

A Novel Hybrid Architecture for Image Captioning

Tran Thi Hue^{1,2}, Le Thi Ha^{3,5}, Pham Van Hai⁴, Do Hoang Long⁴, Luong Thi Hong Lan⁵

¹ Vietnam Academy of Science and Technology, Hanoi, Vietnam

² Dai Nam University, Hanoi, Vietnam

³ Hanoi College of Industrial Economics, Hanoi, Vietnam

⁴ Hanoi University of Science and Technology, Hanoi, Vietnam

⁵ Hanoi University of Industry, Hanoi, Vietnam

Correspondence: Luong Thi Hong Lan, lanlhbk@hau.edu.vn

Received: 31/07/2024, revised: 05/09/2024, accepted: 11/11/2024

Digital Object Identifier: 10.32913/mic-ict-research-vn.v2024.n2.1295

Abstract: Providing descriptions for things, visual objects, or image metadata is called image captioning. It involves using computer vision and natural language processing (NLP) methods to extract image features. In this work, we introduced SPNERSAN, a novel hybrid architecture that synergizes the Enhanced Relational Self-Attention Network (ER-SAN), which follows a Transformer-like structure, with our newly developed Subgraph Proposal Network (SPN). The primary objective of this combination is to enhance the image captioning task by leveraging the unique capabilities of both components. Our experiments on the Coco dataset demonstrate that the SPNERSAN model significantly improves the quality of generated captions compared with related methods. Our proposed model performs better in five standard metrics: BLEU, METEOR, ROUGR, CIDER, and SPICE. It indicates its effectiveness in producing coherent and contextually appropriate captions.

Keywords: Image captioning, transformer model, semantic graph, geometry graph

Tiêu đề: Một kiến trúc tổng hợp mới cho mô tả ảnh

Tóm tắt: Quá trình tự động sinh ra lời mô tả về sự vật, đối tượng trực quan trong ảnh được gọi là chú thích hình ảnh. Nó liên quan đến việc sử dụng cả thị giác máy tính và xử lý ngôn ngữ tự nhiên (NLP) để trích xuất các đặc điểm của hình ảnh. Trong nghiên cứu này, chúng tôi phát triển mô hình SPNER-SAN kết hợp giữa Mạng tự chú ý quan hệ nâng cao (ER-SAN), có dạng cấu trúc giống như Transformer, với Mạng đề xuất đồ thị con (SPN) mới được phát triển của chúng tôi. Mục tiêu chính của sự kết hợp này là cải thiện kết quả của quá trình sinh chú thích cho hình ảnh bằng cách tận dụng các ưu điểm của cả hai thành phần. Thử nghiệm trên bộ dữ liệu Coco chứng minh rằng mô hình SPNER-SAN cải thiện đáng kể chất lượng của câu mô tả sinh ra so với với các phương pháp liên quan. Mô hình đạt hiệu suất cao hơn trên cả 5 chỉ số đo so sánh như BLEU, METEOR, ROUGR, CIDER, và SPICE. Điều này cho thấy tính hiệu quả của nó trong việc sinh tự động các chú thích rõ ràng và phù hợp với ngữ cảnh.

Từ khóa: Chú thích ảnh, mô hình transformer, đồ thị ngữ nghĩa, đồ thị hình học

I. INTRODUCTION

Image captioning, generating descriptive text for visual content, is a crucial yet challenging task in computer vision. It involves extracting features from images using a combination of computer vision and Natural Language Processing (NLP) techniques, identifying objects, actions, and their relationships within the image, and ultimately creating meaningful descriptions. The problem of automatically generating image captions has found applications in diverse fields, such as human-computer interaction [1], medical imaging [2], traffic analysis [3], and, notably,

assistive technologies for the visually impaired [4]. As a multidisciplinary field, automatic image annotation bridges the gap between natural language processing and computer vision. Consequently, with the rapid advancements in deep learning algorithms, researchers have been actively exploring the potential of these methods to enhance the performance of image annotation tasks.

In image captioning, the transformer-based self-attention model has consistently demonstrated superior performance. While the inherent focus of the self-attention mechanism lies in implicit comparisons between a single region and

all other regions within an image, recent research has explored enhancing this capability by incorporating geometric information into the self-attention operator [5]. Leveraging geometric features like relative size and distance, which explicitly capture spatial relationships between objects, has led to state-of-the-art results in numerous approaches [6]. These findings strongly suggest that a promising avenue for future research involves developing visual encoders capable of accurately modeling the complex interplay of objects within an image.

Li et al. [7] recently introduced a groundbreaking visual relation representation and semantic fusion prediction method named ERSAN with notable advancements. Their approach leverages explicitly the visual graph topology as an inductive structure-enhanced relation bias, effectively differentiating the significance of various nodes within the graph. Moreover, in the language prediction stage, they utilize explicit semantic relations in the decoder to adaptively enhance semantic word generation. However, while groundbreaking, ERSAN's approach overlooks the potential overlap and calculation of objects' importance weights while synthesizing graphs from semantic and geometric feature graphs. This limitation is the primary motivation for our research to enhance the ERSAN model.

To overcome these limitations, we introduce the SPNER-SAN model, an innovative solution that combines the strengths of ER-SAN and SPN. SPNER-SAN is designed to handle complex image relationships better using geometric and semantic features. Specifically, the contributions and novelty of this paper are summarized as follows:

- Propose two novel metrics to determine the importance of an object in the image based on the semantic and geometric graph of objects.
- Propose a shortened subgraph that only needs to focus on representing nodes with high-importance measures.
- Develop the novel SPNER-SAN model that integrates the subgraph proposal network into the ERSAN model to reduce the model's computational complexity.
- Prove the effectiveness of the proposed model through a comparative study with the original ERSAN model and Transformer using the Coco data set on evaluation metrics: BLEU, Meteor, ROUGR, cider, and spice.

