Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.054841

ARTICLE

A Concise and Varied Visual Features-Based Image Captioning Model with Visual Selection

Alaa Thobhani[1,*], Beiji Zou[1], Xiaoyan Kui[1], Amr Abdussalam[2], Muhammad Asim[3], Naveed Ahmed[4] and Mohammed Ali Alshara[4,5]

[1]School of Computer Science and Engineering, Central South University, Changsha, 410083, China

[2]Electronic Engineering and Information Science Department, University of Science and Technology of China, Hefei, 230026, China

[3]EIAS Data Science Lab, College of Computer and Information Sciences, Prince Sultan University, Riyadh, 11586, Saudi Arabia

[4]College of Computer and Information Sciences, Prince Sultan University, Riyadh, 11586, Saudi Arabia

[5]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University, Riyadh, 11432, Saudi Arabia

*Corresponding Author: Alaa Thobhani. Email: althobhanialaa@gmail.com

Received: 09 June 2024 Accepted: 29 September 2024 Published: 18 November 2024

ABSTRACT

Image captioning has gained increasing attention in recent years. Visual characteristics found in input images play a crucial role in generating high-quality captions. Prior studies have used visual attention mechanisms to dynamically focus on localized regions of the input image, improving the effectiveness of identifying relevant image regions at each step of caption generation. However, providing image captioning models with the capability of selecting the most relevant visual features from the input image and attending to them can significantly improve the utilization of these features. Consequently, this leads to enhanced captioning network performance. In light of this, we present an image captioning framework that efficiently exploits the extracted representations of the image. Our framework comprises three key components: the Visual Feature Detector module (VFD), the Visual Feature Visual Attention module (VFVA), and the language model. The VFD module is responsible for detecting a subset of the most pertinent features from the local visual features, creating an updated visual features matrix. Subsequently, the VFVA directs its attention to the visual features matrix generated by the VFD, resulting in an updated context vector employed by the language model to generate an informative description. Integrating the VFD and VFVA modules introduces an additional layer of processing for the visual features, thereby contributing to enhancing the image captioning model's performance. Using the MS-COCO dataset, our experiments show that the proposed framework competes well with state-of-the-art methods, effectively leveraging visual representations to improve performance. The implementation code can be found here: https://github.com/althobhani/VFDICM (accessed on 30 July 2024).

KEYWORDS

Visual attention; image captioning; visual feature detector; visual feature visual attention

1 Introduction

In image captioning, the model faces the formidable task of accurately discerning the salient objects within an image, comprehending their inherent characteristics and attributes, and effectively conveying the intricate interactions between these detected objects. Image captioning networks typically adhere to the encoder-decoder framework. Ingrained in Convolutional Neural Network (CNN), the encoder module diligently extracts the visual features and representations embedded within the input image. In parallel, the decoder module, founded on Recurrent Neural Networks (RNN), assumes the crucial role of generating a coherent textual description that encapsulates the essence of the image's content. This structured approach enables the model to seamlessly connect the visual and linguistic realms, transforming visual data into meaningful and interpretable textual descriptions, thus bridging the gap between computer vision [1–3] and natural language processing.

Despite the notable progress made in previous studies on image captioning, there are still inherent limitations in existing approaches. Specifically, conventional visual-based image captioning methods have a tendency to rely on the same set of visual features throughout all time steps. This uniform treatment persists even when many object features in the input image may not be pertinent to the linguistic context required for generating the subsequent word in the caption. This reliance on irrelevant visual features poses a significant challenge. The inclusion of unrelated objects in the visual input has the potential to divert the attention of the image captioning model towards incorrect visual elements. This, in turn, can result in the generation of inaccurate words within the captions produced by the model. Therefore, it becomes imperative to delve into and discern the most relevant visual features at each time step. This exploration aims to augment the visual attention module, with the ultimate goal of refining the performance of caption generators. By identifying and focusing on the most contextually relevant visual cues during each stage of caption generation, we seek to enhance the accuracy and overall quality of the generated captions.

In our research work, we aim to design image captioning model that can develop image descriptors capable of efficiently exploiting the visual features of images. This proposed method can contribute to boosting the performance of the image captioning models and generating high-quality descriptions. The proposed new method, Visual Features Detection Based Image Captioning Model (VFDICM), aims to exploit the visual features of the input image effectively to enhance the performance of the image captioning models and generate more informative descriptions with higher quality. The proposed image captioning model is essentially built on the UpDown [4] framework and incorporates two additional modules that help leverage the visual features of the input image. These two modules are the visual feature detector module (VFD) and the visual feature visual attention module (VFVA). The VFD module is used to dynamically select the most related features from the visual features to generate a new visual matrix, which consists of the top-k most related visual features to the current linguistic context. Meanwhile, the VFVA module is used to attend to the selected visual features and generate a new visual context vector. This generated vector is fed into the language Long Short-Term Memory (LSTM) layer and used to generate the next word of the partial caption. An illustration of the newly proposed image captioning model is shown in Fig. 1.

The primary objective of our algorithm is to capture the visual attributes within the input image and emphasize their importance, as well as enhancing the performance and scoring of the model for the captioning of the input image. In order to determine the effectiveness of our captioning network, a comprehensive evaluation was conducted of thorough assessment by using MS-COCO [5] dataset. The captions generated by our model display a level of quality that is on par with those produced by numerous state-of-the-art methods when it comes to evaluation metrics, demonstrating a level of

quality that is comparable with the results of our model. As a result of our extensive experiments, we have conclusively demonstrated that the visual feature matrix predicted by the VFD serves as a great guide for the decoder network, resulting in the generation of captions that are much more informative as a result. This significantly outperforms numerous recent image captioning models, marking a significant achievement in the field.

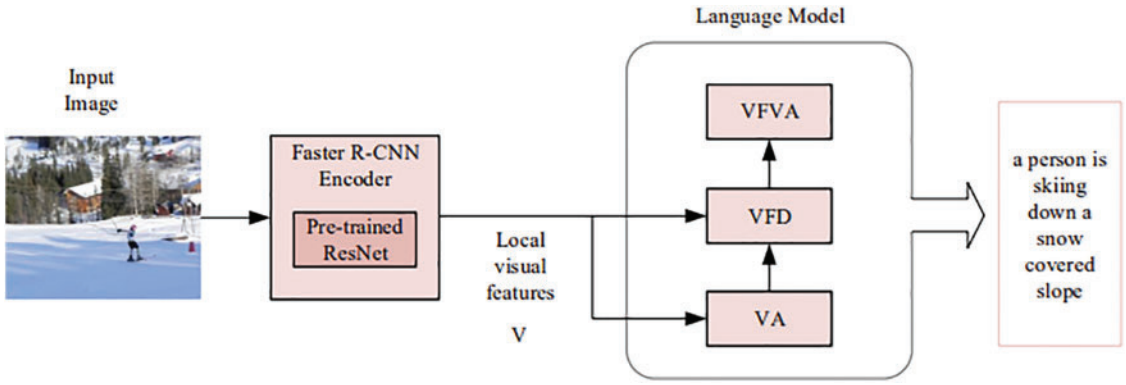

Figure 1: An overview of the proposed pipeline for VFDICM model

This work makes several notable contributions, which can be summarized as follows:

- We explore the impact of utilizing various sets of visual features at each time step to enhance the quality of the generated captions.
- We propose a new method, Visual Features Detection based Image Captioning Model (VFDICM), which aims to exploit the visual features of the input image effectively to enhance the performance of the image captioning models and generate more informative descriptions with higher quality.
- We propose the VFD module for dynamically predicting the most top related visual features, and the VFVA module for addition attention of visual features processing.
- We evaluate our model VFDICM on the MS-COCO dataset, and the results show that the proposed method performs comparably to the latest state-of-the-art techniques according to evaluation metrics.

