Al-Mustansiriyah Journal of Science

Al-Mustansiriyah Journal of Science
ISSN: 1814-635X (print) 2521-3520 (electronic)

ORIGINAL ARTICLE *OPEN ACCESS*

A Transfer Learning Approach for Arabic Image Captions

Haneen Siraj Ibrahim [a], Narjis Mezaal Shati [a], and AbdulRahman A. Alsewari [b]

[a] *Department of Computer Science, Mustansiriyah University, Baghdad, Iraq*
[b] *College of Computing and Digital Technology, Birmingham City University, Birmingham, United Kingdom*

CORRESPONDANCE
Haneen Siraj Ibrahim
haneenserag9@uomustansiriyah.edu.iq

ARTICLE INFO
Received: September 18, 2023
Revised: February 11, 2024
Accepted: February 19, 2024
Published:September 30, 2024

ABSTRACT: ***Background:*** *Arabic image captioning (AIC) is the automatic generation of text descriptions in the Arabic language for images. Applies a transfer learning approach in deep learning to enhance computer vision and natural language processing. There are many datasets in English reverse other languages. Instead of, the Arabs researchers unanimously agreed that there is a lack of Arabic databases available in this field.* ***Objective:*** *This paper presents the improvement and processing of the available Arabic textual database using Google spreadsheets for translation and creation of AR. Flicker8k2023 dataset is an extension of the Arabic Flicker8k dataset available, it was uploaded to GitHub and made public for researches.* ***Methods:*** *An efficient model proposed using deep learning techniques by including two pre-training models (VGG16 and VGG19), to extract features from the images and build (LSTM and GRU) models to process textual prediction sequence. In addition to the effect of pre-processing the text in Arabic.* ***Results:*** *The adopted model outperforms better compared to the previous study in BLEU-1 from 33 to 40.* ***Conclusions:*** *This paper concluded that the biggest problem is the database available in the Arabic language. This paper has worked to increase the size of the text database from 24,276 to 32,364 thousand captions, where each image contains 4 captions.*

KEYWORDS: *CNN; LSTM; GRU; NLP; Computer vision*

INTRODUCTION

Arabic image captioning is the process of attributing an image automatically into text to a comprehensive description of the objects in the image and the relationships between them. This topic is interesting due to the development of deep neural networks today, and the enhancement of computer vision and natural language processing [1]. In several areas, recognizing fake news often known as information that is incorrect or misleading [2]. It is also important in Text Detection in Natural Images [3], Image Retrieval too[4]. The importance of the topic appears in many areas, including answering visual questions [5], people who have vision problems [6], indexing and retrieving images [7], robot vision systems, and describing medical images [8]. as well as Kid's games [9].

Figure 1 shows a model based on two neural networks. Encoder is the process of extracting features from an image based on a transfer learning approach, using two VGG16, VGG19 pre-trained models of Convolutional Neural Network (CNN) to extract features from the image. The decoder introduces a sequence of words to form a meaningful sentence based on the features extracted from the image, we built a model based on Long-Short Term Memory (LSTM) and GRU one of the types of Recurrent Neural Network (RNN).

The researchers focused on image captioning in English because of its ease of processing. There is a clear weakness in the research captioning for images in the rest of the languages. Among the languages that suffered is the Arabic language. The evidence is that there is no large and comprehensive database in the Arabic language, except for the Arabic Flicker8k database. It is the first publicly available Arabic dataset developed by [9] for image captions in Arabic and serves as the basis for the English Flickr8K dataset. Arabic Flicker8k containing 8092 images with 24276 captions. Each paired with

three different captions 6000 training images, 1000 test images, and 1000 verification images this data set suffers from its small size. From related studies, Sabri Monaf Sabri [6], and Hani Daoud Hejazi [10], [11] explained that this Arabic Flickr8k database is inefficient because of its small size and was used as it. Further, its the only database available to the public.

