captioning systems.

Keywords: Image captioning, cross-attention mechanism, transformer, ConceptNet knowledge base, relationship graph.

Classification numbers: 4.7.4, 4.8.3.

1. INTRODUCTION

Creating captions for images ranks among the crucial and demanding activities in artificial intelligence. This involves a multi-modal learning procedure that integrates visual computing with linguistic analysis. The goal is to produce significant text-based narratives derived from given images [1]. The process of automatically crafting precise and semantically abundant descriptions from visuals necessitates a deep comprehension of the image's components, coupled with the model's proficiency in identifying semantic connections among entities, surroundings, and activities shown in the visual [2]. Furthermore, image captioning has numerous practical applications, such as image captioning systems for assisting visually impaired individuals in perceiving their surroundings [3], medical image captioning to aid doctors in diagnosing diseases [4], and human-robot interaction [5], image captioning for explainable visual question answering [6], etc.

At present, most image captioning approaches adopt an encoder–decoder architecture with attention mechanisms [1]. In this framework, visual features are extracted by an encoder—typically a pre-trained convolutional neural network (CNN) or an object detection model such as Faster R-CNN or YOLO—and then decoded into natural language descriptions by a long short-term memory (LSTM) network [7, 11]. However, CNN-based encoders often suffer from information bottlenecks due to compressing the entire image into fixed-length representations, while region-based detectors fail to capture object interactions. Consequently, modelling relationships between image elements becomes essential to provide a more holistic representation and improve caption accuracy.

To overcome these limitations, several studies have incorporated object relationships into image captioning by representing images as graphs [12-15]. These graph-based methods enrich the encoder with relational information, which is then utilised by an LSTM decoder with attention mechanisms to generate captions. Despite their effectiveness, LSTM-based decoders remain limited by sequential computation, slow training, and optimisation issues such as vanishing or exploding gradients. Motivated by the success of Transformers in natural language processing [16], recent image captioning models increasingly replace LSTM with Transformer-based decoders, benefiting from parallelisation and more expressive attention mechanisms. In particular, self-attention and cross-attention enable Transformers to model contextual relationships between objects more effectively than traditional attention designs.

Another fundamental challenge in image captioning arises from the limited diversity of training datasets, which typically provide only one to five captions per image. This constraint limits a model's ability to describe novel objects or implicit semantic attributes that are not explicitly represented in the image. To address this issue, prior work has explored integrating external knowledge sources into the captioning process, including object-level knowledge from external datasets [17], knowledge graphs [18], and semantic resources such as ConceptNet combined with attention mechanisms [19]. These studies demonstrate that incorporating external knowledge beyond the training data is both feasible and effective in enhancing caption generalisation and semantic richness.

Based on these observations, this paper proposes **RGTranCNet**. This novel image captioning framework integrates three key components: a relationship graph to model object interactions, a Transformer-based decoder with a unified cross-attention mechanism, and semantic knowledge extracted from the ConceptNet knowledge base. Unlike prior dual-attention approaches, RGTranCNet fuses object-region features and relational graph embeddings within a single cross-attention block, reducing architectural complexity while improving information integration. Moreover, only the decoder is trained, while the object detection, relationship graph construction, and semantic knowledge extraction modules are reused and kept fixed. This design significantly reduces training cost while maintaining accuracy and scalability.

The main contributions of this paper include:

- Improving image captioning performance using a transformer decoder as the language model in place of LSTM networks and employing cross-attention mechanisms instead of traditional attention mechanisms to integrate multimodal information between the encoder and decoder.

- Integrating semantic knowledge from the ConceptNet knowledge base into the decoder to leverage external knowledge beyond the training dataset, thereby enhancing the accuracy of the generated captions, particularly for novel objects. This approach can be easily applied to other image captioning models.

- Extensive experiments on the benchmark MS COCO dataset demonstrate that the proposed model achieves higher accuracy than previous methods (including LSTM-based ones) across most evaluation metrics while maintaining low training costs by training only the decoder.

The remainder of this paper is organised as follows. Section 2 reviews related work and discusses remaining challenges in image captioning. Section 3 presents the proposed RGTranCNet framework in detail. Section 4 describes the experimental setup and reports the evaluation results. Finally, Section 5 concludes the paper and outlines directions for future research.

2. RELATED WORKS

Many recent image captioning works have been published based on the encoder-decoder framework with attention mechanisms, employing pre-trained CNN networks, object detection models, and models that predict relationships between objects in the image, as well as utilizing external data sources beyond the training dataset, such as:

Patwari *et al.* [20] introduced a method for describing images that relies on an encoder-decoder structure, employing a pre-trained Inception-v3 convolutional neural network to derive visual features. A GRU-equipped decoder, augmented by an attention system, then produces the descriptions. This approach yielded encouraging results on the MS COCO dataset, as reflected in its BLEU-1 through BLEU-4 metrics. However, it is still hampered by its heavy reliance on pre-trained CNNs for extracting visual elements, which leads to challenges in identifying fine-grained object specifics and the connections between them, ultimately hindering a more profound grasp of the image's semantic essence.

Xie *et al.* [21] presented a framework designed to enhance the effectiveness of image description generation by combining bidirectional LSTM architectures and attention systems. Their methodology involves deriving features from object areas in the input visuals via Faster R-CNN, followed by their handling in a Bi-LSTM setup to produce explanatory text. This system underwent experimental testing on the Flickr30k and MS COCO benchmarks, where it exhibited better outcomes than standard references and various contemporary works. That said, a significant shortcoming lies in its narrow focus on isolating object regions, overlooking the interconnections between them - a factor that could enhance the image's semantic depiction and improve the precision of the generated descriptions.