2 Related Work

A dynamic visual attention mechanism was introduced in [6–8], attention was integrated into the captioning process [7], and adaptive attention was adopted [9]. Other developments encompass merging top-down and bottom-up attention mechanisms [4], enhancing attention through memory mechanisms [10], as well as introducing task-adaptive attention to non-visual words through new models of attention [11]. The quality of captions can also be improved if two attentions are focused on pyramid images simultaneously [12] as well the grounding networks based on clusters are considered [13]. A new metric like Proposal Attention Correctness (PAC) [13] provide a bridge between evaluation of the proposal's performance and its visual grounding. Transformer, which is a multimodal model that uses multiple modalities [14], in conjunction with a multiview feature learning system [15], extends image captioning capabilities, collectively refining attention mechanisms and diversifying captioning techniques in the field. Previous work [16] presents an innovative technique for generating captions for images based on wavelet decomposition and convolutional neural networks in order to achieve comprehensive information extraction. Reference [17] introduces a novel Image Captioning

HANs (Hybrid Attention Networks) combine human captioning attention with machine attention mechanisms to address issues like "object hallucination" and enhance caption diversity. Reference [18] introduces novel attention mechanisms (LSA and LSF) to enhance local visual modeling with leveraging the grid features. The authors in [19] present a framework for refined visual attention (RVA), in which the internal reweighting of visual attention is dependent upon the language context. Reference [20] introduces a GVA-based approach to image caption generation, enhancing the quality of captions by re-adjusting attentional weights. Reference [21] introduces JRAN, an image captioning approach that enhances caption coherence by investigating the relationships between features by incorporating semantic features and region. Despite the effectiveness of these methods, they still grapple with challenges. In particular, these approaches often use the same visual features at all time steps, even when many features are irrelevant to the context needed for the next word. This reliance on irrelevant features can misdirect the model's attention, resulting in inaccurate words in the captions. This highlights the necessity of developing a mechanism to investigate and identify the most relevant visual features at each time step. Such an effort aims to refine the visual attention module, ultimately enhancing the performance of caption generators. By pinpointing and emphasizing the most contextually relevant visual cues during each stage of caption generation.

Image attributes have been used alongside image features to enhance caption quality [22], which involved multimodal attribute detectors trained together with captioning models [23]. Additionally, PoS information has been incorporated into models, guiding information flow and caption generation [24–26]. Topics extracted from caption corpora have been integrated into captioning tasks, influencing sentence generation [27–30]. Some models adopt saliency mechanisms that enhance image representations based on visual, semantic, and sample-related saliency [31]. Attention components, such as semantic and text-guided attention, have also been employed to identify semantic attributes associated with image representations [32]. Multi-stage image descriptors like Stack-VS have been designed to efficiently exploit semantic and visual information through top-down and bottom-up techniques [33]. These diverse approaches collectively contribute to improving image captioning by integrating various sources of information and enhancing model performance. Prior work [34] introduces FUSECAP, enriching captions with visual expert insights and a large language model, creating 12 million improved caption pairs. These enhanced captions improve image captioning models and benefit image-text retrieval. Reference [35] presents a novel semantic-guided attention network for image captioning, integrating external knowledge into a Transformer-based model. Reference [36] introduces the Face-Att model, focusing on generating attribute-centric image captions with a special emphasis on facial features. However, semantic attention in image captioning faces limitations such as incomplete coverage of relevant image attributes, reliance on predefined attributes limiting adaptability to new visual elements, sensitivity to noise in part-of-speech (PoS) information, and challenges in effectively capturing diverse topics while avoiding biases from saliency mechanisms.

In tackling the challenge of encapsulating comprehensive visual content within a single caption, certain image captioning methods have opted for generating multiple descriptions encompassing various facets of an image. One such approach involves a multi-sentence image captioning model utilizing conditional Generative Adversarial Networks (GAN) [37]. This model takes an input image and a random vector, facilitating caption diversity through the joint training of a generator and an evaluator. The objective is to describe various image details by leveraging multiple sentences. Another innovative approach in this domain is the introduction of a multi-caption image captioning network based on topics [38]. In this model, an image and a topic are used as input to generate a topic-related caption while maintaining topical consistency. This is achieved through the fusion gate unit and the utilization of a topic classifier for accurate topic prediction. These approaches collectively contribute

to effectively addressing the complexity of image content representation and diversity in captioning. More recently, a novel model presented in [39] introduces a unique approach that takes into account the number of ground truth captions available for an image during training. This model learns from the numbers associated with these captions and utilizes them to generate diverse captions for the image. Rather than solely relying on the semantic information provided by ground truth captions, this model capitalizes on the quantitative availability of multiple captions to create a varied set of captions for images. This innovative strategy contributes to a significant advancement in the domain of image captioning. However, generating multiple descriptions introduces complexity, and evaluating performance requires adapted metrics. Obtaining diverse training data is essential but may be limited. User preferences for the number and style of sentences are subjective, computational resources, risk of redundancy, interpretability issues.

Cross-entropy loss functions are used to predict the next word in ground truth captions, with evaluation metrics used post-generation. These indistinguishable evaluation metrics have been used to optimize image captioning models using reinforcement learning methods in recent years [40–42]. Self-critical sequence training (SCST) [43] leverages the CIDEr metric for optimization, demonstrating significant improvements in model performance, particularly in CIDEr. Based on the global-local discrimination objective, Reference [40] introduces a reinforcement learning-based optimization approach, incorporating local and global constraints to generate more descriptive captions with finer visual details. Another model [44] incorporates a Kullback-Leibler (KL) divergence term in order to differentiate between accurate and inaccurate predictions, leveraging knowledge graphs to enhance description quality. Hierarchical Attention Fusion (HAF) [45] serves as a reinforcement learning baseline for image captioning, incorporating feature mapping for a number of levels and a revaluing scheme for word and sentence-level rewards. Vocabulary-Critical Sequence Training (VCST) [41] uses a word replacement-based vocabulary critic as a means of providing nuanced credit to words, with efficient algorithms for BLEU and CIDEr-D metric computation. Collectively, these approaches better the quality of descriptive captions through the optimization of models based on evaluation metrics and directly improving the accuracy of the captions.

3 Methodology

Our model focuses on optimizing the use of visual features extracted from input images in order to improve the performance of image captioning models. By leveraging advanced techniques, we aim to capture more nuanced details and context from images, which significantly contributes to more accurate and descriptive captions. As illustrated in Fig. 1, the comprehensive workflow begins with a Faster Region-based Convolutional Neural Network (Faster-RCNN) network employed to extract visual features from the input image. Subsequently, the visual features collected are input into the visual attention module and then into the visual Feature Detector Module (VFD) to predict the most pertinent visual features matrix. Then, the generated visual matrix is utilized in the Visual Feature Visual Attention (VFVA) module. This module attends to the selected visual features and generates a visual context vector, which is crucial to the prediction of the partially generated caption of the next word.

3.1 Input Image Visual Features

Within our captioning network, the visual features are initially extracted from the input image to enable their utilization in subsequent processing by the language model. The initial phase of generating a description for an image involves acquiring the visual representations of the input image. A Faster-RCNN network utilizing ResNet-101 extracts object features from input images, producing an object

feature matrix, represented as V, which contains N object feature vectors.

$$V = \{v_1, v_2, \ldots, v_N\} \tag{1}$$

In this representation, $v_i \in \mathbb{R}^d$ refers to a vector that represents the features of an object, in which i ranges from 1 to N, and $V \in \mathbb{R}^{N\times d}$ represents the matrix of features of the objects. In addition, a mean-pooled object features $\bar{v}$ of the input image is also used as an extra input to the image captioning system, with the following definition:

$$\bar{v} = \frac{1}{N}\sum_{i=1}^{N} v_i \tag{2}$$

Here, $\bar{v} \in \mathbb{R}^d$.