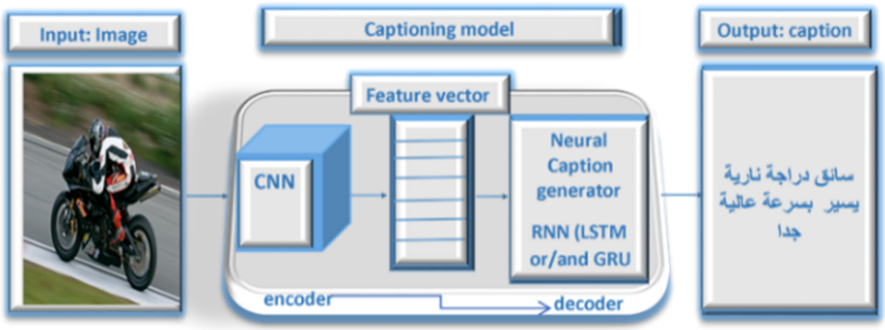

Figure 1. A comprehensive scheme of the encoder and decoder model in Arabic Image Captioning models

The contribution of this paper, developing the AIC model to obtain the best results with two databases. The only available database was improved and processed, in addition to building 16 models based on several methods, two methods VGG16 and VGG19 to extract features from the image. With deep learning techniques based on LSTM and GRU one of the types of RNN, with/without Farasa Lemmatization Preprocessing for text. The paper is organized as follows: Section 2 summarizes the related work of Arabic image captioning. Section 3 includes the processing and methods used in the image and text. Our experiments and results are in section 4. Finally, section 5 for conclusions and future work.

RELATED WORKS

Previous studies, on the subject of AIC, unanimously that there is a significant shortage of databases for the Arabic language. Most of the studies applied the architecture of Encoder–Decoder with the addition of the attention mechanism to the LSTM network or without.

In 2018, Al-Muzaini [12] presented a model based on the architecture of Encoder–Decoder using CNN as an encoder, and RNN as a decoder on two databases in Arabic. The first database contains 1176 images with 5358 captions. The second database consists of 150 images with 750 captions. During this study, the problem of the model is clarified that the database is small and confirms obtaining higher BLEU scores. When working on a larger database in the future and making it available to the public. In [13], presented a model based on deep learning models (CNN) as an encoder and not as a decoder. The model was applied to the Arabic database Flickr8.

In 2020, a small Arabic caption database was announced by [9] extracted from the flickr2 database and translated by Google translated, containing 8,091 images and 24,273 captions. It is Divided into 3 sets of training, testing, and verification, they are 6000, 1000, and 1000 respectively. He used the architecture of Encoder-Decoder and extracted the features from the images using one of the Transfer Learning models which is in VGG16. To extract the string of words, he used both LSTM and GRU and achieved a BLEU1=33 score. He explained that the problem is in the complexity of the Arabic language morphology and presented future solutions, which is the work on a solid database.

In 2021 [6] proposed a new model based on Transformer architecture and use EfficientNet and MobileNetV2 to extract features from images. In addition, they use LSTM and GRU with the inclusion of attention mechanism to get the word sequence for caption formation. Then, they work on natural language processing using AraBERT to segment the words and reduce the existing complexity in the Arabic language. The model was applied on two databases, Flicker8k and Objects Shared in Context (COCO), the results was BLEU-1 = 44.3 and a BLEU-4 = of 15.6 scores. It explained the future optimization mechanism by working on more powerful architectures with the innovation of natural language processing methods to reduce the complexity in the structure of the Arabic language.