This work is arranged according to the following structure. Section II overviews some related works in image captioning. The novel SPNER-SAN model that is being developed to improve the outcomes of image captioning is shown in Section III. To illustrate the efficacy and contextual improvements of our proposed approach, Section IV compares the results of our suggested model with the original ERSAN model and Transformer using the Coco

data set. Finally, some conclusions are summed up in the final part of this study.

II. RELATED WORKS

As a multi-modal task that integrates language and vision, automatically producing image descriptions is essential. The caption generation model is anticipated to identify the noteworthy objects in the picture and explain how they relate to other objects in the composition. To enhance the quality of image captioning, it is crucial to precisely and efficiently model the relationships between identified objects. Using a Faster Region-based Convolutional Neural Network (R-CNN) [8] with picture regions as a bottom-up path in combination with grid-level features, Anderson et al. [9] introduced the bottom-top attention mechanisms for visual areas in an effort to improve the visual features. After that, some researchers [10] use pooled vectors for the input image to encode visual areas. [11] proposed a novel encoder-decoder architecture called Adaptive Text Denoising Network (ATD-Net), which adds the denoising process into the encoder to reduce semantic inconsistency between image and caption.

Image captioning has been effectively achieved with the transformer model, especially concerning the self-attention process. Nevertheless, managing object interactions remains a challenge for most models. By including geometric information in the model, some earlier works tried to solve this problem [12]. Previous research has concentrated on visual graphs representing item connections in pictures. Yao et al. [13] used visualizations based on the semantic relations graph—obtained by a predetermined semantic classifier—and the spatial relations graph—captured between bounding boxes of object pairs. After this study, Yang et al. [14] suggested using a graph-based model of both sentences and images to integrate semantic priors acquired from the text in the image encoder stage.

Subsequently, Shi et al. [15] learned their relational classifier from ground-truth description using a multi-instance learning strategy, giving rise to a reliable visual relation for the given image. The dual-attention interaction technique, which depicts the visual correlations at the object and region levels, was introduced by Liu et al. [16]. In recent years, research [17] recommended a deep neural network architecture based on combining the spatial characteristic of the image with the context information to use visual relationships in caption output.

These findings show that creating a visual encoder that accurately models the relationships between image elements is a potential avenue for future research. Furthermore, we contend that, in many situations, things nearer the present object should be given greater importance than objects

farther away, establishing a contrast in priority. This is also our motivation to learn this work.

III. PROPOSED NOVEL SPNER-SAN ARCHITECTURE FOR IMAGE CAPTIONING

1. Main ideas of the proposed model

The objective of image captioning, given a picture I , is to automatically produce a sentence S based on the visual elements in the image. We aim to fully utilize the explicit relation as a guide to provide more accurate captions.

Figure 1 illustrates the general structure of our suggested SPNER-SAN model. The proposed model comprises four sub-modules: the Feature Extraction Module (Sec III.2), the Refined Subgraph Module (Sec III.3), the Enhanced-Adaptive Relation Attention (Sec III.4), and the Language Prediction Module (Sec III.5). In order to model image representation collaboratively, we first extract geometric, semantic, and visual information. Semantic and geometrical graphs are built based on the features extracted from the objects in the image. To reduce the computational complexity of the model, the graph reduction process is proposed to prune unimportant information in the graph. This module determines the importance of each object in the image at the regional level. Then, the semantic and geometric information of the objects in the image is modeled and embedded into the transformer model to calculate the attention level of each object in the image. Finally, our model produces credible captions that accurately anticipate relationships.

The proposed SPNER-SAN model is designed to enhance image understanding by building upon the capabilities of the ER-SAN model [7]. SPNER-SAN leverages pre-trained data from the WeakVRD model [15] to extract object relationship trees and utilize image features from the bottom-up attention model. Additionally, it employs Faster R-CNN [8] with knowledge training using the ResNet-101 architecture [18] on the Visual Genome dataset. The extracted data is then processed through the subgraph proposal network layer, which utilizes a GAT network [19] to pinpoint the relevant segment of the graph, effectively reducing the graph's size and accelerating training for the ER-SAN model. The following section will provide a more comprehensive explanation of the model.

2. The Feature Extraction Module

a) Contextual Feature Extraction in image

In this section, we will elaborate on the topic of feature extraction to generate the semantic and geometric graph features of objects in image [7]. Within the domain of

object detection, a prominent technique involves the combination of Faster R-CNN with ResNet-101. Faster R-CNN leverages deep convolutional neural networks to perform object detection with improved speed. The incorporation of ResNet-101 enhances the model's ability to recognize and understand complex features, improving both accuracy and performance.

