

A Multi-Caption Fusion Approach for Enhanced Image Captioning

Alaa Thobhani

althobhanialaa@gmail.com

Hunan University of Traditional Chinese Medicine

Yan Junfeng

Hunan University of Traditional Chinese Medicine

Mohanned Abduljabbar Hael

Zhaoqing University

Amr Abdussalam

University of Science and Technology of China

Muhammad Asim

Prince Sultan University

Gauhar Ali

Prince Sultan University

Sajid Shah

Prince Sultan University

Research Article

Keywords: Image Captioning, Multi-Caption Fusion, Multimodal Fusion, Explainable Image Captioning, Image-Number Vectors, Generated Caption Vectors

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11237647/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Multi-Caption Fusion Approach for Enhanced Image Captioning

Abstract

Generating diverse captions for an image enriches descriptions and fosters creative expressions, accommodating varied interpretations that can be performed through various methods. Leveraging the plentiful semantic information in these diverse descriptions for an image could be useful for producing a richer and more interpretable ultimate description of the image. To harness this abundance of information, we propose a two-phased, transparent multi-caption fusion approach with an encoder to bolster the image captioning network, offering an innovative method for enhanced caption generation. In the initial phase, multiple captions are generated for each image using the Image-Number Vectors (INVs). The second phase involves an extra encoder processing generated captions from the first phase, alongside a decoder to produce the final caption for further performance enhancement. The resulting caption vectors, combined with attended visual features, are utilized to generate a comprehensive caption describing the entire image. By exploiting the semantic information within the generated sentences from phase one, the image captioning model's ability to generate high-quality captions is improved while providing a more interpretable fusion process. Experimental evaluations on the MS-COCO dataset demonstrate the potential improvement in image captioning models through our novel approach. Our framework performs competitively with state-of-the-art image captioning models, highlighting its effectiveness for responsible and explainable multimodal fusion.

Keywords: Image Captioning, Multi-Caption Fusion, Multimodal Fusion, Explainable Image Captioning, Image-Number Vectors, Generated Caption Vectors

1 Introduction

image captioning [1–3], which is the process of automatically producing a description of an image in natural language, has become a topic of great interest in computer vision. This is not only due to its practical applications such as aiding visually impaired individuals, but also because it presents a significant challenge for image comprehension, a fundamental issue in computer vision. Generating a meaningful description of an image in natural language requires a deeper level of image comprehension beyond mere object detection and classification. Furthermore, it is noteworthy that this problem links computer vision and natural language processing, which are two of the major

fields in artificial intelligence. Several recent works have adopted the conventional encoder-decoder framework, which has demonstrated notable success in the domain of image captioning [4–6]. This method entails the utilization of a language model that accepts the features of image as input and subsequently generates a descriptive caption illustrating contents of the image. However, the predominant paradigm in current image captioning methodologies typically revolves around generating a lone sentence to encapsulate the entire visual content. Nevertheless, this approach may prove insufficient, particularly for images characterized by intricate and detailed features.

Generally, it is common for image captioning models to generate only one sentence per image, and this generated sentence should cover input image details. In reality, though, it is difficult to describe the entire content of an image with one generated sentence, mainly when the image includes rich details. Alternatively, some previous works, such as [7–9], developed models that can generate multiple captions for an image, each covering a different aspect. Typically, the image captioning benchmark datasets usually provide several ground truth captions per image. These ground truth captions describe the image differently and cover different aspects of its content. However, in many cases, there remains a demand for a single and informative caption with a rich description. While diversity is valuable in certain contexts, the simplicity and clarity of a singular, high-quality caption often better serve the user’s need for a clear and concise understanding of the image content. Balancing the diversity of describing an image with the coherence of a single and informative caption is a significant challenge in image captioning. One feasible way to achieve this goal is by dividing the captioning process into two phases where in one phase multiple descriptions could be produced for an image while in the second phase the information within these generated captions can be fused and a new final singular description will be produced for the image. This way, the diversity represented by the rich image contents could be obtained using the multiple generated captions in the first phase while the coherence of the single final caption is achieved by fusing and leveraging the semantic information found on these generated captions.

To address the aforementioned challenges, we present a novel Two-Phase Multi-Sentence model with an encoder, featuring two distinct phases. This model is given the name Two-Phase Multi-Caption (TPMC) image captioning model. In the first phase, our approach focuses on diversifying and enriching image descriptions by generating multiple captions. This is achieved through adopting the Image-Number Vectors (INVs) [9], generating numerous descriptions for a single image using numeric representations. Moving to the second phase, we explore the integration of the diverse generated captions for enhanced performance. An additional encoder is introduced in this phase, responsible for converting the generated captions from phase one into encoded caption vectors. These vectors serve as input to the decoding part of the second phase, drawing inspiration from machine translation systems [10]. In the decoder of phase two, our goal is to produce a singular, detailed caption that offers a comprehensive description of the entire image. To achieve this, we leverage insights gathered from all generated captions in the preceding training phase, coupled with attention to the visual features. This approach ensures that the generated caption encapsulates

a nuanced understanding of the image, combining textual and visual cues for a more detailed and accurate description.

This work makes substantial contributions in several key areas, which can be summarized as follows:

- We introduce a novel framework, the Two-Phase Multi-Sentence framework with a caption encoder, designed to optimize image caption generation through multi-caption fusion.
- We propose an image captioning framework consisting of two phases, where INVs are utilized to produce multiple diverse captions for an image in the first phase, while these generated captions are fused and leveraged as input in the second phase, enabling interpretable visual-linguistic alignment.
- We encode the generated multiple captions from phase one into vectors and fuse them using various mechanisms to exploit their semantic representations efficiently for generating singular informative final descriptions.
- Our proposed framework underwent testing on the MS-COCO dataset. The achieved scores underscore that the proposed approach is comparable to recent state-of-the-art methods based on evaluation metrics.

2 Related Work

Several studies have investigated multi-phase learning for image captioning. Gu et al. [11] proposed a coarse-to-fine framework using stacked visual attention and multiple LSTMs, where captions are progressively refined across stages. However, the second phase starts only after the first phase is completed, making the learning process strictly sequential. Ling and Fidler [12] incorporated human natural-language feedback into reinforcement learning through a hierarchical phrase-based RNN. Although effective, this approach depends on costly human feedback, which can limit efficient model optimization. Similarly, Zhou et al. [13] introduced a two-phase saliency-enhanced re-captioning model that transfers visual, semantic, and sample saliency from the first stage to improve image representation in the second. However, it mainly performs self-improvement within the same model and uses only one caption from the first phase. In contrast, our method utilizes multiple diverse captions generated in the first phase. These captions provide complementary descriptions of the image, offering richer contextual information for the subsequent phase and improving caption generation.

Image captioning faces the challenge of crafting a single, comprehensive caption that encapsulates the intricacies of a complex image. To address this, certain techniques opt for a more inclusive strategy by generating multiple descriptions for a single image. One innovative approach employs Conditional Generative Adversarial Networks (GANs), utilizing a random vector and the input image to guide the caption generation [7]. This introduces diversity in captions, capturing different facets of the image. The model includes an evaluator, assessing caption quality and promoting contextually relevant and high-quality descriptions. Another approach leverages topics to generate multiple captions related to a specified topic [8]. A fusion gate unit integrates topic embeddings into the LSTM network, ensuring thematic consistency. In text-based image captioning, the Anchor-Captioner approach excels in fine-grained

description tasks by generating multiple captions from various viewpoints [14]. Lastly, a numbering-based approach considers the number of available ground truth captions during training, utilizing quantitative availability to create a varied set of captions for images [9].

