Data Augmentation to Stabilize Image Caption Generation Models in Deep Learning

Hamza Aldabbas[1], Muhammad Asad[2], Mohammad Hashem Ryalat[3],
Kaleem Razzaq Malik[4], Muhammad Zubair Akbar Qureshi[5]
Prince Abdullah bin Ghazi Faculty of Information and Communication Technology,
Al-Balqa Applied University, Salt 19117. Jordan[1,3]
Department of Computer Science
Air University, Multan Campus Multan, Pakistan[2,4,5]

***Abstract*—Automatic image caption generation is a challenging AI problem since it requires utilization of several techniques from different computer science domains such as computer vision and natural language processing. Deep learning techniques have demonstrated outstanding results in many different applications. However, data augmentation in deep learning, which replicates the amount and the variety of training data available for learning models without the burden of collecting new data, is a promising field in machine learning. Generating textual description for a given image is a challenging task for computers. Nowadays, deep learning performs a significant role in the manipulation of visual data with the help of Convolutional Neural Networks (CNN). In this study, CNNs are employed to train prediction models which will help in automatic image caption generation. The proposed method utilizes the concept of data augmentation to overcome the fuzziness of well-known image caption generation models. Flickr8k dataset is used in the experimental work of this study and the BLEU score is applied to evaluate the reliability of the proposed method. The results clearly show the stability of the outcomes generated through the proposed method when compared to others.**

***Keywords*—Convolutional Neural Networks (CNN); image caption generation; data augmentation; deep learning**

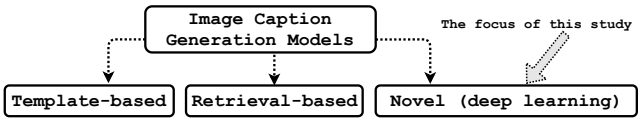

Fig. 1. The main categories of Image Caption generation models

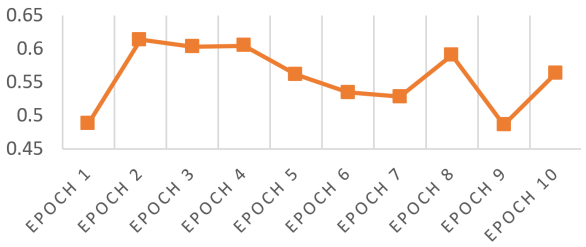

Fig. 2. The BLEU-evaluation-score vs. ten generated epochs where data augmentation is never used.

I. Introduction

Auto generation of captions for images is quite a complex task for computers. Many big names like Google, Microsoft, Apple, etc. are working on the improvement of image analysis. Understanding the objects in images is not the only task for computers but also understanding the relation of these objects in order to to translate this relation in natural language to mimic like a human. This task is quite expensive in terms of computational cost. The story began in 2010 when [1] proposed a method to describe an image into a sentence. By 2011, the GPU speed becomes super fast due to enhancement in technology. That enables to dive into Deep Learning capabilities.

There are three main approaches to generate image captions as displayed in Fig. 1. The first approach is the template-based image caption generation method. In this type we have some templates of captions with missing words, and those missing words are filled according to the objects in an image. For example, [1] use triplet of the scene to fill blank spaces in a template and [2] extract phrases related to the objects detected in an image and define a relationship among objects to create a sentence. The generated sentence in this approach is not a variable length. The second approach is called retrieval-based, in which we have an existing set of captions, which are called candidate captions. In this approach, the query image is matched with a visually similar image in training dataset to produce a caption [3], [4]. The third approach, which is deeply investigated in this paper, is known as novel caption generation, in which the deep learning techniques are utilized to automatically generate image captions. This approach implies the analysis of the objects in images and mapping the analyzed data on the language model to generate captions [5], [6]. This approach produces unique captions in a variable length.

In deep learning based image caption generation models, we extract image features and feed them to a neural network. A CNN (Convolutional Neural Network) is used in this approach to extract features as presented in [7]. After getting features of a dataset, we train a model of neural network and then use it for further prediction tasks. The initial configuration in this process includes specifying the number of epochs, and the loss function. The loss function is an integral part of ANNs which is used to measure the inconsistency between predicted label ($\hat{y}$) and the actual label (y). The models created could be variant and lose stability.

Fig. 2 shows to which extent the outcomes normally differ

from epoch to the other. The BLEU score [8] is employed for evaluation in this study. Fig. 2 asserts the fact that there are obvious variations and instability in the outcomes when data augmentation is never used. However, this research presents a methodology by which we can create stable image caption generation model.

The remaining of this paper is organized as follows. Section II presents literature review and background information. The details of the proposed method are provided in Section III. The experimental work is presented in Section IV and the results is discussed in Section V. The conclusion together with some suggested future work are drawn in Section VI and Section VII respectively.