The two studies [10], [11] were presented a model based on the architecture of Encoder-Decoder. The VGG16 and Inception V3 models were used to extract features from images. The GRU and LSTM for caption generation by using the database provided by [9]. The results were announced, Blue1 = 36.5. It worked on 4 natural language processing methods for text databases. It explained the problem

in the smallness of the available and only database. It clarifies the future improvement mechanism by using a larger and more accurate database. In 2022 [14], an efficient deep-learning model for Arabic image captioning has been suggested. It uses the Arabic Flickr8k dataset for training and relies on the design of the encoder-decoder to use RESNet-101 in an encoder and (LSTM) in a decoder. The Back-propagation has been applied in a thorough manner to build soft attention, and he saw success. The BLEU-1 /2/3/4 is equal to 58.708/46.523/35.712/27.12 respectively. The study did not increase the use of transformers or bigger training data sets based on the results of BLEU-N. They describe how better models may be investigated in the future to improve the results in this field utilizing Generative Adversarial Networks, as arabic is a morphologically complex language, hence requiring new text preprocessing methods and additional techniques. In [15], the author uses GigaBERT and AraBERT as pre-trained models, and presents an encoder-decoder architecture (CNN and RNN) model that was constructed and tested on a number of Arabic picture captions. The training of Arabic uses two publicity datasets (COCO and Flickr8k). The model receives a score of relation to the picture caption standard. The results are 0.39, 0.25, 0.15, and 0.092 for BLEUs 1 through 4. There is a need to describe strategies for building a robust and comprehensive Arabic database that can be made publicly accessible in the future, similar to the COCO database, through translation and verification.

In [16], the researcher presented three models for the improvement of the caption of Arabic image, the first is based on detecting and dealing with one or more objects. The second is based on using a pipeline with the attention mechanism to detect the objects in the image. Two databases, i.e., Flickr30k and COCO, were used on the three models and tested on the dataset in Arabic. The system outperformed and obtained the results compared to the English and Arabic languages.

MATERIALS AND METHODS

In this section, we explain our database, how natural language processing works, and the transfer learning models that extract features from images.

Arabic Captioned Dataset

In this project, we expanded the Arabic Flicker8k dataset to increase the amount of data processed by the recurrent neural network, which enhances learning and leads to more accurate image caption predictions. Each image in the dataset contains four captions. We performed spelling and grammar checks on the texts, and the updated Arabic database, named AR. Flickr8k, was uploaded to GitHub and made publicly available for future research [17]. The AR. Flickr8k database contains 8,092 images with 32,364 captions. After completing the specialized database, the vocabulary size increased. To manage this, we used Farasa Lemmatization to reduce the vocabulary, successfully lowering the number of tokens from 11,387 to 4,357. Figure 2 shows an example from the AR. Flickr8k2023 dataset.

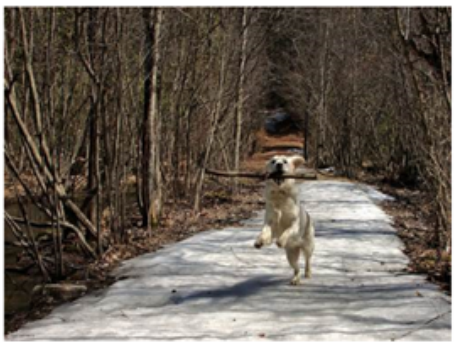

Image sample

1. كلب يقفز إلى الأمام يحمل عصا كبيرة
2. يركض كلب في درب مع عصا في فمه
3. كلب يركض بعصا كبيرة في فمه
4. كلب بعصا في فمه يجري في الغابة

1. A dog jumps forward carrying a large stick.
2. A dog runs down a trail with a stick in his mouth.
3. a dog runs with a large stick in its mouth.
4. A dog with a stick in his mouth runs in a forest.
5. a white dog running down a path with a long stick in his mouth.

Figure 2. an example from the AR. Flickr8k2023dataset

Transfer Learning

This paper utilizes pre-trained CNN weights instead of initializing the decoder CNN weights randomly and training them from scratch. This approach, known as transfer learning, involves applying knowledge gained from one task to another. Transfer learning is frequently employed in the literature to improve performance. In our work, we use three pre-trained CNN models, including VGG16 and VGG19 [18], [19].