3.2 Visual Attention Mechanism

According to the framework we propose, local visual features of the input image are essential for boosting the model's performance. By accurately capturing these details, we can significantly enhance the overall effectiveness of the image captioning process. To focus on these local visual features, our model relies on a conventional visual attention module, which enables the model to selectively emphasize important aspects of the image. This visual attention process can be described by a set of formulas that enable precise attention to the most important visual features as a result of this operation.

$$\alpha_t^i = W_c \cdot \tanh\left(W_a \cdot h_t^a + W_b \cdot v_i\right) \tag{3}$$

$$\beta_t = \text{softmax}\ (\alpha_t) \tag{4}$$

$$\hat{v}_t = \sum_{i=1}^{N} \beta_t^i \odot v_i \tag{5}$$

where $\hat{v}_t \in \mathbb{R}^d$, $\beta_t \in \mathbb{R}^N$, and $\alpha_t \in \mathbb{R}^N$. $W_c \in \mathbb{R}^e$, $W_a \in \mathbb{R}^{g\times e}$, and $W_b \in \mathbb{R}^{d\times e}$ are trainable weights. $h_t^a \in \mathbb{R}^g$ is the hidden state of the attention LSTM.

3.3 Language Model

A diagram illustrating the language model architecture is shown in Fig. 2. As a basis for the approach we propose, the UpDown framework is used as the baseline structure, known for its effectiveness in image captioning tasks. There are two LSTM layers in the framework: one for language LSTMs, denoted $LSTM_{lan}$, and another for attention LSTMs, denoted $LSTM_{att}$. As a result of leveraging these LSTM layers, the model can better capture and integrate sequential information from the input data. The hidden states of the attention LSTM, represented as $h_t^a \in \mathbb{R}^g$, and the language LSTM, represented as $h_t^l \in \mathbb{R}^g$, can be determined by the following equations, which ensure that the interaction between visual features and language generation is precise and dynamic.

$$h_t^a = LSTM_{att}\ \left(h_{t-1}^a;\ \left[h_{t-1}^l, E \cdot y_{t-1}, \bar{v}\right]\right) \tag{6}$$

$$h_t^l = LSTM_{\text{lan}}\left(h_{t-1}^l; \left[h_t^a, \tilde{v}_t, \hat{v}_t\right]\right) \tag{7}$$

Here, $E \in \mathbb{R}^{m\times q}$ represents the word embeddings matrix, and $y_{t-1} \in \mathbb{R}^m$ denotes the token generated in the previous time step. These embeddings play a crucial role in capturing the semantic meaning of the tokens. $\tilde{v}_t \in \mathbb{R}^d$ is an updated context vector which will be explaned in Section 3.5. To predict

the next token, the hidden state of the language LSTM, h_t^l, is fed into a fully connected layer with a softmax activation function. This setup allows the model to generate a probability distribution p_t over the entire vocabulary, ensuring that the most likely subsequent token is selected based on the context provided by h_t^l. The process is described by the following equation:

$$p_t = \text{softmax}\ \left(h_t^l \cdot W_g\right) \tag{8}$$

where $p_t \in \mathbb{R}^m$. $W_g \in \mathbb{R}^{g \times m}$ represents the weights to be trained.

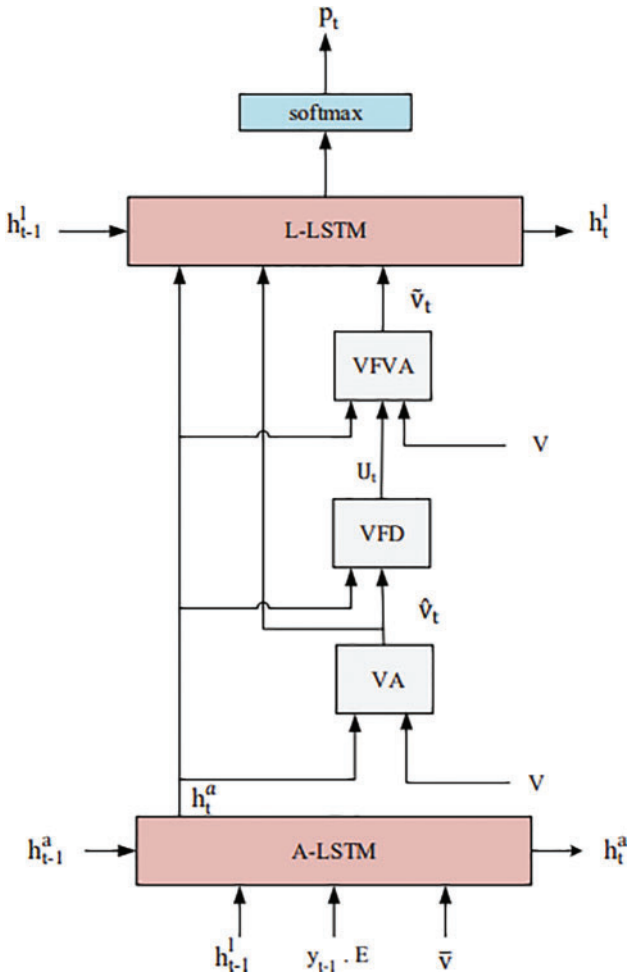

Figure 2: A description of the internal architecture of VFDICM's language model for image captioning

The input word to the attention LSTM at each time step during the training phase of our models is taken from the ground truth annotation. This approach ensures that the model learns from accurate data. Conversely, the input word for the attention LSTM in the testing phase, as defined in Eq. (6), comes from the word predicted in the previous time step. This method allows the model to generate sequences based on its learned predictions. As an initial input in both training and testing, a special "begin-of-sequence" token is used. As the generation of the description's words continues, a special "end-of-sequence" token will be predicted or a maximum description length will be reached indicating the end of the generated description, ensuring a coherent and contextually relevant output.

3.4 Visual Feature Detector (VFD)

The visual feature detector module (VFD) is an ordinary neural network that serves as a detector to dynamically select the most relevant visual features at each time step, in our case the number of top features is referred to as k. The VFD module consists of a concatenation layer, a fully connected layer (FC), and a softmax layer. The input to VFD comprises the output of the conventional visual attention module $\hat{v}_t$ and the hidden state of the attention LSTM h_t^a. The VFD module generates an output matrix known as the selected top-related features matrix (STF), denoted as U_t, which consists of the top-k selected local visual features of the input image. The internal structure of the VFD module is illustrated in Fig. 3. Given the object features matrix V, the context vector $\hat{v}_t$, and the hidden state of the attention LSTM $h_t^a \in \mathbb{R}^g$ as input to the VFD module, we first concatenate $\hat{v}_t$ and h_t^a as follows:

$$x_t = \left[h_t^a, \hat{v}_t\right] \tag{9}$$

where [ ] refers to the concatenation operation. Then, a fully connected layer is used to map the resulted vector $x_t \in \mathbb{R}^{d+g}$ into another vector $\bar{x}_t \in \mathbb{R}^N$ whose length is equal to the number of local visual features N as follows:

$$\bar{x}_t = W_x \cdot x_t \tag{10}$$

where $W_x \in \mathbb{R}^{(d+g)\times N}$ is a learnable parameter matrix. Next, we apply the softmax activation function on $\bar{x}_t$ to generate a new vector $\hat{x}_t \in \mathbb{R}^N$ which represents a probability distribution over N as follows:

$$\hat{x}_t = \text{softmax}\ (\bar{x}_t) \tag{11}$$

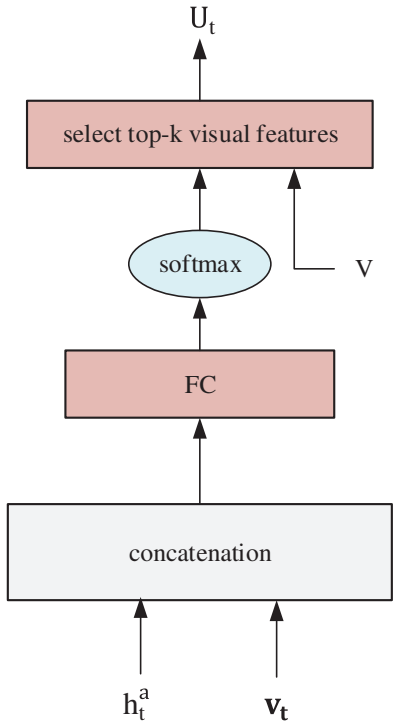

Figure 3: The internal structure of the VFD

The elements of $\hat{x}_t$ are probability values and each element of $\hat{x}_t$ represents the probability of its corresponding local visual feature. After that, the indexes of the k elements of $\hat{x}_t$ with the highest probabilities are determined and their corresponding local visual features in $V \in \mathbb{R}^{N \times d}$ are selected to form the matrix U_t which is given by:

$$U_t = \{u_t^1,\ u_t^2,\ \ldots, u_t^k\} \tag{12}$$

where $u_t^i \in \mathbb{R}^d$ and $U_t \in \mathbb{R}^{k \times d}$. U_t represents a subset of the local visual features the most related to the current linguistic context at the current time step t. Fig. 4 provides an illustration of the selection of the local visual features according to their corresponding probabilities in the VFD module.

