ICIC Express Letters
Part B: Applications
Volume **16**, Number **1**, January **2025**

ICIC International ©2025 ISSN 2185-2766
pp. **73–82**

ENHANCING IMAGE CAPTION PERFORMANCE WITH IMPROVED VISUAL ATTENTION MECHANISM

Zagon Bussabong[1], Praetawan Jarutan[2], Emmanuel Okafor[3] and Olarik Surinta[1,*]

[1]Multi-agent Intelligent Simulation Laboratory (MISL) Research Unit
Department of Information Technology
Faculty of Informatics
Mahasarakham University
Khamriang Sub-District, Kantarawichai District, Mahasarakham 44150, Thailand
64011293003@msu.ac.th; *Corresponding author: olarik.s@msu.ac.th

[2]Computer Department
Science and Technology Faculty
Sakon Nakhon Rajabhat University
680, Nittayo Rd., That Choeng Chum Sub-District, Mueang District, Sakon Nakhon 47000, Thailand
praetawan9925@snru.ac.th

[3]SDAIA-KFUPM Joint Research Center for Artificial Intelligence
King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia
emmanuel.okafor@kfupm.edu.sa

Received April 2024; accepted June 2024

Abstract. *Image captioning analyzes and translates images into text, requiring extensive data and often facing challenges in comprehending the diverse contents of images during text generation. This research enhances image captioning using a visual attention mechanism to improve image-to-text translation performance. We propose a neural network architecture comprising an encoder, decoder, and beam search. The encoder uses either dual convolutional neural networks (Dual-CNN) or a single CNN to extract visual features, which are then passed to the decoder. The decoder employs long short-term memory (LSTM) to learn temporal and sequential patterns, converting visual features into output probabilities. The resulting outputs are then processed by the beam search algorithm to generate the best captions. Three experiments were conducted. First, single CNN architectures (ResNet-101, EfficientNet-B0, and ResNeXt-101) were evaluated with visual attention mechanisms on the Flickr8K dataset using BLEU scores. ResNet-101 achieved the highest performance. Second, three Dual-CNNs combined with attention mechanisms were tested, with ResNet-101 and EfficientNet-B0 outperforming other combinations. Third, early stopping was used to determine the optimal training epoch, revealing that the Dual-CNN with visual attention mechanism yielded the best results. The proposed framework, tested on the Flickr8K dataset, achieved BLEU scores of 68.76%, 49.15%, 35.46%, and 24.71% in different scenarios, demonstrating superior performance compared to other approaches.*

Keywords: Image caption, Visual attention mechanism, Convolutional neural network, Dual-CNN, Beam search

1. **Introduction.** An image caption is a technology at the intersection of computer vision and natural language processing (NLP). It is used for generating a sentence that describes the content information in an image [1]. However, this technology is very challenging due to the complexity involved in learning spatial and semantic features from images and

DOI: 10.24507/icicelb.16.01.73

then creating a descriptive text distribution. Previous research has demonstrated the application of these techniques (attention learning schemes, graph models, hybridization of attention and graph, convolutional neural network, and transformer models) for approaching image captioning tasks. Furthermore, it is pertinent to point out that many researchers in the field of language modeling and machine translation have successfully applied various attention-learning mechanisms [2] to improve neural network generative and translative capabilities. A broader aspect of this article will investigate the impact of ensemble feature extraction integrated with attention learning mechanisms for performing image captioning.

The investigation by Zhou et al. [3] successfully integrated NLP, computer vision, and visual attention mechanisms to generate captions that describe images. Their study demonstrated the potential of hybridizing ResNet-101 with a residual attention block for spatial and global feature extraction. The resulting features were then passed to an attention learning scheme with two layers of a long short-term memory (LSTM) network, which generated sentence sequences. The probability of text distribution was computed using the softmax function. Furthermore, encoder-decoder models were developed by training different encoding schemes based on convolutional neural network (CNN) models, such as InceptionV3, DenseNet-169, ResNet-101, and VGG16, to extract rich feature information from input images [4]. The resulting features from the encoder were transmitted to the decoder, which utilized a gated recurrent unit to generate text sequence captions. Their experimental results showed that the InceptionV3-based encoder-decoder model outperformed other CNN models. Cornia et al. [5] presented an M2 transformer built from a memory-augmented encoder and a meshed decoder for creating captions. The M2 transformer has the significant advantage of learning from the relationships between different components of an image through multi-level encoding at scale. To highlight specific objects within an image, certain models employ the concept of focal points. Focal points direct the caption generation process to prioritize an object as the central subject of the description. In the research by Yanagimoto and Hashimoto [6], focal points in the image were determined and directly linked to the image features extracted by a VGG16 network. Their technique highlights the potential for guiding caption generation towards specific objects of interest.

