International Journal on
"Technical and Physical Problems of Engineering"
(IJTPE)
Published by International Organization of IOTPE

ISSN 2077-3528
IJTPE Journal
www.iotpe.com
ijtpe@iotpe.com

December 2023 | Issue 57 | Volume 15 | Number 4 | Pages 100-106

IMAGE CAPTIONING SCHEME BASED ON DEEP LEARNING WITH VISUAL ATTENTION

Y.H. Zaidan J.W. Salih

Department of Computer Science, College of Science, University of Diyala, Baqubah, Iraq
scicompms2132@uodiyala.edu.iq, jumanawaleed@uodiyala.edu.iq

Abstract- Image captioning represents a complicated task to comprehend multi-modal scenes via automatically creating explanations or captions for the image's salient significations. Hence, any image captioning scheme should involve the capability of analyzing and comprehending the image's visual significations via perceiving salient components and their interaction/association. Depending on these perceptions, these schemes should also involve the capability of accurately explaining these organized visual details using natural language. Furthermore, the performance of image captioning schemes can be boosted by utilizing the mechanism of visual attention. Therefore, this proposed scheme utilizes the EfficientNet B7 pre-trained model for image feature extraction and long short-term memory (LSTM) with an attention mechanism to generate captions (word by word) while focusing on the most relevant parts of the image. This proposed scheme was trained on the MSCOCO dataset using metrics of METEOR and BLEU (B_1-B_4), and the attainable results were (Meteor=0.698, B_1=0.888, B_2=0.875, B_3=0.857 and B_4=0.666).

Keywords: Image Captioning Scheme, EfficientNet B7, LSTM, Visual Attention.

1. INTRODUCTION

Automatically producing captions or descriptions for photographs is a challenging subject that involves a visual combination and language input. In other words; it requires both comprehensive visual understanding and advanced natural language creation. That is why the natural language processing and computer vision communities have embraced it as an enthralling challenge [1]. Image captioning is the technique of producing a natural language description for an image automatically using a computer; As a bridge of mentioned communities research areas, besides the requirement of a high-level comprehension of the image's semantic contents, image captioning requires the ability to articulate the information into phrases like a human. Recognizing the existence; qualities; and connections of items in an image is tough enough. The complexity of structuring a phrase to convey such information adds to the task's difficulty [2].

Researchers have been striving to clarify the images' content using a plausible sentence in any language; which has become a hot subject in the computer vision field in recent years. To clarify an image; an automatic image caption creation model frequently incorporates many objects; connections of object; semantic properties, and probable actions included inside it into a representation vector. On top of that; the caption may be produced word through word utilizing a word generator. As a consequence of recent developments in deep learning techniques; a great number of approaches to that challenge have been published employing that paradigm [3].

In recent times, the schemes of deep learning have considerably delivered to the significant progresses in every field [4-6]. Convolution Neural Network (CNN) represents a computer vision deep learning network that can identify and categorize image features. The structure and operations of the visual cortex had an impact on CNN architecture. It's created to resemble how neurons link in the human brain. The pre-processing needed for CNN is less than for other methods. So, CNNs are the best learning algorithms for understanding visual data. It has also demonstrated exceptional image classification, recognition, segmentation, and retrieval capabilities [7-10]. In the field of image captioning, the research community has made significant strides in model design over the last few years: from the first proposals based on deep learning to the advances of self-attention, transformers, and techniques that have been enhanced with attentive schemes and the learning of reinforcement in Recurrent Neural Networks (RNNs) supplied with global visual descriptors. Simultaneously, researchers in the disciplines of computer vision and natural language processing (NLP) have developed assessment processes and criteria for comparing outcomes to human-generated ground facts. Despite years of research and advancements; image captioning is still far from a solved problem [11].

In this paper, we present a scheme consisting of an Encoder-decoder model in an encoder using Convolutional Neural Network (CNN) and an encoder using RNNs with visual attention for sentence generation. The fundamental contributions of the proposed scheme are provided as follows:

1) Apply an end-to-end deep learning-based image captioning scheme using pre-trained CNN (EfficientNet B7) and Long-Short Term Memory (LSTM) with visual attention. The model of LSTM with visual attention processes the vector of an image feature in order to fine-grain and further abstract visual depiction. The language-LSTM is capable of keeping salient objects and predicting the following word on the track of context words and objects.
2) Integrate the accommodative attention model for extracting vision features, mining language structure, and generating image descriptions with more details.