Attention mechanisms have significantly advanced image captioning [15–17]. Adaptive attention allows models to decide on language or visual features [18]. Other approaches combine bottom-up and top-down visual attention, integrating object proposals [19]. Further enhancements included spatio-temporal memory mechanisms in visual attention (STMA) [20], a task-adaptive attention module [21], and dual attention for pyramid image feature mapping [22]. Innovations like cluster-based grounding networks [23], Proposal Attention Correctness (PAC) metric [23], multimodal Transformer [24], region-aware interaction learning [25], and methods from [26–28] have introduced diverse improvements, challenging traditional interpretations and enhancing various aspects of attention mechanisms in image captioning. [29] presents a Hybrid Attention Network (HAN) merging mechanisms to address issues like "object hallucination". [30] introduces attention mechanisms (LSA and LSF) to enhance local visual modeling. [31] presents a refined visual attention (RVA) framework internally reweighting visual attention. [6] propose the Guided Visual Attention (GVA) method, which improves caption quality by adjusting attentional weights. Meanwhile, [32] presents JRAN, aiming to enhance caption coherence by incorporating both region and semantic features. However, these approaches encounter challenges with comprehensively capturing global context, exhibiting bias towards local information, and lacking robustness to variations in object scale, position, and occlusions.

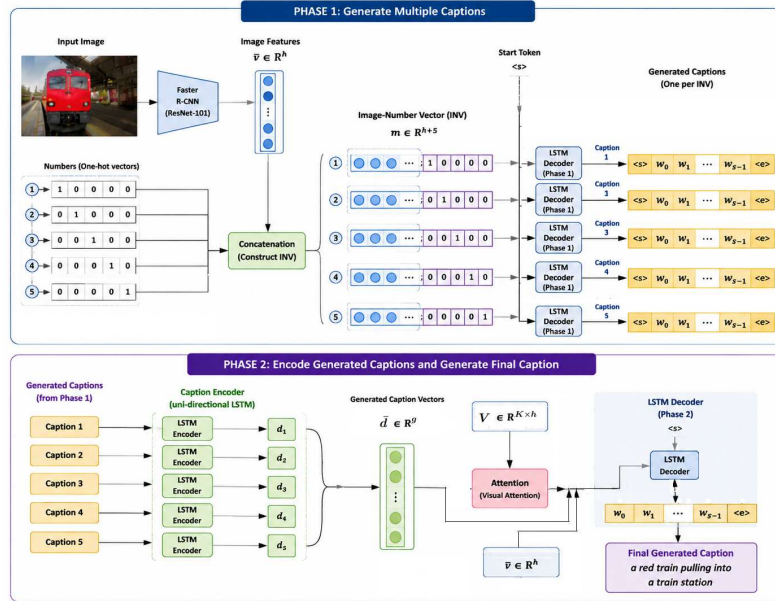


Fig. 1 Description of main concept and pipeline of our proposed method.

Semantic information enriches captions in image captioning [33–35]. Attribute-based methods predict attributes from images, improving synergy between visual and semantic elements [34]. Stack-VS utilizes multi-stage descriptors to optimize attention weights for detailed descriptions [36]. FUSECAP integrates visual insights and a large language model, benefiting image captioning and retrieval [37]. Part-of-speech (PoS) integration controls information flow [38], combining PoS with word embeddings to influence caption generation [38–40]. Topics-based image captioning integrates topics with captioning algorithms, enhancing sentence generation through topic detection and re-weighting [5, 41]. The two-phase learning re-captioning model utilizes saliency mechanisms like visual, semantic, and sample saliency to refine image representations and improve caption quality [13]. These approaches contribute to the evolution of image description generation techniques. However, challenges remain in semantic attention, including inadequate coverage of pertinent image attributes and difficulties in capturing diverse topics without bias from saliency mechanisms.

In contrast to other models that produce a single sentence per image, which may not fully capture all the details, our model has the ability to generate multiple captions for an image in the first phase, each focusing on a different aspect. Subsequently, we input these generated captions into a separate language network to serve as input. This approach is designed to offer more informative descriptions and leverage the diverse range of image descriptions in order to create a single-sentence description that is enriched with comprehensive information.

3 Methodology

Figure 1 illustrates the overall captioning process, comprising two phases. In the first phase, multiple captions are generated for each image. The second phase involves an additional uni-directional LSTM layer acting as an encoder to produce final encoded vectors for the captions generated in phase one. These encoder vectors, along with the attended image features, are inputted into the decoder of phase two to generate the final caption, thereby enhancing the performance of the image captioning models.

3.1 Phase One

Figure 2 illustrates the first phase of the proposed framework, which is designed as an extension of the conventional encoder–decoder model. The visual content of each image is first encoded using Faster R-CNN [19] with a ResNet-101 backbone to obtain image features. These features are then enriched with an additional numerical signal during training.

More specifically, the extracted image features are concatenated with a one-hot number vector. Since each image in the dataset is associated with five ground-truth captions, five distinct number vectors are constructed, with each one linked to a different reference caption. By combining the same image features with each of these number vectors, five different Image-Number Vectors (INVs) are obtained for every image.

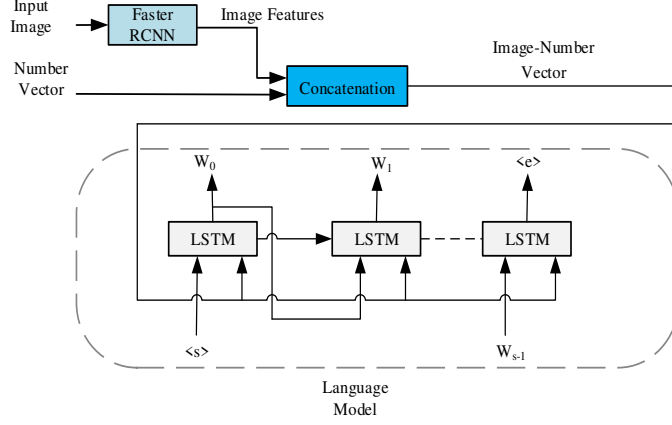


Fig. 2 Description of the phase one of our algorithm methodology, distinguished by the special starting word [s] and the special ending word [e] , while W_0 is the first word in the generated caption and W_{s-1} is the last word of the generated caption.

The resulting INVs are used as inputs to the LSTM decoder. For each INV, the decoder is trained to generate the caption associated with that specific number vector. Through this design, the captioning model can produce diverse descriptions for the same image by varying the numerical part of the input while keeping the visual features unchanged. In this paper, the INV is denoted as $m \in \mathbb{R}^{h+5}$, representing the concatenation of the image feature vector and the one-hot number vector. The formulation of INV is given as follows:

$$m = [n; \bar{v}] \quad (1)$$

A unique INV is constructed for every ground-truth caption in the dataset by combining two components: the global image feature vector $\bar{v} \in \mathbb{R}^h$, extracted using Faster R-CNN, and the number vector $n \in \mathbb{R}^5$, represented in one-hot form.