II. Literature Review

This section presents an overview on the research papers that are related to the image caption generation models. The authors of [3] adopt the retrieval based protocol in which image query is checked to get a sentence from a pool of reference sentences associated with the image. This technique can be used in any system that stores both, images and sentences. The main idea in that study is the retrieval of image-to-sentence. To establish the results, they use the kernel canonicals correlation analysis along with many linguistic and visual Kernels to map images and sentence into space where the similarity between them can be computed directly. They train the model using 6,000 images with real-world captions. For further enhancements in the results, they introduce an algorithm which is called Stacked Auxiliary Embedding that can transfer tens of thousands of annotated images every week to improve the accuracy of the retrieval-based Image caption.

In [5], they use the encoder-decoder pipeline to learn based on the work of generating the descriptions of the images using the Multimodel. A multimodel is based on the embedding space of the images and text and a novel language model that matches distributed representation of the text and images. This multimodel consists of two trained models, which are as follows.

- Neural Language Model
- Image-Text Model

The function of the encoder in their pipeline is to rank the images and sentences. On the other hand, the function of the decoder is generating the description for images from scratch. They use LSTM [9] to encode the sentence, and match their results with Flickr8K and Flickr30K without using object detection techniques. They get their best result using the convolution neural network with 19x layers. However, they use three different methods to generate descriptions of the images.

- Template-based method
- Composition-based method
- Neural-network-based method

The following methods are used in their pipeline to enable the encoder and the decoder to rank generation:

- Long short term memory RNNs
- Multimodel distributed representation
- Log-bilinear NL models
- Multiplicative NL models
- Structure-content NL models

Long-term Recurrent Convolution Network (LRCNs) is applied in [10]. LRCNs consists of the convolution layers, and temporal recursion with long-range. It is considered as end to end trainable method. They train their model for the specific video activity recognition and image caption generation. The LRCNs model is both spatially and temporally deep and flexible enough to be applied in vision-based tasks. Their results consistently demonstrate learning sequential dynamics with a deep sequence model. They use the deep neural network like CNN for capturing the features from the images, and then they add another model LSTM, which is used to generate the sequence of words based on the natural language. They combined both CNN and RNN, and under these, they use LSTM to generate the description of images and videos. The whole system contains the features of CNN and RNN and also a sequence generator.

In [11], they use CNN model instead of traditional RNN model. They use CNN as image "encoder" and then they use the last hidden layer of the network as an input to the RNN "decoder" that generates the sentence. They call this model natural image caption (NIC). It is a neural network which is fully trainable using well-known techniques such as stochastic gradient decent (SGD) [12]. Their model also combines the state of the art sub-networks that perform subtasks like vision and natural language processing. Using these sub-models, they take advantages of pre-training these model on large datasets. The performance of their system compared to the state of the art models is very good. For example, on the Pascal dataset, NIC BLUE score is 59%, and the current state of the art model score is 25%, while human performance reaches 69%. On Flickr30k they improve from 56% to 66%.

The authors of [13] describe a new approach to the image caption generation that tries to generate the caption using a form of attention with two variants.

- A mechanism of hard attention
- A mechanism of soft attention

They generate two attention based model for the image caption generator under a common framework. The first one is a soft deterministic attention-based model which is trained through the back-propagation method, and the second one is a hard attention-based model which is trained by maximizing the lower bound. Flickr8k [14], Flickr30k and the MS COCO [15] dataset are used in their study. For evaluation, they use BLEU as well as METEOR metrics. They also present how the learned attention can be used into model generation process and demonstrate that learned alignments correspond very well to the human intuition. This model is not very simple, but the result of this model is satisfying.

In [16], the authors use a different method for caption generation. This technique is different from the previous approaches. They suppose in their work that description can be represented by collections of nouns, verbs, and prepositional

phrases. The object in the image is described as a Noun phrase. The interaction between the object in the image is encoded as a verb phrase or maybe a preposition. Flickr30k and MS COCO are used as dataset in their experiments. In both datasets, every image has a five (or six) sentence descriptions. When combining both datasets they can have 559,113 sentences so they propose the simplest model that can infer different phrases from image samples. From the phrases predicted, their model can automatically generate sentences using a statistical language model. Their algorithm, despite being simpler than state-of-the-art models, achieves similar results on this task. Also, their model generates some new sentences which are not generally present in the dataset. They measure the quality of the generated sentences using BLEU score.

The authors of [17] use a new method of embedding the visual and language data. Their proposed model is trained using the following parameters:

- **Learning and inference**: they try to retrieve images based on the given query sentence. They train the model on a set of N Images and N correspondent statements that describe their content. After the completion of training process, they discard the training data and evaluate the results of the model on the unseen data.
- **Fragment Embedding**: it is another variant they use in their work. They break down the image and sentence into fragments and embed these fragments into a vector for validation.

Their model has some limitations when it assigns a simple phrase like "A cat is black and white" into multiple relations; it fails to relate them together. On the other hand, from the image side, it counts many persons in the picture as one person. However, the overall results are satisfying.