Preprocessing for Image

The data set is divided into two parts, images and captions. The image file consists of 8091 images of unequal sizes and in RGB color format. We apply the same pre-processing steps for the image that ElJundi *et al.* [8] followed. The process starts by applying pre-trained models such as VGG19 and Inception-ResNet-v2 that receives a fixed size colored image of dimension $(224 \times 224 \times 3)$. Thereafter, the model is trained regarding the extracted features from the image and The model is saved to minimize overall processing time.

Text Preprocessing

The preprocessing of text is a very important topic in image captioning because of the complexity of the Arabic language.

1 General Preprocessing

This part mentions the basic stages for text preprocessing:

1. Normalizing the Arabic text by converting it to the original, (e.g., convert " إ أ آ " to " ","ى" ,"ا ء" ,"ؤ" ,"ي"" etc.).
2. Removing diacritics from letters, for example("
   ّ| # Tashdid
   َ| # Fatha
   ً| # Tanwin Fath
   ُ| # Damma
   ٌ| # Tanwin Damm
   ِ| # Kasra
   ٍ| # Tanwin Kasr
   ْ| # Sukun
   - # Tatwil/Kashida ")
3. Remove punctuation like ('÷×;<> () *&%] [-, /:'?)
4. Remove the English letters (a-z, A-Z)
5. Remove repeated letters
6. Add the word "start" to the beginning of the sentence and the word "end" to the end of the sentence.
7. I used Farasa Lemmatize to pre-process the text.

2 Farasa

The Arabic language is a language with a complex and rich morphology and there is difficulty in dealing with it. To reduce this complexity, we used Farasa, which is a complete and modern software toolkit for processing texts in the Arabic language provided by the Qatar Computing Research Institute (QCRI), the Arabic Machine Processing Department. It is considered the best, most accurate and fastest package compared to other competing packages [20]. It provides several tasks, the most important of which are Segmentation and Lemmatization.

2.1 Farasa Segmentation

Farasa Segmentation is a tool that uses advanced algorithms to remove prefixes and su ixes from words, reducing them to their root form. It analyzes the structure of words based on basic units called (morphemes). For example, The word "يذهبون" is made up of three units, the first being "ي" to indicate that the verb It was done by a third party, and the second "ذهب" is the verb, and it represents the basic unit for the word, and the third "ون" to refer to the masculine or plural. Another example show in Figure 3, the word(فبالكتاب) returns to the root (كتاب) [21].

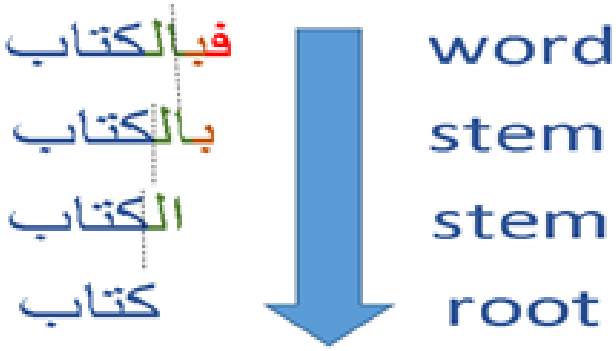

Figure 3. An example for Farasa Segmentation

2.2 Farasa Lemmatization

The Lemmatization is similar to the stemming in the idea, but it is more powerful and effective. It is not sufficient to remove the extra letters in the words, but the search for their meaning depends on a dictionary to remove prefixes and suffixes from words while taking into account the meaning, which is more accurate but slower compared to Segmentation. Table 1 illustrates an example of Farasa Lemmatization from (AR. Flickr8k) dataset. By using this method, Farasa Lemmatization the vocabulary of 11387 tokens were reduced to 4357 tokens.

