Automatic Generation of Image Caption Based on Semantic Relation using Deep Visual Attention Prediction

M. M. EL-GAYAR
Department of Information Technology-Faculty of Computers and Information, Mansoura University
Mansoura 35516, Egypt
Faculty of Computer Science and Engineering, New Mansoura University, New Mansoura, Egypt

***Abstract*—While modern systems for managing, retrieving, and analyzing images heavily rely on deriving semantic captions to categorize images, this task presents a considerable challenge due to the extensive capabilities required for manual processing, particularly with large images. Despite significant advancements in automatic image caption generation and human attention prediction through convolutional neural networks, there remains a need to enhance attention models in these networks through efficient multi-scale features utilization. Addressing this need, our study presents a novel image decoding model that integrates a wavelet-driven convolutional neural network with a dual-stage discrete wavelet transform, enabling the extraction of salient features within images. We utilize a wavelet-driven convolutional neural network as the encoder, coupled with a deep visual prediction model and Long Short-Term Memory as the decoder. The deep Visual Prediction Model calculates channel and location attention for visual attention features, with local features assessed by considering the spatial-contextual relationship among objects. Our primary contribution is to propose an encoder and decoder model to automatically create a semantic caption on the image based on the semantic contextual information and spatial features present in the image. Also, we improved the performance of this model, demonstrated through experiments conducted on three widely used datasets: Flickr8K, Flickr30K, and MSCOCO. The proposed approach outperformed current methods, achieving superior results in BLEU, METEOR, and GLEU scores. This research offers a significant advancement in image captioning and attention prediction models, presenting a promising direction for future work in this field.**

Keywords*—Semantic image captioning; deep visual attention model; long short-term memory; wavelet driven convolutional neural network

I. Introduction

One of the active research topics in the field of computer vision is the creation of captions for images automatically. Image captioning refers to generate a text-based description or caption for a given image. This task unites computer vision and natural language processing methodologies to produce an easily understandable narrative that concisely conveys the image's content. The primary objective is to offer a brief and precise depiction of the elements, settings, actions, and occurrences depicted in the image [1-3]. There is an increasing daily demand for image retrieval and analysis systems because they are used in many fields and on a large scale via the Internet, social media, and various search engines [4] [5]. Some ways in which image captioning benefits daily life include the following:

- For individuals with visual impairments: Image captions offer crucial details about an image's content, allowing them to gain a better understanding of the context surrounding the image.
- Cross-lingual communication: By translating image captions into various languages, individuals from different language backgrounds can better grasp the content of an image, promoting intercultural communication and appreciation.
- Search Engine Optimization (SEO): By offering pertinent textual data connected to an image, image captions can enhance SEO. This allows search engines to index and rank the content more precisely, improving online visibility and discoverability.
- User engagement on social media: Image captions contribute to a better user experience on social media platforms by supplying contextual information and additional details about images. This results in increased interaction and improved communication among users.
- Educational and e-learning contexts: Image captions play a supportive role in learning environments by making visual content more explicit and accessible for students, especially those facing learning disabilities or language challenges. This assistance leads to enhanced learning outcomes and better understanding of diverse topics.
- Data management and retrieval: Image captioning assists in organizing and locating visual information within extensive databases, simplifying the process of searching for particular images or content based on their descriptions.

In summary, image captioning serves as a crucial tool that enhances accessibility, comprehension, and distribution of visual content, providing advantages to a broad spectrum of users and applications in everyday life. This type of research is a vital topic that researchers are attracted to because it

combines three main areas: machine learning, natural language processing, and computer vision. It also serves a wide range of practical applications. Fusing these components result in sophisticated systems capable of autonomously interpreting the situation depicted in the image and generating coherent sentences to describe it.

The conventional method for image captioning consists of two key components: feature extraction and language modeling. In the feature extraction stage, an input image is processed by a pre-trained Convolutional Neural Network (CNN) model, such as VGG, Inception, or ResNet, to derive high-level visual features. Subsequently, during the language modeling phase, these extracted features are input into a language model, typically a Recurrent Neural Network (RNN) or a Long Short-Term Memory (LSTM) network, which produces a word sequence that forms the caption. By training the language model on a vast dataset of images and their associated captions, it learns to associate visual features with right words and phrases. Automatic image caption creation can be used in many practical systems and applications such as image retrieval through search engines, video labels, answering visual questions, assisting visually impaired people, biomedical imaging, robotics, etc. Recently, multiple approaches have been developed to automatically generate image captions, reducing many computer vision challenges [6].

The first method used in the image retrieval process relies on comparing the input image with a similar template to create a caption for the image through the matching or comparison process. However, the effectiveness of this method remains unproven, and it yields imprecise outcomes when dealing with intricate images containing multiple targets. Consequently, an alternative strategy relies on the development of a deep neural network, where the image is encoded, and captions are produced using a language model. In this process, the visual content of the image is analyzed in depth, and then this information is translated into natural language text descriptions. Nevertheless, there is a need to enhance CNN-based attention models by effectively utilizing multi-scale features in this model.