In [14] authors focus on the work of associating images with sentences drawn from a big predefined pool of the images' descriptions. These descriptions are written by people who were asked to describe the images. They provide an alternate method for describing the image which is best suitable to that image. They use a rank system rather than generating the description for the image. The new rank system works based on the nearest-neighbour search for the image description. The representation in this paper is straightforward. They only rely on three types of low-level pixels perceptual features that capture shape, colour and text in the form of SIFT descriptor. They use two different kernels, the histogram kernel and the pyramid kernel. In both cases, they compute separate kernel for each of three types of images feature and average their results. They then draw similarities between these kernels. These kernels are string kernels with lexical similarities, the Lin similarity kernel, and the distributional similarity kernel.

In [18], the authors propose a new query expansion approach, which is used in auto image captioning. The main idea behind this is just to convert the visual query into distributed semantics. It is generated by the average of sentence vectors that are generated from the captions of the visual images that are similar to the input images. In their study, they use three image captioning standard datasets and show that their technique leads to more accurate results. Automatic image captioning is very popular in computer vision and language processing. The data-driven method, automatic metrics, and subjective evaluation are the techniques which are discussed and compared in this research paper. The first approach generates novel captions from images directly. In this approach, computer vision techniques like object detection and classification use output to extract the visual contents of the input image and generate captions. These studies combine CNN with RNN to generate captions for images. The second approach uses joint representations of images and captions. They use ML techniques to form a common embedding space for textual as well as visual data and accomplish image to sentence in that intermediate space to find the most appropriate captions for a query image. The third technique is to use data-driven approach, by this all image can be treated for captions as a caption transfer problem [19].

In [20], the authors present a model that generates a description of images in natural language. Their proposed model is based on a novel combination of CNNs over image regions and bidirectional-RNN over sentences. The authors also describe the Multi-model RNN architecture that uses the inferred alignments to generate novel descriptions of images. Flickr8K, Flickr30K, and MSCOCO are used in their experiments. The previous work in visual recognition was labeling images with visual categories. They evaluate the output of multimodel RNN architecture on both full-frame and region-level experiments and show that in both cases, multimodel RNN output forms the retrieval baselines.

III. Data Augmentation to Stabilize Image Caption Generation Models

As presented in VGG16 [7], our proposed method skips the last layer in the pre-trained model in order to utilize that layer for feature extraction. We remove the last layer of the model which gives the raw output in the form of 4096 vector size (i.e. the raw features of the image). After feature extraction, we feed these features to train our neural network. The network, which is shown in Fig. 3, is implemented using Keras API under Tensorflow back-end environment. The contribution that this research presents is the employing of data augmentation to generate models in more stable manners. We generate 10 models from both techniques, the augmented-images technique and non-augmented-images technique. After that, we pick a picture and generate a description with all 20 models (10 from augmented-images and 10 from non-augmented-images). Finally, we accept the best description generated by both categories and make a reference description to calculate BLEU score [8] as shown in Fig. 4.

The following represents the relation between P, m, and W_t where m is a number of words which belongs to the candidate text which exists in the reference text, and W_t represents the total words in the candidate text. r is the effective length of the reference corpus, and c is the total length of the translation corpus.

$$P = \frac{m}{w_t}$$

$$\exists p = \begin{cases} 1 & if\, c > r \\ e^{(1-\frac{r}{c})} & if\, c \leq r \end{cases}$$