Moreover, in constructing scene graphs, the methodology known as Caption-Led Visual Relationship Graphs (CGVRG) is employed [15]. This approach harnesses weakly supervised visual relationship detection (WeakVRD) to build scene graphs where nodes represent identified objects, and their corresponding node embeddings encapsulate the object features. The establishment of edges connects objects with intersecting bounding boxes as determined by the object detection algorithm, and the features of these edges are determined through the computation of metrics associated with the overlapping areas between objects. Such an approach contributes to elucidating object contexts and relationships within images. We first utilize Faster R-CNN to encode the region-level RoI features into n -region features $o^v = [o_1^v, o_2^v, \dots, o_n^v]$. Also, we obtain the object category labels c , and then we adopt the learnable object embedding layer to map the object category labels c into feature embedding vectors o^c . To obtain the contextual representation of the object node feature, we concatenate the visual feature o^v and embedding vectors o^c as:

$$H_{O_i} = \phi_o ([o^v; o^c]) + o^v \quad (1)$$

Where $H_{O_i} \in \mathbb{R}^{N \times d_k}$ is the following input node region feature representation passing in the transformer encoder layer, ϕ_o is feed-forward networks using ReLU activation, $[:]$ means concatenation operation.

b) Build the Semantic and Geometric graph

Caption-Guided Visual Relationship Graph (CGVRG) builds visual relationship graphs. A significant challenge in utilizing predicate triples from captions to construct CGVRG lies in the ambiguity of links between predicates and their corresponding object regions. To address this problem, weakly supervised multi-instance learning is employed. This method helps to establish more explicit connections between predicates and object regions, thereby improving the accuracy and relevance of the visual relationship graphs constructed for image captioning.

Once the trained module, a CGVRG graph $G = (V, E)$ can be built for a given image I , where the node set V includes two types of nodes: object nodes and predicate nodes. Denote o_i as the i -th object node and r_{ij} as a predicate node that connects o_i and o_j . The edges in E

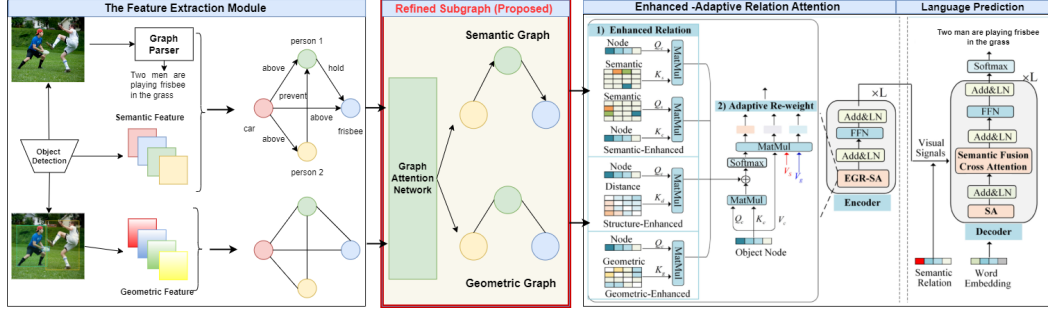


Figure 1. Overview of the enhanced SPNER-SAN architecture.

are added based on triples; i.e., (o_i, r_{ij}, o_j) will assign two directed edges from node o_i to r_{ij} and from r_{ij} to o_j , respectively.

Build Semantic Graph : A combination of the Weakly-VRD Graph Parser and the Object feature is used to construct the semantic graph. We employ the graph parser to extract the semantic relation triples (e.g., <subject-predicate-object>) to represent the precise visual semantic associations, as described in [15]. This approach creates caption-guided visual relationship graphs that employ weakly supervised multi-instance learning to introduce advantageous inductive bias. The semantic relation matrix between subject area I and object region J is computed using the semantic relation triples and has $R^{N \times N}$ dimensions.

Build Geometric Graph : In contrast to the semantic graph, the geometric graph is undirected. First, we compute the relative geometric features between two boxes, i and j , as follows.

$$\left(\log \left(\frac{|x_i - x_j|}{w_i} \right), \log \left(\frac{|y_i - y_j|}{h_i} \right), \log \left(\frac{w_i}{w_j} \right), \log \left(\frac{h_i}{h_j} \right) \right) \quad (2)$$

where x , y , w , and h represent the box's center, width, and height. We then use the PE_{pos} functions [20] to compute high-dimensional representations, $G \in R^{N \times N \times d_g}$, to obtain the geometric embedding. Additionally, to get a sparse geometric graph, the geometric relation between two objects is assigned if the relative distance $dis(b_i, b_j)$ and Intersection over Union $IoU = (b_i, b_j)$ of both objects are within a threshold.

3. Refined subgraph module

The results of the attribute extraction module of objects in the image are semantic and geometric attribute maps. However, the graph is quite bulky and has many duplicates, so this section aims to prune redundant or unimportant nodes. Additionally, we presume that things closer to the image's center require more attention than those farther

away, and vice versa. Consequently, we investigate how to use visual graph layouts to enhance model performance in image annotation.

a) *Propose a measure of the importance of an object in an image*

A well-known guideline that aids photographers and artists in producing visually appealing images is the rule of thirds [21]. The rule of thirds allows for more precise identification of the main value and the main subject or focus of interest of the image. That's why in images, in this section, we have proposed two measurements to only focus on the objects of interest in the image. This helps eliminate redundant objects and reduce the computational complexity of the model, which is especially beneficial in situations where computational resources are limited or real-time processing is required.

Global Importance weight (GI) is a metric that reflects the overall prevalence of an object within the entire dataset. Initially, we calculate the total occurrences of the object across the dataset. For each image given $C = \{c_1, c_2, \dots, c_N\}$ with N is the number of detected objects, c_i is the number of occurrences of an object i^{th} in the whole dataset's caption. Then, we compute the global importance ratio by applying Minmax scaling to obtain C_{scaled} , kindly forward it to a sigmoid function $\sigma()$, and receive $\tilde{C}$.

$$C_{scaled} = \frac{c - c_{min}}{c_{max} - c_{min}} \quad (3)$$

$$\tilde{C} = \sigma(C_{scaled}) \quad (4)$$

The primary motivation for employing the encoding method based on the occurrence frequency vector of objects in an image stems from the need to address the varying frequencies at which different objects appear in the annotation dataset. In many cases, some objects occur frequently, while others are rarer. When a simple linear transformation is applied to such a dataset, the distances between objects

may become distorted, leading to the overrepresentation of common objects and the underrepresentation of rare ones. This imbalance can distort the model's learning process and result in skewed outcomes.