2. RELATED WORKS

The previous relevant works concerning image captioning generation are provided in this section. In recent times, various schemes have been presented for image description generation. Wang, et al. [12], 2020, presented a method for image captioning in which a deep CNN was utilized for extracting a visual representation of an entered image considering the regions of interest as nodes, and building a graph of relationships where the whole nodes are completely connected in a non-directed manner. The messages are propagated via Graph Neural Network (GNN) over the whole edges in a recurrent way and the whole representations concerning graph nodes are output, these representations can be considered implied relation-aware visual depictions between image's objects. The model of visual context-aware attention selects a significant relationship of "160" representations and the model of LSTM language generates sentences. This presented method was tested using MS COCO and Flickr30K. Yan, et al. [13], 2020, proposed a hierarchy attention technique; taking into account the identification of object features and global image features for resulting in better performance. In such a paradigm, the processes of object detection and CNN-encoding extract local and global information, correspondingly. Global and local attention techniques are used to pass these features to the models of LSTM. The models of LSTM concatenate and decode outputs into words. The MS COCO dataset was used to test that model.

Cao, et al. [14], 2020, proposed interactions-guided generative adversarial network which is an effective cycle-consistent technique that uses multiple scales features depiction and object-object interaction to train the model for unsupervised image captioning without the usage of labeled image caption pairings. This technique is made up of three fundamental components: encoding image, extracting features from object-object interaction, and adversarial cycle-consistent creation. The achieved results offered an encouraging performance using MSCOCO dataset. Zhang, et al. [15], 2021, introduced a unique visual connection attention model based on parallel attention and learning spatial restrictions for the first time. In that model, the image encoder is a Faster R-CNN, and the language decoder is a two-layer LSTM. Between the LSTM-1 (coarse decoder) and the LSTM-2 (fine decoder), both attention models are present (fine decoder). The MSCOCO dataset was used to test this strategy.

3. PRELIMINARIES

3.1. EfficientNet

The primary goal of computer vision and deep learning is to find more trustworthy and accurate methods using smaller models. EfficientNet produces more effective outcomes by consistently scaling depth, width, and resolution while shrinking the model. There are 8 models total, ranging from B_0 to B_7, and while the number of parameters stays the same as the number of models rises, the model's accuracy unexpectedly rises [16]. In Figure 1, the EfficientNet B0 model's schematic depiction is shown.