To generate image descriptions in the first phase, the model employs an LSTM-based decoder. As illustrated in Figure 3, the decoder receives the INV as input and produces the corresponding caption sequence according to the following equations:

$$h_t^{Ph1} = LSTM_{Ph1}(h_{t-1}^{Ph1}; [m, E_1 \cdot y_{t-1}]) \quad (2)$$

$$p_t = \text{softmax}(W_a \cdot h_t^{Ph1}) \quad (3)$$

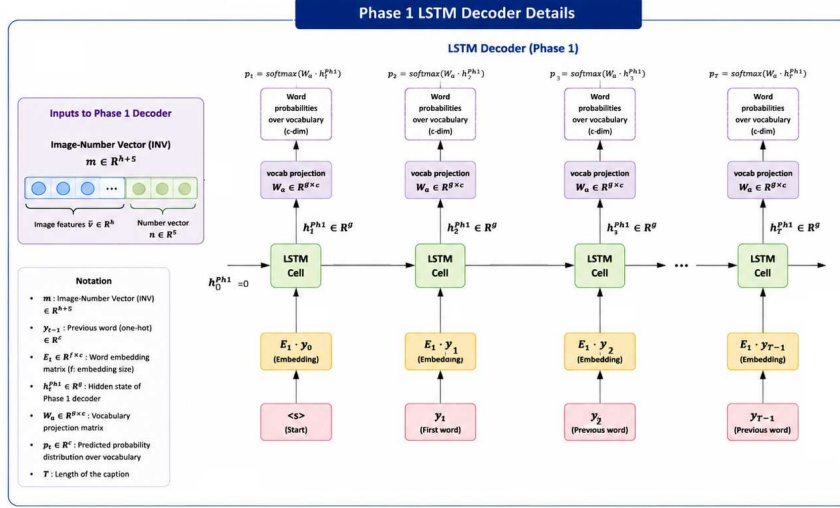


Fig. 3 Description of the LSTM component in phase one of the proposed model

The hidden representation of the first-phase decoder is denoted by $h_t^{Ph1} \in \mathbb{R}^g$, where $LSTM_{Ph1}$ models the caption sequence over time. Word representations are obtained from the embedding matrix $E_1 \in \mathbb{R}^{f \times c}$. At each time step, the model predicts $p_t \in \mathbb{R}^c$, which gives the probability distribution over the full vocabulary, based in part on the previously generated word $y_{t-1} \in \mathbb{R}^c$. The matrix $W_a \in \mathbb{R}^{g \times c}$ contains the trainable parameters used for the vocabulary projection. Model training is performed with the standard cross-entropy (XE) objective, computed between the predicted distribution and the ground-truth caption, as defined below:

$$\text{Loss}_{Ph1}^{XE} = \frac{1}{T} \sum_{t=1}^T -\log(p_t(y_t | y_{t-1}, m)) \quad (4)$$

During training, image features extracted by Faster R-CNN are combined with Image-Number Vectors (INVs) to generate multiple captions in Phase 1, while Phase 2 encodes these captions and incorporates visual attention to generate the final caption. During inference, the learned model constructs INVs from the test image, generates multiple candidate captions, and applies caption encoding with visual attention to produce the final descriptive caption. As illustrated in Fig. 4, the proposed framework consists of two stages: a training phase and an inference phase.

3.2 Phase Two

After training the model in phase one, we generate five captions for each image in the dataset, with one caption corresponding to each INV. These generated captions serve as input to the model in phase two, where it operates both as an encoder and a decoder. In the encoder component, a uni-directional LSTM is employed to encode the generated captions from phase one into vectors, which are then utilized as input

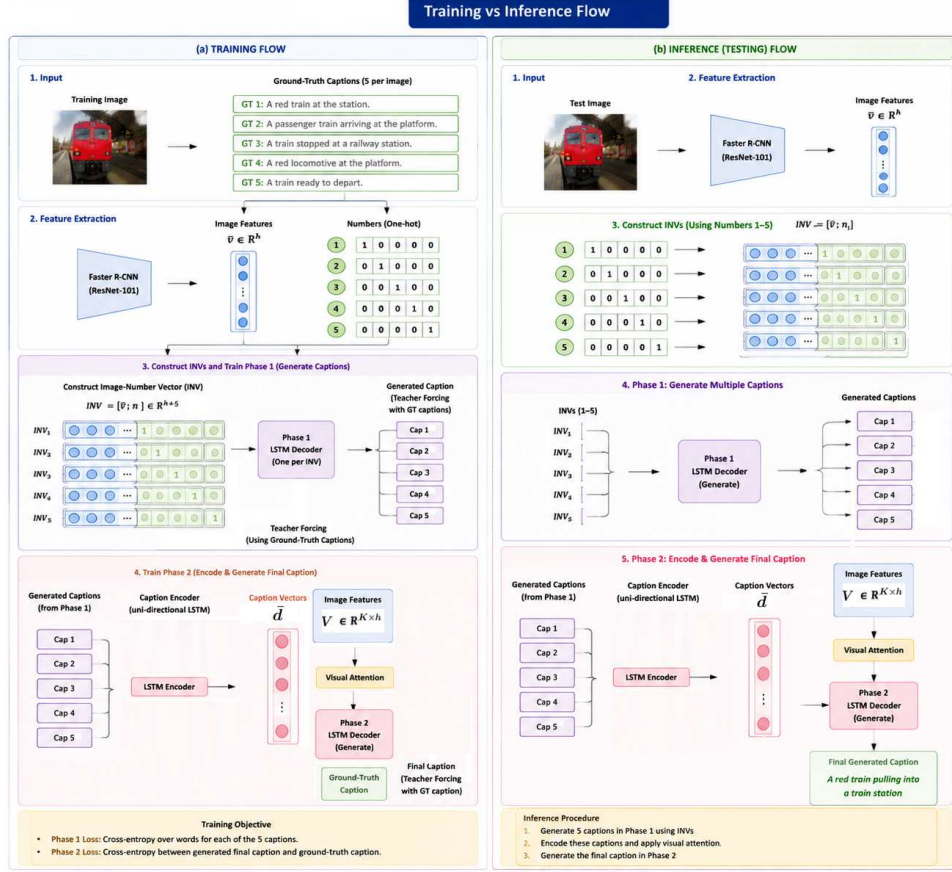


Fig. 4 Overview of the proposed training and inference framework.

in the decoder component. This process results in the generation of a final caption that comprehensively describes most aspects of the input image.