This paper introduces an automatic image captioning framework that generates semantically meaningful captions. The approach uses a deep neural network architecture, comprising a CNN that encodes the visual features and RNN that decodes and generates the text [7], [8]. It then employs LSTM [9], [10] and gated recurrent units (GRU) [11] to derive significant insights. The key contributions of this article can be summarized as follows:

- Propose an encoder and decoder framework to automatically create a semantic caption on the image based on the semantic contextual information and spatial features present in the image.
- A Deep Visual Prediction Model (DVPM) is proposed by enabling the extraction of further semantic information from the image to utilize the convolutions on the feature maps generated using the Wavelet-driven Convolutional Neural Network (WCNN). Both channel and spatial attention are calculated using this approach, which are derived from the resulting feature maps.
- A semantic spatial contextual connection derived from the WCNN model is established to predict area proposals between distinct objects within the image.
- The feature maps produced by the WCNN model are leveraged by a Semantic Relation Extractor (SRE) to predict region proposals to determine the spatial relationships among diverse objects in the image.

The effectiveness of the suggested framework is assessed by employing three widely recognized benchmark datasets: Flickr8K, Flickr30K, and MS-COCO. Furthermore, a comparison with existing studies is conducted using various evaluation metrics, including Bilingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit Ordering (METEOR), and Consensus-based Image Description Evaluation (CIDEr).

This manuscript is organized into five sections. Section II reviews relevant previous work. Section III describes the proposed approach in detail. Experiments and results are presented in Section IV. Section V concludes the paper.

II. Related Works

This section will review related studies regarding recent modalities in image captioning using attention mechanisms. A previously trained CNN (Encoder) would generate a hidden state (HS) in classic image captioning. Next, decoding this hidden state utilizes an LSTM (as the decoder) to frequently generate each word from the state. However, when the model attempts to produce the next word of the caption, there is an issue that this word typically only describes part of the image.

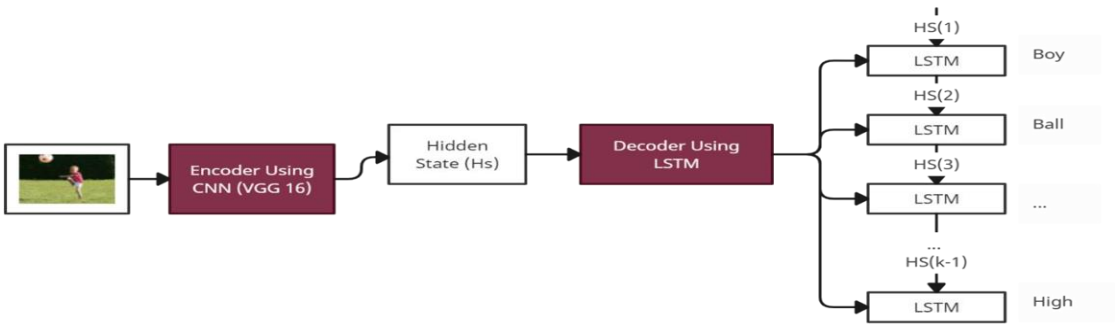


Fig. 1. A traditional image captioning model.

It is also unable to capture the essence of the entire input image. The model cannot efficiently generate different words for different parts of the image. Therefore, the attention mechanism is useful for representing the image [12]–[14]. Thus, generating an appropriate textual description requires a deeper understanding of the image's spatial and semantic content. As previously mentioned, initial efforts to create image translations involve extracting visible features using RAM (CRF) and converting these features into text through holistic or consensus-based improvements. Later recovery methods generate translations of single or multiple sentences from predefined phrases based on visual similarities. Developing a deep neural network architecture aids in achieving more advanced visual and natural language modeling by producing more insightful descriptions of the image.

Karpathy et al. [15] proposed a multimodal RNN for producing better descriptions using an alignment of replacement between the segments of the image and the sentence. Deng et al. [16] proposed an adaptive attention model with a visual sentinel. This model is presented to extract the global image characteristics of the encoding phase.

Zha et al. [17] proposed a context-aware visual policy network for better caption generation, reducing the dependency on previously predicted words using fine-grained image sentence captioning. Yu et al. [18] proposed a model that used dual attention (P and D) feature maps in the hierarchical image to explore the visual semantic connections and improve the quality of the sentences created.

Yang et al. [19] proposed a CaptionNet model for improved caption generation, which decreased the reliance on previously predicted words. This model only allows attended image features to be input into the memory of CaptionNet through input gates. Cornia et al. [20] proposed an innovative image captioning technique utilizing memory vectors and connecting the encoder and decoder sections of the transformer model.

Li X et al. [21] proposed an innovative technique for aligning the image and language modalities to gain more reasonable semantic extraction from images using anchor points. Jiang et al. [22] introduced a Multi-Gate Attention model to enhance caption generation, expanding upon the conventional self-attention mechanism by integrating an extra Attention Weight Gate. Wang et al. [30] proposes an automatic architecture search method for neural networks focused on cross-modality tasks like image captioning. The method approximates the associative connection between visual and language models through the internal structure of RNN cells. Over 100 generated RNN variants exceed performance of 100 on CIDEr and 31 on BLEU4, with the top model achieving 101.4 and 32.6, respectively. Wu et al. [31] introduced a novel global-local discriminative objective built on a reference model to generate more detailed descriptive captions. Evaluated on MS-COCO, the method outperforms baselines significantly and competes well with top approaches. Self-retrieval experiments demonstrate its ability to generate discriminative captions. Wang et al. [32] introduced a visual attention layer for low-level visual information and a semantic attention layer for high-level semantic attributes. The margin-based loss encourages more discriminative captions. Extensive experiments on COCO and Flickr30K datasets validate the approach, demonstrating superior performance in captioning. The method achieves state-of-the-art 70.6 CIDEr-D on Flickr30K and competitive 123.5 CIDEr-D on COCO.