Chen *et al*. [22] proposed a technique that generates an abstract scene graph from authentic captions to guide the production of image captions, offering increased variety and better alignment with user preferences. Drawing from this foundation, Yan *et al.* [23] advanced the system by incorporating transformer elements in conjunction with a two-level LSTM design to enhance smoothness and consistency. Within this setup, the primary LSTM layer integrates inputs across modalities, merging visual and linguistic signals, while the secondary layer constructs the captions. The transformer component oversees the weighting of diverse feature types during the decoding stage. Although these works adeptly utilise abstract scene depictions obtained from labelled captions to elevate the quality of descriptions, they both suffer from a shared limitation: the inadequate exploitation of the image's intrinsic elements, especially the semantic interconnections among objects. This deficiency leads to their somewhat inferior outcomes on multiple conventional evaluation measures.

Ramos *et al.* [24] integrated the ConvNeXt architecture with a Long Short-Term Memory (LSTM) system, augmented by a visual attention component, to boost the effectiveness of generating image descriptions. This framework underwent testing on the MS COCO dataset, where its performance was measured via the BLEU score, revealing greater precision relative to approaches that employ pre-trained CNNs as encoders. While ConvNeXt demonstrated advantages over established pre-trained CNN frameworks, it nonetheless encounters difficulties in thoroughly grasping the interconnections among elements within the images. The LSTM element also presents shortcomings, as noted in prior discussions. Moreover, assessing outcomes based exclusively on the BLEU metric falls short of providing a thorough analysis of the model's strengths, as supplementary evaluation criteria are necessary to encompass the diverse dimensions of the image description generation task.

Thinh *et al.* [12] introduced an image captioning framework that incorporates both object detection and relationship prediction to capture the semantic structure of an image. Initially, objects are identified using a detection model enhanced with a graph convolutional network, followed by the inference of inter-object relationships informed by contextual cues and predefined relational knowledge. These relationships are then categorised to form a structured graph that semantically represents the image. To support the caption generation process, a dual-attention mechanism is employed, allowing the model to selectively attend to both visual object regions and corresponding nodes in the relationship graph. Caption generation is carried out by an LSTM network equipped with this dual-attention design, utilising both the extracted image features and reference captions. Experimental evaluation on the MS COCO dataset confirms the model’s effectiveness. Nonetheless, the approach encounters limitations: the two independent attention modules within the dual-attention design do not sufficiently integrate visual and semantic features, and the reliance on LSTM networks introduces sequential processing constraints. These limitations suggest the need to replace the LSTM component with a transformer architecture and adopt a cross-attention mechanism to align heterogeneous features better and improve captioning performance.

Wang *et al.* [25] developed a system for generating image descriptions using a transformer architecture, aiming to address the shortcomings inherent in CNN-LSTM configurations. Yang *et al.* [26] presented a transformer-oriented framework dedicated to context detection, aimed at boosting the precision of generated captions. Li *et al.* [27] suggested integrating a transformer with supplementary external data to capitalise on inter-object connections, thereby elevating the quality of image captioning results. These techniques were subjected to empirical testing on the MS COCO dataset and demonstrated their efficacy across typical performance indicators for image description tasks, including BLEU, METEOR, ROUGE, and CIDEr.

Zhou *et al.* [18] suggested an approach to boost the precision of generating image descriptions by utilising data from the ConceptNet repository. They incorporated semantic insights associated with image elements into the encoder section of the NIC framework for captioning [7], achieving improved outcomes over methods that rely solely on visual characteristics. That said, a notable drawback of this technique is that incorporating an overload of input data might create interference in the training phase, ultimately diminishing the model's overall efficiency. Hafeth *et al.* [19] presented a semantic attention-oriented network designed to embed supplementary knowledge (derived from ConceptNet) into the transformer's attention components, which in turn enhances the performance of image description generation.

From the survey and analysis of related works, it is evident that the image captioning problem, mainly using deep learning networks such as transformers, has garnered significant attention from the research community and has proven effective. Moreover, integrating semantic knowledge from external data sources beyond the training dataset is also feasible and practical. Building on the foundation of existing research and addressing the limitations of previously published methods, the proposed image captioning approach utilizes a relationship graph, a transformer decoder with cross-attention mechanisms, and the integration of semantic knowledge from ConceptNet into the decoder, aiming to improve accuracy and enhance the model's generalization capability.

3. PROPOSED METHOD

In this study, we introduce an image captioning model built upon the encoder-decoder architecture, as illustrated in Figure 1. The model comprises three core modules: (i) an image encoder responsible for learning visual representations from the input image; (ii) a semantic knowledge extractor that retrieves relevant object-level semantics from the ConceptNet knowledge base; and (iii) a Transformer-based decoder that generates image captions by utilizing the visual features from (i) in conjunction with the semantic information from (ii), thereby improving caption quality and semantic alignment.

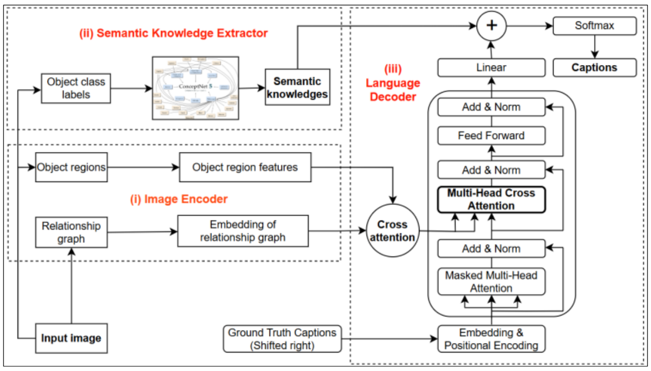


Figure 1. Architecture of the image captioning model integrating semantic knowledge and cross-attention mechanism.

3.1. Image Encoder

The image encoder comprises two main processes: (3.1.1) identifying object regions within the image and extracting their visual features, and (3.1.2) generating and embedding a relationship graph that captures interactions among the detected objects. These two types of features are integrated through a cross-attention mechanism, which provides the contextual input for the decoder to generate descriptive captions.