To address this issue, a non-linear transformation is introduced as a corrective measure. The transformation can compress the distances between objects with varying occurrence frequencies by applying a non-linear function. This ensures that less frequent objects in the dataset are given higher priority, allowing them to be more fully represented in the transformed space. Consequently, rare objects are more likely to be retained during processes such as semantic graph pruning, which typically focuses on preserving the most significant objects.

This method helps balance the representation of frequently occurring and rare objects, ensuring that the model learns accurately and comprehensively from the dataset. The non-linear transformation enhances the model's ability to capture the most essential features of the dataset while minimizing the risk of losing or excluding important yet rare objects.

Local Importance weight (LI) is a metric that assesses the significance of an object's position within a specific image. According to the rule of thirds [21], a photograph's focal point or major element should be situated roughly at or inside the four intersections. Based on this rule, we divided the entire image into nine equal regions, as shown in Figure 2. Objects appearing at the intersections of these regions or near the center of the image are considered more important.

0.25	0.5	0.25
0.5	1	0.5
0.25	0.5	0.25

Figure 2. The weights correspond to each region in the image

The formula for calculating the local importance of an object j in image k is:

$$\text{Local Score} = \sum_{k=1}^9 S_k \times \text{score}_k \quad (5)$$

where S_k represents the percentage of the detected object's area in the image corresponding to the k^{th} part. As the example below, we have an object that lies in many parts, so the Local score of this object is calculated as:

$$\text{Local score} = 0.15 \times 0.5 + 0.35 \times 0.25 + 0.35 \times 0.5 + 0.15 \times 1 = 0.4875$$

The concept presented here elaborates on how the formula evaluates objects in an image based on their position

	0.15	0.35
	0.15	0.35

Figure 3. Visual example about the object's location in image

and size. First, objects that occupy a large portion of the image, such as the background, are generally assigned a lower rating. This is because large objects located on the periphery of the image are often less critical than those positioned centrally, as they do not draw the viewer's attention directly. Another key point is the dispersion of objects across multiple areas of the image. When an object is spread out rather than concentrated in a specific location, its weight in the assessment decreases. This is because dispersed objects are less likely to stand out or play a significant role in conveying the main content of the image.

Ultimately, the primary idea is that our attention is typically directed toward objects of moderate size positioned at the center of the image. These objects are more easily recognizable, attract more focus, and tend to carry more important information. Most visual models and image evaluation methods operate on this principle, where centrally located objects of appropriate size are given priority. Applying this formula thus ensures that attention is focused on key objects while minimizing the influence of unnecessary background elements.

b) Build a Subgraph Proposal Network (SPN)

This section proposes a novel subgraph-based network architecture utilizing Graph Attention Networks (GAT). The primary objective of our approach is to effectively capture and leverage the intricate relationships and structural information present in subgraphs for various tasks such as node classification, link prediction, and graph classification.

Initialization: Each node in the graph is initialized with an initial feature vector, denoted as h_i for each node i . Additionally, the output is initialized as a matrix where the number of rows corresponds to the number of objects detected in the image, and the number of columns corresponds to the number of objects detected in the image.

Message Calculation: During the message calculation phase, each node i sends a message to its neighbouring nodes j . This message is computed based on the feature vectors of nodes i and j . GAT uses an attention mechanism to assign weights to these messages. The amount of information sent from node i to node j must be multiplied by the free prior coefficient according to the following formula (6).

$$m_{ij} = \epsilon_j W h_j \quad (6)$$

$$\epsilon_j = \sqrt{0.3 \times GI_j + 0.7 \times LI_j} \quad (7)$$

where the GI is Global Importance and LI is Local Importance. e_{ij} is the attention score between node i and node j , typically determined by a small neural network as follows:

$$e_{ij} = \alpha(W h_i, W h_j) \quad (8)$$

The attention weight α_{ij} is computed using the formula:

$$\alpha_{ij} = \text{Softmax}(e_{ij}) \quad (9)$$

Aggregation: Each node i receives messages from its neighboring nodes j in the aggregation phase. These messages are aggregated using a function such as summation or averaging:

$$h'_i = \sigma \left(\sum_{j \in N(i)} \epsilon_j W h_j \right) \quad (10)$$

Where $N(i)$ is the set of neighboring nodes of i , and σ is an activation function such as ReLU.

Update Graph: The feature vector of each node is updated using the aggregated feature vector from the aggregation phase, resulting in a new set of feature vectors for the nodes in the graph:

$$h_i \leftarrow h'_i \quad (11)$$

The updated output matrix is passed through a softmax layer to refine the classification process further. This layer converts the continuous output values into probability scores, providing the final classification probabilities for each node. The softmax function ensures that the output probabilities sum to one, facilitating precise and interpretable classification into the critical and non-important categories.

Prune node from Graph: During the Prune phase, nodes classified as non-important are systematically removed from the graph. This process is not merely a refinement of the structure but a critical step in improving the overall model performance. Removing non-important nodes makes the graph more streamlined, reducing noise and irrelevant data. This allows the model to allocate computational resources more effectively, focusing on the most significant nodes.

An essential part of this process is determining the importance of each node. To achieve this, we compute the Importance Index (IS), which quantifies the role of a node based on two factors: its local influence (ϵ_i) and its connections to neighboring nodes (through the sum of weights α_{ij}). The Importance Index is calculated using the following formula:

$$IS_i = \sqrt{\epsilon_i + \frac{1}{N(i)} \sum_j \alpha_{ij}} \quad (12)$$

where IS_i is the Importance Score for node i , ϵ_i represents the local influence of node i , α_{ij} is the weight between node i and node j , and $N(i)$ is the number of neighboring nodes connected to node i .