Table 1. for processing texts in the Arabic language

Word from AR. Flickr8k2023dataset	Farasa lemmatization
اثنان	اثنان
اثنتان	اثنتان
اثنين	اثنان
اثنتين	اثنتان

2.3 Stemming vs Lemmatizing

1. Both aim to explicitly correlate words with similar meanings and reduce the corpus of words to which the model is exposed.
2. When choosing one method over the other, you must compromise on both accuracy and speed. The distinction is that stemming employs a cruder approach by simply removing a word's ending using heuristics, without considering the context in which the word is used. As a result, stemming may or may not produce a real word from the dictionary. Additionally, while it is typically less precise, it is faster because the rules are so straightforward.
3. Lemmatizing makes use of more thorough analysis to group words with comparable meanings based on the context in which they are used, their part of speech, and other elements.
4. A dictionary word is always returned by lemmatizers, making this approach generally more accurate due to the additional context considered. The drawback, however, is that it may require more computational resources.

5. The decision to use stemming or lemmatization depends purely on the project's requirements, such as for essential projects and project's sentence structure and language applications, lemmatization is required.

RESULTS AND DISCUSSION

This section test and validate the suggested AIC model using two datasets. We presented a caption system and some experiments were conducted on it, including 36 variables classified as follows: 4 databases, 3 models for extracting features from images using CNN models, and 2 Deep Learning Methods. Figure 4 shows the stages through which each experiment begins, starting with the ENCODER stage. Passing the image database to one of the pre-trained CNN models (VGG16/ VGG19/ Inception-ResNet-v2). At this stage, a vector of features is obtained for each image. On the other hand, the text database was entered to work on NLP natural language processing.

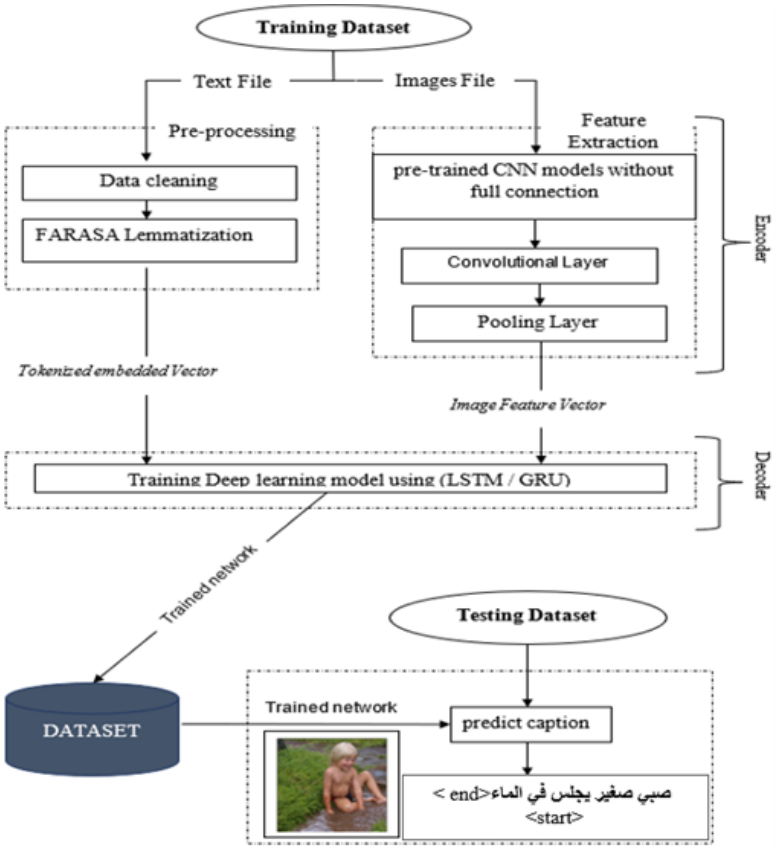

Figure 4. Block diagram of the proposed system in general

Cleaning the texts from English letters include: Al-tanween, removing repetitive letters, making normalization, and punctuation. After that, the results are passed to the second stage, which is decoding. Recurrent Neural Networks (RNN) is a type of recurrent deep learning network introduced by Michael I. Jordan [22]. It works as a decoder to produce the words that characterize the image in order. The recurrent neural network suffers from the length of the input being equal to the length of the output, and this is illogical in translation because its not keeping a long-term memory. It suffers from forgetfulness and the fading of words over time. Furthermore, in order to solve this problem, a kind of repetitive game network is used, the Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU). At the end, the predicted results of the text are obtained and performance measures are used. All the results of BLUE 1/2/3/4 are saved for the experiments. The computer to perform these experiments needs a 64-bit operating system Intel(R) Core(TM) i7-4600M CPU @ 2.90GHz 2.90 GHz and 16 GB RAM.