Moreover, a novel image caption method using a graph attention network with global context has been developed, significantly enhancing caption quality by integrating both local and global image features. This method utilizes a grid feature interaction graph and a transformer-based decoder, achieving a 133.1% CIDEr score on the Microsoft COCO dataset. It effectively produces detailed and context-aware captions [7]. Additionally, Chen et al. introduced the Abstract Scene Graph (ASG2Caption) model, which employs abstract graphs to control the granularity of generated captions, enhancing intention-aware caption diversity. Tested on VisualGenome and MSCOCO, this model has demonstrated improved controllability over traditional methods [8].

Many studies have shown improvements in generating image descriptions. However, creating accurate and detailed descriptions of complex scenes involving multiple objects or unusual relationships is still a problem. This research aims to address these gaps by using a dual convolutional neural network architecture within the encoder. This approach allows for a richer and more nuanced extraction of visual features, potentially improving the accuracy of the generated captions.

This research investigates the development of an enhancement encoding scheme and training of visual attention mechanisms that are integrated into a caption generator, as illustrated in Figure 1, with the objective of performing image captioning. The visual attention mechanisms encompassed encoder and decoder architectures, whose probabilistic outputs from the decoder were passed to a beam search algorithm that determined the best sequential text description of the input images that were fed to the visual attention

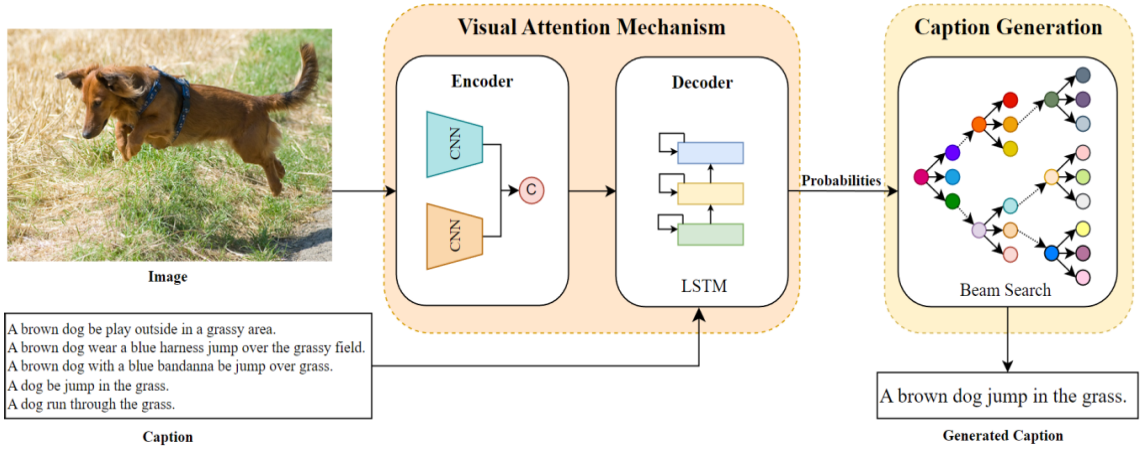

FIGURE 1. Illustration of the proposed visual attention mechanism framework

systems. To actualize enhancement in the encoder, we propose a Dual-CNN that concatenates two CNN towers to generate rich image feature representations. In the investigated image captioning models, within the encoders, we compared Dual-CNNs with a single CNN for performing feature extraction and then the outputs were fed to the decoder that was built from the LSTM network that was used for learning long-term sequence of visual features. Finally, the beam search (caption generation technique) was employed to search for the optimal caption from various potential paths. The modified visual attention mechanism based on our Dual-CNN outperformed the original visual attention mechanism using a single CNN.