3.1.1. Detecting and extracting features from object regions

Pre-trained object detection models such as SSD, Faster R-CNN, and YOLO have achieved strong performance in image captioning applications. However, they mainly focus on individual object attributes and often fail to capture contextual dependencies among objects, particularly in complex scenes. To alleviate this limitation, ODwGCN was proposed in our previous work [12], which enhances object detection via a two-stage process. First, a Graph Convolutional Network (GCN) is employed to model object co-occurrence patterns. These patterns are then used to refine the outputs of pre-trained detectors. Experimental results on the MS COCO dataset demonstrated notable improvements in detection accuracy.

In the current study, ODwGCN is utilised to identify object regions. The visual features of these regions are extracted using ResNet101, yielding a set of feature vectors denoted as F_I for a given input image I. These features are later integrated with the embedding of the relationship graph and provided as input to the decoder for caption generation.

3.1.2. Constructing and representing the relationship graph of the image

To explicitly model object interactions, a relationship graph is constructed for each input image. Following [12], the relationship graph, referred to as the **R-Graph**, is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of detected object regions and $\mathcal{E}$ denotes directed edges encoding relationships between object pairs. Each edge $e_{ij} = (v_i, v_j, r_{ij})$ corresponds to a semantic relation r_{ij} drawn from a predefined relationship set $\mathcal{R}$.

3.1.2.1. Creating the relationship graph

The relationship graph is built using the VRP+RK model proposed in [12], which formulates relationship prediction as a multi-class classification problem. Given a pair of object regions and their semantic labels, the model predicts one of $N_{\mathcal{R}} + 1$ classes, including $N_{\mathcal{R}}$ predefined relations and a "$none\ relation$" class. After training on the Visual Genome dataset, VRP+RK is jointly applied with ODwGCN to construct a relationship graph for an input image.

Specifically, all detected object regions are paired, and the classifier outputs a probability distribution over relationship classes for each pair. If the probability of the "$none\ relation$" class is below a threshold γ, a directed edge is established between the corresponding nodes and labelled with the most probable relationship. An example of the resulting relationship graph is illustrated in Figure 2.

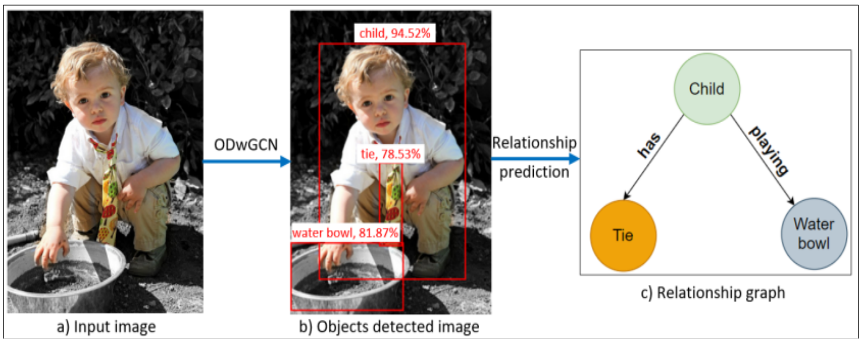


Figure 2. Creating a relationship graph from an input image: (a) the input image, (b) the result after applying the improved object detection model ODwGCN, and (c) the relationship graph obtained after predicting the relationships between the objects.

3.1.2.2. Representation the relationship graph

Although the relationship graph provides a structured description of image content, its heterogeneous nature makes direct integration into neural language models challenging [26]. To address this issue, we adopt the enriched relationship graph representation, **R-Graph***, proposed in [12]. The transformed graph is defined as $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$, where the vertex set $\mathcal{V}^*$ includes both object class labels and relationship (predicate) labels. For each edge (v_i, v_j, r_{ij}) in the original graph, two directed edges are created in $\mathcal{G}^*$: one from v_i to r_{ij}, and another from r_{ij} to v_j. This transformation preserves semantic structure while making the graph compatible with language-oriented neural architectures. Figure 3 illustrates an example of this conversion.

To obtain vector representations for vertices in R-Graph*, we employ the GraphSAGE framework [28] with unsupervised training. Initial node features are derived from pre-trained word embeddings (GloVe). For each node, information from incoming and outgoing neighbours is aggregated and combined with the node's own representation through multiple GCN layers. The final embeddings capture both semantic labels and relational context. For input image I, the resulting set of graph embeddings is denoted Z_I and is subsequently used as input to the decoder.

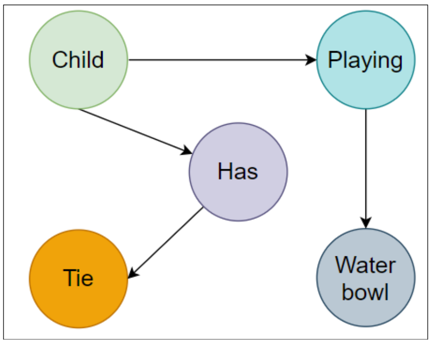


Figure 3. Result of converting the relationship graph R-Graph in Figure 2(c) into the extended relationship graph R-Graph*.

3.2. Semantic knowledge extractor

ConceptNet is a multilingual knowledge base describing words and phrases commonly used by humans and their typical relationships. The knowledge in ConceptNet is collected from various sources, including community-contributed resources such as Wiktionary and Open Mind Common Sense, expert-curated resources like WordNet and JMDict, and many other open data sources [29]. This creates a knowledge base that links concepts through semantic relationships such as “IsA”, “PartOf”, “UsedFor”, and many others. These relationships enable the model to understand better the context and semantic links between words and phrases. As a result, it aids artificial intelligence systems in understanding context, reasoning, and interacting with humans. In this study, the ConceptNet knowledge base $\mathcal{K}$ is defined as a graph as follows:

Definition 1. Graph **CK-Graph** $\mathcal{K} = (V, E, W)$ is a directed graph consisting of:

- Vertex set $V = \{v_i \in C, \forall i = \overline{1, N_C}\}$, N_C is the number of concepts in ConceptNet,

- Edge set $E = \{e_{ij} = (v_i, v_j, w), \forall i, j = \overline{1, N_C}, i \neq j\}$,

- Weight set $W = \{w_i \in \mathbb{R}^+, \forall i = \overline{1, N_E}\}$, N_E is the number of edges in E.