3.2.1 LSTM Encoder

In the encoder part, the five generated captions are sequentially encoded using an additional LSTM layer, as illustrated in Figure 5. We refer to each of generated captions as $C_i = \{c_0^i, c_1^i, \dots, c_{T'}^i\}$ where $i \in \{0, 1, \dots, n-1\}$ and T' represents the maximum caption's length. These generated captions are fed into the encoding LSTM, denoted as $LSTM_E$. The hidden state of the encoding LSTM $h_t^i \in \mathbb{R}^g$ at time step t is defined as follows:

$$h_t^i = LSTM_E(h_{t-1}^i; [E_2 \cdot c_t^i]) \quad (5)$$

where $E_2 \in \mathbb{R}^{c \times f}$ is a word embedding matrix, $c_t^i \in \mathbb{R}^c$ is the input word at time step t , and $t \in \{0, 1, \dots, T'\}$. The initial hidden state of the encoding LSTM is set to zero. Thereafter, the hidden state of the encoding LSTM at the final time step ($t = T'$) is

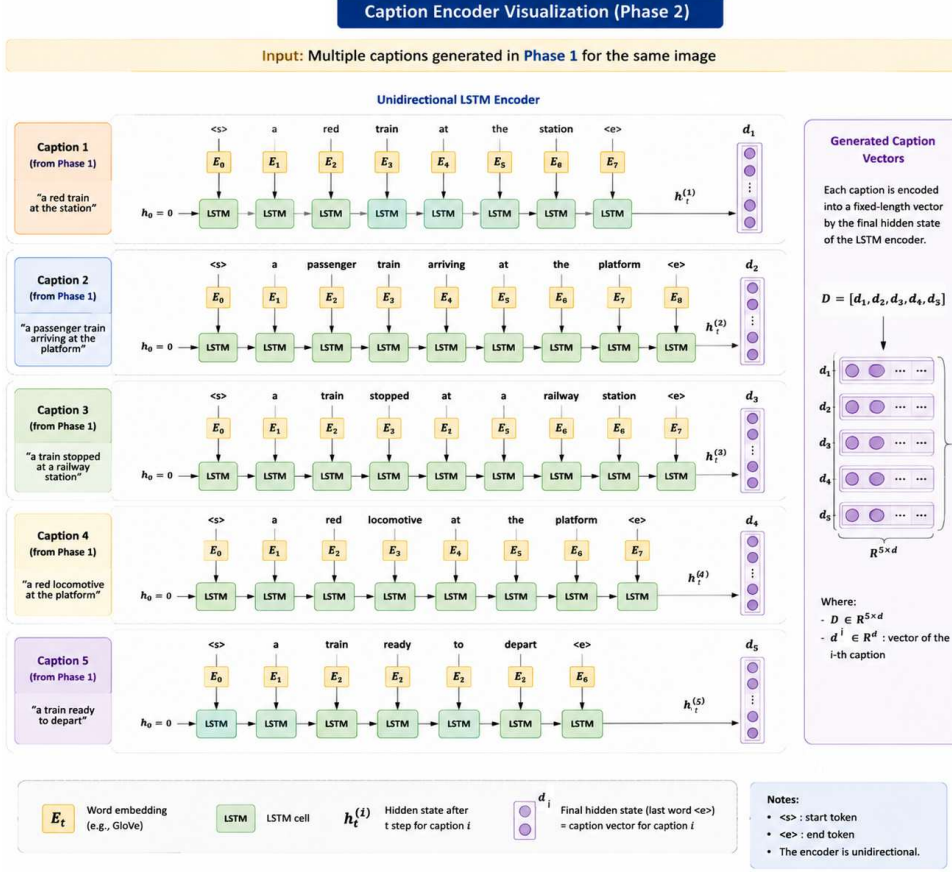


Fig. 5 Description of the LSTM encoder component in the second phase of the proposed model.

used as the encoding vector for the generated caption. That is, $h_{t=T}^i$ represents the encoded vector of the corresponding generated caption C_i . For simplicity, we refer to the encoded vector as d_i as follows:

$$d_i = h_{t=T}^i \quad (6)$$

where $d_i \in \mathbb{R}^g$. The encoded vectors can be arranged in a single matrix named the generated vectors matrix $D \in \mathbb{R}^{g \times n}$ as follows:

$$D = \{d_0, d_1, \dots, d_{n-1}\} \quad (7)$$

where n is the number of generated captions defined in equation (1) and is selected to be five.

3.2.2 Decoder

After obtaining the generated vectors matrix D , we take the mean of its vectors and use as input to the language model.

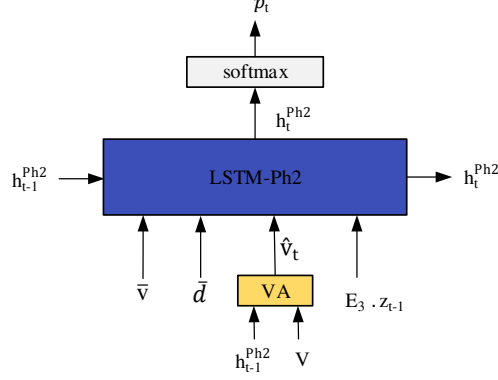


Fig. 6 Explanation of the LSTM decoder component in the second phase of the proposed model.

$$\bar{d} = \frac{1}{n} \sum_{i=0}^{n-1} d_i \quad (8)$$

where $\bar{d} \in \mathbb{R}^g$.

Local object features are extracted from the input image using Faster R-CNN with a ResNet-101 backbone. These features are then arranged into an object feature matrix V , which contains K object-level feature vectors and is expressed as follows:

$$V = \{v_1, v_2, \dots, v_K\} \quad (9)$$

The matrix $V \in \mathbb{R}^{K \times h}$ is formed by K object feature vectors, where each vector $v_i \in \mathbb{R}^h$ represents the feature embedding of an individual detected object for $i = 1, \dots, K$. From these object-level representations, the global image feature $\bar{v} \in \mathbb{R}^h$ is computed as follows:

$$\bar{v} = \frac{1}{K} \sum_{i=1}^K v_i \quad (10)$$

To capture the most informative local regions, the model further incorporates a visual attention mechanism over the object feature matrix V . The corresponding attention computations are defined as follows:

$$\alpha_t^i = W_3 \cdot \tanh(W_1 \cdot h_{t-1}^{Ph2} + W_2 \cdot v_i) \quad (11)$$

$$\beta_t = \text{softmax}(\alpha_t) \quad (12)$$

$$\hat{v}_t = \sum_{i=1}^K \beta_t^i \odot v_i \quad (13)$$

where $\hat{v}_t \in \mathbb{R}^h$, $\beta_t \in \mathbb{R}^K$, and $\alpha_t \in \mathbb{R}^K$. $W_3 \in \mathbb{R}^e$, $W_1 \in \mathbb{R}^{g \times e}$, and $W_2 \in \mathbb{R}^{h \times e}$ are trainable weights.

The hidden state $h_t^{Ph2} \in \mathbb{R}^g$ of the decoder LSTM $LSTM_{Ph2}$ at time step t is given by:

$$h_t^{Ph2} = \text{LSTM}_{Ph2}(h_{t-1}^{Ph2}; [E_3 \cdot z_{t-1}, \bar{d}, \bar{v}, \bar{v}]) \quad (14)$$

where $E_3 \in \mathbb{R}^{c \times f}$ is a word embedding matrix and $z_{t-1} \in \mathbb{R}^c$ is the word generated in the previous time step. Next, h_t^{Ph2} is fed into a linear layer succeeded by a softmax layer for producing the probability distribution over the vocabulary words as follows:

$$p_t = \text{softmax}(W_p \cdot h_t^{Ph2}) \quad (15)$$

where $p_t \in \mathbb{R}^c$ represents the probability distribution along all the words in the vocabulary and $W_p \in \mathbb{R}^{g \times c}$ refers to trainable parameters. For training our model, we adopt the traditional cross-entropy (XE) loss between the predicted word distribution and the ground truth caption as follows:

$$\text{Loss}_{Ph2}^{XE} = \frac{1}{T} \sum_{t=1}^T -\log(p_t(z_t | z_{t-1}, V)) \quad (16)$$

In the decoder part of phase two, illustrated in Figure 6, we utilize the vectors of the five generated captions as input for our framework. Additionally, we input the attended visual features and the word embeddings to generate a single final caption with the most details, describing the entire image by leveraging all generated captions from the preceding training phase. Consequently, our model is capable of producing a single caption for a single image, thereby enhancing its performance in standard evaluation criteria.