The remaining components of this article are as follows. Section 2 provides a brief explanation of the concepts of visual attention mechanism and caption generation. The description of the experimental setups and significant results for the various techniques are reported in Section 3. A brief description of lessons learned from the experimental investigation is reported in Section 4. Finally, a summary of the article and the possible directions for future works are described in Section 5.

2. Visual Attention Mechanism Framework.

In this section, we implement the visual attention mechanism framework to generate the caption, as shown in Figure 1. The details of the proposed visual attention mechanism framework are presented herein.

2.1. Attention mechanism.

The visual attention mechanism is a system that integrates encoder and decoder modules. The encoder module is designed based on the end goal of utilizing convolutional operators in a single CNN [9] or Dual-CNN [10] for performing feature extraction from input images. Then, the effective feature maps from the encoder module are passed to a decoder module which employs a recurrent neural network based on LSTM for learning sequential pattern abstractions. Further details of the attention mechanism are described as follows.

Encoder architecture. We implemented Dual-CNN and single CNN to encode the visual features from the input image. In the single CNN, we trained three well-known and successful CNN models on image captioning tasks: ResNet-101, EfficientNet-B0, and ResNeXt-101. While for the Dual-CNN (ResNet-101+EfficientNet-B0, ResNet-101+ResNeXt-101, and EfficientNet-B0+ResNeXt-101) architecture, the resulting outputs for each of the CNN were merged with the aid of a concatenation operator. Consequently, the resulting feature maps were passed to the decoder architecture. Hence, the decoder architecture could learn the long-term dependencies or patterns of the visual features using the LSTM network. We further provide insights about the concatenation operation. This was done by using a fusion strategy to merge the visual features of two CNN architectures through

a concatenation operation. Using the concatenation technique to merge two CNN models has previously been demonstrated to enhance predictive performance [11]. The feature maps of the last layer are transformed to a varying dimension based on the nature of the CNN. As a result, concatenating visual features from two CNN models increases the size of the feature maps.

a) *ResNet-101.* A typical ResNet101 is a kind of CNN in which the feature representation of the deep learning is computed using skip connections which involves merging identity mappings with the corresponding neural network layer outputs. ResNet architecture [12], has a hierarchical structure with multiple stages, each containing a set of residual units. The network starts with a convolutional layer, followed by a max pooling operation to reduce the spatial dimensions. Then, it goes through four stages, each composed of multiple residual units. The number of residual units in each stage may vary, but in ResNet101, the first stage has three units, and subsequent stages have 4, 23, and 3 units, respectively. Hence, the last convolutional layer feature maps are transferred to the average pooling layer, and then passed to the fully connected layer. These characteristics make ResNet-101 a robust choice for extracting intricate features from images, which is crucial for generating accurate captions. According to a comparison study, ResNet-101 outperforms VGG in image captioning tasks, achieving higher BLEU-4 scores and demonstrating faster convergence during training [13].

b) *ResNeXt-101.* The ResNeXt architecture is the extended version of the ResNet architecture, which applies the principle of cardinality [14], a paradigm that allows several independent pathways or groups within a network layer. Each pathway comprises a set of convolutional filters that extract different features from the input data. A typical ResNeXt can be described as a multi-branch architecture inside the residual block. In our work, we employed similar numbers of convolutional blocks as in ResNet-101, but the sizes of its feature maps were double. ResNeXt models can be scaled to different complexities and sizes, enabling them to adapt to various tasks and datasets. The network depth is often modified by stacking multiple residual blocks together. The effectiveness of ResNeXt-101 in generating high-quality image captions has been demonstrated, showing superior performance metrics such as Top-5 Accuracy and BLEU-4 compared to other models [15].

c) *EfficientNet-B0.* EfficientNet-B0 is a CNN architecture and is the baseline model of the EfficientNet model variants. This technique was earlier proposed by [16] and is considered as the least computationally expensive variant. A typical architecture of an EfficientNet-B0 relies on the principle of compound scaling that optimizes the depth and width of a neural network architecture. Moreover, this method involves the stacking of convolutional layers, followed by batch normalization, activation functions, and pooling operations. The EfficientNet-B0 network model involves a combination of skip connections and squeeze-and-excitation blocks, which aid in improving information flow. According to a study on efficient image captioning methods, EfficientNet-B0 demonstrated a balance between computational efficiency and performance, making it a suitable choice for tasks requiring efficient yet effective image feature extraction [17].