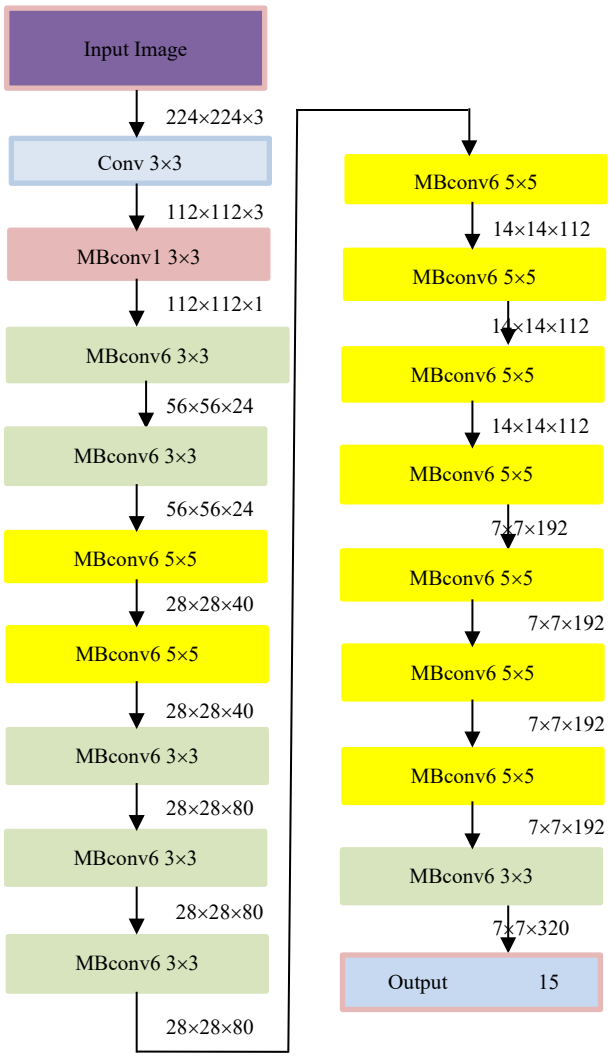

Figure 1. The schematically depiction of EfficientNet B0

EfficientNet utilizes a novel activation function called Swish, it is a reduplication of linear and sigmoid activation functions, in contrast to other cutting-edge CNN models that use ReLU as an activation function. The bottleneck that is reversed EfficientNet relies heavily on MBConv, which connects many fewer channels than the expansion layer since direct connections are employed to bypass bottlenecks. Blocks of MBConv are made up of the layer that compresses the channel after they have been expanded. The EfficientNet models outperform current state-of-the-art CNN models in terms of accuracy and

efficiency, and EfficientNet B7 achieves state-of-the-art accuracy on ImageNet of 84.4 percent for the top-1 and 97.1 percent for the top-5.

The architecture of EfficientNet uses in-depth separable convolutions as opposed to conventional layers, which reduces calculation by almost k times. The kernel size is $k2$, this indicates the 2D convolution window's height and weight. To scale depth (D), width (W), and resolution (R) uniformly using the concepts ($\alpha \geq 1, \beta \geq 1, \Upsilon \geq 1$) the compound coefficient Φ is used in compound scaling [17]. ($D = \alpha^{\Phi}, W = \beta^{\Phi}, R = \Upsilon^{\Phi}$), grid search can be used to compute α, β, Υ constants, and, a user-defined coefficient (Φ), controls how many resources are available for model scaling. On the other hand, α, β, Υ dictate how these extra resources are allotted to the architectural width, depth, and resolution, respectively. In order to scale baseline EfficientNet B0, the compound scaling approach does so in two parts. In the first step, it is assumed that there are twice as many resources available, and the best values for α, β, Υ and are identified using a grid search with $\Phi = 1$. After establishing the obtained, values as constants, the baseline network is enlarged in order to get EfficientNet B1 through B7 with various values utilizing ($\alpha \geq 1, \beta \geq 1, \Upsilon \geq 1$) [18].

3.2. LSTM

The development of LSTM was made in order to overcome the problem of vanishing and exploding gradients. To do this, gates are placed among each of its hidden points and the rest of its layers, protecting the hidden activation. The cell state is the protected hidden activation. The three LSTM gates-forget, input, and output-take on responsibility of guarding cell state [19].

The forget gate is the 1st gate that affects the cell during the forward pass. It determines which cell activations are forgotten and how much. It accomplishes this by multiplying each element of the cell by a vector, ft (0, 1) mh. The corresponding element in the state of the cell will be erased and set to zero if the forget gate emits a value that is close to zero, but it will fully maintain its value if it emits a value that is close to one. The input gate controls how much fresh information is introduced to the protected state. This also occurs at the same time as the determination of a new candidate cell state. Like the forget gate, it $\in$ (0, 1) mh input gate is multiplied by the candidate state and adds it to the cell state. As a result, the cell state is avoided from additions in an unneeded manner. The output gate is the last component and is crucial for backpropagation. It chooses which aspects of the cell state propagated forward and is included in the network's output [15]. The LSTM can be represented mathematically as follows [20]:

$$i_t = \text{sigmoid}\left(W_i X \left[h_{t-1}, x_t\right] + b_i\right) \quad (1)$$

$$f_t = \text{sigmoid}\left(W_f X \left[h_{t-1}, x_t\right] + b_f\right) \quad (2)$$