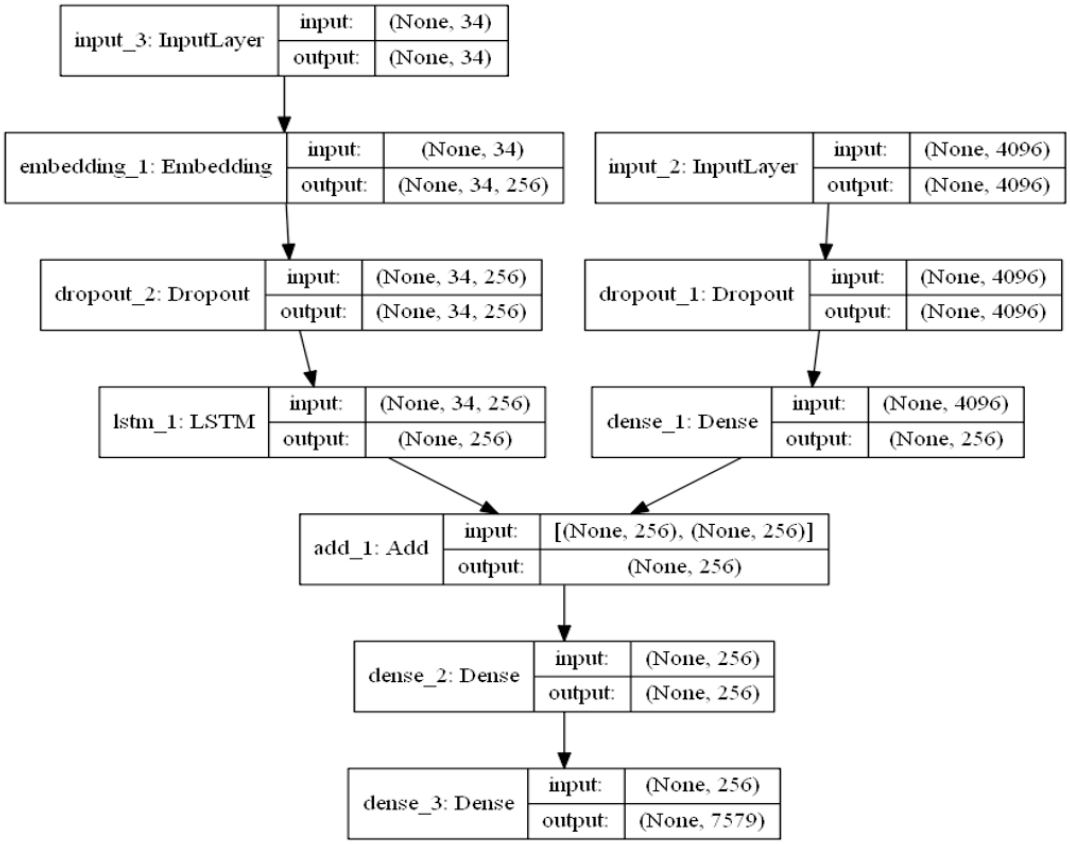

Fig. 3. Proposed Network Diagram to Generate Image Caption Models

$$BLEU = p.e^{\sum_{n=1}^{N} (\frac{1}{n} * \log Pn)} \quad (1)$$

In (1), P_n is the geometric average of the modified n-gram precision, and N is the length of n-grams used to compute P_n. The BLEU score, which is used in this paper to evaluate the results, is calculated using the formula in (1).

$$Avg\_BLEU\_\mathrm{Im}_k = \frac{\sum_{n=1}^{n=10} BLUE(j)}{n} \quad (2)$$

By using (1), we develop a formula represented in (2) where n is the number of epochs (every epoch have its model), k is the ID of the image, j is to identify the BLEU score grams, and j follows the following inequality $1 \leq j \leq 4$.

IV. Experiments

This section presents the experimental work that is designed to evaluate the proposed approach in this study. We apply data augmentation technique on Flickr8k dataset [14]. Flickr8k dataset consists of 8,000 photos and up to 5 captions for each photo. The VGG16 technique presented in [7] is implemented and used as a base reference model to extract features, but we did not classify the output of the VGG16 model into 1000 categories. We use transfer learning concept and remove the prediction layer in VGG16 and save the raw output. Then we implement the proposed model on the raw output of VGG16 model, and present the impact of data augmentation obtained on our dataset using the VGG16 and our Network Diagram along with the proposed methodology. One hundred images were selected from Flickr8k dataset and divided into two groups. The evaluation in this section is evaluated over these two groups. Fig. 5 represents a sample reference image from the first group and Fig. 7 from the second group.

We trained 10 models with the augmented and non-augmented images and tested the reference images shown in Fig. 5 and Fig. 7 from the first and the second group respectively to all 10 models of both types. We picked the highlighted definitions shown in Table I and Table II. To avoid biases, we picked more accurate generated description equals the other generated description and took this description as a candidate text to perform BLEU. The bar chart in Fig. 6 and Fig. 8 confirm that image caption models generated using augmented data are more stable and steady than those generated without augmentation.

Table I presents the BLEU scores for the first group of images, where the image in Fig. 5 is a reference image and Table II presents the BLEU scores for the second group of images, where the image in Fig. 7 is a reference image. The