Decoder architecture. We employed an LSTM to decode the long-term sequence of visual features. The benefit of the LSTM architecture is that the model can link previous information to inform the understanding of the present information. To handle long sequence data, the LSTM contains a memory cell and three gating mechanisms: an input gate, an output gate, and forget gate [18]. Consequently, LSTMs are effective at learning sequential temporal data by retaining key feature representations and filtering out irrelevant information.

2.2. **Caption generation.** We investigated caption generation using a well-known algorithm (beam search); a beam search is widespread due to its ability to generate precise captions. Beam search is a greedy search algorithm that explores a graph of potential

paths from the most profitable node by considering the width of the beam, called beam width [19]. In this research, the beam search algorithm employed the output probabilities after applying the softmax function to generate the best optimal caption. The beam search utilizes the following processes. The beam width (k) parameter is set in the first process and if $k = 2$, then the beam search selects two alternative words as the candidate sequence (called the first position). In the second process, the two best words, which are the two highest probabilities in the first position, are selected and found in the next position. In the third process, the algorithm executes the model twice to generate probabilities of the first and second words by fixing the potential words in the second position. Fourth, the algorithm repeats processes 2 and 3 by constraining the first two positions with the other highest probabilities. Finally, the algorithm computes the most efficient position for the final words and selects the best caption with the highest probability.

3. **Experimental Setup and Results.**

3.1. **Flickr8K dataset.** The Flickr8K dataset consists of 8,000 images, partitioned into 6,000 training images, 1,000 validation images, and 1,000 test images [20]. The images were collected from the Flickr website with the permission of its creators under creative commons licenses. The Flickr8K dataset is a valuable resource in image caption research. Further, each image in the dataset has five unique captions generated by a different person to provide diverse descriptions. An illustration of examples of image captioning is shown in Figure 2.

FIGURE 2. Some examples of images and captions of the Flickr8K dataset

3.2. **Implementation setup.** We implemented the proposed method on the PyTorch deep learning framework. The experiments were conducted on a computer Intel(R) i5-13400F CPU, 32 GB of 3200 MHz RAM, and an RTX 4070Ti GPU with 12 GB RAM. We experimented with a visual attention mechanism framework with only a focus on the encoder mechanism. The experiments were divided into three sections. Firstly, we trained the encoder mechanism with three CNN architectures: ResNet-101, EfficientNet-B0, and ResNeXt-101, on the Flickr8K dataset to evaluate the accuracy performance of each CNN architecture using the BLEU score. Secondly, we experimented with the encoder mechanism by discovering the best combination of two CNNs. Then, two CNNs were combined using the concatenation operation [21] called the Dual-CNN encoder. In the last experiment, we explored a way to curb the overfitting problem. To prevent overfitting, we employed the early stopping strategy [22] to shorten the training process when the performance of the proposed model does not improve within a specific consecutive number of epochs (known as the patient) [23]. We then set the patience parameter to 10. We

trained 50 epochs on the proposed framework, using a learning rate of 1e-4 for the encoder mechanism and 4e-4 for the decoder mechanism. In addition, we utilized a beam search algorithm with a beam size of 3 during the experiments.

3.3. **Evaluation metric.** The bilingual evaluation understudy (BLEU) score was used as the performance evaluation metric to measure the similarity of the machine-translated text to a set of reference captions [24]. In the BLEU score, n defines the size of n-grams [6]. A higher BLEU score indicates a superior quality of the produced text [25]. We used BLEU-1, -2, -3, and -4 to evaluate the proposed framework and other approaches. The BLEU metric is computed as Equation (1).

$$\log(BLEU) = \min\left(1 - \left[\frac{l_r}{l_c}\right], 0\right) + \sum_{n}^{N} W_n \log(P_n) \tag{1}$$

where l_r and l_c represent the lengths of the reference and candidate texts from the machine translation. P_n measures how many n-grams precision in the candidate text relative to the reference text. The W_n is weight assigned to each n-gram precision [26]. Note that the component $\min\left(1 - \left[\frac{l_r}{l_c}\right], 0\right)$ accounts for the output from computing the logarithm of the brevity penalty.