$$c_t = f_t X c_{t-1} + i_t X \tanh\left(W_c X \left[h_{t-1}, x_t\right] + b_c\right) \quad (3)$$

$$o_t = \text{sigmoid}\left(W_o X \left[h_{t-1}, x_t\right] + b_o\right) \quad (4)$$

$$h_t = o_t X \tanh\left(c_t\right) \quad (5)$$

where, its f_t, and o_t are the input, forget, and output gates, respectively, c_t is the cell state at time step t, h_t is the hidden state at time step t, x_t is the input to the LSTM at time step t, h_{t-1} is the hidden state of the LSTM at time step t-1, W_i, W_f, W_c, and W_o are the weight matrices for the input, forget, cell, and output gates, respectively, and b_i, b_f, b_c, and b_o are the bias terms for the input, forget, cell, and output gates, respectively.

4. THE PROPOSED CAPTIONING GENERATION SCHEME

The proposed scheme aims to describe the visual content of an image. This work was applied to an effective EfficientNet B7. This scheme involves various stages; input dataset of images, pre-processing of images, splitting the dataset into training and testing, and implementation of EfficientNet B7 model for feature extraction as encoder and utilized LSTM with visual attention as decoder for NLP, as demonstrated in Figure 2.

4.1. Input and Pre-Processing

Initially, images are utilized from the Microsoft Common Objects in Context (MSCOCO 2017) dataset, and this dataset is available at a publicly accessible. In the feature extraction process from images, providing an appropriate size of images represents an extremely complicated aspect. However, in order to provide effective system implementation, the size of input images should be fixed. Therefore, in this stage, after loading the dataset, the input images of diverse sizes are resized to 600×600×3.

4.2. Dataset Splitting

In the dataset splitting stage, the dataset is partitioned into 80% and 20% for "training" and "testing" partitions, respectively.

4.3. Encoder

The Encoder (feature extraction) is done by EfficientNet B7. Firstly, it is based on mobile inverted bottleneck convolution (MBConv), and MBConv layers are designed to be more efficient than standard convolutional layers by using a mixing up of depth-wise separatable convolutions and pointwise convolutions. A single filter is used to transform each input channel into a "depth-wise" convolution in a depth-wise separable convolution. The output of the depth-wise convolutions is then combined in a pointwise convolution. This allows MBConv layers to use significantly fewer parameters and compute resources compared to standard convolutional layers while still achieving good performance.

The Efficientnet-B7 consists of seven blocks the first contains 3 MBconv, the second 7, the third also 7, the fourth 10, fifth also 10, the sixth 13, and the seventh 4 MBconv. The architecture of each block involves multiple operations, including Pointwise Convolution this is a 1×1 convolution that reduces the spatial dimensions of the features maps while increasing the number of channels, Depthwise Convolution is a 3×3 convolution that operates on each channel independently, preserving the spatial information of the feature maps, Squeeze-and-Excitation module is an attention that recalibrates the feature maps based on the channel-wise relationships.