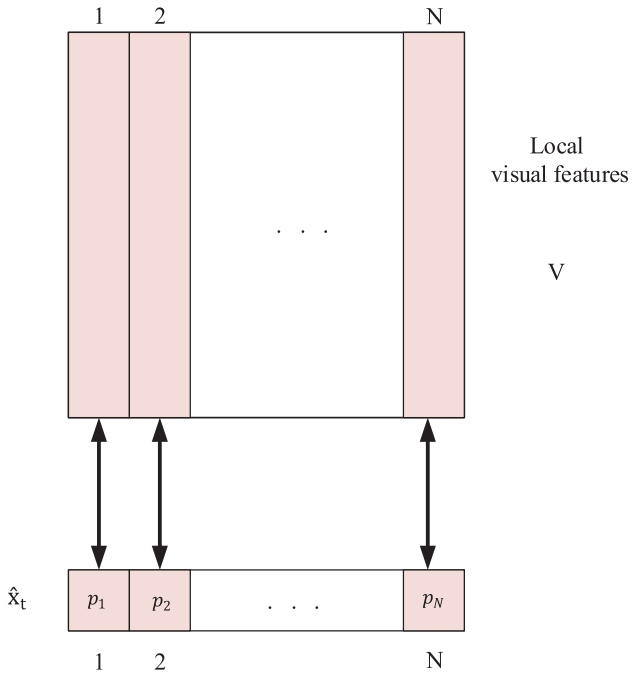

Figure 4: An illustration of the selection of the local visual features according to their corresponding probabilities in the VFD module

3.5 Visual Feature Visual Attention (VFVA)

After generating the STF matrix from the VFD module, we have developed an additional module that takes the STF as its input. This module is originally a visual attention module and is referred to as the Visual Feature Visual Attention (VFVA). The main purpose of this module is to attend to the visual features of the STF matrix, yielding an updated context vector known as the Updated Context Vector (UCV). The UCV is then fed into the language LSTM of the language model to guide the generation of the subsequent word. Considering the updated context vector $\tilde{v}_t$ as defined in the following equations, the VFVA module's formulas demonstrate its role in attending to the importance of each of the various visual feature vectors.

$$\delta_t^i = W_d \cdot \tanh\ \left(W_e \cdot h_t^a + W_f \cdot u_t^i\right) \tag{13}$$

$$\gamma_t = \text{softmax}\ (\delta_t) \tag{14}$$

$$\tilde{v}_t = \sum_{i=1}^{k} \gamma_t^i \odot u_t^i \tag{15}$$

where $\tilde{v}_t \in \mathbb{R}^d$, $\delta_t \in \mathbb{R}^k$, and $\gamma_t \in \mathbb{R}^k$. $W_e \in \mathbb{R}^{g \times e}$, $W_f \in \mathbb{R}^{d \times e}$, $W_d \in \mathbb{R}^e$ are learnable weights.

3.6 Loss Functions

Our image annotating network is trained using two stages: cross-entropy (XE) and CIDEr optimization. In the first stage, standard cross-entropy loss is applied, which is determined as follows:

$$\text{Loss}_{XE} = \frac{1}{T} \sum_{t=1}^{T} -\log \left(p_t \left(y_t \mid y_{1:t-1}, V \right) \right) \tag{16}$$

Second stage, Self-Critical Sequence Training (SCST) is utilized alongside CIDEr-D for optimizing and training the model. The loss function at this stage is specified as follows:

$$\text{Loss}_{RL} = -E_{w_{1:T} \sim \theta} \left[r \left(w_{1:T} \right) \right] \tag{17}$$

In this context, $w_{1:T}$ refers to the sampled annotation, while r indicates CIDEr-D score of sampled annotation. Gradient approximation for Loss_{RL}, symbolized as $\nabla_\theta \text{Loss}_{RL}$, is detailed in Eq. (18). Here, $r\left(\hat{w}_{1:T}\right)$ refers to the CIDEr reward for the maximally sampled annotation, and $r\left(w_{1:T}^s\right)$ denotes the CIDEr reward corresponding to the randomly sampled annotation. This approach enables the model to refine the generated annotations by aligning them more closely with human references. Utilizing the CIDEr-D score helps the model to effectively grasp the intricacies and relevance within the sampled annotations. Consequently, this method enhances the overall quality and accuracy of the generated captions. Additionally, it provides a robust framework for training models in tasks that require a nuanced understanding of content similarity.

$$\nabla_\theta \text{Loss}_{RL} = -\left(r\left(w_{1:T}^s\right) - r\left(\hat{w}_{1:T}\right) \right) \nabla_\theta \log \left(p\left(w_{1:T}^s\right) \right) \tag{18}$$

Algorithm 1 outlines the training process for the VFDICM model. It begins by extracting feature representations from the input images using Eq. (1). Then, for each word position in the caption, the algorithm computes the output of the conventional visual attention module, the selected top-related feature matrix from the VFD module, and the output of the VFVA module using Eqs. (5), (12), and (15), respectively. These computations aid in generating the probability distribution for the next word using Eq. (8). The algorithm continues updating the cross-entropy loss and reinforcement learning loss based on Eqs. (16) and (17) until convergence is achieved. Finally, it returns the generated word probabilities and the model parameters.

Algorithm 1: VFDICM training steps

Require: Training data of image and GT caption pairs *(I, C)* where I is the input image and $C = \{y_1 y_2 \ldots y_T\}$.
Require: Initialize batch size Z and learning rate ψ.
Ensure: Probability of generated word y_t, VFDICM Model parameters W
1: **repeat**
2: Compute feature representations V using Eq. (1)
3: **for** $t = 1; t < T + 1; t++$ **do**

(Continued)

Algorithm 1 (continued)

Compute the output of conventional visual attention module $\hat{v}_t$ using Eq. (5);
Compute the selected top-related feature matrix U_t from the VFD module using Eq. (12);
Compute the output of VFVA module $\tilde{v}_t$ using Eq. (15);
Compute y_t using Eq. (8);
end for
Update $loss_{XE}$ and $loss_{RL}$ according to Eqs. (16) and (17);
until convergence
return y_t, W

4 Experiments and Results

This section discusses various crucial aspects related to our conducted experiments, including the evaluation metrics, the dataset used, the model's configurations, and the training process for our image annotation networks. Additionally, we present and analyze the experimental results and comparisons of our networks. Furthermore, we offer an assessment of the generated text's quality in detail.

4.1 Datasets and Evaluation Metrics

Extensive evaluations were conducted using the MS-COCO dataset [5] in image captioning. As a popular dataset, due to its diversity and numerous caption-image pairs, this dataset served as a solid base for evaluating the effectiveness of the models. The MS-COCO dataset, frequently employed in tasks involving image annotation, is comprised of 123,287 images. Following Karpathy's well-established data splitting method [46], 113,287 images were allocated for training, 5000 images for validation, and another 5000 images for testing. Notably, most of the MS-COCO images are associated with a total of five ground truth captions, resulting in a substantial collection of about 616,747 different ground truth captions for all the MS-COCO images. These captions vary in length, spanning a length of 5 to 49 words, offering a wide spectrum of scenarios and contexts for model training and evaluation.

To assess our model's performance, the model was tested using a series of evaluation metrics, including CIDEr [47], METEOR [48], BLEU [49], ROUGE-L [50], and SPICE [51]. As a result, each of these metrics offers unique insight into a variety of different characteristics of the performance of the model. BLEU, a precision-based metric originally intended for machine translation, has been shown to be highly correlated with human evaluation, emphasizing n-gram precision. METEOR evaluates machine-generated translations by employing a generalized concept of unigram matching with human reference translations, which allows for a balanced assessment of precision and recall. As part of its similarity calculation, CIDEr uses cosine similarity between words in candidate and reference captions based on Term Frequency-Inverse Document Frequency weighted n-grams, thereby taking both recall and precision into account, which is particularly useful for image captioning tasks.