3.4. **Experiments with single-based CNN encoder.** We trained visual attention mec-hanisms with three single-based CNN models: ResNet-101, EfficientNet-B0, and ResNeXt-101 evaluated on Flickr8K dataset. In this experiment, the models were trained for 20 epochs. The experimental results are presented in Table 1.

TABLE 1. The generated-caption performances (BLEU-1, -2, -3, and -4) of the visual attention mechanism that considered a single-based CNN encoder architectures evaluated on the Flickr8K dataset

CNN models	**BLEU scores (↑)**			
	BLEU-1 (%)	**BLEU-2 (%)**	**BLEU-3 (%)**	**BLEU-4 (%)**
ResNet-101	**63.72**	**42.37**	**28.73**	**18.87**
ResNeXt-101	62.92	41.66	28.08	18.13
EfficientNet-B0	60.93	40.01	26.81	17.75

The generated-caption performance of the visual attention mechanism using each of the single CNN encoder architectures evaluated on the used dataset yielded a BLEU-1 metric score $> 60\%$. According to the experiments, the performance in terms of BLEU-1 score for ResNet-101, ResNeXt-101, and EfficientNet-B0 was 63.72%, 62.92%, and 60.93%, respectively. Note that the highest BLEU scores indicate the best performance on the generated captions. However, the performance of the generated captions drops to approximately 42% for BLEU-2, 28% for BLEU-3, and 18% for BLEU-4. Overall, the visual attention mechanism with ResNet-101 as an encoder outperforms other visual attention models that inherently integrate single-based CNN architectures (others) during caption generation when evaluated on the Flickr8K dataset. Figure 3 shows the captions generated by ResNet-101, ResNeXt-101, and EfficientNet-B0.

3.5. **Experiments with Dual-CNN encoder.** Our experiments emphasized using the Dual-CNN encoder feature extraction schemes in the visual attention mechanism. We investigated three dual CNN encoders: ResNet-101+EfficientNet-B0, ResNet-101+ResNeXt -101, and EfficientNet-B0+ResNeXt-101. A summary of the visual attention mechanisms factoring Dual-CNN encoders is presented in Table 2. As shown in Table 2, the experimental results showed that the combination between ResNet-101 and EfficientNet-B0 achieved the best BLEU scores. The Dual-CNN encoder of ResNet-101+EfficientNet-B0

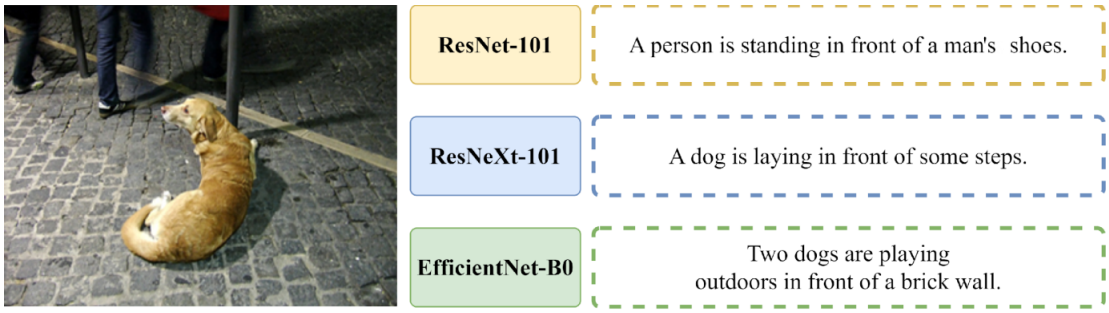

FIGURE 3. Illustration of caption generation using the visual attention mechanism with different CNN encoders

TABLE 2. The percentage of caption generation performance (BLEU-1, -2, -3, and -4) of the visual attention mechanism with Dual-CNN encoder evaluated on the Flickr8K dataset

Dual-CNN encoders	**BLEU scores (↑)**			
	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**
ResNet-101+EfficientNet-B0	**65.75**	**45.19**	**31.88**	**21.84**
ResNet-101+ResNeXt-101	65.64	44.59	31.05	20.97
EfficientNet-B0+ResNeXt-101	63.59	42.14	28.75	19.04

architecture achieved the following BLEU metric scores: 65.75% for BLEU-1, 45.19% for BLEU-2, 31.88% for BLEU-3, 21.84% for BLEU-4 surpassing all other visual attention mechanism integrated with different dual encoding schemes when evaluated on the same BLEU metrics. An interesting inference can be drawn from the experiments; a visual attention mechanism factoring Dual-CNN encoders significantly outperforms a visual attention mechanism factoring single CNN architectures yielding a significant difference of approximately 2% across the examined BLEU metrics.