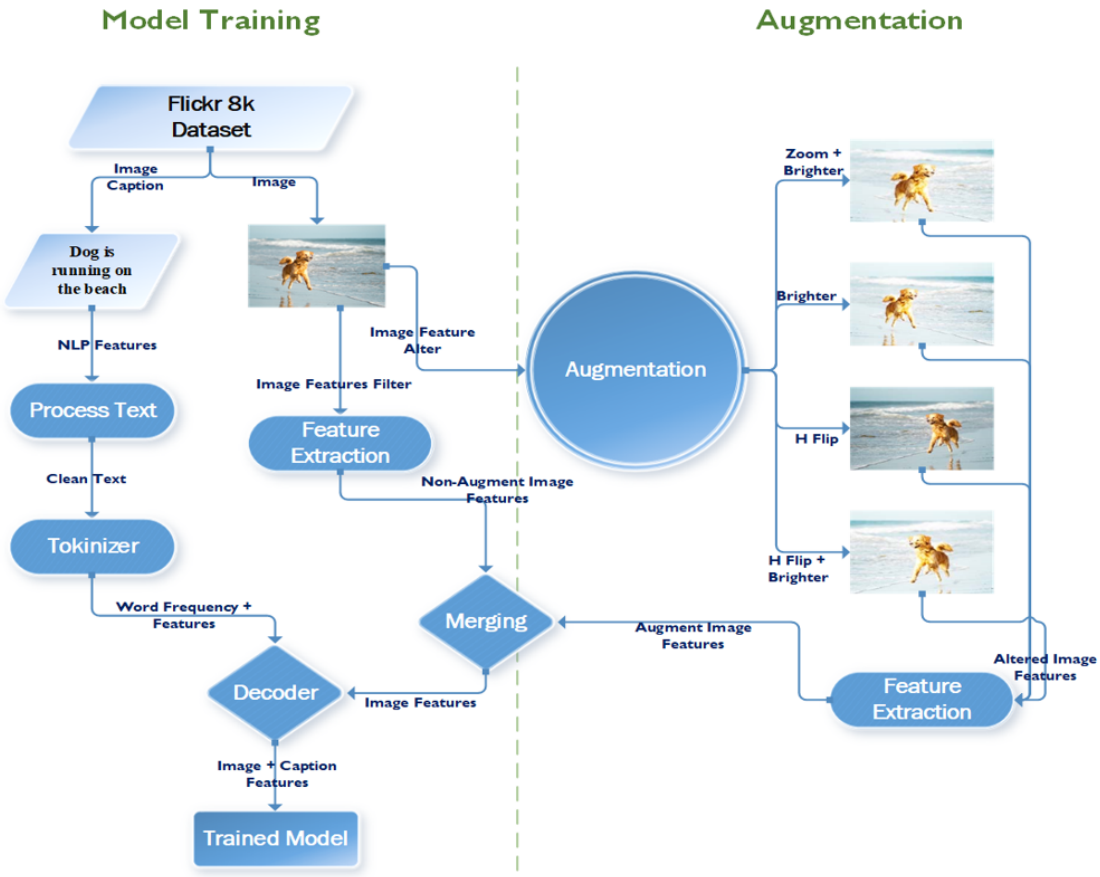

Fig. 4. The Proposed Methodology

TABLE I. BLEU SCORES FOR THE FIRST GROUP, WHERE THE IMAGE IN FIG. 5 IS A REFERENCE IMAGE (AUGMENTED VS. NON-AUGMENTED).

Models	Augmentation	BLEU 1	BLEU 2	BLEU 3	BLEU 4	No Augmentation	BLEU 1	BLEU 2	BLEU 3	BLEU 4
epoch1	dog is running through the grass	0.846	0.846	0.846	0.846	dog is running	0.5643	0.5079	0.4232	0.2822
epoch2	dog is running through the grass	0.846	0.846	0.846	0.846	dog is running	0.5643	0.5079	0.4232	0.2822
epoch3	white dog is running through the grass	1	1	1	1	dog is running	0.5643	0.5079	0.4232	0.2822
epoch4	dog runs through the grass	0.536	0.335	0.223	0	white dog is running	0.5714	0.5	0.4	0.25
epoch5	white dog is running through the grass	1	1	1	1	dog is running	0.5643	0.5079	0.4232	0.2822
epoch6	white dog runs through the grass	0.705	0.508	0.212	0	two dogs are running	1	1	1	1
epoch7	dog runs through the grass	0.536	0.335	0.223	0	the brown dog is running	0.5	0.4286	0.3333	0.2
epoch8	white dog is running through the grass	1	1	1	1	white dog running	0.5643	0.5079	0.4232	0.2822
epoch9	white dog is running through the grass	1	1	1	1	dog running	0.5363	0.5027	0.4469	0.3352
epoch10	white dog is running through the grass	1	1	1	1	dog running	0.5363	0.5027	0.4469	0.3352
	Average	**0.847**	**0.787**	**0.735**	**0.669**	**Average**	**0.5966**	**0.5473**	**0.4743**	**0.3531**

Fig. 5. Sample reference image from the first group

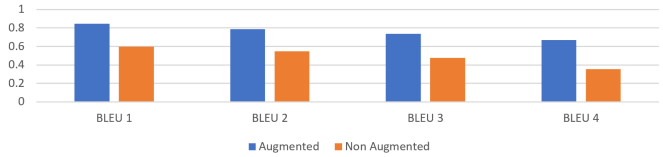

Fig. 6. Bar Chart shows the average BLEU Score of the 10 Models on reference image in Fig. 5

Fig. 7. Sample reference image from the second group

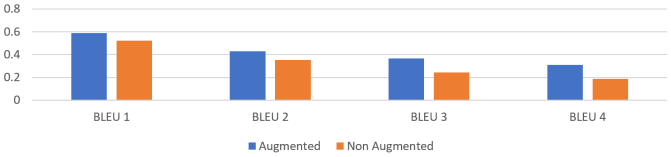

Fig. 8. Bar Chart shows the average BLEU Score of the 10 Models on reference image in Fig. 7

BLEU scores in both tables are calculated for the augmented and the non-augmented dataset. The numbers displayed in Table I and Table II are evaluated using the equation presented in (2) in Section III.

V. Results and Discussions

The results indicate clearly that the proposed method outperforms the other approaches in experiments carried out at benchmark subsets. The introduced method reports a stable prediction model generation.

Fig. 9, Fig. 10, Fig. 11, and Fig. 12 display the values of BLEU1, BLEU2, BLEU3, and BLEU4 scores, respectively. The blue line in these figures shows the value of BLEU score of the generated models when data augmentation is used, whereas the orange line shows the value of BLEU score when data augmentation is never used. The stability of generating captions can be seen clearly in these figures.