Although these methods effectively generate image captions, they fail to incorporate refined semantic components and the contextual spatial connections between various objects within the image. Therefore, expansions in network structure are essential to remedy these deficiencies. Furthermore, when several objects exist in the image, it is critical to properly consider the optical contextual connection between them to produce a more detailed and representative caption. This problem can be resolved by integrating attention mechanisms and assessing the spatial relations among elements within the instance. Table I shows the summary of recent related works.

TABLE I. SUMMARY OF RECENT PREVIOUS RELATED WORKS

Ref.	Dataset	BLEU	METEOR	CIDEr
7	FLICKER 8K	21.5	20.8	-
16	FLICKER 8K	25.7	22.6	52.6
19	FLICKER 8K	21.3	20.4	-
27	FLICKER 8K	16	-	-
16	FLICKER 30K	22.3	19.6	-
19	FLICKER 30K	19.8	18.5	-
30	FLICKER 30K	24.9	20.9	59.7
31	MS-COCO	37.2	28.4	123.4
32	MS-COCO	36.2	27.8	121.1

III. PROPOSED FRAMEWORK

As illustrated in Fig. 1, a typical image captioning model would use a pre-trained convolutional neural network (CNN), such as VGG-16, to encode the input image and generate image features (HS) [23], [24]. Then, it would decode this HS using a Long Short-Term Memory (LSTM) and recursively render each caption word. The downside of this approach is that when the model tries to generate the following word in the caption, it fails to fully comprehend the overall meaning or essence of the entire input image. Therefore, a semantic deep visual attention mechanism can be helpful. With a semantic deep attention mechanism, the image is separated into n regions, and we calculate with CNN representations of each region HS(1),…, HS(n). When the RNN-decoder generates a further word, the attention procedure concentrates on the appropriate region of the image, so the decoder only uses exact areas. Attention could be considerably distinguished into two types [25]–[29]:

- Global Attention is positioned on all origin positions, as shown in Fig. 2.
- Local Attention is positioned just on a few sources' places, as shown in Fig. 3.

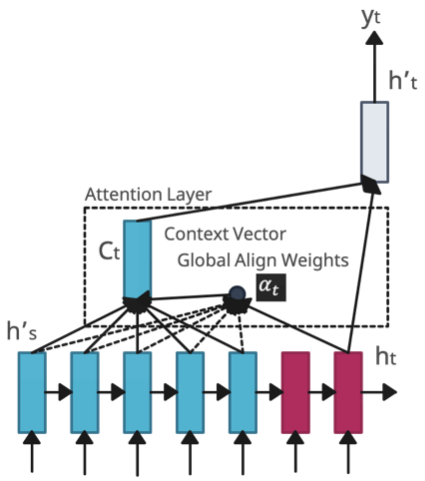


Fig. 2. Global attention model.

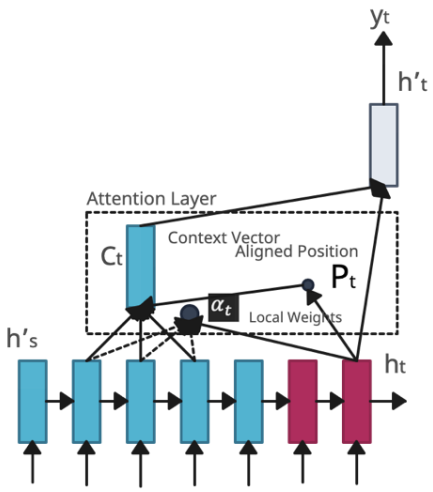


Fig. 3. Local attention model.

The global attention considers each HS coding for the excitation of the context vector. However, focusing global attention on all the main collateral terms of all destiny words is computationally expensive. In addition, it is not valid to use long phrases. To address this limitation, we can employ local attention to focus only on a small, relevant subset of the image features (HS) for generating each word in the caption. The proposed framework consists of an encoder- decoder model with a Deep Visual Prediction Model (DVPM), that transforms an input image (IMG) into a series of encoded expressions and words, T = [T1, T2, …, TL], with Ti ∈ RM, depicting the image, where L is the rendered caption's size, and M is the terminology size. The architecture of the proposed framework is illustrated in Fig. 6. The proposed framework is divided into four main phases. The first phase is the encoder using WCNN. The second phase is DVPM. The third phase is the semantic relation extractor. The final phase is the decoder using the LSTM model.

A. Encoder Phase

Comprising the WCNN model, the encoder merges two tiers of different wavelet decomposition alongside convolutional neural network layers to extract the image's visual characteristics, as illustrated in Fig. 4. The Level 1 and Level 2 features obtained from the CNN layers are bilinearly downsampled and fused into a 32x32x960 feature map. In the first phase, the input image (I) is first resized to 256x256 dimensions. The image is then separated into RGB color components. Each color component is decomposed into specifics and approximations using low-pass (LP) and high-pass (HP) discrete wavelet filters. The implementation of dual-phase discrete wavelet decomposition generates {LP, LF}, {HP, LF}, {LP, HF}, and {HP, HF} sub-bands, where LF and HF represent the low-frequency and high-frequency sections of the input image, respectively. In the second phase, only the {LP, LF} sub-band encounters further disassembly for each of the three elements. These components are combined and fused at every tier with the initial dual CNN stage outputs, encompassing four layers featuring numerous convolutional and pooling layers with a 2x2 kernel dimension, as shown in Fig. 5.

Table II offers detailed information on the different convolutional layers. By incorporating the DWT stage alongside CNN, we aim to improve the visual modeling of the input image and extract some unique spectral characteristics. This method assists in capturing finer details of objects, including spatial orientation and color information, which allows for the identification of visually salient features or regions within the image. These features draw more attention, much like the human visual system, due to their distinct characteristics compared to other areas.