This score determines whether a node should be pruned from the graph if its value is below a certain threshold. Our subgraph-based GAT architecture leverages the strengths of attention mechanisms, allowing the model to focus on the most critical parts of the graph dynamically. By integrating global and local importance measures, the model evaluates the significance of nodes individually and in relation to the overall graph structure. After the pruning phase, a softmax layer is utilized for final classification, ensuring that only the most essential nodes are preserved. The pruning phase enhances model efficiency by concentrating on high-priority nodes and minimizing distractions from irrelevant elements. As a result, this approach demonstrates substantial improvements over traditional GAT architectures, particularly in the accuracy of node classification tasks.

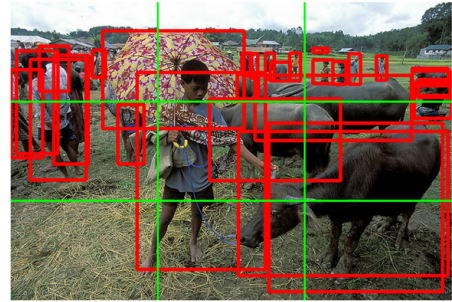


Figure 4. The illustration example for input image of the Refined Subgraph Module

A visual example to demonstrate the effectiveness of the proposed model is shown in Figures 4 and 5. Figure 4 presents over 30 the bounding boxes of objects in the image that Faster RCNN detected. It's large and may not be mentioned in the image's caption. Figure 5 represents the outcome after a refined graph process is applied. The number of objects has been reduced by about half, leaving only 15. These objects are considered the most important

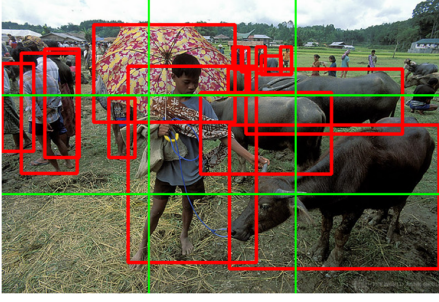


Figure 5. The illustration image after finishing the Refined sub-graph module using two proposed metrics

in the image, as determined through a training process. As you can observe, the remaining objects are positioned in easily noticeable areas, with many occupying central locations within the image. This refined selection highlights the critical elements deemed most relevant and visually impactful in the scene.

4. Enhanced-Adaptive Relation Attention

This section aims to model relations between objects leveraging graphs' semantic and geometric embedding information into the Transformer model. Our current research aligns with the fundamental principles and concepts presented in the study [7].

Enhanced Relation Attention: ER-SAN incorporates explicit visual semantic associations into the Transformer. Specifically, the enhanced attention score, including the effect of content objects H , semantic relation embedding S , and geometric relation embedding G , is computed as:

$$Q_C = HW_{q,c}, K_C = HW_{k,c} \quad (13)$$

$$\tilde{A} = \underbrace{Q_C K_C^T}_{CW} + \underbrace{\phi(Q_C, K_C, S)}_{SE} + \underbrace{\varphi(Q_C, G)}_{GE} \quad (14)$$

where CW is Content Weight, SE is Semantic-Enhanced, GE is Geometric-Enhanced, $\tilde{A} = \mathbb{R}^{N \times N}$, the scaling factor $\sqrt{d}$ is omitted for simplicity. ϕ is the semantic-enhanced attention function. φ is the geometric-enhanced attention function. Q_C and $K_C \in \mathbb{R}^{N \times d}$ are the projected content objects vectors generated using projection matrices $W_{q,c}$, and $W_{k,c} \in \mathbb{R}^{d_k \times d}$, respectively. Eq.4 considers implicit relations (content weight) and explicit relations guidance (semantic and geometric).

Direction-Sensitive Semantic-Enhanced Attention : ER-SAN uses semantic embedding features to enhance attention score when there is a specific directional relationship between two objects. It considers content objects

to semantic relations ($c2s$) and semantic relations to content objects ($s2c$) attention.

$$Q_S = SW_{q,s}, K_S = SW_{k,s} \quad (15)$$

$$\phi(Q_C, K_C, S) = \underbrace{Q_C K_S^T}_{c2s} + \underbrace{Q_S K_C^T}_{s2c} \quad (16)$$

where Q_C and K_S are the projected content vectors and semantic embedding, and subscripts c and s represent content and semantics, respectively. $W_{q,s}$ and $W_{k,s} \in \mathbb{R}^{d_s \times d}$ are the projection matrices.

Geometric-Enhanced Attention: Different from semantic relation, the geometric relation matrix is symmetric. ERSAN conducts a dynamical geometric enhanced attention between content Q_C and geometric relation K_g as:

$$K_g = GW_{k,g}, \varphi(Q_C, G) = Q_C K_g^T \quad (17)$$

Adaptive Re-weight Relation: ER-SAN contains an adaptive learning re-weight module in the attention aggregation operation to get different levels of enhanced semantic and geometric relation information. It introduces the gated fusion module and conducts sigmoid function $\sigma()$ as:

$$V_c = HW_{v,c}, V_s = SW_{v,s}, V_g = GW_{v,g} \quad (18)$$

$$\beta = \sigma(W_\beta [V_c, V_s, V_g]) \quad (19)$$

where $\beta = [\beta^s, \beta^g] \in \mathbb{R}^{N \times 2}$ are learnable parameters. The gating value indicates which feature information is more critical for the current state. The output of enhanced adaptive relation attention is computed as:

$$Z = \text{Softmax}(\tilde{A}) (V_c + \beta^s \times V_s + \beta^g \times V_g) \quad (20)$$

Then, ER-SAN accumulates the final explicit exploration representation followed by an FFN operation as Fig.3.9 (b).