3.6. **Experiments with different training epochs.** When training the deep learning model using a large number of images, the issue of overfitting is encountered. In this experiment, we trained the proposed framework using 20, 50, 100, 150, and 200 epochs. The experimental results are presented in Table 3.

TABLE 3. The generated-caption performances of visual attention mechanism with Dual-CNN encoder (proposed framework) at different numbers of training varying epochs

Epochs	**BLEU scores (↑)**			
	BLEU-1 (%)	**BLEU-2 (%)**	**BLEU-3 (%)**	**BLEU-4 (%)**
20	**65.75**	**45.19**	**31.88**	**21.84**
50	63.92	43.61	30.24	20.39
100	62.47	41.8	28.49	18.95
150	62.09	41.61	28.48	18.96
200	61.77	41.26	28.19	18.74

The experimental results reported in Table 3 show that training the visual attention mechanism with Dual-CNN encoders obtained its best performance at 20 epochs and experienced a decrease in performance across the BLEU metric scores at higher-order epochs. The performance disparity (BLEU metrics) of the proposed framework was marginal when trained for 100-200 epochs.

3.7. **Experiments with early stopping.** We examined the effectiveness of training the proposed visual attention mechanism framework with and without the early stopping method for 200 epochs. Therefore, the framework was trained over a duration of 20 hours for 200 epochs. Subsequently, we trained the proposed framework with the early stopping method and set the patience parameter of 10. The training process terminates if there is no improvement within ten consecutive epochs. Importantly, with the early stopping method, the proposed framework finished training within only 18 epochs. The best performance was attained at 8 epochs, as shown in Figure 4.

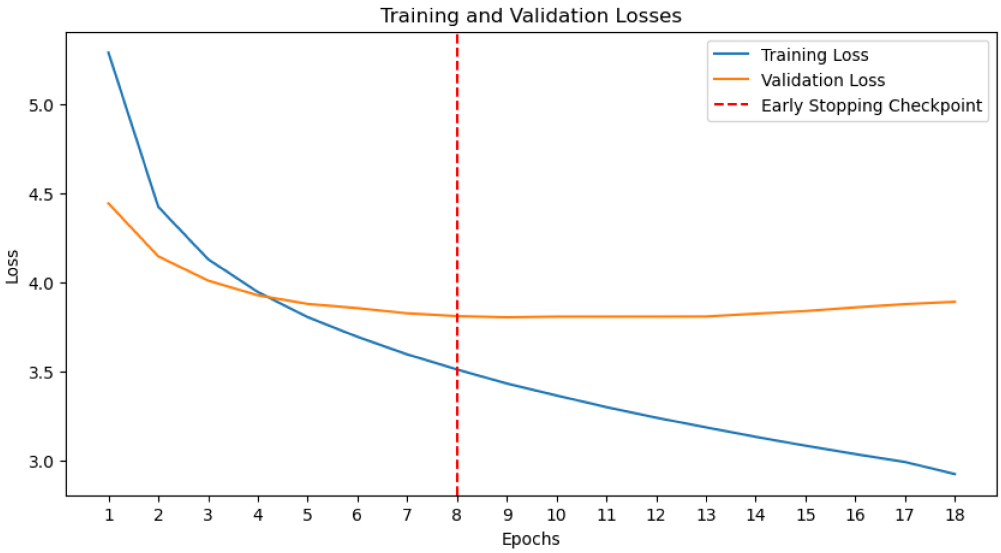

FIGURE 4. Illustration of the training and validation losses when training using the early stopping method

Figure 5 illustrates example images and their generated captions using the proposed visual attention mechanism with early stopping. The proposed method, utilizing a visual attention mechanism with Dual-CNNs, was trained with early stopping techniques. This approach resulted in BLEU scores of 68.76% for BLEU-1, 49.15% for BLEU-2, 35.46% for BLEU-3, and 24.71% for BLEU-4. In comparison, training for 20 epochs without early stopping produced BLEU scores of 65.75% for BLEU-1, 45.19% for BLEU-2, 31.88% for BLEU-3, and 21.84% for BLEU-4. Implementing early stopping not only enhanced performance but also reduced training time.