To integrate semantic knowledge from ConceptNet into the caption generation process to improve accuracy, particularly in describing novel objects not present in the training dataset, we first represent ConceptNet as a knowledge graph $\mathcal{K} = (V, E, W)$, as defined in Definition 1. Then, object class labels in the image are used to query semantically similar knowledge from this graph. Figure 4 illustrates the result of querying information for the object "laptop" from $\mathcal{K}$. Notably, each object corresponds to a probability value, representing the degree of correlation with the queried term.

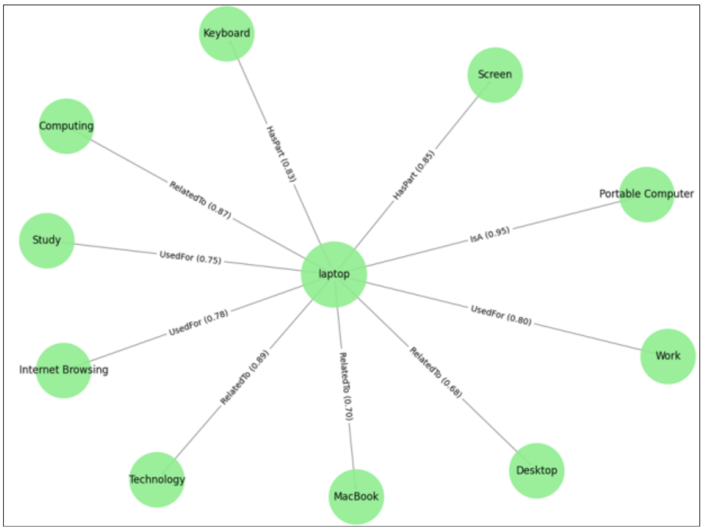


Figure 4. Illustration of knowledge extraction from ConceptNet for the object class label "laptop".

In this paper, the top-k related objects for each detected object in the image are used to enhance the information for the decoder during the caption generation process. This set, denoted as O, is utilized when generating the next word of the decoder. The process of extracting knowledge for the related objects is described in Algorithm 1.

Algorithm 1. ExtractRelatedObjectCNet$(L_I, \mathcal{K})$

```
Input: The object label set of the image $I$, $L_I = \{l_1, l_2, .., l_{N_I}\}$, ConceptNet knowledge base $\mathcal{K}$
Output: The list of related objects and their corresponding weights $O$
Begin
    # Initialize the empty set O
    $O \leftarrow \emptyset$
    foreach$l_i \in L_I$do
        # Retrieve the edge set $E_i \in E$ from ConceptNet $\mathcal{K}$
        $E_i = \{(v_s, v_t, w) | v_s = l_i\}$
        # The set of related objects of $l_i$
        $O_i = \{(v_t, w) | (l_i, v_t, w) \in E_i\}$
        # Update $O$
        $O \leftarrow O \cup O_i$
    end
End
```

Algorithm 1 extracts semantically related objects from the ConceptNet knowledge base based on the class labels of objects detected in an input image. For each object label in L_I, the algorithm retrieves the corresponding outgoing edges from ConceptNet and identifies the associated related objects. These objects, together with their relevance weights, are aggregated into a unified set. Given an input image I with object label set L_I, Algorithm 1 outputs the semantic knowledge set O_I, which consists of related objects and their corresponding weight values.

3.3. Language Decoder

In this paper, the decoder component of the transformer is utilized as a language model for image caption generation. The features extracted from the image in the encoder, including object region features and embeddings of the relational graph, are combined and input into the decoder through a cross-attention mechanism to train the caption generation model. Semantic knowledge extracted from ConceptNet is also integrated to enhance the model's performance. The training and caption generation process of the encoder is described through two algorithms: Algorithm 2, which trains the transformer decoder by integrating semantic knowledge to create a model for image caption generation, and Algorithm 3, which generates captions for input images using the model trained in Algorithm 2.

To focus on two main improvements: (1) the multi-head cross-attention mechanism, which integrates information from object region features and relational graph embeddings, and (2) the integration of semantic knowledge from ConceptNet to adjust the predicted probabilities during generation, thereby aiding the model in producing more accurate, diverse, and meaningful captions, especially for objects not present in the training dataset, we omit the details of other layers in the transformer decoder, such as masked multi-head attention, feed forward layer, and add & norm, as these components are retained according to the original design.

In Algorithm 2, N_T is the number of data samples in the training dataset, F_i, Z_i, S_i and O_i represent the object region features, embeddings of the nodes in the relational graph, ground truth captions, and the related knowledge object set for the i^{th} data sample (image), respectively. The ground truth captions for each image are denoted as $S = \{s_1, s_2, \ldots, s_{N_S}\}$, where s_i represents the i^{th} word in the sentence, $\forall i = \overline{1, N_S}$, with N_S being the number of words in sentence S. H is the hidden state, and W_Q, W_K, W_V and W_O are the weight matrices of the transformer decoder. These weight matrices are randomly initialised and will be learned and updated during training.