5. Language Prediction Module

After encoding, the model uses the Transformer decoder for caption generation. The inputs of the language prediction module are features from the last encoder layer, which contain aligned relations between region features and words, while the outputs are descriptive sentences. The enhanced adaptive relation attention is not used in the language prediction module because the decoder is autoregressive and inputs variable-length sentences. After the decoding layer, it outputs H^D . Then, the confidence of the word distribution is:

$$P(S) = \text{Softmax}(WH^D + b) \quad (21)$$

After formulating the ultimate softmax layer equation, the subsequent step involves employing beam search [22] to generate the conclusive caption. Beam Search, a heuristic search algorithm, concurrently explores multiple sequences to ascertain the most probable sequence of words. This technique holds particularly significant value in domains such as image captioning, where the objective is to produce coherent and contextually appropriate sentences.

IV. EXPERIMENTS

1. Experimental environment

Dataset: In order to test our proposed framework, we performed thorough experiments using the MS-COCO dataset, which is widely utilized for image captioning. This dataset comprises 123,287 images, each annotated with at least five captions. Based on the Karpathy splits [23], 5,000 images were allocated for validation, 5,000 for testing, and 113,287 for training.

Experimental tools: To develop and test the model's source code, it will be programmed and tested on a robust workstation operating on the 64-bit Linux operating system. The workstation is equipped with an NVIDIA QUADRO RTX 6000 GPU graphics card. For running the proposed Deep Learning model SPNER SAN on the GPU, the workstation also has a compatible version of the CUDA driver that can be seamlessly integrated with the NVIDIA card. This configuration provides ample computational resources for the image processing and prediction tasks.

Experimental Setup and Hyperparameter Selection: The proposed GAT model was meticulously designed to optimize task performance. The input dimension was set to 2048, with the output dimension corresponding to the number of detected objects in the image. The hidden dimension was maintained at 2048 to ensure the model can learn from high-dimensional features. An edge feature dimension of 8 was incorporated to effectively capture the edges' characteristics within the graph structure. The model employed the Adam optimizer, configured with a learning rate of 0.01 and a weight decay of 0.0005 to mitigate the risk of overfitting. Given the relatively small size of the graph, the entire graph was processed in each iteration rather than utilizing mini-batches. This approach allowed for the simultaneous consideration of all nodes and edges, which is feasible due to the manageable graph size. During the initial cross-entropy training phase, the learning rate was linearly increased to $3e-4$ over a warm-up period of 3 epochs, followed by a decay rate of 0.5 every three epochs.

The model underwent training for 50 epochs using cross-entropy loss, ensuring thorough learning and performance optimization.

2. Experimental results

In the experimental part, we designed two experimental scenarios: scenario 1 for the first purpose is to test the effectiveness of the Subgraph proposal network in reducing computational complexity on graphs; Experimental scenario 2 aims to demonstrate the effectiveness of the proposed model through experimental comparison with the original ERSAN model. The experimental results obtained correspond to two scenarios described as follows:

a) Experimental scenario 1: Results using Subgraph Proposal Network

This test aims to illustrate the effectiveness in the case using SPN to reduce the computational complexity of both semantic and geometric graphs that are the output of the Feature extraction module in Figure 1. The experimental results are clearly described in Table I with two cases: using and not using SPN corresponds with the initial and final value of Node/Edge of Graph.

The proposed subgraph-based GAT model demonstrates significant efficiency in reducing the complexity of the graph while retaining relevant information. The average initial number of nodes in the graph is 42, which is reduced to 16 nodes after applying the pruning phase, resulting in a node reduction rate of 61.9%. This indicates that the model is highly effective in identifying and removing non-important nodes, substantially simplifying the graph structure.

Table I
PERFORMANCE RESULTS OF THE PROPOSED
SUBGRAPH-BASED GAT ON MS COCO DATASET

Metric	Value
Average Initial Number of Nodes	42
Average Final Number of Nodes	16
Node Reduction rate	61.9%
Average Initial Number of Edges	63
Average Final Number of Edges	27
Edge Reduction Rate	57.1%

In terms of edges, the average initial number of edges is 63, which is reduced to 27 edges after the pruning phase. This corresponds to an edge reduction rate of 57.1%, showing that the model not only reduces the number of nodes but also significantly decreases the number of edges. This further simplifies the graph, making it easier to process and analyze. These results highlight the effectiveness of the proposed subgraph-based GAT.

Architecture in pruning the graph to retain only the most significant nodes and edges. The model can operate more efficiently by reducing the graph's complexity, improving overall performance, and reducing computational costs. The substantial reduction in both nodes and edges demonstrates the model's ability to focus on the most relevant parts of the graph, thereby enhancing its capability to perform various graph-based tasks on the MS COCO dataset.

b) Experimental scenario 2: Experimental comparison with related models

To demonstrate the effectiveness of the proposed model, we experimentally compare it with related models, specifically the original ERSAN model and the Transformer model. Five evaluation metrics are used to test the quality of image captioning, including BLEU, METEOR (M), ROUGR (R), CIDER (C), and SPICE (S) [24]. The experimental results compared to the measurements are presented in Table II. In addition, we also represent the generated sentences corresponding to the input images in Figure 6. In Figure 6, the underlined words are marked as word/phrase that improves compared to the Transformer model. The bold words represent the enhanced words when comparing the proposed model's sentences with Transformer and ERSAN.