4. **Discussion.** The visual attention mechanism framework requires a significant amount of computation time. It takes around 20 hours to train the model with 200 epochs. Also, the overfitting problem occurs after around epoch 15, as shown in Figure 4. We then applied the early stopping method while training the proposed visual attention mechanism framework to address the overfitting and reduce the computational time simultaneously. Therefore, by applying early stopping, the model was trained for only 3 hours, reducing computation time by more than six times. The optimal model was achieved within only nine epochs with high BLEU scores. Furthermore, the proposed visual attention mechanism with a Dual-CNN encoder improved performance, reduced computational time, and prevented overfitting.

5. **Conclusions.** In this research, we propose a visual attention mechanism framework with a dual convolutional neural network (Dual-CNN) encoder. The proposed framework was evaluated on the Flickr8K dataset using BLEU scores (BLEU-1, BLEU-2, BLEU-3, and BLEU-4). The experimental results were divided into three experiments. Firstly, we trained the visual attention mechanism framework with three CNN architectures:

FIGURE 5. Examples of captions generated using the proposed visual attention mechanism with Dual-CNN encoder compared to ground truth captions

ResNet-101, EfficientNet-B0, and ResNeXt-101. Hence, the ResNet-101 achieved the highest BLEU scores. Secondly, we fused two CNN architectures in the encoding scheme of the visual attention mechanism. We report that the concatenation of ResNet-101 and EfficientNet-B0 yielded the best BLEU scores. Finally, the early stopping method was employed to discover the optimal training epoch. We discovered that the proposed visual attention mechanism was trained for just 20 epochs when using the early stopping method. Our experimental results suggest that at the eighth epoch, our proposed framework was able to attain the highest performance yielding a BLEU-1 score of 68.76%.

In future work, to improve image caption performance, we will employ the object detection method to extract the valuable attributes in the image [27] before sending the entire image and extracting attributes to the encoder mechanism.

Acknowledgment. This research project was financially supported by Mahasarakham University.

REFERENCES

[1] Y. Zhou, Z. Hu, D. Liu, H. Ben and M. Wang, Compact bidirectional transformer for image captioning, *ArXiv*, DOI: 10.48550/arXiv.2201.01984, 2022.

[2] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan and A. Mian, Visual attention methods in deep learning: An in-depth survey, *ArXiv*, DOI: 10.48550/arXiv.2204.07756, 2022.

[3] J. Zhou, Y. Zhu and H. Pan, Image caption based on visual attention mechanism, *International Conference on Image, Video and Signal Processing (IVSP)*, pp.28-32, DOI: 10.1145/3317640.3317660, 2019.

[4] R. khan, M. Shujah Islam, K. Kanwal, M. Iqbal, M. Imran Hossain and Z. Ye, A deep neural framework for image caption generation using GRU-based attention mechanism, *ArXiv*, DOI: 10.48550/arXiv.2203.01594, 2022.

[5] M. Cornia, M. Stefanini, L. Baraldi and R. Cucchiara, Meshed-memory transformer for image captioning, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.10578-10587, DOI: 10.1109/CVPR42600.2020.01059, 2020.

[6] H. Yanagimoto and K. Hashimoto, Topic-bound image caption generation: Amulti-modal encoder-decoder model of neural networks with transformer, *ICIC Express Letters, Part B: Applications*, vol.14, no.6, DOI: 10.24507/icicelb.14.06.587, 2023.

[7] J. Sui, H. Yu, X. Liang and P. Ping, Image caption method based on graph attention network with global context, *The 7th International Conference on Image, Vision and Computing, ICIVC 2022*, pp.480-487, DOI: 10.1109/ICIVC55077.2022.9886239, 2022.

[8] S. Chen, Q. Jin, P. Wang and Q. Wu, Say as you wish: Fine-grained control of image caption generation with abstract scene graphs, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.9959-9968, DOI: 10.1109/CVPR42600.2020.00998, 2020.

[9] T. Ghandi, H. Pourreza and H. Mahyar, Deep learning approaches on image captioning: A review, *ArXiv*, DOI: 10.48550/arXiv.2201.12944, 2022.