4 Experiments and Results

In this segment, we delve into key elements of the conducted experiments, covering evaluation criteria, the dataset employed, and model configurations. Furthermore, we offer an examination of experimental findings and comparisons between networks.

4.1 Datasets and Evaluation Metrics

The MS-COCO dataset is widely recognized as the benchmark for image captioning, with numerous previous studies leveraging its extensive annotations [42, 43]. Comprising 123,000 images, each annotated with five reference sentences, MS-COCO sets the standard in the field. To evaluate our framework, we conducted experiments on this dataset, following Karpathy’s data split [44], which divides it into training, validation, and testing sets of 113,000, 5,000, and 5,000 images, respectively. We utilize a beam search strategy with a beam size of 3 during inference and assess our model’s performance using standard metrics, including BLEU, CIDEr, SPICE, METEOR, and ROUGE-L, abbreviated as B-n (n=1, 2, 3, 4), C, S, M, and R, respectively.

4.2 Image Captioning Model

For visual representation, object-level features are extracted using Faster R-CNN with a ResNet-101 backbone. Each image is represented by $K = 36$ regions, with every region described by a 2048-dimensional feature vector ($h = 2048$). For MS-COCO, words appearing more than five times are retained, resulting in a vocabulary

Table 1 The cross-entropy (XE) scores achieved by our models on the MS-COCO dataset.

MODEL	B1	B2	B3	B4	M	R	C	S
Baseline1	73.3	56.7	43.0	32.8	25.8	54.3	101.1	18.5
TPMC(GRU)	73.7	56.9	43.2	32.9	25.9	54.4	101.5	18.6
TPMC(LSTM)	73.6	57.0	43.4	33.2	26.0	54.4	101.9	18.7
Baseline2	76.7	60.7	47.1	36.4	27.6	56.6	114.1	20.5
TPMC(GRU+VA)	77.0	61.6	47.4	36.5	27.7	56.7	114.6	20.7
TPMC(LSTM+VA)	77.2	61.3	47.7	36.9	27.8	56.9	115.5	20.8

of $c = 9487$ tokens. Both the word embedding dimension f and the LSTM hidden-state dimension g are fixed at 1000, while the attention hidden dimension is set to $e = 512$. Since each image contains five reference captions, the number of number-vectors is chosen as $n = 5$.

The TPMC model is optimized for 50 epochs using Adam and cross-entropy loss in both training phases. The initial learning rate is 5×10^{-4} and is multiplied by 0.8 after every five epochs. Training uses a batch size of 40, dropout of 0.5, and gradient clipping with a maximum absolute value of 0.1. Scheduled sampling is raised by 5% every five epochs until it reaches 25%. During inference, captions are generated through beam search with a beam width of 3. All experiments are implemented in PyTorch.

4.3 Preparing Image-Number Vectors for the First Phase

The image-number vector comprises image global features and a number-vector. The image global features are derived from the Faster RCNN based on ResNet-101 architecture, while number-vectors are encoded using one-hot encoding. For each image, which is associated with five ground truth captions, we create corresponding number-vectors using the numbers from 0 to 4. The assignment of numbers to the ground truth captions can be done arbitrarily in a one-to-one manner. To generate the image-number vectors (INVs), we create five duplicates of the image global features and concatenate each duplicate with a number-vector. This results in five INVs per image, serving as distinct clones of the image features, with each clone having a unique ground truth description.

4.4 Quantitative Evaluation

Table 1 displays the MS-COCO evaluation results. Baseline1 represents the results of phase one only, obtained using the highest beam search score strategy as introduced in [9] for evaluating multiple generated captions. The TPMC(LSTM) model, which is the same model as phase two but without visual attention, surpasses Baseline1 in all evaluation metrics by a noticeable margin. Additionally, TPMC(LSTM+VA) outperforms Baseline2 by a significant margin in all metrics. Baseline2 is identical to the phase two model but without the use of generated captions from phase one and without the encoder LSTM in phase two.

The TPMC(GRU) represents another version of our proposed model, utilizing GRU units instead of LSTM in the encoder part and excluding visual attention mechanisms.

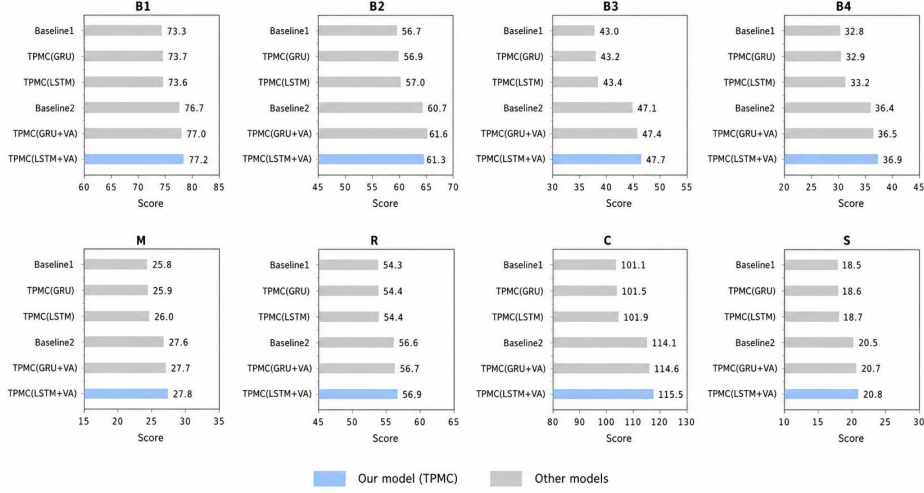


Fig. 7 Cross-entropy scores of the baseline and TPMC-based models on the MS-COCO dataset across different evaluation metrics.

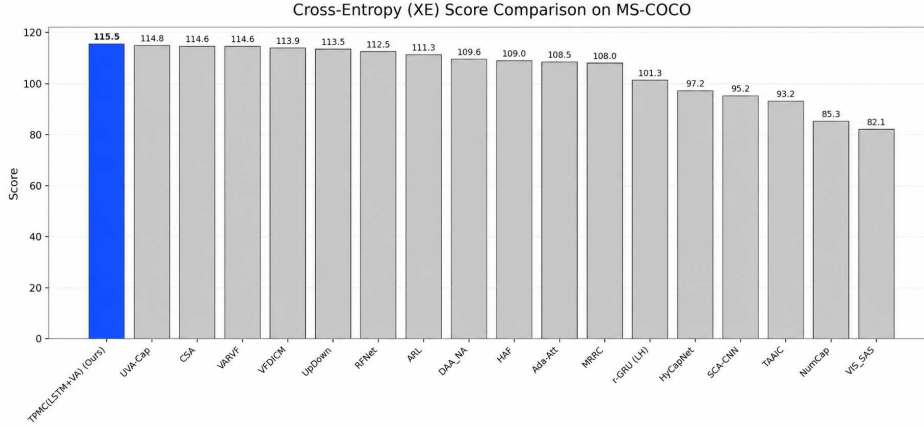


Fig. 8 Comparison of cross-entropy (XE) scores of TPMC(LSTM+VA) with existing image captioning methods on the MS-COCO dataset.