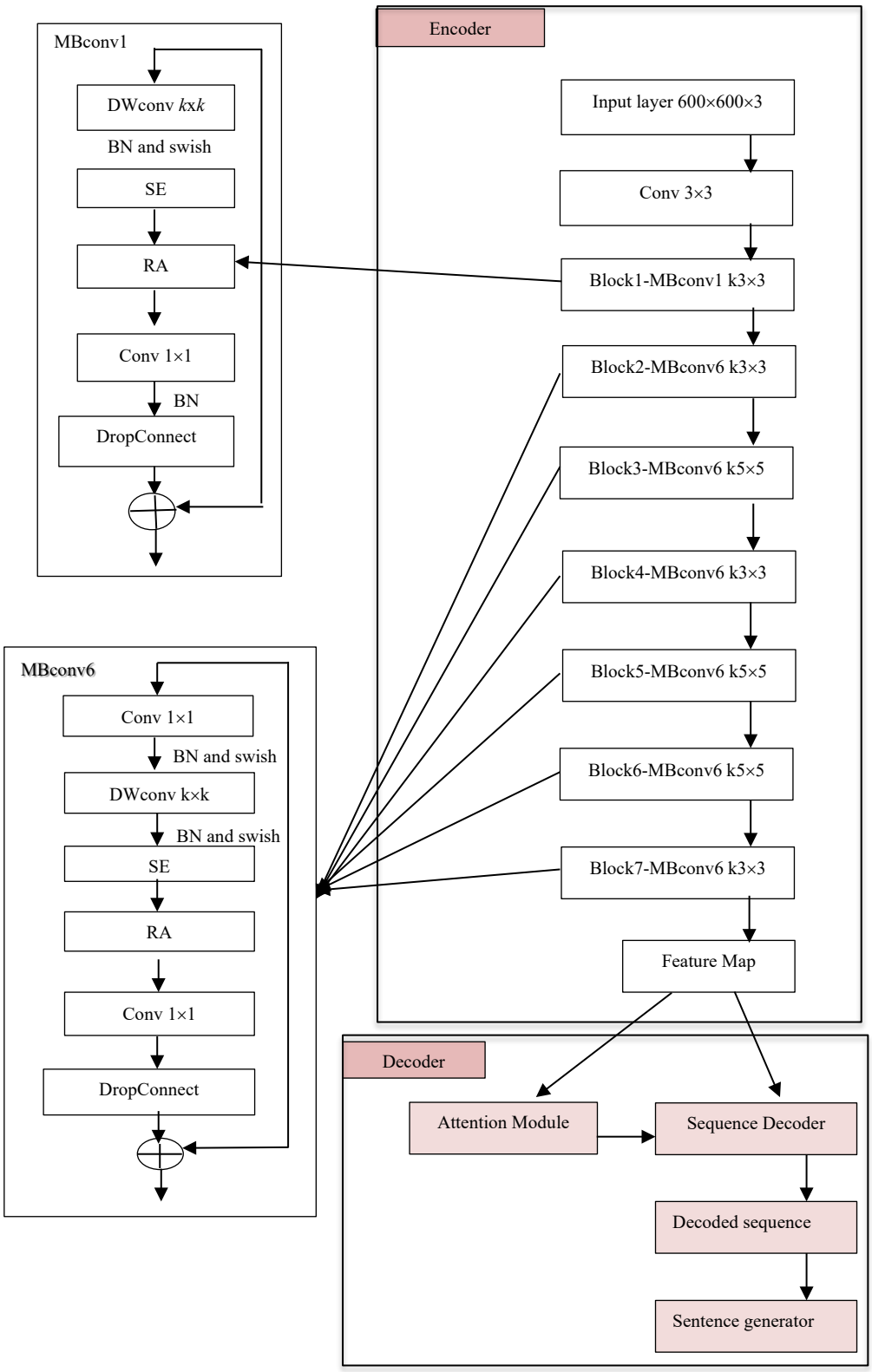

Figure 2. Architecture of Proposed System with EfficientNet B7

Batch Normalization normalizes the activations to reduce the internal covariate shift and improve training stability.

ReLU activation is to apply an element-wise non-linear activation to the output of each block. These details are illustrated in Table 1.

Table 1. The layer's structure of EfficientNet B7

No.	Operator	Resolution	Channels	Layers
1	Image input	600×600	3	
2	Conv 3×3	600×600	64	1
3	Block1-MBconv1 3×3	600×600	32	3
4	Block2-MBconv6 3×3	300×300	48	7
5	Block3-MBconv6 5×5	300×300	80	7
6	Block4-MBconv6 3×3	75×75	80	10
7	Block5-MBconv6 5×5	75×75	244	10
8	Block6-MBconv6 5×5	35×35	384	13
9	Block7-MBconv6 3×3	35×35	640	4