Comparative results with the original ER-SAN model

The results from our experiments, as illustrated in Table II, reveal several critical insights into the performance of the original ER-SAN model on the image captioning task.

Note that the new contribution of the research is to propose adding a subgraph proposal network module in the original ERSAN model to focus on highly important objects in the input image. Therefore, to evaluate the relative contributions of the components in the image encoding and decoding stages of SPNER-SAN, we only assess experimentally on the training model with cross-entropy and without the support of Reinforcement Learning. This is a supervised training process at the word level; the training model will predict the next word based on the previous accurate word.

Table II
COMPARISON RESULTS OF THE PROPOSED MODEL
WITH OTHER MODELS ON THE SAME METRICS

Model	Cross Entropy					
	B@1	B@4	M	R	C	S
Transformer	75.2	34.0	27.6	55.9	113.3	20.8
ER-SAN	78.2	38.8	29.2	58.5	122.9	22.2
SPNER-SAN (ours)	79.3	39.1	30.1	59.2	124.5	22.5

Our proposed SPNER-SAN model shows competitive performance, with a BLEU- 1 score of 79.3 and a BLEU-4 score of 39.1. It achieves a METEOR (M) score of 30.1, indicating good precision and recall in the generated captions. The ROUGE and CIDEr indexes of the SPNER-SAN model



Figure 6. Examples of corresponding captions of the ground-truths captions, Transformer, ERSAN, and proposed SPNER-SAN

are 59.2 and 124.5, respectively, still demonstrating the model's robustness. The SPICE (S) score of 22.5 reflects SPNER-SAN's capability to capture the semantic content of images effectively. This suggests that our enhancements to the ER-SAN baseline have significantly improved its performance, making SPNER-SAN a strong contender in the field. Meanwhile, ER-SAN (Base) [7] presents a solid baseline with BLEU-1 and BLEU-4 scores of 78.2 and 38.8, respectively. Its METEOR score of 29.2 and ROUGE score of 58.5 are somewhat lower than those of SPNER-SAN, suggesting that the enhancements in SPNER-SAN have contributed to improved performance. Specifically, as shown in 6, the bold letters in the sentences generated from our proposed model have shown an improvement in context in the sentence compared to the original ERSAN.

Comparative results with the original Transformer model

The findings indicate that SPNER-SAN outperforms the Transformer (Base) [20] model in various metrics, such as B@1, B@4, and CIDEr scores. Specifically, SPNER-SAN achieves scores of 79.3 and 39.1 in B@1 and B@4, respectively, compared to the Transformer's scores of 75.2 and 34.0. This indicates that SPNER-SAN has a slight edge in predicting individual words and longer sequences, likely due to better handling of context and dependencies in the data. Additionally, SPNER-SAN scores 124.5 in the CIDEr metric, significantly higher than the Transformer's score of 113.3, indicating SPNER-SAN's enhanced capability to generate outputs aligned with human-annotated references.

Furthermore, SPNER-SAN demonstrates superior performance in M, R, and S scores, with scores of 30.1, 59.2, and 22.5, respectively, compared to the Transformer's scores of 27.6, 55.9, and 20.8. These results highlight SPNER-SAN's more robust ability to capture semantic meaning, recall im-

portant information, and understand complex relationships within the data, suggesting that its architecture is better suited for tasks requiring deep semantic comprehension and structural understanding. As shown in 6, the captions generated by the proposed SPNER-SAN illustrate a more accurate semantic relation, e.g., "sand" and "parked". It indicates that SPNER-SAN has a comprehensive understanding of visual relations.

Overall, the results highlight that our proposed SPNER-SAN model performs competitively across various metrics, validating enhancements to the original ER-SAN and Transformer mode. Compared to the original model, the proposed model is slightly higher than the ERSAN model in terms of comparative metrics. Our idea of adding the SPN block is that in many situations, things nearer the present object should be given greater importance than objects farther away, establishing a contrast in priority in the context image. Furthermore, While the performance gain may appear modest, it is important to note that our SPN approach significantly reduces the number of nodes (by 61.9%) while still improving the overall performance. This result underscores the efficiency of our method, as we are able to maintain or increase model accuracy even after discarding a substantial amount of data. The added SPN improves the baseline model's ability to focus on the most relevant aspects of the graph, which is particularly beneficial in situations where computational resources are limited or real-time processing is required. The comparative analysis of these models underscores the importance of architectural advancements and optimization strategies in achieving superior performance in image captioning tasks.

V. CONCLUSION

In this work, we introduced SPNERSAN, a novel hybrid architecture that synergizes the Enhanced Relational Self-Attention Network (ER-SAN), which follows a Transformer-like structure, with our newly developed Subgraph Proposal Network (SPN). The primary objective of this combination is to enhance the image captioning task by leveraging the unique capabilities of both components. ERSAN's structure effectively handles complex dependencies and relationships within the image data. At the same time, SPN excels in proposing meaningful subgraphs through a graph attention network, thus providing enriched context for caption generation.

The ER-SAN component of SPNERSAN focuses on capturing intricate and high-level features from images. Its multi-head self-attention mechanism enables the model to weigh the importance of different image parts, facilitating a comprehensive understanding. This mechanism is particularly beneficial for image captioning, where recognizing

fine-grained details and their interrelationships is crucial for generating accurate and descriptive captions. By capturing these intricate relationships, ER-SAN ensures the generated captions are contextually rich and semantically coherent.