H. D. HEJAZI [10] proposed 4 methods for preprocessing text in Arabic and reached results using the BLEU-1= 36, while J. Emami [15] presented a model based on transformers with the use of

AraBERT and GigaBERT as pre-trained models and achieved results superior than El-Jundi [9]. in addition, they proposed an Arabic caption database extracted from the flickr2 English database and translated by Google that containing 8,091 images and 24,273 captions. In this study, we expanded the Arabic Flicker8k dataset to increase the amount of data processed by the recurrent neural network, which enhances learning and leads to more accurate image caption predictions. Our Arabic database (AR. Flickr8k) includes 8,092 images with 32,364 captions.

All of the experiments we conducted exceeded the results presented. Table 2 shows the results compared with other related work. Our proposed model, which achieved the highest results, includes our own database, AR. Flickr8k2023, processed with FARASA. The VGG19 model is used to extract features from images, and LSTM is employed to generate a sequence of meaningful words.

Table 2. BLEU score comparison of our models and previous work

model	**BLEU-1**
Obeida ElJundi12020 [8]	33
Hani Hejazi 2022 [10]	36
Jonathan Emami 2022 [14]	39
Ours: VGG19+AR. Flickr8k2023+ FARASA	40

Figure 5 shows the BLEU-1 score results using four variables and two techniques, VGG16 and VGG19, with the GRU deep learning method. Figure 6 displays the image caption results produced using the encoder-decoder model with our own database, AR. Flickr8k2023, processed through FARASA. The VGG19 model was used to extract features from the image, and LSTM was employed to generate a sequence of meaningful words. To compare results between datasets, I present four variables, testing the datasets with and without FARASA, using two models, VGG16 and VGG19, to extract features from the image in combination with one of the deep learning methods, GRU or LSTM.

Figure 7 shows the BLEU-1 scores using four variables with two techniques, VGG16 and VGG19, in combination with the LSTM deep learning method. Figure 8 presents the BLEU-1/2/3/4 scores using our AR. Flickr8k2023 database with VGG19 and the LSTM deep learning method. Our results outperformed those of El-Jundi [9], with the highest BLEU-1 score of 44. The variables were as follows:

(A) represents the model test, Arabic Flickr8k database without FARASA.

(B) represents the model test, the Arabic Flickr8k database available was processed using FARASA.

(C) represents the model test, AR. Flickr8k2023 database without FARASA.

(D) represents the model test, AR. Flickr8k2023 database WITH FARASA.