ROUGE, on the other hand, measures the quality of a summary by comparing it with a human-generated summary, which helps in understanding how well the model can generate concise and relevant descriptions. SPICE assesses how effectively captions represent attributes, objects, as well as their relationships, offering a more semantic assessment of captions generated. To simplify the representation of these metrics, we denote METEOR, CIDEr, ROUGE-L, SPICE, and BLEU-n (n = 1, 2, 3, 4), as M, C, R, S, and B-n (n = 1, 2, 3, 4), respectively.

4.2 Experimental Settings

To extract features of objects from images, Faster-RCNN is used based on ResNet-101, resulting in object features of dimensions 36×2048. To manage longer sentences, those exceeding 16 tokens are truncated. In constructing our vocabulary, only words occurring more than five times are included, yielding a vocabulary of 9487 words for MS-COCO. Word embeddings are 1000-dimensional, and hidden states are set for both LSTMs to 1000. During our experiments, we select k to be equal to 10 since it gives us the best scores.

In our image captioning model, we utilize a visual features vector of size $d = 2048$ to represent the input image. LSTM hidden state sizes $g = 1000$ allow for the capture of complex linguistic patterns during caption generation. Words in our vocabulary are embedded in vectors of length $q = 1000$. The features of $N = 36$ objects in the image are used for data extraction. The vocabulary for our captioning task consists of $m = 9487$ unique words for MS-COCO. Additionally, we incorporate an internal hidden attention mechanism of size $e = 512$ to improve the model's ability to generate captions for relevant sections of an image. The number of selected most relevant visual features is set to $k = 10$.

For training our annotation networks, we employ the Adam optimizer and conduct training for 50 epochs in the cross-entropy stage, followed by 100 epochs in the CIDEr optimization stage. Initially, the learning rate is set at 0.0005 and subsequently decreases by a factor of 0.8 every 5 epochs during cross-entropy training and every 10 epochs during CIDEr optimization. Regarding the batch size, a batch size of 40 is chosen. The scheduled sampling percentage increases by 5% every 5 epochs until it reaches 25% during cross-entropy training. The gradients are clipped to an absolute maximum of 0.1. We employ a dropout ratio of 0.5 in our model. Testing is conducted with a beam size of 3 with the beam search strategy. Our networks are developed using the PyTorch framework.

4.3 Building the Vocabulary and Preparing Captions

To construct the vocabulary for our proposed model, we underwent several processing steps with the ground truth caption corpus. Initially, we analyzed all the ground truth sentences, totaling 6,454,115 words, and identified 27,929 unique words. Following word counting, we retained words that appeared more than five times, resulting in a vocabulary of 9486 unique words, while discarding 18,443 unique words. These discarded words constituted approximately 66% of the total unique words in the caption corpus. However, in terms of overall word count, they represented merely 0.5%. While the percentage of discarded unique words may seem high, the actual impact on the model's performance is negligible due to their small contribution to the total word count. All discarded words were replaced with the 'Unknown' special token, which was subsequently added to the vocabulary, expanding it to 9487 words. Additionally, we tokenized all ground truth sentences, replacing each word with its corresponding unique index or integer value assigned from the vocabulary. Sentences longer than 16 words were truncated, and those shorter than 16 words were padded using a '0' digit, serving as zero padding. Furthermore, we prepended the beginning-of-sequence token to the start of each ground truth caption and appended the end-of-sequence token to the end. These preprocessing steps prepared the ground truth captions for training our models.

4.4 Quantitative Scores

As shown in Table 1, our model performed in cross-entropy and CIDEr optimizations using MS-COCO data. The cross-entropy results indicate that our proposed image captioning model surpasses the baseline in most evaluation metrics, notably excelling in METEOR, BLEU-4, SPICE, CIDEr, and ROUGE. Furthermore, experimental results underscore the superiority of our model over the baseline

in terms of CIDEr optimization scores across various evaluation metrics, with significant differences observed in BLEU-4, BLEU-1, ROUGE, METEOR, and CIDEr metrics. These scores unequivocally demonstrate the superior performance of our proposed framework, highlighting the effectiveness and value of our approach.

Table 1: The scores of our model on the MS-COCO dataset for cross-entropy (XE) and CIDEr optimization (RL)

Model	B1	B4	M	R	C	S
Baseline (XE) [4]	76.6	36.2	27.0	56.4	113.5	20.3
VFDICM (XE)	76.4	36.3	27.7	56.6	113.9	20.6
Baseline (CIDEr) [4]	79.8	36.3	27.7	56.9	120.1	21.4
VFDICM (CIDEr)	80.8	37.2	28.3	57.9	122.4	21.5

VFDICM demonstrated significant enhancement during the CIDEr phase, where optimization considers caption quality and diversity. Significant advancements were observed in the BLEU-4, BLEU-1, METEOR, CIDEr, and ROUGE metrics. The integration of VFD and VFVA within our network significantly improves the image captioning algorithm's ability to capture rich visual representations and features, resulting in improved overall performance and description generation. The VFD module updates the visual features matrix, while the VFVA module directs attention to this matrix, generating an updated context vector for the language model to produce informative descriptions.

4.5 Comparison Results

The comparison of other models with our model using the MS-COCO dataset is detailed in Tables 2 and 3, covering cross-entropy optimization and CIDER stages. Table 2 focuses on the cross-entropy results, where our model demonstrates remarkable performance. Specifically, our model outperforms most other models in several key metrics, including ROUGE-L, BLEU-4, BLEU-3, and SPICE. Additionally, it achieves the second-highest scores in CIDEr, BLEU-1, and BLEU-2. It is noteworthy that the r-GRU model secured METEOR's highest score in this phase. Significant differences in scores between our model and others across the majority of metrics underline its greater capabilities during the cross-entropy training phase. Moving on to Table 3, which displays the results from the CIDEr optimization stage, our model continues to excel. It achieves top scores in several metrics such as ROUGE-L, BLEU-3, BLEU-4, BLEU-2, BLEU-1, and METEOR. However, it ranks second in the CIDEr and SPICE metrics. The significant variance between our model's performance and that of other models in metrics like METEOR, BLEU-3, BLEU-2, and BLEU-1 further highlights the effectiveness of our model at this optimization stage.

In more detail, the cross-entropy phase results in Table 2 reveal that our model is not only competitive but often leading in many critical areas. For example, its performance in BLEU-3 and BLEU-4 indicates strong capabilities in generating accurate and contextually appropriate sequences. The high scores in ROUGE-L and SPICE metrics suggest that our model excels in generating linguistically rich and semantically relevant annotations. Despite the r-GRU model's achievement of the highest METEOR score, our model's close performance in this metric indicates its robustness and reliability. Similarly, the CIDEr optimization results shown in Table 3 confirm our model's strong performance across multiple evaluation criteria. Its leading scores in BLEU-1 through BLEU-4

metrics indicate consistent and high-quality output. The model's top performance in METEOR and ROUGE-L underscores its ability to produce both precise and contextually appropriate annotations. While the Stack-VS model attains the highest CIDEr score, our model's second-place ranking in this metric and others like SPICE demonstrates its overall superior performance across the board.