[10] R. Li, H. Liang, Y. Shi, F. Feng and X. Wang, Dual-CNN: A convolutional language decoder for paragraph image captioning, *Neurocomputing*, vol.396, pp.92-101, DOI: 10.1016/j.neucom.2020.02.041, 2020.

[11] O. E. Gannour, S. Hamida, B. Cherradi, M. Al-Sarem, A. Raihani, F. Saeed and M. Hadwan, Concatenation of pre-trained convolutional neural networks for enhanced Covid-19 screening using transfer learning technique, *Electronics*, vol.11, no.1, pp.1-26, DOI: 10.3390/electronics11010103, 2022.

[12] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770-778, DOI: 10.1109/CVPR.2016.90, 2016.

[13] V. Atliha and D. Šešok, Comparison of VGG and ResNet used as encoders for image captioning, *2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, Vilnius, Lithuania, pp.1-4, DOI: 10.1109/eStream51002.2020.9108880, 2020.

[14] S. Xie, R. Girshick, P. Dollár, Z. Tu and K. He, Aggregated residual transformations for deep neural networks, *ArXiv*, DOI: 10.48550/arXiv.1611.05431, 2016.

[15] R. Castro, I. Pineda, W. Lim and M. E. Morocho-Cayamcela, Deep learning approaches based on transformer architectures for image captioning tasks, *IEEE Access*, vol.10, pp.33679-33693, DOI: 10.1109/ACCESS.2022.3161428, 2022.

[16] M. Tan and Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, *ArXiv*, DOI: 10.48550/arXiv.1905.11946, 2020.

[17] Z. Lin, F. Feng, X. Yang and C. Ding, An efficient image captioning method based on generative adversarial networks, *2021 4th International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, Xiamen, China, pp.1-6, DOI: 10.1145/3488933.3488941, 2021.

[18] T. Singkhornart and O. Surinta, Multi-language video subtitle recognition with convolutional neural network and long short-term memory networks, *ICIC Express Letters*, vol.16, no.6, pp.647-655, DOI: 10.24507/icicel.16.06.647, 2022.

[19] R. Biswas, M. Barz and D. Sonntag, Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking, *KI-Künstliche Intelligenz*, vol.34, no.4, pp.571-584, DOI: 10.1007/s13218-020-00679-2, 2020.

[20] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj and R. K. Mishra, Image captioning: A comprehensive survey, *International Conference on Power Electronics and IoT Applications in Renewable Energy and Its Control (PARC)*, pp.325-328, DOI: 10.1109/PARC49193.2020.236619, 2020.

[21] S. Takkar, A. Jain and P. Adlakha, Comparative study of different image captioning models, *The 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pp.1366-1371, DOI: 10.1109/ICCMC51019.2021.9418451, 2021.

[22] X. Ying, An overview of overfitting and its solutions, *Journal of Physics: Conference Series*, vol.1168, no.2, 022022, DOI: 10.1088/1742-6596/1168/2/022022, 2019.

[23] S. Paguada, L. Batina, I. Buhan and I. Armendariz, Being patient and persistent: Optimizing an early stopping strategy for deep learning in profiled attacks, *ArXiv*, DOI: 10.48550/arXiv.2111.14416, 2021.

[24] H. Wang, Y. Zhang and X. Yu, An overview of image caption generation methods, *Comput. Intell. Neurosci.*, DOI: 10.1155/2020/3062706, 2020.

[25] F. H. Dahri, A. A. Chandio, N. A. Dahri and M. A. Soomro, Image caption generator using convolutional recurrent neural network feature fusion, *Journal of Xi'an Shiyou University, Natural Science Edition*, vol.19, no.3, pp.1088-1095, 2023.

[26] V. Jain, F. Al-Turjman, G. Chaudhary, N. Devang, V. Gupta and A. Kumar, Video captioning: A review of theory, techniques and practices, *Multimed. Tools Appl.*, DOI: 10.1007/s11042-021-11878-w, 2022.

[27] S. Dubey, F. Olimov, M. A. Rafique, J. Kim and M. Jeon, Label-attention transformer with geometrically coherent objects for image captioning, *Inf. Sci.*, vol.623, pp.812-831, DOI: 10.1016/j.ins.2022.12.018, 2023.