Algorithm 2 trains the Transformer decoder to generate image captions by integrating object-region features and relational graph embeddings through a multi-head cross-attention mechanism, while simultaneously incorporating semantic knowledge from ConceptNet. For each training instance, the decoding process starts with masked multi-head self-attention to ensure autoregressive prediction, followed by cross-attention that fuses visual and relational representations. The resulting decoder outputs are projected to vocabulary logits, which are adjusted using ConceptNet-derived knowledge when applicable. A softmax function normalises the adjusted logits into word probability distributions, and the model parameters are updated over N_T samples using the cross-entropy loss.

```
Algorithm 2. TrainingTransDecCNet$(\mathcal{D}, \varphi)$
Input: Training dataset $\mathcal{D} = \{(F_i, Z_i, \mathcal{S}_i, O_i), \forall i = \overline{1, N_T}\}$.
Output: The parameters of the model $\varphi$ have been optimized.
Begin
    $\varphi \leftarrow$ Random Initialization
    # Process each data sample in the training set
    for $i = 1\ to\ N_T$ do
        $Loss \leftarrow 0$
        $H_{init} = Embedding(S_i)$
        $Q_{masked} = H_{init}W_q^{masked}, K_{masked} = H_{init}W_k^{masked}, V_{masked} = H_{init}W_v^{masked}$
        $Att_{masked} = Softmax\left(\frac{Q_{masked}K_{masked}^T}{\sqrt{d}}\right)V_{masked}$
        $H_{masked} = Add\&Norm(H_{int} + Att_{masked})$
        # Multi-Head cross-attention between the outputs of the encoder and decoder.
        $Q_F = H_{masked}W_q^F, K_F = F_iW_k^F, V_F = F_iW_v^F$
        $Att_F = Softmax\left(\frac{Q_FK_F^T}{\sqrt{d}}\right)V_F$
        $Q_Z = H_{masked}W_q^Z, K_F = ZW_k^F, V_Z = Z_iW_v^Z$
        $Att_Z = Softmax\left(\frac{Q_ZK_Z^T}{\sqrt{d}}\right)V_Z$
        $CombinedAtt = \alpha.Att_F + (1 - \alpha).Att_Z$
        $H_{cross} = Add\&Norm(H_{masked} + CombinedAtt)$
        $H_{final} = FeedForward(H_{cross})$
        $H_{final} = FeedForward(H_{final})$
        $Logits = W_oH_{final}$
        # Adjusting the logits with semantic knowledge from ConceptNet
        foreach $(l, w) \in O_i$ do
                $Logits'[l] = Logits[l] + \beta w$
        endfor
        $P = Softmax(Logits')$
        # Calculating the loss for the i$^{th}$ sample
        $Loss_i = -\sum_{t=1}^{N_{S_i}} logP(s_t^i|S_{<t}^i, F_i, Z_i, O_i)$
        # Updating the model parameters
        $\varphi = \varphi - \eta\frac{\partial Loss_i}{\partial \varphi}$
    endfor
End
```

In Algorithm 3, captions for an input image are produced via a pre-trained transformer decoder (φ). Utilising the features from object regions and embeddings of the relationship graph, the process begins with the token. For every iteration, the ongoing sequence undergoes encoding through masked multi-head attention, followed by fusion with image features using multi-head cross-attention. The resulting output passes through a linear layer to yield logits across the vocabulary, which are then normalised by $softmax$ into probabilities. The term with the peak probability becomes the subsequent addition. This loop persists, adding each new word to the sequence, until it reaches the token or the predefined maximum length. Ultimately, this yields a precise and relevant caption drawn from the image's characteristics.

```
Algorithm 3. GenerateCaption($F_I, Z_I, \varphi$)
Input:$F_I, Z_I$, the trained transformer model $\varphi$
Output: the captions of image $\hat{S}_I$
Begin
    $\hat{S}_I \leftarrow [< start >]$
    $X = Embedding(\hat{S}_I)$
    while$(last\ token\ \neq < end >)$ $and$ $\left(length\ of\ \hat{S}_I < \max length\right)$do
        $H_{init} = Embedding\left(\hat{S}_I\right)$
        $H_{masked} = MaskedMultiHeadAttention(H_{init})$
        # Combine features
        $H_{cross} = MultiHeadCrossAttention(H_{masked}, F_I, Z_I)$
        # Predict the next word
        $Logits(s_t) = W_o.FeedForward(H_{cross})$
        $P(s_t) = Softmax(Logits(s_t))$
        $s_t = argmax_{s \in Vocab} P(s_t)$
        # Update the caption
        $\hat{S}_I = \hat{S}_I \cup s_t$
    end
End
```

4. EXPERIMENTS AND RESULTS

Grounded in the theoretical framework and model architecture described earlier, this section presents the experimental setup and performance evaluation using standard metrics commonly adopted in image captioning tasks. It also provides an analysis of the results, along with a comparative assessment against baseline methods and recent state-of-the-art models, in order to emphasise both the strengths and potential limitations of the proposed approach.

4.1. Data and Experimental setup

This section describes the experimental data, parameters, and configuration settings for implementing the proposed method. It also presents performance evaluation metrics.

4.1.1. Experimental Data

The proposed framework for generating image captions underwent evaluation on the MS COCO dataset [30], a standard reference widely utilized for object recognition, segmentation, and image captioning tasks. This collection features 82,783 training images and 40,504 validation images, each provided with a minimum of five captions authored by people. For uniformity in our analyses, we limited usage to the first five captions per image. To facilitate equitable comparisons with existing research, we adhered to the established partitioning scheme from [31], assigning 82,783 images to training, 5,000 to validation, and a further 5,000 to testing. In the preparation phase, words occurring fewer than five times were omitted from the lexicon, culminating in 10,010 unique terms and a maximum of 16 tokens per caption.

4.1.2. Implementation Details

The proposed model was developed using Python version 3.9 and implemented with the PyTorch deep learning framework version 2.0. All experiments were conducted on the Google Colab Pro platform, utilising the following computational settings and hyperparameters:

The process of creating and embedding the relationship graph was carried out according to the setup in [12].

Transformer Decoder: The decoder consists of N=6 blocks with 8 heads. The vector dimension for word representation is 512. The Adam optimizer is used with a learning rate of 0.00004 and a batch size of 32.