In contrast, TPMC(GRU+VA) integrates visual attention. As illustrated in Table 1, the TPMC(GRU) model consistently outperforms Baseline1 across all evaluation metrics, showcasing a noticeable performance gap. Additionally, the results highlight that TPMC(GRU+VA) surpasses Baseline2 by a considerable margin across all metrics, as shown in Fig. 7.

In this research, our primary goal is to demonstrate the impact of utilizing generated caption vectors to enhance the performance of an image captioning model. Our approach is based on the conventional encoder-decoder architecture, which is simpler than some existing models, allowing us to focus on the contribution of generated

Table 2 Comparison of performance percentages between our models and others on MS-COCO data trained using cross-entropy (XE) optimization. The highest and second-highest scores are highlighted in bold and underline, respectively.

MODEL	B1	B2	B3	B4	M	R	C	S
SCA-CNN [45]	71.9	54.8	41.1	31.1	25.0	53.1	95.2	-
Ada-Att [18]	74.2	58.0	43.9	33.2	26.6	54.9	108.5	-
RFNet [46]	76.4	60.4	46.6	35.8	27.4	56.5	112.5	20.5
UpDown [19]	77.2	-	-	36.2	27.0	56.4	113.5	20.3
VIS.SAS [13]	72.5	52.6	38.2	28.1	23.7	55.4	82.1	-
ARL [47]	75.9	60.3	46.5	35.8	27.8	56.4	111.3	-
HAF [48]	75.9	59.5	45.4	34.4	26.8	-	109.0	-
MRRC [49]	75.5	59.8	46.0	35.2	26.5	55.9	108.0	19.7
DAA.NA [50]	76.2	60.0	46.0	35.0	26.7	55.9	109.6	20.0
TAAIC [21]	71.0	-	-	27.7	23.8	51.1	93.2	18.3
r-GRU (LH) [51]	76.8	60.7	45.9	35.1	29.2	54.9	101.3	-
CSA [52]	77.2	59.8	46.0	36.2	<u>27.9</u>	56.4	114.6	-
NumCap [9]	66.9	49.4	36.5	27.3	24.1	50.7	85.3	17.0
VFDICM [53]	76.4	60.4	46.9	36.3	27.7	56.6	113.9	<u>20.6</u>
VARVF [54]	<u>76.9</u>	<u>60.9</u>	<u>47.4</u>	<u>36.8</u>	27.8	<u>56.8</u>	114.6	20.8
HyCapNet [55]	63.7	-	-	35.8	<u>29.5</u>	58.6	97.2	-
UVA-Cap [56]	76.6	60.6	47.2	36.6	27.8	56.8	<u>114.8</u>	<u>20.6</u>
TPMC(LSTM+VA) (Ours)	77.2	61.3	47.7	36.9	27.8	56.9	115.5	20.8

caption vectors to image captioning effectiveness. Nonetheless, the incorporation of generated caption vectors significantly improves the performance of our model, making it competitive with contemporary image captioning approaches. The comparative analysis on the MS-COCO dataset between our TPMC(LSTM+VA) model and prior approaches is presented in Table 2. According to the results, our model achieves the highest scores in BLEU-2, BLEU-3, BLEU-4, and CIDEr, while tying for the highest performance in BLEU-1 with UpDown [19] and CSA [52], and in SPICE with VARVF [54]. In contrast, HyCapNet [55] attains the highest scores in METEOR and ROUGE. Overall, these results demonstrate the strong competitiveness of our model against recent approaches, including VFDICM [53], VARVF [54], HyCapNet [55], and UVA-CapF [56], highlighting the important role of generated caption vectors in improving image captioning performance, as shown in Fig. 8

4.5 Qualitative Evaluation

In addition to the quantitative evaluation, the caption generation ability of TPMC(LSTM+VA) is further investigated through qualitative comparisons in Figure 9. For each image, the figure presents the caption produced by baseline2 (B), the caption generated by our model (T), and five human-annotated reference captions (GT).



Fig. 9 Illustration of sample outcomes depicting the generated captions from our various modules. B represents the caption generated by Baseline2, while T denotes the captions generated by the TPMC(LSTM+VA) module. GT signifies the ground truth captions.

For the traffic-light example, baseline2 generates “a traffic light on the side of a road,” which correctly identifies the main object but misses important attributes such as the signal color. In comparison, TPMC(LSTM+VA) produces “a traffic light with a green light on it,” providing a more detailed description that agrees with the ground-truth captions describing a green traffic light above a street.

For the elephant scene, baseline2 describes the image as “a group of elephants standing under a tree,” while our model generates “a couple of elephants standing under a tree.” The proposed caption better reflects the number of animals and is closer to the human descriptions mentioning two elephants standing near trees.

In the zebra example, baseline2 provides a general description of zebras grazing in a field, whereas TPMC(LSTM+VA) adds contextual details by describing them as grazing on a lush green field, which better captures the surrounding environment. For

Table 3 Lists of diverse versions of the proposed image captioning model in phase two.

MODEL	V/W	B/G/L	R/S	A/C/T
TPMC(WLRA)	W	L	R	A
TPMC(WLSA)=TPMC(LSTM)	W	L	S	A
TPMC(WLRC)	W	L	R	C
TPMC(WLSC)	W	L	S	C
TPMC(WGSA)=TPMC(GRU)	W	G	S	A
TPMC(VGSA)=TPMC(GRU+VA)	V	G	S	A
TPMC(VLSA)=TPMC(LSTM+VA)	V	L	S	A
TPMC(WBRT)	W	B	R	T

the river-surfing image, baseline2 incorrectly describes the man as riding a wave on a river, while our model identifies the surfboard, generating a more accurate description of the activity.

For the horse example, baseline2 only mentions two white horses, while TPMC(LSTM+VA) recognizes that the horses are pulling a carriage, capturing additional semantic information from the image. Similarly, for the bathroom image, baseline2 identifies a toilet and sink but lacks scene context, whereas our model describes the object arrangement within a bathroom environment.

Overall, these examples demonstrate that TPMC(LSTM+VA) generates captions with richer visual details and stronger semantic alignment with human annotations. The qualitative improvements shown in Figure 9 are consistent with the performance gains reported in Table 1.

4.6 Ablation Study

We have conducted extensive experiments to assess the effectiveness of various methods in improving the performance of our proposed approach. These methods include investigating the impact of different components in our image captioning model, such as the presence or absence of visual attention, the use of different types of Recurrent Neural Network (RNN) for caption embedding, and the arrangement of generated vectors in the subsequent LSTM for caption generation. These modifications were thoroughly evaluated in the second phase of our model, as depicted in Table 3 and Table 4.