Table 2: The performance of other models and VFDICM trained on MS-COCO using cross-entropy optimization (XE) is compared, with scores for the top two positions highlighted in bold and underline, respectively

Model	B1	B2	B3	B4	M	R	C	S
RFNet [52]	<u>76.4</u>	<u>60.4</u>	<u>46.6</u>	35.8	27.4	<u>56.5</u>	112.5	<u>20.5</u>
UpDown [4]	**77.2**	–	–	<u>36.2</u>	27.0	56.4	113.5	20.3
RecallNet [53]	73.4	–	–	32.2	25.9	53.9	101.6	–
VIS_SAS [31]	72.5	52.6	38.2	28.1	23.7	55.4	82.1	–
SCST [43]	–	–	–	30.0	25.9	53.4	99.4	–
HAF [45]	75.9	59.5	45.4	34.4	26.8	–	109.0	–
MRRC [54]	75.5	59.8	46.0	35.2	26.5	55.9	108.0	19.7
Vis-to-lang [32]	73.9	56.4	41.7	30.9	27.1	–	–	–
TAAIC [11]	71.0	–	–	27.7	23.8	51.1	93.2	18.3
r-GRU [55]	**77.2**	**61.3**	46.3	35.6	**30.2**	55.7	109.2	–
NumCap [39]	66.9	49.4	36.5	27.3	24.1	50.7	85.3	17.0
CSA [56]	**77.2**	59.8	46.0	36.2	27.9	56.4	**114.6**	–
VFDICM (ours)	<u>76.4</u>	<u>60.4</u>	**46.9**	**36.3**	27.7	**56.6**	<u>113.9</u>	**20.6**

Table 3: comparison of the performance of other models and VFDICM trained on the MS-COCO dataset using CIDEr optimization (RL). Scores for the top two places are bolded and underlined, respectively

Model	B1	B2	B3	B4	M	R	C	S
RFNet [52]	79.1	63.1	48.4	36.5	27.7	57.3	121.9	21.2
UpDown [4]	79.8	–	–	36.3	27.7	56.9	120.1	21.4
RecallNet [53]	75.8	–	–	33.1	24.7	54.9	103.7	–
HAF [45]	<u>80.5</u>	62.9	47.7	35.5	27.3	–	116.4	–
Stack-VS [33]	79.4	<u>63.6</u>	<u>49.0</u>	**37.2**	<u>27.9</u>	<u>57.7</u>	**122.6**	**21.6**
SCST [43]	–	–	–	34.2	26.7	55.7	114.0	–
TDA+GLD [40]	78.8	62.6	48.0	36.1	27.8	57.1	121.1	**21.6**
TAAIC [11]	78.6	–	–	<u>37.1</u>	27.5	57.2	119.6	21.2
MRRC [54]	75.3	59.7	46.0	35.3	26.6	55.7	108.2	19.7
VFDICM (ours)	**80.8**	**64.2**	**49.3**	**37.2**	**28.3**	**57.9**	<u>122.4</u>	<u>21.5</u>

These findings strongly underscore the significant improvements our proposed method brings to model performance. The incorporation of the Visual Feature Decoder (VFD) and Visual Feature Visual Attention (VFVA) modules plays a crucial role in this enhancement. By integrating these modules, our model can more effectively harness visual features from input images, which leads to the generation of more detailed and informative descriptions. The VFD module is designed to dynamically identify and extract relevant features from the visual input. This process results in the creation of a new visual matrix that encapsulates the most important aspects of the image. In turn, the newly formed visual matrix is used by the VFVA module, focusing its attention on these critical features. In this way, VFVA updates the contextual data that the language model uses to predict the next word.

This sophisticated interplay between the VFD and VFVA modules significantly boosts model performance on a range of standard evaluation metrics. The VFD module's ability to distill relevant visual information ensures that the model captures essential details from the input images. Meanwhile, the VFVA module's attention mechanism allows the model to maintain a contextual understanding of these visual features, thereby enhancing the accuracy and relevance of the generated descriptions. A further advantage of this method is that it is designed to ensure that each word prediction is informed by a comprehensive understanding of the visual context, which in turn contributes to the model's robust performance. This approach not only improves the model's descriptive capabilities but also ensures consistency and coherence in the generated text. As a result, the model excels across multiple evaluation metrics, demonstrating superior performance compared to methods that do not leverage such advanced visual feature integration.

4.6 Ablation Studies

We conducted several experiments using MS-COCO dataset, refer to Table 4. In experiment VFDICM_sig2, we employed a sigmoid activation function within the VFD component, instead of the softmax activation function of the VFD. In the language LSTM, we concatenated only two inputs: h_t^a and $\tilde{v}_t$, removing $\hat{v}$ from the inputs of the language LSTM. In contrast, in experiment VFDICM_soft2, we used the same two inputs in the language LSTM as the VFDICM_sig2 but applied a softmax activation function in the VFD component.

Table 4: Ablation study of the model with different activation functions and various numbers of inputs to the language LSTM using MS-COCO dataset

Model	B1	B2	B3	B4	M	R	C	S
VFDICM_sig2	79.7	62.5	47.4	35.1	27.5	56.9	117.1	20.7
VFDICM_soft2	79.4	62.1	47.1	34.9	27.4	56.8	116.3	20.7
VFDICM_sig3	80.6	63.8	49.0	36.7	28.3	57.9	122.5	21.5
VFDICM (VFDICM_soft3)	80.8	64.2	49.3	37.2	28.3	57.9	122.4	21.5

In experiment VFDICM_sig3, we again used a sigmoid activation function within the VFD component. However, in the language LSTM, we concatenated three inputs: the output of the first attention mechanism ($\hat{v}_t$), along with h_t^a and $\tilde{v}_t$. The VFDICM_soft3 experiment is our proposed model which uses softmax activation function in the VFD with three inputs into the language LSTM.

To demonstrate and analyze the significance of the second attention mechanism and the impact of the absence of $\tilde{v}_t$, we conducted several experiments using MS-COCO dataset, refer to Table 5, where the attention mechanism was replaced with the mean of features ($\tilde{v}_t = \frac{1}{k}\sum_{i=1}^{k} u_t^i$).

Table 5: Ablation study for replacing VFD component with the mean of features with different activation functions and various numbers of inputs to the language LSTM using MS-COCO dataset

Model	B1	B2	B3	B4	M	R	C	S
VFDICM_soft2_m	78.8	61.2	46.2	34.2	27.0	56.1	112.6	20.1
VFDICM_sig2_m	79.0	61.7	46.7	34.6	27.3	56.5	114.0	20.4
VFDICM_soft3_m	80.9	63.9	48.8	36.6	28.1	57.7	121.1	21.5
VFDICM_sig3_m	80.8	63.9	49.0	36.8	28.2	57.9	122.0	21.3
VFDICM	80.8	64.2	49.3	37.2	28.3	57.9	122.4	21.5

In experiment VFDICM_sig2_m, we used a sigmoid activation function in the VFD component and two inputs in the language LSTM: h_t^a and the mean of features ($\tilde{v}_t$). Conversely, in experiment VFDICM_soft2_m, we used the same inputs in the language LSTM but with a softmax activation function.

For experiment VFDICM_sig3_m, we utilized a sigmoid activation function in the VFD component and three inputs in the language LSTM: $\hat{v}_t$, h_t^a, and $\tilde{v}_t$. Similarly, in experiment VFDICM_soft3_m, we used the same inputs in the language LSTM as VFDICM_sig3_m but with a softmax activation function in the VFD component.

To evaluate the impact of selecting the most relevant visual features at every time step (top-k) on VFDICM performance, we performed ablation experiments varying the k parameters using MS-COCO dataset, refer to Table 6. Different versions were compared with top-k values of 1, 10, 15, 20. We observed a significant correlation between parameter k and overall performance. With $k = 10$, the version outperformed the others with $k = 1$, 15, and 20. Therefore, we chose the configuration with $k = 10$ since it yielded the best results for integration.

Table 6: Ablation studies involving different top-k parameters in the phase of CIDEr optimization applied to the MS-COCO dataset

Top-k	B1	B2	B3	B4	M	R	C	S
$k = 1$	80.6	63.8	48.8	36.5	28.1	57.8	120.7	21.3
$k = 10$	80.8	64.2	49.3	37.2	28.3	57.9	122.4	21.5
$k = 15$	81.0	64.2	49.3	37.0	28.2	57.9	122.4	21.5
$k = 20$	80.6	63.8	48.9	36.7	28.1	57.7	121.7	21.5

4.7 Qualitative Evaluation

In addition to conducting quantitative score analysis, we must assess the quality of the captions produced by VFDICM. Fig. 5 displays a selection of sample images from the test dataset, accompanied by their respective captions. Each image in Fig. 5 is associated with two different types of description. First, we describe the image using our proposed model, followed by the ground truth image captions.