The settings which are used to perform data augmentation in this study are:

- zoom + brighter
- brighter
- H-flip
- H-flip + brighter

The BLEU-2, BLEU-3, and BLEU-4 are evaluated using the settings displayed in Fig. 13, Fig. 14, and Fig. 15, respectively. It is obvious that our generated model keeps a stable results regardless using different setting in data augmentation.

VI. Conclusion

This paper presented a new methodology in the form of a pipeline to automatically generate image captions. The proposed method in this study employed deep learning techniques to enhance and stabilize the generated model. By utilizing the power of data augmentation, our method applied CNNs over a set of augmented images to extract their features. Those extracted features are merged with the features which are extracted from a set of non-augmented images and the resulting combination underwent several phases of our proposed pipeline including text processing, tokinizer, decoder, and training model.

This study used the Flickr8k dataset which consists of 8,000 photos and up to 5 captions for each photo, and the VGG16 technique is implemented and used as a base reference model. In order to evaluate the robustness and stability of the proposed method, the BLEU score metric is applied. The outcomes asserted the significant stability of achieved results when data augmentation is used which emphasizes the correctness of the basic contribution of this study.

VII. Future Work

The future directions of this research are exploring the effects of changing data augmentation settings over the stability of the generated image caption, as well as producing a mobile application which automatically generates image captions.

TABLE II. BLEU SCORES FOR THE SECOND GROUP, WHERE THE IMAGE IN FIG. 7 IS A REFERENCE IMAGE (AUGMENTED VS. NON-AUGMENTED).

Models	Augmentation	BLEU 1	BLEU 2	BLEU 3	BLEU 4	No Augmentation	BLEU 1	BLEU 2	BLEU 3	BLEU 4
epoch1	man in green shirt	0.778	0.63	0.571	0.5	man in green shirt	0.67	0.375	0	0
epoch2	man in green shirt is standing beside the river	0.636	0.4	0.333	0.25	man in black hat is sitting beside table	1	1	1	1
epoch3	man in red shirt is wearing red shirt and black hair	0.455	0.4	0.333	0.25	man in black and black and man are sitting on the camera	0.54	0.417	0.273	0.1
epoch4	man in green shirt is wearing green hat	0.552	0.5	0.441	0.35	man in black and black and black and black hair are sitting on the camera	0.47	0.357	0.231	0.08
epoch5	man in green hat is sitting beside river	0.7	0.44	0.375	0.29	man is standing on the beach	0.4	0.121	0	0
epoch6	the man is wearing red shirt and is wearing green hat	0.455	0.1	0	0	two people are playing in the water	0.21	0	0	0
epoch7	man in red shirt is wearing red shirt and black hat	0.455	0.4	0.333	0.25	man is sitting on the beach	0.51	0.364	0.303	0.2
epoch8	man in red shirt is wearing red shirt and black hair	0.385	0.33	0.273	0.2	man wearing black hat and black hat is sitting on the camera	0.58	0.364	0.3	0.22
epoch9	man in green shirt is sitting in the street	1	1	1	1	two people are standing on the river	0.21	0.125	0	0
epoch10	the man is wearing green hat and is wearing green shirt	0.455	0.1	0	0	man wearing black and black hat is sitting on the camera	0.64	0.4	0.333	0.25
	Average	**0.587**	**0.43**	**0.366**	**0.31**	**Average**	**0.52**	**0.352**	**0.244**	**0.19**

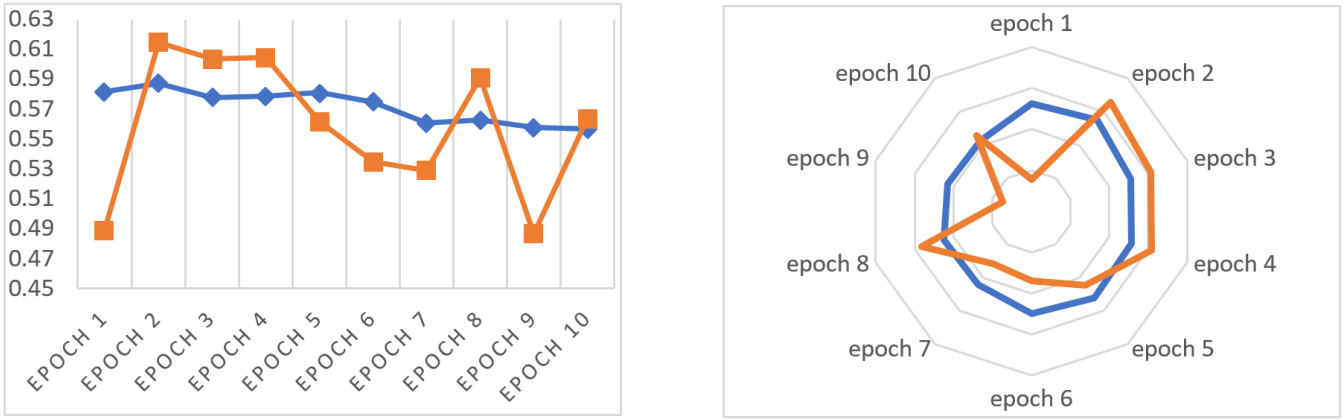

Fig. 9. Impact of Augmentation on Caption Generation Models Evaluated on BLEU-1 Score