Table 3 lists various versions of the proposed model in phase two. The symbols in the first row of Table 3 signifies different meanings. The symbol V indicates the adoption of visual attention while the symbol W signifies the absence of visual attention. The symbols B, G, and L indicate that the adopted encoding RNN used to encode the generated captions is a Bidirectional LSTM, GRU, or LSTM, respectively. The symbol R indicates that the final embedded vector for a generated caption is derived by averaging the h_t from the encoding RNN at each time step. Conversely, the symbol S denotes that the embedded vector of a generated caption comes from the h_t at the final time step of the encoding RNN. The symbol A indicates that the five generated caption vectors are averaged before entering the decoder’s LSTM, while the symbol C shows that the five generated caption vectors are concatenated before they entered

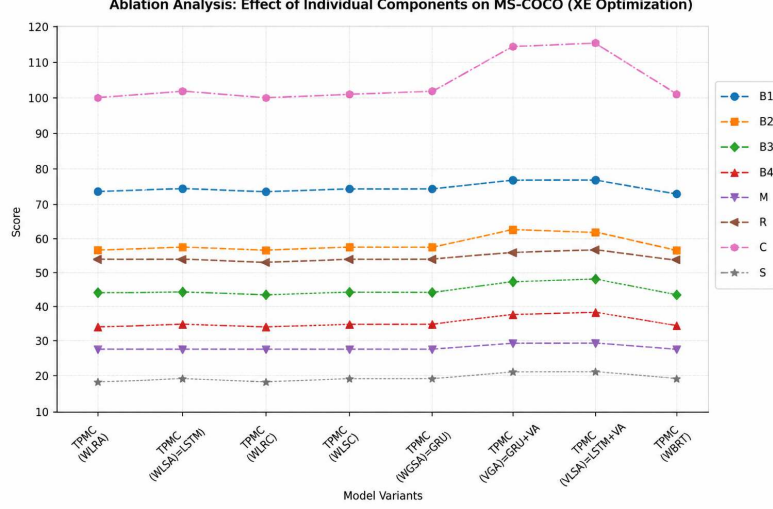


Fig. 10 Ablation analysis of the individual components of the proposed model on the MS-COCO dataset using cross-entropy (XE) optimization.

Table 4 Ablation studies on the distinct components of the proposed model, trained on MS-COCO data under the cross-entropy (XE) optimization.

MODEL	B1	B2	B3	B4	M	R	C	S
TPMC(WLRA)	72.9	56.2	42.6	32.4	25.7	54.0	100.4	18.4
TPMC(WLSA)=TPMC(LSTM)	73.6	57.0	43.4	33.2	26.0	54.4	101.9	18.7
TPMC(WLRC)	72.7	55.8	42.3	32.1	25.6	53.8	99.9	18.3
TPMC(WLSC)	73.5	56.7	43.2	33.0	25.9	54.3	101.4	18.7
TPMC(WGSA)=TPMC(GRU)	73.7	56.9	43.2	32.9	25.9	54.4	101.5	18.6
TPMC(VGSA)=TPCM(GRU+VA)	77.0	61.6	47.4	36.5	27.7	56.7	114.6	20.7
TPMC(VLSA)=TPCM(LSTM+VA)	77.2	61.3	47.7	36.9	27.8	56.9	115.5	20.8
TPMC(WBRT)	73.0	56.3	42.8	32.7	25.8	54.2	100.8	18.5

into the decoder’s LSTM. Finally, the symbol T indicates that the five generated caption vectors are first entered into an attention module and the output of this attention module is fed into the decoder’s LSTM.

As demonstrated in Table 4, the TPMC(WLSA), TPMC(WGSA), TPMC(VGSA), and TPMC(VLSA) models yield impressive results, surpassing the baselines performances. Consequently, these models have been incorporated into our proposed method. Conversely, TPMC(WLRA), TPMC(WLRC), and TPMC(WBRT) models underperform relative to the baselines. However, TPMC(WLSC) stands out by exceeding the baseline in the majority of the evaluation metrics. Fig. 10 presents the ablation analysis of different model components and their impact on the performance of the proposed method on the MS-COCO dataset.

5 Conclusion

This study introduces a two-phase image captioning network with an encoding LSTM, which utilizes generated captions as input in the language model during the second phase. By incorporating generated caption vectors, the model not only expands descriptions of images but also enhances caption quality, resulting in improved performance. Experiments conducted on a publicly available dataset demonstrate that our model performs comparably to advanced image captioning algorithms, significantly surpassing recent state-of-the-art methods. In the future, our goal is to integrate the generated caption vectors into other advanced image captioning frameworks to contribute to enhancing their efficiency and performance.

Declarations

Funding. Not Applicable

Ethics declaration . Not Applicable

References

- [1] Thobhani, A., Zou, B., Kui, X., Abdussalam, A., Asim, M., Shah, S., Elafendi, M.: A survey on enhancing image captioning with advanced strategies and techniques. *Computer Modeling in Engineering & Sciences (CMES)* **142**(3) (2025)
- [2] Hossain, M.A., Ye, Z., Hossen, M.B., Rahman, M.A., Islam, M.S., Abdullah, M.I.: Mssa: memory-driven and simplified scaled attention for enhanced image captioning. *Scientific Reports* (2026)
- [3] Al-Malla, M.A., Hamdoun, O., Ghneim, N.: A comprehensive survey on deep learning approaches for image captioning: a systematic review. *Journal of Big Data* (2026)
- [4] Wu, L., Xu, M., Wang, J., Perry, S.: Recall what you see continually using gridlstm in image captioning. *IEEE Transactions on Multimedia* **22**(3), 808–818 (2019)
- [5] Yu, N., Hu, X., Song, B., Yang, J., Zhang, J.: Topic-oriented image captioning based on order-embedding. *IEEE Transactions on Image Processing* **28**(6), 2743–2754 (2018)
- [6] Hossen, M.B., Ye, Z., Abdussalam, A., Hossain, M.I.: Gva: guided visual attention approach for automatic image caption generation. *Multimedia Systems* **30**(1), 50 (2024)
- [7] Dai, B., Fidler, S., Urtasun, R., Lin, D.: Towards diverse and natural image descriptions via a conditional gan. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2970–2979 (2017)

- [8] Mao, Y., Zhou, C., Wang, X., Li, R.: Show and tell more: Topic-oriented multi-sentence image captioning. In: IJCAI, pp. 4258–4264 (2018)
- [9] Abdussalam, A., Ye, Z., Hawbani, A., Al-Qatf, M., Khan, R.: Numcap: A number-controlled multi-caption image captioning network. *ACM Transactions on Multimedia Computing, Communications and Applications* **19**(4), 1–24 (2023)
- [10] Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014)
- [11] Gu, J., Cai, J., Wang, G., Chen, T.: Stack-captioning: Coarse-to-fine learning for image captioning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018)
- [12] Fidler, S., et al.: Teaching machines to describe images with natural language feedback. *Advances in Neural Information Processing Systems* **30** (2017)
- [13] Zhou, L., Zhang, Y., Jiang, Y.-G., Zhang, T., Fan, W.: Re-caption: Saliency-enhanced image captioning through two-phase learning. *IEEE Transactions on Image Processing* **29**, 694–709 (2019)
- [14] Xu, G., Niu, S., Tan, M., Luo, Y., Du, Q., Wu, Q.: Towards accurate text-based image captioning with content diversity exploration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12637–12646 (2021)
- [15] You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4651–4659 (2016)
- [16] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: *International Conference on Machine Learning*, pp. 2048–2057 (2015). PMLR
- [17] Jiang, W., Wang, W., Hu, H.: Bi-directional co-attention network for image captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **17**(4), 1–20 (2021)
- [18] Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 375–383 (2017)
- [19] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: 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, pp. 6077–6086 (2018)