ConceptNet: ConceptNet 5.7 is employed to retrieve related entity knowledge of objects in the image via the REST API at api.conceptnet.io. Five objects with the highest probability are selected for each image to extract semantic knowledge from ConceptNet. For each object, the top 10 most relevant semantic knowledge entries (with the highest probability) are selected and input into the decoder to enhance performance during caption generation.

The configuration of Google Colab Pro used: Tesla T4 GPU with 15 GB, 51 GB RAM. The image captioning model training time is approximately 12 hours (20 epoches), and the average inference time per image is approximately 1.5 seconds.

It is worth noting that in the proposed framework, only the Transformer decoder is trained, while the object detection, relationship graph construction, and graph embedding modules are reused from prior work and kept fixed during both training and inference. This design significantly reduces training cost and computational overhead while maintaining competitive performance.

4.2. Evaluation metrics

The evaluation metrics used in this study are widely adopted measures for assessing the quality of generated image captions compared to the provided ground truth caption set, including BLEU [32], METEOR [33], ROUGE [34], and CIDEr [35]. Each metric evaluates the captions from distinct perspectives and uses different calculation methods. However, a common characteristic of these metrics is that the higher the score, the better the model's performance.

4.3. Results and Discussion

The experimental results of the proposed image captioning method are presented in Table 1, with the BLEU1, BUEU4, METEOR, ROUGE, and CIDEr scores achieving 77.5, 34.9, 28.3, 55.3, and 98.4, respectively, for the RGTran model (without integrating semantic knowledge from ConceptNet), and 79.8, 36.3, 35.6, 57.2, and 107.8 for the RGTranCNet model (with semantic knowledge integration from ConceptNet). These results indicate that the RGTran model (using transformer and cross-attention mechanism) outperforms the OD-VR-Cap method [12] (which uses LSTM and dual-attention mechanism) across all evaluation metrics, particularly with a significant increase in CIDEr (+13.3 points). This improvement is mainly attributed to the cross-attention mechanism of the transformer, which is capable of integrating information from various sources into a shared space, producing more comprehensive and semantically rich features than independent attention mechanisms as in OD-VR-Cap. Additionally, the transformer decoder is more effective than LSTM in handling complex relationships, thanks to the self-attention mechanism that flexibly and robustly captures the dependencies between words in a sentence. Furthermore, integrating semantic knowledge from ConceptNet into the decoder of the RGTranCNet model leads to an overall improvement across all evaluation metrics, with notable increases in METEOR (up by 7.3 points) and CIDEr (up by 9.4 points). The improvement in METEOR is due to the ability to match synonyms. At the same time, CIDEr reflects the fluency and coherence of the generated captions, which aligns with the integration of semantic knowledge from ConceptNet, resulting in more accurate and meaningful captions.

Table 1. Image captioning performance of the proposed method on the experimental dataset's test set.

Methods	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
RGTran (**without ConceptNet**)	77.5	34.9	28.3	55.3	98.4
RGTranCNet (**with ConceptNet**)	**79.8**	**36.3**	**35.6**	**57.2**	**107.8**

An example of the results from the proposed image captioning model is presented in Figure 5. In this figure, (a) shows the input image, and (b) shows the captions generated by the respective models. The results indicate that the OD-VR-Cap model [12] only captures the primary objects and relationships in the image (the object “man” and action “fixing” on the left, and the objects “person”, “skateboard” and action “jumping” on the right), with details and context not fully represented. RGTran enhances the model's ability to identify specific locations and contexts; however, it still struggles with detailed contexts or objects and actions not present in the training dataset. In contrast, RGTranCNet incorporates additional semantic knowledge from ConceptNet into the decoder, allowing the model to handle unseen objects by substituting them with semantically relevant concepts or enriching the relationships between objects. As a result, the generated captions are more accurate and meaningful.

These qualitative improvements are a direct result of integrating structured semantic knowledge into the captioning process. Specifically, the relationship graph enables the model to encode pairwise spatial and functional relations between detected objects. At the same time, ConceptNet provides external semantic associations that help enrich the meaning of individual concepts. For instance, replacing “fixing” with “repairing” or “jumping” with “performing a skateboard trick” reflects a deeper understanding of the functional context of the scene, not just object labels. This combination allows RGTranCNet to generalize better to unseen situations and produce captions that are not only accurate in terms of object recognition but also more human-like in their semantic expressiveness.

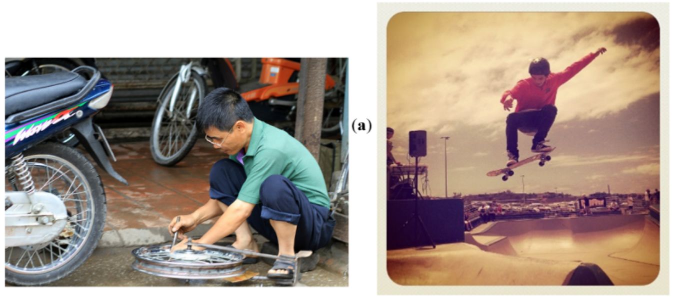
To further demonstrate the effectiveness of RGTranCNet, Table 2 presents a comparative evaluation against several baseline models [9] and recent state-of-the-art approaches on the MS COCO dataset. RGTranCNet achieves the highest performance across all reported metrics, including BLEU-1 (79.8), BLEU-4 (36.3), METEOR (35.6), ROUGE (57.2), and CIDEr (107.8). Compared to CNet-NIC - a model that also incorporates ConceptNet but relies on a conventional NIC architecture - RGTranCNet outperforms it significantly in BLEU-4 (+6.4), METEOR (+10.0), and CIDEr (+0.6), indicating substantial improvements in both expressiveness and semantic representation. Similarly, in comparison to ConvNeXt, a model built upon a modern visual encoder architecture, RGTranCNet also achieves superior results in BLEU-1 and BLEU-4.