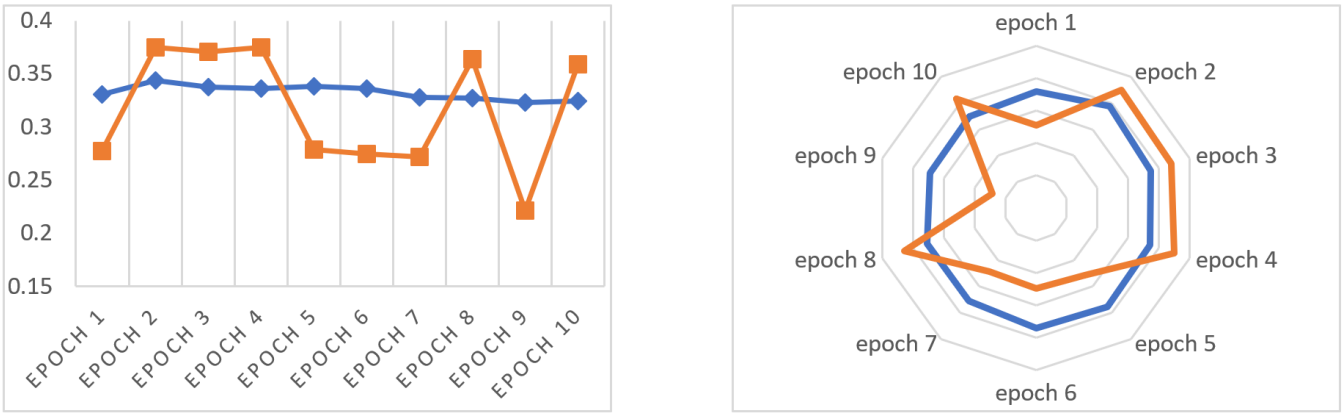

Fig. 10. Impact of Augmentation on Caption Generation Models Evaluated on BLEU-2 Score

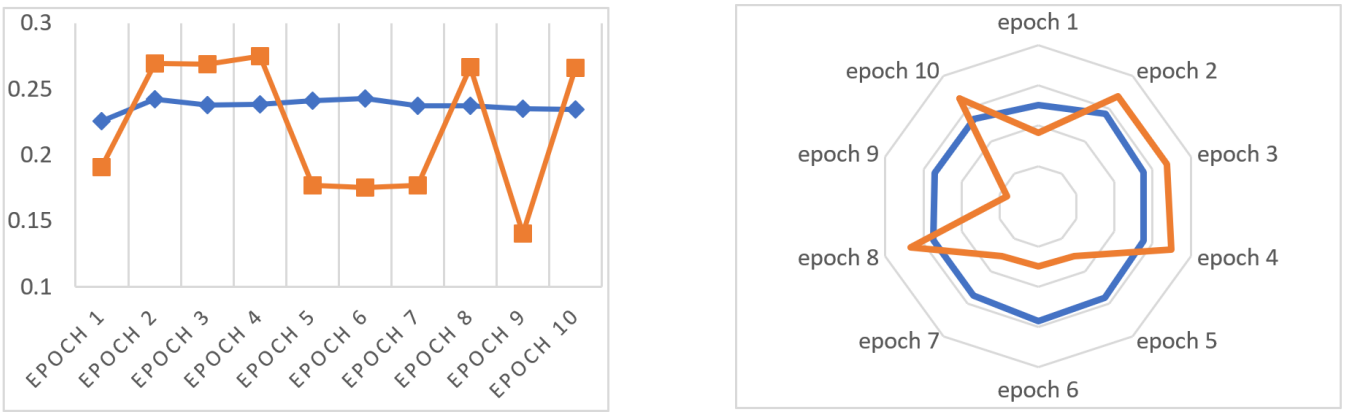

Fig. 11. Impact of Augmentation on Caption Generation Models Evaluated on BLEU-3 Score