- [20] Ji, J., Xu, C., Zhang, X., Wang, B., Song, X.: Spatio-temporal memory attention for image captioning. *IEEE Transactions on Image Processing* **29**, 7615–7628 (2020)
- [21] Yan, C., Hao, Y., Li, L., Yin, J., Liu, A., Mao, Z., Chen, Z., Gao, X.: Task-adaptive attention for image captioning. *IEEE Transactions on Circuits and Systems for Video technology* **32**(1), 43–51 (2021)
- [22] Yu, L., Zhang, J., Wu, Q.: Dual attention on pyramid feature maps for image captioning. *IEEE Transactions on Multimedia* **24**, 1775–1786 (2021)
- [23] Jiang, W., Zhu, M., Fang, Y., Shi, G., Zhao, X., Liu, Y.: Visual cluster grounding for image captioning. *IEEE Transactions on Image Processing* **31**, 3920–3934 (2022)
- [24] Yu, J., Li, J., Yu, Z., Huang, Q.: Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology* **30**(12), 4467–4480 (2019)
- [25] Liu, A.-A., Zhai, Y., Xu, N., Nie, W., Li, W., Zhang, Y.: Region-aware image captioning via interaction learning. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(6), 3685–3696 (2021)
- [26] Guo, M.-H., Lu, C.-Z., Liu, Z.-N., Cheng, M.-M., Hu, S.-M.: Visual attention network. *Computational Visual Media* **9**(4), 733–752 (2023)
- [27] Mehrani, P., Tsotsos, J.K.: Self-attention in vision transformers performs perceptual grouping, not attention. *arXiv preprint arXiv:2303.01542* (2023)
- [28] Shi, B., Darrell, T., Wang, X.: Top-down visual attention from analysis by synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2102–2112 (2023)
- [29] Jiang, W., Li, Q., Zhan, K., Fang, Y., Shen, F.: Hybrid attention network for image captioning. *Displays* **73**, 102238 (2022)
- [30] Ma, Y., Ji, J., Sun, X., Zhou, Y., Ji, R.: Towards local visual modeling for image captioning. *Pattern Recognition* **138**, 109420 (2023)
- [31] Al-Qatf, M., Hawbani, A., Wang, X., Abdusallam, A., Alsamhi, S., Alhabib, M., Curry, E.: Rvaic: Refined visual attention for improved image captioning. *Journal of Intelligent & Fuzzy Systems* (Preprint), 1–13 (2024)
- [32] Wang, C., Gu, X.: Learning joint relationship attention network for image captioning. *Expert Systems with Applications* **211**, 118474 (2023)

- [33] Zhang, M., Yang, Y., Zhang, H., Ji, Y., Shen, H.T., Chua, T.-S.: More is better: Precise and detailed image captioning using online positive recall and missing concepts mining. *IEEE Transactions on Image Processing* **28**(1), 32–44 (2018)
- [34] Huang, Y., Chen, J., Ouyang, W., Wan, W., Xue, Y.: Image captioning with end-to-end attribute detection and subsequent attributes prediction. *IEEE Transactions on Image processing* **29**, 4013–4026 (2020)
- [35] Wu, Q., Shen, C., Wang, P., Dick, A., Van Den Hengel, A.: Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence* **40**(6), 1367–1381 (2017)
- [36] Cheng, L., Wei, W., Mao, X., Liu, Y., Miao, C.: Stack-vs: Stacked visual-semantic attention for image caption generation. *IEEE Access* **8**, 154953–154965 (2020)
- [37] Rotstein, N., Bensaid, D., Brody, S., Ganz, R., Kimmel, R.: Fusecap: Leveraging large language models to fuse visual data into enriched image captions. *arXiv preprint arXiv:2305.17718* (2023)
- [38] Bae, J.-W., Lee, S.-H., Kim, W.-Y., Seong, J.-H., Seo, D.-H.: Image captioning model using part-of-speech guidance module for description with diverse vocabulary. *IEEE Access* **10**, 45219–45229 (2022)
- [39] Zhang, J., Mei, K., Zheng, Y., Fan, J.: Integrating part of speech guidance for image captioning. *IEEE Transactions on Multimedia* **23**, 92–104 (2020)
- [40] Al-Qatf, M., Hawbani, A., Wang, X., Abdusallam, A., Zhao, L., Alsamhi, S.H., Curry, E.: Nposc-a3: A novel part of speech clues-aware adaptive attention mechanism for image captioning. *Engineering Applications of Artificial Intelligence* **131**, 107732 (2024)
- [41] Al-Qatf, M., Wang, X., Hawbani, A., Abdusallam, A., Alsamhi, S.H.: Image captioning with novel topics guidance and retrieval-based topics re-weighting. *IEEE Transactions on Multimedia* (2022)
- [42] Li, X., Yuan, A., Lu, X.: Vision-to-language tasks based on attributes and attention mechanism. *IEEE transactions on cybernetics* **51**(2), 913–926 (2019)
- [43] Li, L., Tang, S., Zhang, Y., Deng, L., Tian, Q.: Gla: Global–local attention for image description. *IEEE Transactions on Multimedia* **20**(3), 726–737 (2017)
- [44] Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137 (2015)
- [45] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.-S.: Sca-cnn:

- Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5659–5667 (2017)
- [46] Jiang, W., Ma, L., Jiang, Y.-G., Liu, W., Zhang, T.: Recurrent fusion network for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 499–515 (2018)
 - [47] Wang, J., Wang, W., Wang, L., Wang, Z., Feng, D.D., Tan, T.: Learning visual relationship and context-aware attention for image captioning. Pattern Recognition **98**, 107075 (2020)
 - [48] Wu, C., Yuan, S., Cao, H., Wei, Y., Wang, L.: Hierarchical attention-based fusion for image caption with multi-grained rewards. IEEE Access **8**, 57943–57951 (2020)
 - [49] Sur, C.: Mrrc: multiple role representation crossover interpretation for image captioning with r-cnn feature distribution composition (fdc). Multimedia Tools and Applications **80**(12), 18413–18443 (2021)
 - [50] Xiao, F., Xue, W., Shen, Y., Gao, X.: A new attention-based lstm for image captioning. Neural Processing Letters **54**(4), 3157–3171 (2022)
 - [51] Carmo Nogueira, T., Vinhal, C.D.N., Cruz Júnior, G., Ullmann, M.R.D., Marques, T.C.: A reference-based model using deep learning for image captioning. Multimedia Systems **29**(3), 1665–1681 (2023)
 - [52] Zhao, D., Yang, R., Wang, Z., Qi, Z.: A cooperative approach based on self-attention with interactive attribute for image caption. Multimedia Tools and Applications **82**(1), 1223–1236 (2023)
 - [53] Thobhani, A., Zou, B., Kui, X., Abdussalam, A., Asim, M., Ahmed, N., Alshara, M.: A concise and varied visual features-based image captioning model with visual selection. Computers, Materials, & Continua **81**(2), 2873 (2024)
 - [54] Thobhani, A., Zou, B., Kui, X., Abdussalam, A., Asim, M., ELAffendi, M., Shah, S.: A novelty framework in image-captioning with visual attention-based refined visual features. Computers, Materials & Continua **82**(3) (2025)
 - [55] Bhambu, P., Kumar, S., Arora, K., Chandra, S., Sharma, V., Mankame, D., Singla, B.S.: A hybrid deep learning framework to generate language for image captioning. International Journal of Information Technology **17**(8), 4899–4906 (2025)
 - [56] Thobhani, A., Junfeng, Y., Hael, M.A., Abdussalam, A., Derea, Z., Shakil, K.A., Wani, M.A., Asim, M., Obaid, E.M.: An image captioning framework for therapeutic description generation of traditional chinese medicine herbs. Journal of