Notably, the Caption TLSTMs model achieves a CIDEr score of 101.8, surpassing both OD-VR-Cap and RGTran, which demonstrates its ability to produce highly informative captions. This performance stems from its architectural design, which integrates an abstract scene graph and a Transformer block inserted between two LSTM layers. Such a configuration enables the model to generate captions that are diverse in content and coherent in structure, thereby contributing to a notable improvement in CIDEr. However, its CIDEr score remains lower than that of RGTranCNet, suggesting that while the architecture is effective, it does not reach the level of semantic and structural richness provided by the proposed model. Moreover, Caption TLSTMs underperforms on other metrics such as BLEU and METEOR, indicating that its semantic expressiveness is still less comprehensive compared to RGTranCNet, which achieves consistently high performance across both syntactic and semantic dimensions.

Table 2. Comparison of image captioning performance across methods on the experimental dataset.

Methods	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
Show, attend and tell (Hard-ATT) [9]	71.8	25.0	23.0	-	-
Show, attend and tell (Soft-ATT) [9]	70.7	24.3	23.9	-	-
CNet-NIC [18]	73.1	29.9	25.6	53.9	107.2
En-De-Cap [20]	70.6	24.3	-	-	-
Caption TLSTMs [23]	-	22.9	25.2	50.9	101.8
Bi-LS-AttM [21]	68.8	25.2	21.5	-	41.2
ConvNeXt [24]	74.8	34.8	-	-	-
OD-VR-Cap [12]	72.6	28.1	24.8	53.4	85.1
RGTran (ours)	77.5	34.9	28.3	55.3	98.4
RGTranCNet (ours)	**79.8**	**36.3**	**35.6**	**57.2**	**107.8**

In summary, the experimental results validate that integrating relationship graphs, a Transformer decoder with cross-attention, and semantic knowledge from ConceptNet constitutes a practical approach to image captioning. RGTranCNet not only exploits visual and relational structural information but also demonstrates the ability to synthesise semantics and infer meaning from external knowledge sources. Consequently, it generates captions that are more accurate, fluent, and semantically enriched than those produced by prior methods.



OD-VR-Cap: a man **fixing** a motorcycle on the ground
RGTran: A man **sitting on the ground** fixing the **wheel of a motorcycle**
RGTranCNet: A man sitting on the ground **repairing** the wheel of a motorcycle **outdoors**.

(b)

OD-VR-Cap: a person **jumping** a skateboard in the air
RGTran: A person jumping on a skateboard **in a skate park**
RGTran-CNet: A person **performing** a skateboard trick in the air at a skate park

Figure 5. Example results from the test image set for the proposed method and OD-VR-Cap.

5. CONCLUSIONS

In this research, we present RGTranCNet, an innovative model for image captioning grounded in the encoder-decoder paradigm. It employs a transformer-based decoder equipped with cross-attention functionality, while integrating semantic insights from ConceptNet directly into the decoding phase. By harnessing the transformer's strength in modelling contextual dependencies and drawing on outside semantic resources, the system boosts both the precision and variety of the generated descriptions. Evaluations performed on the benchmark MS COCO dataset indicate that our framework surpasses contemporary efforts in terms of overall effectiveness. The addition of ConceptNet-derived semantics enhances the model's ability to craft more precise and contextually aligned descriptions, thereby improving the capabilities of automated captioning solutions. Importantly, this technique can be seamlessly incorporated into the decoding components of alternative encoder-decoder-based captioning systems to amplify their results. As such, the proposed solution proves both viable and applicable, laying the groundwork for advancing captioning technologies across diverse practical sectors. While our tests were confined to the MS COCO dataset, the methodology lends itself to adaptation for other collections (such as Flickr8k or Flickr30k), given that training is limited to the decoder with other elements held constant.

Furthermore, the external knowledge infusion from ConceptNet operates independently of particular image traits, ensuring the method avoids over-reliance on the format of any given training set. Looking ahead, we intend to deploy and assess the model on these additional datasets to confirm its versatility. We also aim to incorporate inter-concept relationships from ConceptNet within the encoder to provide richer context and refine caption quality.

Acknowledgements. We acknowledge the support of the Institute of Mechanics and Applied Informatics - VAST, the Faculty of Information Technology and Telecommunications - Graduate University of Science and Technology - VAST, and the Faculty of Information Technology - Ho Chi Minh City University of Education (HCMUE) for this study.

CRediT authorship contribution statement. Nguyen Van Thinh: Methodology, Software, Investigation, Writing – original draft. Tran Van Lang: Formal analysis, Supervision, Writing – review & editing. Van The Thanh: Supervision, Writing – review & editing.

Declaration of competing interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. Jamil A., Saif Ur R., Mahmood K., Villar M. G., Prola T., Diez I. D. L. T., Samad M. A., Ashraf I. - Deep Learning Approaches for Image Captioning: Opportunities, Challenges and Future Potential. IEEE Access, **12** (2025) 1-1. https://doi.org/10.1109/access.2024.3365528.
2. Verma A., Yadav A. K., Kumar M., Yadav D. - Automatic image caption generation using deep learning. Multimed. Tools Appl., **83** (2023) 5309-5325. https://doi.org/10.1007/s11042-023-15555-y.
3. Kavitha R. - E3S Web of Conferences, EDP Sciences, (2023) 04005. https://doi.org/10.1051/e3sconf/202339904005.
4. Pavlopoulos J., Kougia V., Androutsopoulos I. - Proceedings of the Second Workshop on Shortcomings in Vision and Language, Association for Computational Linguistics, (2019) 26-36. https://doi.org/10.18653/v1/W19-1803.