4.4. Decoder with Visual Attention

A Decoder in image captioning is a component of a deep learning model that generates a textual description of an image. It takes the feature representations from an image encoder as input and outputs a sequence of words, typically in natural language. It's done by using LSTM with a visual attention, the input to the model is a sequence of numerical values representing the words or sub-words in the text. The input is passed through an embedding layer, which converts the numerical values to a dense vector representation. This allows the model to process words as continuous vectors rather than discrete symbols, which is more efficient and effective for learning patterns in the text. The input is then passed through LSTM layers. An LSTM layer has a set of weights and biases that are adjusted during training to learn the patterns and relationships between words in the text. On top of the LSTM layers, an attention mechanism is included to enable predictions to be made while focusing just on particular areas of the input. Instead of processing all data equally, the attention mechanism learns to prioritize different components of the input. When creating predictions, it enables the model to concentrate on the input's most pertinent data. The output from the LSTM layers and attention mechanism is passed through a dense layer, which maps the outputs to probability distributions over the vocabulary. Predicted words are the word with the highest probability. An output layer is a fully connected layer with a softmax activation function, which converts the output to a probability distribution over the vocabulary.

5. EXPERIMENTS AND DISCUSSION

We employ the MSCOCO dataset to assess the effectiveness of our system. Common item images from real-world and natural situations make up the MS COCO collection. MSCOCO is more difficult since it has the traits of a complicated background, numerous object types and instances, and small object sizes. With 123,287 images, the MSCOCO dataset is bigger than the Flickr30k collection.

After the splits, it uses 113,287 photos for training, 5000 photos for validation, and 5000 photos for testing. The MS COCO dataset also includes 5 ground-truth words of varying lengths for each image [21].

Most image caption schemes utilize similarity-based measures between ground truth and machine-generated sentences such as BLEU, and METEOR measurements. The score of BLEU can be utilized as an evaluation measure, and it provides m-gram precision between reference and candidate sentences, and the greater m indicates a more suitable understanding at the sentence level instead of a greater similarity between words within the sentence. The METEOR measure identifies the whole matches between sentences using specific criteria of matching, like paraphrase, synonym, and exact word matching.

During this stage, training is done using 80% of the MSCOCO dataset. This separation procedure is not random; rather, whole data rates for testing and training take explored, starting with 50% just for testing data collection and 50% for training up to 90% of training data against 10% to testing data. The separation of 80 percent of the training and 20 percent of the testing dataset make the proposed scheme obtained the highest percentages and best outcome. Table 2 illustrates these values of BLEU and METEOR scores. The generated captions not only include more details about the objects and their relationships, but they also define the semantic linkages between the target object and the scene in the image.

Traditional approaches to image caption generation use a simple LSTM-based model that generates captions based solely on the features extracted from the input image. These models typically do not consider the spatial relationships between different regions of the image or attend to specific regions of the image when generating the caption. While, Attention-based image caption generators (for example, this proposed system) use a more complex model that attends to different regions of the image at each time step, allowing the model to focus on the most relevant regions of the image when generating the caption. This attention mechanism can improve the quality of the generated captions by enabling the model to generate more detailed and accurate descriptions of the image as Figure 3.

Table 2. A comparison of the various ratios of the data pertained

No.	Training and testing datasets, expressed as a percentage	BLEU				Meteor
		B1	B2	B3	B4	
1	50% training 50% testing	0.611	0.411	0.25	0.06	0.516
2	60% training 40% testing	0.75	0.69	0.64	0.54	0.632
3	70% training 30% testing	0.538	0.25	0.18	0.10	0.50
4	80% training 20% testing	0.888	0.875	0.857	0.666	0.698
5	90% training 10% testing	0.73	0.70	0.642	0.542	0.581