B. DVPM Phase

Extracting semantic attributes from input image feature maps, including aspects like an object's scale, shape, and texture features, is crucial. Differences in these characteristics within an image can create obstacles to accurate identification or recognition. To generate a semantic feature map of dimensions 32x32x256, four multi-receptive filters are employed: one consisting of 64 filters with a 1x1 kernel size and the other three featuring 3x3 kernel sizes, each containing 64 filters with dilation rates of 3, 5, and 7, respectively. An example of attention changes to reflect the relevant parts of the image is shown in Fig. 7.

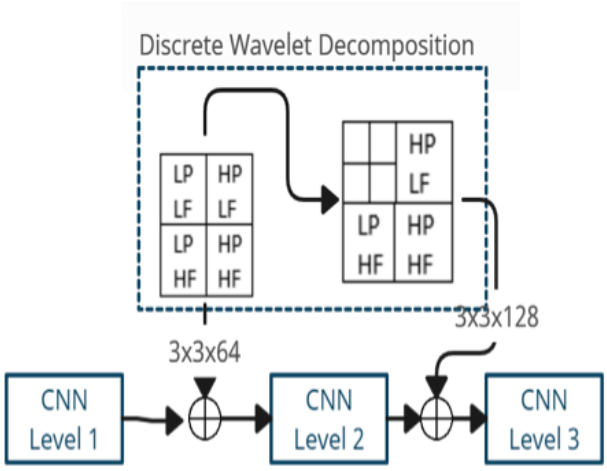


Fig. 4. Proposed encoder using WCNN.

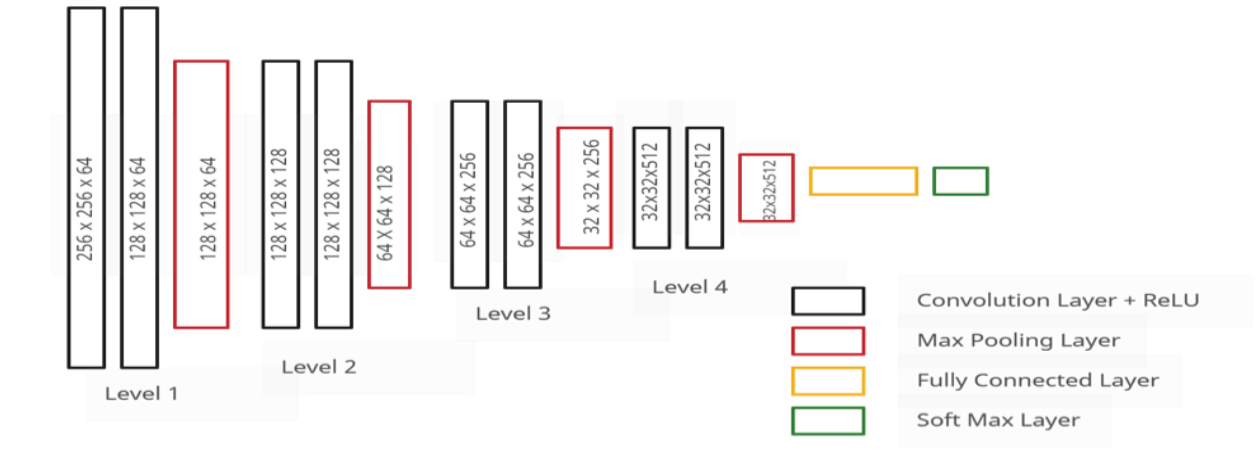


Fig. 5. Proposed CNN architecture.

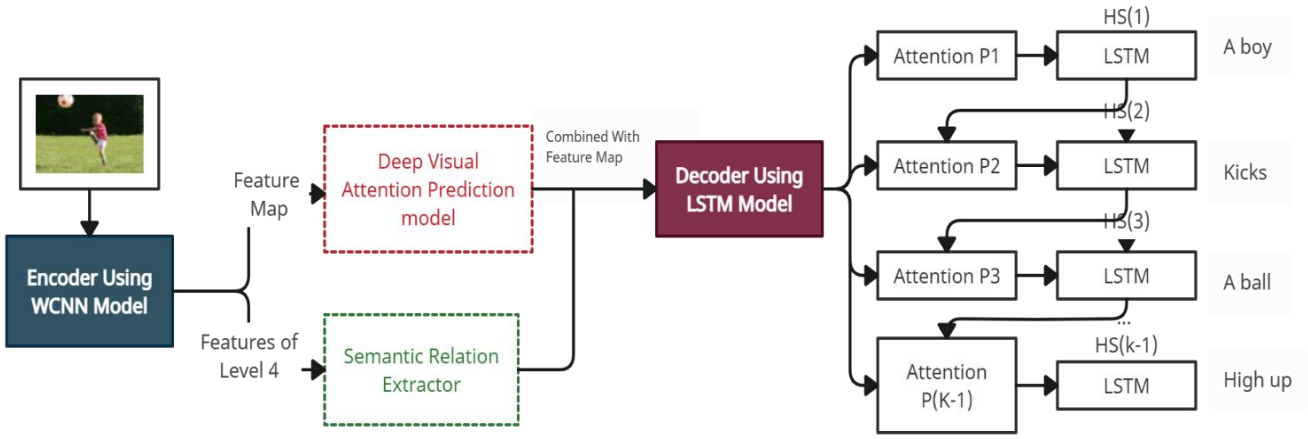


Fig. 6. Proposed framework.