5. Szafir D., Szafir D. A. - Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, Association for Computing Machinery, (2021) 281-292. https://doi.org/10.1145/3434073.3444683.
6. Lin Y.-J., Tseng C.-S., Hung Y.-K. - Relation-Aware Image Captioning with Hybrid-Attention for Explainable Visual Question Answering. J. Inf. Sci. Eng., **40**(3) (2024) 479-494.
7. Vinyals O., Toshev A., Bengio S., Erhan D. - Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, (2015) 3156-3164. https://doi.org/10.1109/CVPR.2015.7298935.
8. Huang L., Wang W., Chen J., Wei X.-Y. - Proceedings of the IEEE/CVF International Conference on Computer Vision, IEEE, (2019) 4634-4643. https://doi.org/10.1109/ICCV.2019.00473.
9. Xu K., Ba J., Kiros R., Cho K., Courville A., Salakhutdinov R., Zemel R., Bengio Y. - Proceedings of the 32nd International Conference on Machine Learning, PMLR, (2015) 2048-2057. https://doi.org/10.48550/arXiv.1502.03044.
10. Thinh N. V., Lang T. V., Thanh V. T. - The 16th National Conference on Fundamental and Applied IT Research (FAIR'2023), Natural Science and Technology Publishing House, (2023) 395-404. https://doi.org/10.15625/vap.2023.0063.
11. Anderson P., He X., Buehler C., Teney D., Johnson M., Gould S., Zhang L. - Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, (2018) 6077-6086. https://doi.org/10.1109/CVPR.2018.00636.
12. Nguyen Van T., Lang T. V., Van V. T. T. - OD-VR-Cap: Image captioning based on detecting and predicting relationships between objects. J. Comput. Sci. Cybern., **40**(4) (2024) 327-346. https://doi.org/10.15625/1813-9663/20929.
13. Xu N., Liu A.-A., Liu J., Nie W., Su Y. - Scene graph captioner: Image captioning based on structural visual representation. J. Vis. Commun. Image Represent., **58** (2019) 477-485. https://doi.org/10.1016/j.jvcir.2018.12.027.
14. Thinh N. V., Lang T. V., Thanh V. T. - The 15th National Conference on Fundamental and Applied IT Research (FAIR'2022), Natural Science and Technology Publishing House, (2022) 431-439.
15. Li Z., Wei J., Huang F., Ma H. - Modeling graph-structured contexts for image captioning. Image Vis. Comput., **129** (2023) 104591. https://doi.org/10.1016/j.imavis.2022.104591.
16. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. - Advances in Neural Information Processing Systems, Neural Information Processing Systems Foundation, (2017) 5998-6008. https://doi.org/10.48550/arXiv.1706.03762.
17. Hendricks L. A., Venugopalan S., Rohrbach M., Mooney R., Saenko K., Darrell T. - Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, (2016) 1-10. https://doi.org/10.1109/CVPR.2016.8.
18. Zhou Y., Sun Y., Honavar V. G. - IEEE Winter Conference on Applications of Computer Vision (WACV 2019), IEEE, (2019) 283-293. https://doi.org/10.1109/WACV.2019.00036.
19. Hafeth D. A., Kollias S., Ghafoor M. - Semantic Representations With Attention Networks for Boosting Image Captioning. IEEE Access, **11** (2023) 40230-40239. https://doi.org/10.1109/access.2023.3268744.
20. Patwari N., Naik D. - 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), IEEE, (2021) 1206-1211. https://doi.org/10.1109/ICCMC51019.2021.9418414.
21. Xie T., Ding W., Zhang J., Wan X., Wang J. - Bi-LS-AttM: A Bidirectional LSTM and Attention Mechanism Model for Improving Image Captioning. Appl. Sci., **13**(13) (2023) 7916. https://doi.org/10.3390/app13137916.

22. Chen S., Jin Q., Wang P., Wu Q. - Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, (2020) 9959-9968. https://doi.org/10.1109/CVPR42600.2020.00998.
23. Yan J., Xie Y., Luan X., Guo Y., Gong Q., Feng S. - Caption TLSTMs: combining transformer with LSTMs for image captioning. Int. J. Multimed. Inf. Retr., **11**(2) (2022) 111-121. https://doi.org/10.1007/s13735-022-00228-7.
24. Ramos L., Casas E., Romero C., Rivas-Echeverría F., Morocho-Cayamcela M. E. - A Study of ConvNeXt Architectures for Enhanced Image Captioning. IEEE Access, **12** (2024) 13711-13728. https://doi.org/10.1109/access.2024.3356551.
25. Wang Y., Xu J., Sun Y. - Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, (2022) 2585-2594. https://doi.org/10.1609/aaai.v36i3.20160.
26. Yang X., Wang Y., Chen H., Li J., Huang T. - Context-aware transformer for image captioning. Neurocomputing, **549** (2023) 126440. https://doi.org/10.1016/j.neucom.2023.126440.
27. Li Z., Su Q., Chen T. - External knowledge-assisted Transformer for image captioning. Image Vis. Comput., **140** (2023) 104864. https://doi.org/10.1016/j.imavis.2023.104864.
28. Hamilton W. L., Ying Z., Leskovec J. - Advances in Neural Information Processing Systems, Neural Information Processing Systems Foundation, (2017) 1024-1034. https://doi.org/10.48550/arXiv.1706.02216.
29. Speer R., Chin J., Havasi C. - Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, (2017) 4444-4451. https://doi.org/10.1609/aaai.v31i1.11164.
30. Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C. L. - European Conference on Computer Vision, Springer, (2014) 740-755. https://doi.org/10.1007/978-3-319-10602-1_48.
31. Karpathy A., Fei-Fei L. - Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, (2015) 3128-3137. https://doi.org/10.1109/CVPR.2015.7298932.
32. Papineni K., Roukos S., Ward T., Zhu W.-J. - Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, (2002) 311-318. https://doi.org/10.3115/1073083.1073135.
33. Banerjee S., Lavie A. - Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, (2005) 65-72.
34. Lin C.-Y. - Proceedings of the ACL-04 Workshop, Association for Computational Linguistics, (2004) 74-81.
35. Vedantam R., Lawrence Zitnick C., Parikh D. - Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, (2015) 4566-4575. https://doi.org/10.1109/CVPR.2015.7299087.