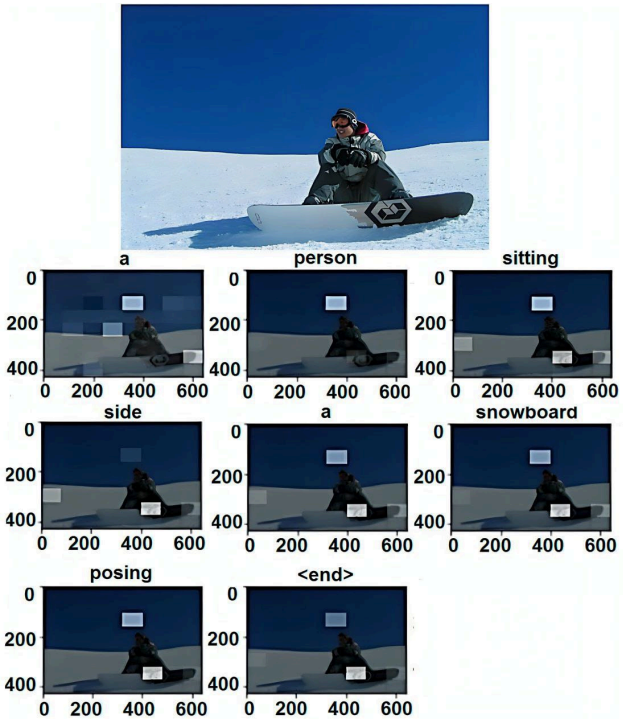

Figure 3. Example of image captioning, the attention mechanism can be used to generate high-quality captions by allowing the model to focus on the most relevant regions of the image at each time step, resulting in more informative and accurate descriptions of the image

6. CONCLUSION

In this paper, an effective scheme was proposed in which the EfficientNet B7 pre-trained model and LSTM with an attention mechanism were utilized for accurately identifying the contents of an image and generating a caption that describes it. The utilization of EfficientNet B7 as the model's backbone (which is a cutting-edge deep neural network architecture for image classification tasks) worked on improving the scheme's performance in image captioning tasks. In addition, the exploitation of visual attention with LSTM allows the scheme to emphasize the significant portions of the image, resulting in highly insightful and precise captions, and hence, leading to a more efficient and accurate image captioning scheme. Besides the generated captions including more details about the objects and their relationships, these captions define the semantic linkages between the target object and the scene in the image. In the future, it is potential to explore the utilization of transfer learning or meta-learning techniques, where the scheme learns from multiple related tasks or datasets, to improve the model's ability to generalize to new tasks and datasets. Furthermore, we can explore the use of different pre-trained models for image feature extraction to compare their performance with EfficientNet-B7.

REFERENCES

[1] Y. H. Tan, C.S. Chan, "Phrase-Based Image Caption Generator with Hierarchical LSTM Network", Neurocomputing, Vol. 333, pp. 86-100, 2019.

[2] S. Bai, S. An, "A Survey on Automatic Image Caption Generation", Neurocomputing, Vol. 311, pp. 291-304, 2018.

[3] X. He, B. Shi, X. Bai, G.S. Xia, Z. Zhang, W. Dong, "Image Caption Generation with Part of Speech Guidance", Pattern Recognit. Lett., Vol. 119, pp. 229-237, 2019.

[4] T.M. Hasan, S.D. Mohammed, J. Waleed, "Development of Breast Cancer Diagnosis System Based on Fuzzy Logic and Probabilistic Neural Network", Eastern-European Journal of Enterprise Technologies, Vol. 4, No. 9-106, pp. 6-13, 2020.

[5] J. Waleed, T. Abbas, T.M. Hasan, "Facemask Wearing Detection Based on Deep CNN to Control COVID-19 Transmission", The 2022 Muthanna International Conference on Engineering Science and Technology (MICEST), pp. 158-161, 2022.

[6] E.H.A. Ameer, Z.F.H. Shouman, Z.G. Abdul Hasan, "Comparison between X-Ray and CT-Scans Images for Infected People Based on Deep Learning", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 53, Vol. 14, No. 4, pp. 9-16, December 2022.

[7] N. Remzan, K. Tahiry, A. Farchi, "Brain Tumor Classification in Magnetic Resonance Imaging Images Using Convolutional Neural Network", Int. J. Electr. Comput. Eng., Vol. 12, No. 6, 2022.

[8] J. Waleed, S. Albawi, H.Q. Flayyih, A. Alkhayyat, "An Effective and Accurate CNN Model for Detecting Tomato Leaves Diseases", The 4th International Iraqi Conference on Engineering Technology and Their Applications (IICETA), pp. 33-37, 2021.

[9] E.H. Hssayni, M. Ettaouil, "Generalization Ability Augmentation and Regularization of Deep Convolutional Neural Networks Using $l^{1/2}$ Pooling", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 48, Vol. 13, No. 3, pp. 1-6, September 2021.