TABLE II. SUMMARY OF RECENT PREVIOUS RELATED WORKS

Levels	Name	Kernel Size	Filter Size	Output Size
L1	Convolution L1,1 Convolution L1,2 Max Pool L1,1	3x3 3x3 2x2	64 64 64	256x256x64 256x256x64 128x128x64
L2	Convolution L2,1 Convolution L2,2 Max Pool L2,1	3x3 3x3 2x2	128 128 128	128x128x128 128x128x128 64x64x128
L3	Convolution L3,1 Convolution L3,2 Max Pool L3,1	5x5 5x5 2x2	256 256 256	64x64x256 64x64x256 32x32x256
L4	Convolution L4,1 Convolution L4,2	7x7 7x7	512 512	32x32x512 32x32x512

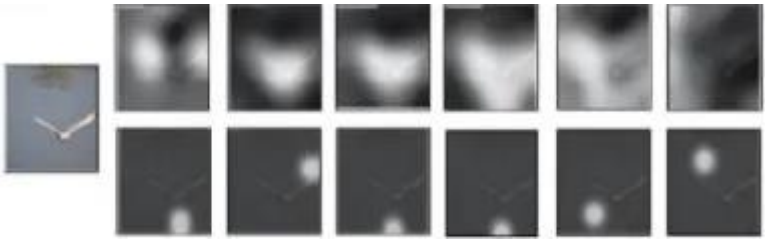


Fig. 7. Example of attention changes over time.

C. Semantic Relation Extractor Phase

Generating rich image translations requires leveraging the contextual spatial relationships between multiple objects in the image and their semantic details. In the WCNN model, the ultimate tier's feature map serves as input for the Semantic Relation Extractor (SRE) to identify object regions in the image. Subsequently, objects are paired, and numerous uniformly sized 32x32 sub-images are created through resizing. Each sub-image is fed into the CNN layers that contain 64 filters, with each filter having a receptive field of 3x3. This process generates features that represent the spatial relationships between pairs of objects. The feature maps of individual objects are combined to form a 32x32x64-dimensional contextual spatial relation feature map. This feature map is then merged with the output from a feed-forward neural network of local attention (fa) and supplied to the LSTM to produce the next word in the caption.

D. Decoder Phase

To generate more precise image captions, the channel attention weights W_c and spatial attention weights W_s are computed based on $H_{t-1} \in R^n$. D is the dimension of the hidden state. By using this method, additional contextual data is integrated into the image while generating captions. The feature map, denoted as F_{map} has dimensions $F_{map} \in R^{h \times w \times c}$ where h, w, and c represent the height, width, and a total number of channels of the feature map, respectively. The initial step involves average pooling on a per-channel basis, resulting in a channel feature vector, $F_v \in R^c$.

Since global attention focuses on all the words of the secondary origin of all objective words, This process becomes expensive and impractical for translating lengthy sentences. So instead, local attention concentrates on a small subset of the encoder's hidden states for each target word to address this limitation. So, we do softmax to get the input probability distribution of the channel attention weights (W_c). W_c can be calculated as follows:

$$W_c = Softmax\ (E_{it}) \quad (1)$$

$$E_{it} = X_c\ (P_{t-1}, H_i) \quad (2)$$

$$X_c = V_{att}^T * \tanh(U_{att} *\ H_i +\ W'_c *\ P_t) \quad (3)$$

Where:

- E_{it} means at every t^{th} time steps of decoder, how important i^{th} is the pixel location in the input image.
- P_{t-1} is the pervious state of decoder.
- H_i is the state of encoder.
- X_c is simple feed forward neural network which is a linear transformation of input ($U_{att} *\ H_i +\ W'_c * P_t$) and then a non-linearity (tanh) on the top of that.
- $V_{att}^T \in R^D$, $U_{att} \in R^{K*D}$, $W'_c \in R^{K*D}$, $H_i\ \in\ R^D$ and $P_t \in\ R^D$.

Now, we need to feed weighted sum combination to decoder. So, the weighted sum of input (context vector C_t) is calculated from Eq. (4).

$$C_t = \sum_{i=1}^{T} W_c\ H_i\ such\ that\ \sum_{i=1}^{T} W_c = 1 \quad (4)$$

E. Summary

In this part, we summarize the steps of the proposed system and link them to the proposed algorithms.

1) Clean data (as discussed in algorithm-I), i.e., clearing punctuations and numeric values from the text.

2) Preprocessing the images and captions (as discussed in algorithm-II and algorithm-III, respectively, by appending '<start>' and '<end>' labels to every caption) so that the proposed model understands the starting and stopping of each caption.

3) We have to reshape every image before feeding it to the WCNN model.

4) The captions will be tokenized, and a vocabulary of words in our data corpus will be established.

5) Producing Encoder Hidden States — The encoder employs a WCNN model that integrates dual-level discrete wavelet decomposition with CNN layers, efficiently extracting an image's visual features.

6) Applying DVMP to output the semantic feature map of size 32×32×256, we use four multi-receptive filters.

7) Applying SRE to find the object regions in the image by entering the feature map from the last level of the WCNN model.

8) As described in Algorithm 4, the RNN Local Decoder utilizes the hidden state (HS) from the previous decoder and the current decoder output.

The Decoder RNN processes these inputs to generate a new hidden state.

9) The alignment scores are calculated as in algorithm IV.

10)Softmaxing the previous scores.

11)A context vector is calculated.

12)The context vector is merged with the decoder's hidden state (HS), produced in Step 8, resulting in a new output.