For instance, consider the picture located in row one, column one, top left. The caption for our proposed model, "two men are skiing on a snowboard in the snow," describes the scene in greater detail by identifying "two men." This caption is very similar to the human-generated caption, "Two men

use their snowboards to go down a snowy incline," demonstrating our model's capability to generate human-like captions. In another example, look at the image in the first column and first row from the right. According to our model, "a teddy bear sitting on top of a wooden table," accurately depicts the image. In this caption, the human-annotated ground truth caption is very similar, "A cake shaped as a Teddy Bear on a wooden table." VFDICM's performance and quality of generated descriptions remain high, significantly exceeding that of the baseline on standard evaluation metrics, as shown by the scores in Table 3. These examples illustrate the model's ability to understand and describe various contexts effectively. Moreover, the consistency in generating accurate captions across different images underscores the robustness of our model.

Figure 5: Examples of the generated captions from VFDICM model. The VFDICM generates caption V, while the GT represents the ground truth captions

4.8 Discussion

In this work, we have introduced an innovative image captioning model that utilizes the visual characteristics of the input image to produce informative sentences. This model places a strong emphasis on the local visual features of the input image, enabling it to guide the prediction of the next word in the evolving caption. Through the integration of the Visual Feature Detector (VFD) and Visual Feature Visual Attention (VFVA) modules, the model effectively harnesses the visual representations of the input image, leading to a significant enhancement in the performance of the image captioning model.

VFDICM exclusively relies on the visual attributes of the input image, without depending on semantic information such as topics, attributes, or Parts of Speech (PoS). This reliance solely on the visual features empowers the model to exploit the richness of the visual content within the input image, ultimately generating high-quality text. Additionally, the proposed framework operates independently of any external data sources, other than the adopted dataset, allowing the model to concentrate entirely on the available visual data within the image content. As well as its simplified architecture and fewer hyperparameters, our model is a good choice as a baseline for more complicated models and architectures.

Furthermore, the VFDICM model, as proposed, makes a substantial contribution to improving the efficiency of image captioning algorithms and producing high-quality descriptions. This novel approach effectively exploits the visual features of the input image to improve the performance of image captioning models and generate more accurate and informative descriptions. Our model incorporates two additional modules, VFD and VFVA, to facilitate this. The VFD module dynamically selects the most relevant features of the input, creating a new visual matrix that consists of these relevant visual features in the current linguistic context. Meanwhile, the VFVA module attends to these selected visual features and generates a new visual context vector, which is then utilized for generating a new word in the evolving caption.

5 Conclusion

An innovative approach for improving image captioning is presented in this paper. VFDICM model comprises the Visual Feature Detector (VFD) and Visual Feature Visual Attention (VFVA) modules, which dynamically select and emphasize relevant visual features at each time step. Through rigorous experimentation to substantiate the advantages of VFD and VFVA modules, the outcomes illustrate that the methodology attains state-of-the-art performance of the model trained with both cross-entropy loss and RL-based loss. By selectively emphasizing pertinent visual features, we aim to enhance caption accuracy and relevance, contributing to the advancement of image captioning technology. This work deepens the understanding of how selective attention to visual features can improve image captioning quality. In the future work, we aim to explore the integration of Transformer architectures, specifically by incorporating VFD and VFVA within the Transformer decoder and multi-head self-attention into our model architecture. This research also offers promising directions for future studies in this field, ultimately leading to more accurate and informative image descriptions for real-world applications.

Acknowledgement: The authors would like to thank Prince Sultan University for their support.

Funding Statement: This work is supported by the National Natural Science Foundation of China (Nos. U22A2034, 62177047), High Caliber Foreign Experts Introduction Plan funded by MOST, and Central South University Research Programme of Advanced Interdisciplinary Studies (No.

2023QYJC020). Also, the authors would like to thank Prince Sultan University for paying the APC of this article.

Author Contributions: Alaa Thobhani: Software, Project administration, Investigation, Conceptualization, Visualization. Beiji Zou: Supervision. Xiaoyan Kui: Writing review and editing. Amr Abdussalam: Validation. Muhammad Asim: Resources, Formal analysis. Naveed Ahmed: Investigation. Mohammed Ali Alshara: Writing review. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: We used the well-known MS COCO dataset, which is publicly available.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] M. H. Bashir, M. Ahmad, D. R. Rizvi, and A. A. A. El-Latif, "Efficient CNN-based disaster events classification using UAV-aided images for emergency response application," *Neural Comput. Appl.*, vol. 36, no. 18, pp. 1–14, 2024. doi: 10.1007/s00521-024-09610-4.

[2] H. Ibrahim *et al.*, "Efficient color image enhancement using piecewise linear transformation and gamma correction," *J. Opt.*, vol. 53, pp. 2027–2037, 2024. doi: 10.1007/s12596-023-01171-4.

[3] S. R Waheed, N. M. Suaib, M. S. M. Rahim, A. R. Khan, S. A. Bahaj and T. Saba, "Synergistic integration of transfer learning and deep learning for enhanced object detection in digital images," *IEEE Access*, vol. 12, pp. 13525–13536, 2024. doi: 10.1109/ACCESS.2024.3354706.

[4] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[5] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Comput. Vis.–ECCV 2014: 13th Eur. Conf.*, Zurich, Switzerland, Springer, pp. 740–755.

[6] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4651–4659.

[7] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Int. Conf. on Mach. Learn.*, PMLR, 2015, pp. 2048–2057.

[8] W. Jiang, W. Wang, and H. Hu, "Bi-directional co-attention network for image captioning," *ACM Trans. Multimedia Comput., Commun., Appl. (TOMM)*, vol. 17, no. 4, pp. 1–20, 2021. doi: 10.1145/3460474.

[9] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 375–383.

[10] J. Ji, C. Xu, X. Zhang, B. Wang, and X. Song, "Spatio-temporal memory attention for image captioning," *IEEE Trans. Image Process.*, vol. 29, pp. 7615–7628, 2020. doi: 10.1109/TIP.2020.3004729.

[11] C. Yan *et al.*, "Task-adaptive attention for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 43–51, 2022. doi: 10.1109/TCSVT.2021.3067449.

[12] L. Yu, J. Zhang, and Q. Wu, "Dual attention on pyramid feature maps for image captioning," *IEEE Trans. Multimedia*, vol. 24, pp. 1775–1786, 2021. doi: 10.1109/TMM.2021.3072479.

[13] W. Jiang, M. Zhu, Y. Fang, G. Shi, X. Zhao and Y. Liu, "Visual cluster grounding for image captioning," *IEEE Trans. Image Process.*, vol. 31, pp. 3920–3934, 2022. doi: 10.1109/TIP.2022.3177318.

[14] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4467–4480, 2019. doi: 10.1109/TCSVT.2019.2947482.

[15] A. A. Liu, Y. Zhai, N. Xu, W. Nie, W. Li and Y. Zhang, "Region-aware image captioning via interaction learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3685–3696, 2021. doi: 10.1109/TCSVT.2021.3107035.

[16] M. H. Guo, C. Z. Lu, Z. N. Liu, M. M. Cheng, and S. M. Hu, "Visual attention network," *Comput. Vis. Media*, vol. 9, no. 4, pp. 733–752, 2023. doi: 10.1007/s41095-023-0364-2.

[17] W. Jiang, Q. Li, K. Zhan, Y. Fang, and F. Shen, "Hybrid attention network for image captioning," *Displays*, vol. 73, 2022, Art. no. 102238. doi: 10.1016/j.displa.2022.102238.

[18] Y. Ma, J. Ji, X. Sun, Y. Zhou, and R. Ji, "Towards local visual modeling for image captioning," *Pattern Recognit.*, vol. 138, 2023, Art. no. 109420. doi: 10.1016/j.patcog.2023.109420.

[19] M. Al-Qatf *et al.*, "RVAIC: Refined visual attention for improved image captioning," *J. Intell. Fuzzy Syst.*, vol. 46, pp. 1–13, 2024. doi: 10.3233/JIFS-233004.