[10] F.K. Al Jibory, O.A. Mohammed, M.S.H. Al Tamimi, "Age Estimation Utilizing Deep Learning Convolutional

Neural Network", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 53, Vol. 14, No. 4, pp. 219-224, December 2022.
[11] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, R. Cucchiara, "From Show to Tell: A Survey on Deep Learning-Based Image Captioning", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 45, No. 1, pp. 539-559, 2022.
[12] J. Wang, W. Wang, L. Wang, Z. Wang, D.D. Feng, T. Tan, "Learning Visual Relationship and Context-Aware Attention for Image Captioning", Pattern Recognit., Vol. 98, p. 107075, 2020.
[13] S. Yan, Y. Xie, F. Wu, J.S. Smith, W. Lu, B. Zhang, "Image Captioning via Hierarchical Attention Mechanism and Policy Gradient Optimization", Signal Processing, Vol. 167, p. 107329, 2020.
[14] S. Cao, G. An, Z. Zheng, Q. Ruan, "Interactions Guided Generative Adversarial Network for Unsupervised Image Captioning", Neurocomputing, Vol. 417, pp. 419-431, 2020.
[15] Z. Zhang, Q. Wu, Y. Wang, F. Chen, "Exploring Region Relationships Implicitly: Image Captioning with Visual Relationship Attention", Image Vis. Comput., Vol. 109, p. 104146, 2021.
[16] M. Tan, Q. Le, "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks", International Conference on Machine Learning, pp. 6105-6114, 2019.
[17] R.H. Hridoy, F. Akter, A. Rakshit, "Computer Vision-Based Skin Disorder Recognition Using EfficientNet: A Transfer Learning Approach", International Conference on Information Technology (ICIT), pp. 482-487, 2021.
[18] C. Akyel, N. Arici, "Linknet-b7: Noise Removal and Lesion Segmentation in Images of Skin Cancer", Mathematics, Vol. 10, No. 5, p. 736, 2022.
[19] A. Mosavi, S. Ardabili, A.R. Varkonyi-Koczy, "List of Deep Learning Models", Engineering for Sustainable Future: Selected Papers of the 18th International Conference on Global Research and Education Inter-Academia 2019, pp. 202-214, 2020.
[20] A. Tsantekidis, N. Passalis, A. Tefas, "Recurrent Neural Networks", Deep Learning for Robot Perception and Cognition, Elsevier, pp. 101-115, 2022.
[21] C. Zheng, et al., "A Novel Equivalent Model of Active Distribution Networks Based on LSTM", IEEE Trans. neural networks Learn. Syst., Vol. 30, No. 9, pp. 2611-2624, 2019.
[22] F. Xiao, X. Gong, Y. Zhang, Y. Shen, J. Li, X. Gao, "DAA: Dual LSTMs with Adaptive Attention for Image Captioning", Neurocomputing, Vol. 364, pp. 322-329, 2019.

BIOGRAPHIES

Name: **Yasir**
Middle Name: **Hameed**
Surname: **Zaidan**
Birthday: 17.6.1995
Birthplace: Diyala, Iraq
Bachelor: Computer Science, University of Diyala, Baqubah, Iraq, 2018
Master: Student, Computer Science, University of Diyala, Baqubah, Iraq, 2023
Research Interests: Image Processing, Image Captioning, and Machine and Deep Learning Algorithms

Name: **Jumana**
Middle Name: **Waleed**
Surname: **Salih**
Birthday: 16.12.1982
Birthplace: Baghdad, Iraq
Bachelor: Computer Sciences, Al-Yarmouk University College, Baqubah, Iraq, 2004
Master: Computer Science/Data Security, University of Technology, Baghdad, Iraq, 2009
Doctorate: Computer Application Technology, Central South University, Changsha, China, 2015
The Last Scientific Position: Assist. Prof., Department of Computer Science, College of Science, University of Diyala, Baqubah, Iraq, Since 2018
Research Interests: Image Processing, Home Automation, Forensics Techniques, Optimization Techniques, and Machine and Deep Learning Algorithms
Scientific Publications: 3 Papers, 2 Theses