13)Steps 6 through 13 are iteratively executed for each time step in the decoder until a token is generated.

Algorithm I - Data Cleaning

Input: Original Text (OT)
Output: Cleaned Text (CT)

```
Start Procedure
  OT ← Original Text
  CT ← null
  CT ← OT.translate(string.punctuation)
  TL ← txt_length_more_than_1
  TL ← null

  Foreach word in CT.split():
    IF length(word) >1:
      TL + = " " + word
    End IF
  End Foreach
End Procedure
```

Algorithm II - Image Preprocessing

Input: Data
Output: IMG

```
Start Procedure
  [ ] ← IMG_vector
  Foreach fnames in data["filename"]:
    path ← img_dir + "/" + fnames
    all_img_name_vector.append(path)
    IMG ← tf.io.read(path)
    IMG ← tf.img.decode_jpg(IMG, ch=3)
    IMG ← tf.img.resize(IMG, (224,224))
  End Foreach
End Procedure
```

Algorithm III - Caption Preprocessing

Input: Data
Output: Total Captions (t_cp)

```
Start Procedure
  [ ] ← t_cp
  Foreach cp in data["cp"] astype(str):
    cp ← '>start<' + cp +'>end<'
    t_cp.append(cp)
  End Foreach
End Procedure
```

Algorithm IV - RNN Local Decoder

Input: units, vocab_size, features map (features) and hidden
Output: state, attention weights (att_w), Context Vector (CV)

```
Start Procedure
  Uatten ← tf.keras.layers.Dense(units)
  Wc ← tf.keras.layers.Dense(units)
  Vatten ← tf.keras.layers.Dense(1)
  hidden_time_axis ←tf.expand_dim(hidden, 1)
  score ←use equation 3
  att_w ← tf.softmax(score,axis=1)
  att_w← use equation 1
  CV ← attention_weights * feature
  CV ← use equation 4
End Procedure
```

IV. Experimental Results Analysis and Discussion

This section will discuss and compare the outcomes of our diverse experiments with pertinent prior studies. We implemented the proposed framework using TensorFlow 2.3 and executed it on Google Cloud with the help of Google Colab.

A. Description of Datasets

Numerous open-source datasets, including Flickr 8k, Flickr 30k, and MS COCO, are accessible for this research. The experiments are carried out on the following three benchmark datasets:

- Flickr 8k — 6400 images (training set), 700 images (validation set), and 700 images (testing set).
- Flickr 30k— 24k images (training set), 3k images (validation set), and 3k images (testing set).
- MS-COCO — 128k images (training set), 16k images (validation set), and 16k images (testing set).

B. Hyperparamters

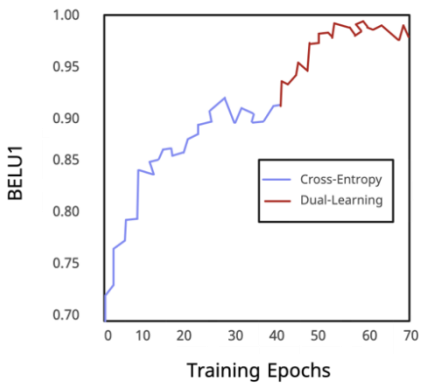
During the model's training stage, we used hyper-parameter settings such as batch size, dropout, etc. The values of a few hyperparameters include the exponential decay rates for ADAM optimizer, learning rate, batch size, and dropout. The number of iterations used is 50. These hyperparameters are changed on a trial-and-error. Finally, the hyperparameters are tuned into our method to improve the results.

- For Flickr8k, Flickr30k, and MSCOCO datasets, *the batch sizes* employed are 16, 32, and 64, respectively.
- *Dropout*: To prevent overfitting, a dropout rate of 0.2 is applied, L2 regularization, and a weight decay value of 0.001.
- *Epochs*: The model begins with 40 epochs of training based on cross-entropy loss. Afterward, an extra 80 epochs of fine-tuning are conducted via dual learning to reach the highest CD score within the validation set.

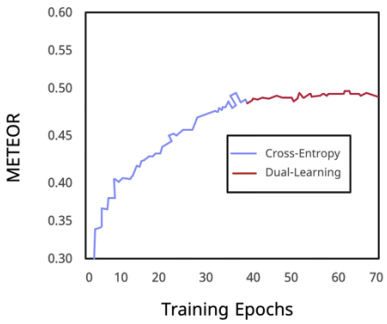
C. Evaluation Metrics

Our experiments use the performance evaluation metrics – BLEU as B score from 1 to 4, METEOR as MR score, and CIDEr as CD score. The BLEU metric is employed to assess the generated captions for the test set. Recognized as a reliable metric, BLEU quantifies the similarity between a single predicted sentence and multiple reference sentences. Table III provides a summary of the metrics featured in this paper. Additionally, the Beam search technique was utilized to evaluate the captions.

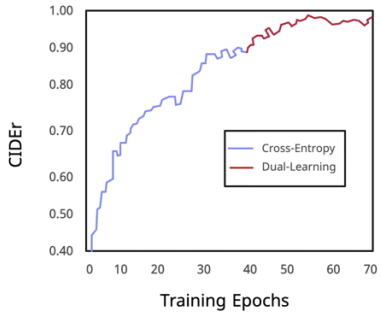
D. Results



(a) BELU-1



(b) METEOR



(c) CIDEr

Fig. 8. Learning curves of the proposed framework according to each metric.

Fig. 9. Samples of RNN-generated captions on test images.