[20] M. B. Hossen, Z. Ye, A. Abdussalam, and M. I. Hossain, "GVA: Guided visual attention approach for automatic image caption generation," *Multimed. Syst.*, vol. 30, no. 1, 2024, Art. no. 50. doi: 10.1007/s00530-023-01249-w.

[21] C. Wang and X. Gu, "Learning joint relationship attention network for image captioning," *Expert. Syst. Appl.*, vol. 211, no. 20, 2023, Art. no. 118474. doi: 10.1016/j.eswa.2022.118474.

[22] M. Zhang, Y. Yang, H. Zhang, Y. Ji, H. T. Shen and T. S. Chua, "More is better: Precise and detailed image captioning using online positive recall and missing concepts mining," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 32–44, 2018. doi: 10.1109/TIP.2018.2855415.

[23] Y. Huang, J. Chen, W. Ouyang, W. Wan, and Y. Xue, "Image captioning with end-to-end attribute detection and subsequent attributes prediction," *IEEE Trans. Image Process.*, vol. 29, pp. 4013–4026, 2020. doi: 10.1109/TIP.2020.2969330.

[24] J. W. Bae, S. H. Lee, W. Y. Kim, J. H. Seong, and D. H. Seo, "Image captioning model using part-of-speech guidance module for description with diverse vocabulary," *IEEE Access*, vol. 10, no. 11, pp. 45219–45229, 2022. doi: 10.1109/ACCESS.2022.3169781.

[25] J. Zhang, K. Mei, Y. Zheng, and J. Fan, "Integrating part of speech guidance for image captioning," *IEEE Trans. Multimedia*, vol. 23, pp. 92–104, 2020. doi: 10.1109/TMM.2020.2976552.

[26] M. Al-Qatf *et al.*, "NPoSC-A3: A novel part of speech clues-aware adaptive attention mechanism for image captioning," *Eng. Appl. Artif. Intell.*, vol. 131, no. 4, 2024, Art. no. 107732. doi: 10.1016/j.engappai.2023.107732.

[27] N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, "Topic-oriented image captioning based on order-embedding," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2743–2754, 2018. doi: 10.1109/TIP.2018.2889922.

[28] H. Wei, Z. Li, F. Huang, C. Zhang, H. Ma and Z. Shi, "Integrating scene semantic knowledge into image captioning," *ACM Trans. Multimedia Comput., Commun., Appl. (TOMM)*, vol. 17, no. 2, pp. 1–22, 2021. doi: 10.1145/3439734.

[29] M. Liu, H. Hu, L. Li, Y. Yu, and W. Guan, "Chinese image caption generation via visual attention and topic modeling," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 1247–1257, 2020. doi: 10.1109/TCYB.2020.2997034.

[30] M. Al-Qatf, X. Wang, A. Hawbani, A. Abdusallam, and S. H. Alsamhi, "Image captioning with novel topics guidance and retrieval-based topics re-weighting," *IEEE Trans. Multimedia*, vol. 25, pp. 5984–5999, 2023. doi: 10.1109/TMM.2022.3202690.

[31] L. Zhou, Y. Zhang, Y. G. Jiang, T. Zhang, and W. Fan, "Re-caption: Saliency-enhanced image captioning through two-phase learning," *IEEE Trans. Image Process.*, vol. 29, pp. 694–709, 2019. doi: 10.1109/TIP.2019.2928144.

[32] X. Li, A. Yuan, and X. Lu, "Vision-to-language tasks based on attributes and attention mechanism," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 913–926, 2019. doi: 10.1109/TCYB.2019.2914351.

[33] L. Cheng, W. Wei, X. Mao, Y. Liu, and C. Miao, "Stack-VS: Stacked visual-semantic attention for image caption generation," *IEEE Access*, vol. 8, pp. 154953–154965, 2020. doi: 10.1109/ACCESS.2020.3018752.

[34] N. Rotstein, D. Bensaid, S. Brody, R. Ganz, and R. Kimmel, "FuseCap: Leveraging large language models to fuse visual data into enriched image captions," 2023, *arXiv:2305.17718*.

[35] D. A. Hafeth, S. Kollias, and M. Ghafoor, "Semantic representations with attention networks for boosting image captioning," *IEEE Access*, vol. 11, pp. 40230–40239, 2023. doi: 10.1109/ACCESS.2023.3268744.
[36] N. Haque, I. Labiba, and S. Akter, "FaceAtt: Enhancing image captioning with facial attributes for portrait images," 2023, *arXiv:2309.13601*.
[37] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional gan," in *Proc. of the IEEE Int. Conf. on Comput. Vis.*, 2017, pp. 2970–2979.
[38] Y. Mao, C. Zhou, X. Wang, and R. Li, "Show and tell more: Topic-oriented multi-sentence image captioning," in *IJCAI*, 2018, pp. 4258–4264.
[39] A. Abdussalam, Z. Ye, A. Hawbani, M. Al-Qatf, and R. Khan, "NumCap: A number-controlled multi-caption image captioning network," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 19, no. 4, pp. 1–24, 2023. doi: 10.1145/3576927.
[40] J. Wu, T. Chen, H. Wu, Z. Yang, G. Luo and L. Lin, "Fine-grained image captioning with global-local discriminative objective," *IEEE Trans. Multimedia*, vol. 23, pp. 2413–2427, 2020. doi: 10.1109/TMM.2020.3011317.
[41] H. Liu, S. Zhang, K. Lin, J. Wen, J. Li and X. Hu, "Vocabulary-wide credit assignment for training image captioning models," *IEEE Trans. Image Process.*, vol. 30, pp. 2450–2460, 2021. doi: 10.1109/TIP.2021.3051476.
[42] N. Xu *et al.*, "Multi-level policy and reward-based deep reinforcement learning framework for image captioning," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1372–1383, 2019. doi: 10.1109/TMM.2019.2941820.
[43] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7008–7024.
[44] Y. Zhang, X. Shi, S. Mi, and X. Yang, "Image captioning with transformer and knowledge graph," *Pattern Recognit. Lett.*, vol. 143, no. 6, pp. 43–49, 2021. doi: 10.1016/j.patrec.2020.12.020.
[45] C. Wu, S. Yuan, H. Cao, Y. Wei, and L. Wang, "Hierarchical attention-based fusion for image caption with multi-grained rewards," *IEEE Access*, vol. 8, pp. 57943–57951, 2020. doi: 10.1109/ACCESS.2020.2981513.
[46] A. Karpathy and L. Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3128–3137.
[47] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2015, pp. 4566–4575.
[48] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Meas. for Mach. Transl. Summarizat.*, 2005, pp. 65–72.
[49] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meet. Assoc. Computat. Linguistics*, 2002, pp. 311–318.
[50] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.
[51] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Comput. Vis.–ECCV 2016*: Amsterdam, The Netherlands, Springer, 2016, pp. 14–398.
[52] W. Jiang, L. Ma, Y. -G. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 499–515.
[53] L. Wu, M. Xu, J. Wang, and S. Perry, "Recall what you see continually using gridlstm in image captioning," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 808–818, 2019. doi: 10.1109/TMM.2019.2931815.
[54] C. Sur, "MRRC: Multiple role representation crossover interpretation for image captioning with R-CNN feature distribution composition (FDC)," *Multimed. Tools Appl.*, vol. 80, no. 12, pp. 18413–18443, 2021. doi: 10.1007/s11042-021-10578-9.

[55] T. do Carmo Nogueira, C. D. N. Vinhal, G. da Cruz Júnior, M. R. D. Ullmann, and T. C. Marques, "A reference-based model using deep learning for image captioning," *Multimed. Syst.*, vol. 29, no. 3, pp. 1665–1681, 2023. doi: 10.1007/s00530-022-00937-3.

[56] D. Zhao, R. Yang, Z. Wang, and Z. Qi, "A cooperative approach based on self-attention with interactive attribute for image caption," *Multimed. Tools Appl.*, vol. 82, no. 1, pp. 1223–1236, 2023. doi: 10.1007/s11042-022-13279-z.