On the other hand, the Subgraph Proposal Network (SPN) within SPNERSAN introduces an innovative approach to identify and propose significant subgraphs within the image. By employing a graph attention network, SPN effectively highlights salient regions and objects, providing a more focused context for the captioning model. This method ensures that the generated captions are contextually rich and relevant to the critical elements within the image. The graph attention network allows SPN to dynamically adjust its focus on different parts of the image, enhancing its ability to generate more meaningful subgraphs that contribute to more accurate and detailed captions.

Our experiments demonstrate that the SPNERSAN model significantly improves the quality of generated captions compared to existing related methods. The proposed model performs better in standard metrics such as BLEU, METEOR, ROUGR, CIDER, and SPICE, indicating its effectiveness in producing coherent and contextually appropriate captions. Moreover, qualitative analysis shows that SPNER-SAN can capture subtle details and generate more human-like descriptions, underscoring its practical applicability in real-world scenarios. The ability to create captions that closely resemble human descriptions highlights the potential of SPNERSAN in various applications, from automated image tagging to assistive technologies.

ACKNOWLEDGEMENT

This work has been supported by Vietnam National University, Hanoi, under the Project, code: QG.23.66

REFERENCES

- [1] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep learning approaches on image captioning: A review," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–39, 2023.
- [2] A. Selivanov, O. Y. Rogov, D. Chesakov, A. Shelmanov, I. Fedulova, and D. V. Dylov, "Medical image captioning via generative pretrained transformers," *Scientific Reports*, vol. 13, no. 1, p. 4171, 2023.
- [3] M. Shoman, D. Wang, A. Aboah, and M. Abdel-Aty, "Enhancing traffic safety with parallel dense video captioning for end-to-end event analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7125–7133.
- [4] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10971–10980.
- [5] K. Safiya and R. Pandian, "A real-time image captioning framework using computer vision to help the visually impaired," *Multimedia Tools and Applications*, vol. 83, no. 20, pp. 59413–59438, 2024.

- [6] Z. Song, X. Zhou, L. Dong, J. Tan, and L. Guo, "Direction relation transformer for image captioning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5056–5064.
- [7] J. Li, Z. Mao, S. Fang, and H. Li, "Er-san: Enhanced-adaptive relation self-attention network for image captioning," in *IJCAI*, 2022, pp. 1081–1087.
- [8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.
- [9] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.
- [10] H. Wu, Y. Liu, H. Cai, and S. He, "Learning transferable perturbations for image captioning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 2, pp. 1–18, 2022.
- [11] M. Yuan, B.-K. Bao, Z. Tan, and C. Xu, "Adaptive text denoising network for image caption editing," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 1s, pp. 1–18, 2023.
- [12] S. Dubey, F. Olimov, M. A. Rafique, J. Kim, and M. Jeon, "Label-attention transformer with geometrically coherent objects for image captioning," *Information Sciences*, vol. 623, pp. 812–831, 2023.
- [13] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.
- [14] X. Yang, H. Zhang, and J. Cai, "Learning to collocate neural modules for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4250–4260.
- [15] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving image captioning with better use of captions," *arXiv preprint arXiv:2006.11807*, 2020.
- [16] A.-A. Liu, Y. Zhai, N. Xu, W. Nie, W. Li, and Y. Zhang, "Region-aware image captioning via interaction learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3685–3696, 2021.
- [17] A. Abedi, H. Karshenas, and P. Adibi, "Multi-modal reward for visual relationships-based image captioning," *arXiv preprint arXiv:2303.10766*, 2023.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [19] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio *et al.*, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48 550, 2017.
- [20] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.
- [21] X. Li, G. Zhang, Y. Wu, X. Li, and Y. Zhang, "Rule of thirds-aware reinforcement learning for image aesthetic cropping," *The Visual Computer*, vol. 39, no. 11, pp. 5651–5667, 2023.
- [22] M. Freitag and Y. Al-Onaizan, "Beam search strategies for neural machine translation," *arXiv preprint arXiv:1702.01806*, 2017.
- [23] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.
- [24] J. Li, Z. Mao, H. Li, W. Chen, and Y. Zhang, "Exploring visual relationships via transformer-based graphs for enhanced

image captioning," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 5, pp. 1–23, 2024.



TRAN THI HUE receive the degree in Computer Science from the Military Technical Academy in 2008. She works at Dai Nam University, Hanoi. Her research interests include artificial intelligence, image processing, machine learning, etc
Email: tranthihue@gmail.com



Email: lethihads468@gmail.com

LE THI HA receive the degree in Computer Science from Hanoi University of Industry in 2011. She works at Hanoi College of Economics and Industry and studied for a master's degree at Hanoi University of Industry. Her research interests include artificial intelligence, image processing, machine learning, soft computing, etc



Do Hoang Long has been graduated from Hanoi University of Science and Technology in 2024. He works at FPT Software and is pursuing a Master's degree at the same university. His research focuses on Data Mining, Fuzzy Computing, and Artificial Intelligence.
Email: dhlng2301@gmail.com



ing, Rule-based, and Fuzzy Systems.
Email: haipv@soict.hust.edu.vn

PHAM VAN HAI receive a Doctor of Engineering degree (Ph.D.) at Ritsumeikan University (Japan). He is an Associate Professor at the School of Information and Communication Technology, Hanoi University of Science and Technology. His major fields include Artificial Intelligence, Knowledge-Based, Big Data, Soft Comput-



ing, Rule-based, and Fuzzy Systems.
Email: lanlhbkc@hau.edu.vn

LUONG THI HONG LAN has been graduated from Hanoi University of Technology in 2002 and received a PhD. Degree in Computer Science in Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam, in 2021. Now, she is working at the Faculty of Information Technology at Hanoi