E. Comparison Between the Proposed Model and the Previous Related Works

This sub-section displays the evaluation outcomes for the comparative analysis between the proposed model and the related previous works on different datasets.

TABLE III. COMPARISON BETWEEN THE PROPOSED MODEL AND RELATED WORKS ON THE FLICKER 8K DATASET

Ref.	B-1	B-2	B-3	B-4	MR	CD
7	64.7	45.9	31.7	21.5	20.8	-
27	57.9	38.3	24.5	16	-	-
19	67.2	45.9	31.4	21.3	20.4	-
16	68.1	49.3	34.9	25.7	22.6	52.6
Our Model	73.4	52.3	36.9	29.2	27.3	68.4

TABLE IV. COMPARISON BETWEEN THE PROPOSED MODEL AND RELATED WORKS ON THE FLICKER 30K DATASET

Ref.	B-1	B-2	B-3	B-4	MR	CD
7	64.6	44.8	30.7	20.5	17.8	-
27	57.3	36.9	24.1	15.7	15.3	-
19	66.9	43.9	29.6	19.8	18.5	-
16	66.2	46.7	32.5	22.3	19.6	-
30	-	-	-	24.9	20.9	59.7
Our Model	72.2	50.3	35.7	27.4	21.9	66.8

TABLE V. COMPARISON BETWEEN THE PROPOSED MODEL AND RELATED WORKS ON THE MS-COCO DATASET

Ref.	B-1	B-2	B-3	B-4	MR	CD
7	67	49.2	35.7	26.3	22.6	80.3
27	62.7	45.3	32.3	23.4	20.2	66.2
19	71.9	50.8	35.8	25.1	23.1	-
31	-	-	-	37.2	28.4	123.4
32	78.9	62.9	48.9	36.2	27.8	121.1
Our Model	79.8	63.4	50.1	39.2	28.8	123.9

F. Discussion

As shown in Fig. 8 and Fig. 9, after about 40 epochs, all the evaluation metrics converge, and the performance of the proposed model evolves better when we fine-tune the model on the unpaired data by employing the dual learning mechanism. The results comparing the proposed model on the Flickr8K and Flickr30K datasets are presented in Table III and Table IV. As seen in Table III, the proposed model shows notable improvements of 2.3%, 2.8%, and 2.1% in B-1, B-4, and MR scores for the Flickr8K dataset. Likewise, the model achieves increases of 2.4% and 0.9% in B-4 and MR scores for the Flickr30K dataset in Table IV. The model also attains a respectable CD value of approximately 66.8. Table V displays the evaluation outcomes for the comparative analysis on the MSCOCO test partition. As indicated in Table IV, the proposed model yields a strong CD score of 123.9 and exhibits relative enhancements of around 0.9%, 0.5%, and 0.7% in B-4, MR, and CD scores, respectively. We can use the proposed model in IoT systems in [33],[34] to ensure controllability, safety and effectiveness as a future work. Unlike other methods, this improvement stems from the proposed model's image feature maps incorporating spectral information alongside spatial and semantic details. The model

can obtain detailed data during object identification by integrating discrete wavelet decomposition into the CNN model. Additionally, the model considers the contextual spatial relationships between objects in the image and employs spatial and channel-specific attention to enhance feature maps resulting from convolution. Using multi-receptive field filters facilitates the detection of visually prominent objects with diverse shapes, scales, and sizes.

V. CONCLUSION

This manuscript introduces a deep visual attention framework for image caption generation, utilizing an encoder-decoder architecture based on semantic relationships. The encoder comprises a WCNN, while the decoder comprises a DVPM and LSTM. The DVPM calculates channel and location attention for visual features, taking into account the spatial-contextual relationship between various objects. Merging wavelet decomposition with the convolutional neural network allows the model to extract spatial, semantic, and spectral data from the input images. In-depth image captions are produced by applying spatial and channel-wise attention to the feature maps generated by DVPM and considering the contextual spatial relationships among objects via the CSE network. Assessments are conducted on three standard datasets—Flickr8K, Flickr16K, and MS-COCO—utilizing evaluation metrics such as BLEU, METEOR, and CIDEr. With the MS-COCO dataset, the model achieves remarkable B-4, MR, and CD scores of 39.2, 28.8, and 123.9, respectively. We believe there are several promising directions for future work. First, the proposed model could be refined and tested with various other attention mechanisms, potentially improving the model's performance even further. Second, the application of this model to other vision-and-language tasks, such as visual question answering and image-based storytelling, IoT systems could be explored. Additionally, the integration of other types of contextual information, such as object-object interaction or more explicit spatial information may enhance the model's ability to generate even more detailed and accurate captions. Finally, while the model has been tested on standard datasets, it would be worthwhile to evaluate its performance on a diverse array of real-world images and scenarios. These future research directions will help to further reinforce and extend the significant contributions of our study.

REFERENCES

[1] M. al Sulaimi, I. Ahmad, and M. Jeragh, “Deep Image Captioning Survey: A Resource Availability Perspective,” Conference of Open Innovation Association, FRUCT, vol. 2021-May, pp. 3–13, May 2021.

[2] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj, and R. K. Mishra, “Image Captioning: A Comprehensive Survey,” 2020 International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control, PARC 2020, pp. 325–328, Feb. 2020.

[3] S. Sukhi, A. Q. Ohi, M. S. Rahman, and M. F. Mridha, “A Survey on Bengali Image Captioning: Architectures, Challenges, and Directions,” 2021 International Conference on Science and Contemporary Technologies, ICSCT 2021, 2021.