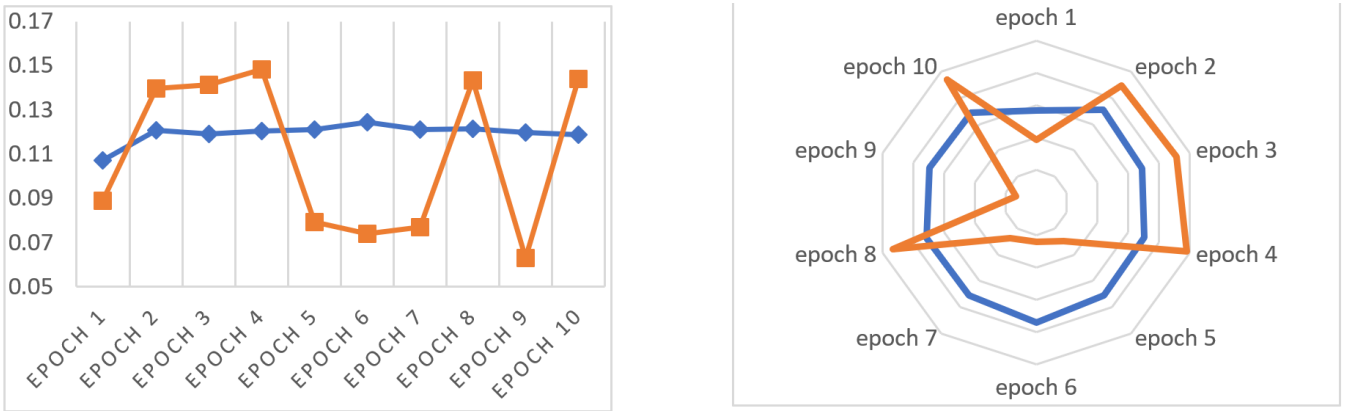

Fig. 12. Impact of Augmentation on Caption Generation Models Evaluated on BLEU-4 Score

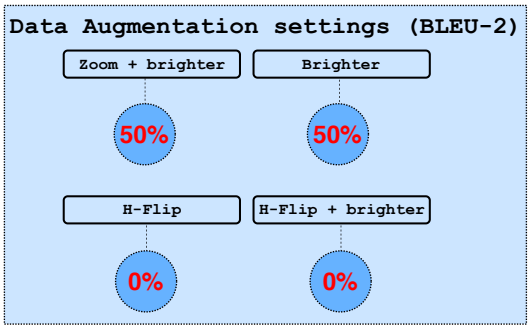

Fig. 13. Data augmentation settings (BLEU-2 Score)

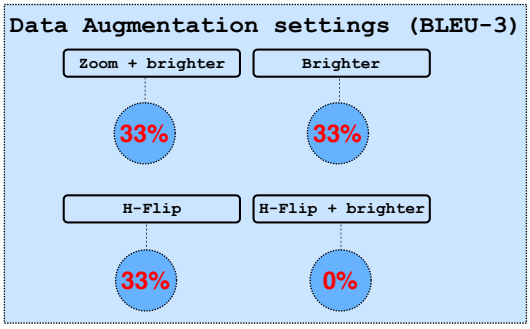

Fig. 14. Data augmentation settings (BLEU-3 Score)

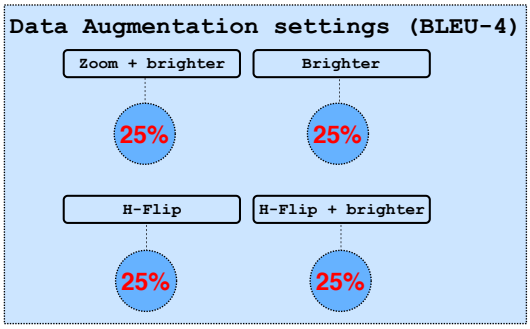

Fig. 15. Data augmentation settings (BLEU-4 Score)

REFERENCES

[1] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *European conference on computer vision*. Springer, 2010, Conference Proceedings, pp. 15–29.

[2] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2011, Conference Proceedings, pp. 220–228.

[3] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *European conference on computer vision*. Springer, 2014, Conference Proceedings, pp. 529–545.

[4] C. Sun, C. Gan, and R. Nevatia, "Automatic concept discovery from parallel text and visual corpora," in *Proceedings of the IEEE international conference on computer vision*, 2015, Conference Proceedings, pp. 2596–2604.

[5] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[6] S. Jabbar, K. R. Malik, M. Ahmad, O. Aldabbas, M. Asif, S. Khalid, K. Han, and S. H. Ahmed, "A methodology of real-time data fusion for localized big data analytics," *IEEE Access*, vol. 6, pp. 24 510–24 520, 2018.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, Conference Proceedings, pp. 311–318.

[9] F. Gers, "Long short-term memory in recurrent neural networks," Thesis, 2001.

[10] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, Conference Proceedings, pp. 2625–2634.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, Conference Proceedings, pp. 3156–3164.

[12] O. Bousquet, U. von Luxburg, and G. Rätsch, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*. Springer, 2011, vol. 3176.

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, Conference Proceedings, pp. 2048–2057.

[14] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, Conference Proceedings, pp. 740–755.

[16] R. Lebret, P. O. Pinheiro, and R. Collobert, "Phrase-based image captioning," 2015.

[17] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in neural information processing systems*, 2014, Conference Proceedings, pp. 1889–1897.

[18] S. Yagcioglu, E. Erdem, A. Erdem, and R. Cakici, "A distributed representation based query expansion approach for image captioning," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, vol. 2, 2015, Conference Proceedings, pp. 106–111.

[19] K. Z. Haider, K. R. Malik, S. Khalid, T. Nawaz, and S. Jabbar, "Deepgender: real-time gender classification using deep learning for smartphones," *Journal of Real-Time Image Processing*, vol. 16, no. 1, pp. 15–29, 2019.

[20] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, Conference Proceedings, pp. 3128–3137.