[4] M. El-Gayar, H. Soliman and N. Meky, "A comparative study of image low level feature extraction algorithms", Egyptian Informat. J., vol. 14, no. 2, pp. 175-181, 2013.

[5] M. M. El-Gayar, N. E. Mekky, A. Atwan and H. Soliman, "Enhanced search engine using proposed framework and ranking algorithm based on semantic relations", IEEE Access, vol. 7, pp. 139337-139349, 2019.

[6] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, “From Show to Tell: A Survey on Deep Learning-based Image Captioning,” IEEE Trans Pattern Anal Mach Intell, Jan. 2022.

[7] A. Hani, N. Tagougui, and M. Kherallah, “Image caption generation using a deep architecture,” Proceedings - 2019 International Arab Conference on Information Technology, ACIT 2019, pp. 246– 251, Dec. 2019.

[8] C. S. Kanimozhiselvi, V. Karthika, S. P. Kalaivani, and S. Krithika, “Image Captioning Using Deep Learning,” 2022 International Conference on Computer Communication and Informatics, ICCCI 2022, 2022.

[9] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, “Guiding Long-Short Term Memory for Image Caption Generation,” Sep. 2015.

[10] S. Wang et al., “Cascade attention fusion for fine- grained image captioning based on multi-layer LSTM,” ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2021-June, pp. 2245–2249, 2021.

[11] L. Gao, X. Wang, J. Song, and Y. Liu, “Fused GRU with semantic-temporal attention for video captioning,” Neurocomputing, vol. 395, pp. 222– 228, Jun. 2020.

[12] P. Shah, V. Bakrola, and S. Pati, “Image captioning using deep neural architectures,” Proceedings of 2017 International Conference on Innovations in Information, Embedded and Communication Systems, ICIIECS 2017, vol. 2018-January, pp. 1–4, Jan. 2018.

[13] A. Hani, N. Tagougui, and M. Kherallah, “Image caption generation using a deep architecture,” Proceedings - 2019 International Arab Conference on Information Technology, ACIT 2019.

[14] I. Hrga and M. Ivašic-Kos, “Deep image captioning: An overview,” 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2019 - Proceedings, pp. 995–1000, May 2019.

[15] A. Karpathy and L. Fei-Fei, “Deep Visual-Semantic Alignments for Generating Image Descriptions,” IEEE Trans Pattern Anal Mach Intell, vol. 39, no. 4, pp. 664–676, Dec. 2014.

[16] M. Yang et al., “Multitask learning for cross- domain image captioning,” IEEE Trans Multimedia, vol. 21, no. 4, pp. 1047–1061, Apr. 2019.

[17] Z. J. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu, “Context-Aware Visual Policy Network for Fine- Grained Image Captioning,” IEEE Trans Pattern Anal Mach Intell, vol. 44, no. 2, pp. 710–722, Jun. 2019.

[18] L. Yu, J. Zhang, and Q. Wu, “Dual Attention on Pyramid Feature Maps for Image Captioning,” IEEE Trans Multimedia, vol. 24, pp. 1775–1786, 2022.

[19] L. Yang, H. Wang, P. Tang, and Q. Li, “CaptionNet: A Tailor-made Recurrent Neural Network for Generating Image Descriptions,” IEEE Trans Multimedia, vol. 23, pp. 835–845, 2021.

[20] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, “Meshed-Memory Transformer for Image Captioning,” Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 10575–10584, Dec. 2019.

[21] X. Li et al., “Oscar: Object-Semantics Aligned Pre- training for Vision-Language Tasks,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12375 LNCS, pp. 121–137, Apr. 2020.

[22] W. Jiang, X. Li, H. Hu, Q. Lu, and B. Liu, “Multi- Gate Attention Network for Image Captioning,” IEEE Access, vol. 9, pp. 69700–69709, 2021.

[23] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-critical Sequence Training for Image Captioning,” Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017-January, pp. 1179–1195, Dec. 2016.

[24] H. Wang, H. Wang, and K. Xu, “Evolutionary recurrent neural network for image captioning,” Neurocomputing, vol. 401, pp. 249–256, Aug. 2020.

[25] K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," 32nd International Conference on Machine Learning, ICML 2015, vol. 3, pp. 2048–2057, Feb. 2015.

[26] 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, Jul. 2020.

[27] M. A. Al-Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," J Big Data, vol. 9, no. 1, pp. 1–16, Dec. 2022.

[28] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 6077–6086, Jul. 2017.

[29] Z. Deng, Z. Jiang, R. Lan, W. Huang, and X. Luo, "Image captioning using DenseNet network and adaptive attention," Signal Process Image Commun, vol. 85, p. 115836, Jul. 2020.

[30] H. Wang , H. Wang, and K. Xu, "Evolutionary recurrent neural network for image captioning," Neurocomputing, 401:249–56, 2020.

[31] 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, 23:2413–27, 2021.

[32] S. Wang, Y. Meng, Y. Gu, L. Zhang, X. Ye, J. Tian, L. Jiao, "Cascade attention fusion for fine-grained image captioning based on multi-layer lstm," 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2245–2249, 2021.

[33] H. Fetooh, M.M.El-Gayar, A. Aboelfetouh, "Detect Technique and Mitigation Against a Phishing Attack". Int J Adv Comput International Journal of Information ManagementSci Appl. The Science and Information (SAI) Organization, 12:177–88, 2021.

[34] N.A. Hikal, M.M. El-Gayar, "Enhancing IoT botnets attack detection using machine learning-IDS and ensemble data preprocessing technique", Lect Notes Networks Syst, Springer, 114:89–102, 2020.