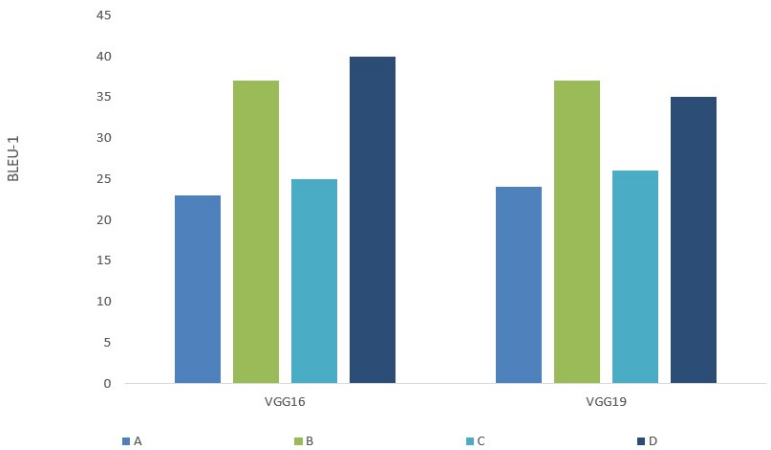

Figure 5. BLUE-1 with 4 variables and GRU deep learning method

Figure 6. The results of testing our AIC model

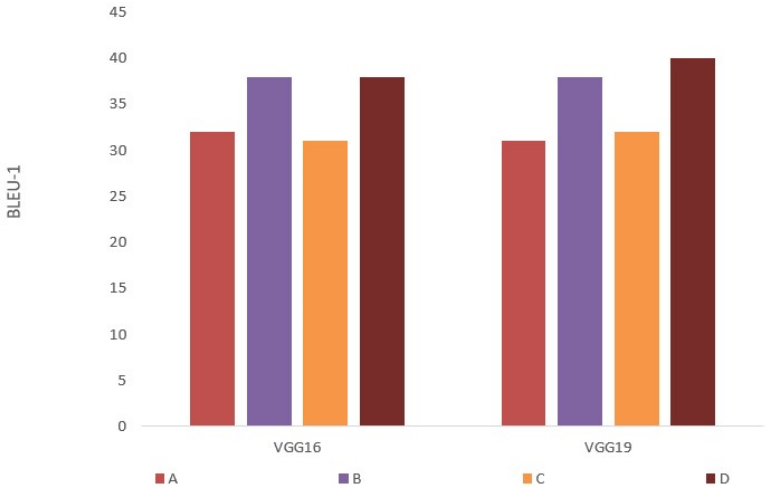

Figure 7. BLUE-1 with 4 variables and LSTM deep learning method

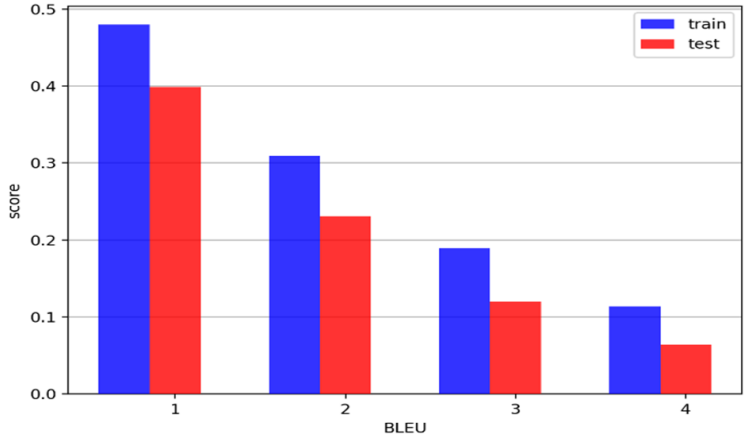

Figure 8. BLUE-Score for training and testing to our dataset(AR. Flickr8k) for LSTM

CONCLUSION

After studying the image caption problems, this paper concluded that the biggest problem is the database available in the Arabic language. We have worked to increase the size of the text database from 24,276 to 32,364 captions, where each image contains 4 captions. The Arabic language was processed using Farasa Lemmatization to reduce vocabulary from 11387 to 4357. The LSTM one-way nonsense and the GRU. This helped to improve the proposed model for image captioning and obtaining BLUE-1=40 using VGG19 and LSTM. In the future, the authors hope to build a database that is linguistically solid and not using translation to avoid errors during translation.

SUPPLEMENTARY MATERIAL

None.

AUTHOR CONTRIBUTIONS

Haneen Siraj Ibrahima: Writing, editing, visualization. Narjis Mezaal Shati: Methodology, software; validation, formal analysis, and investigation. AbdulRahman: Methodology, software; validation, formal analysis, and investigation.

FUNDING

None.

DATA AVAILABILITY STATEMENT

Data is available in the article.

ACKNOWLEDGMENTS

We extend our thanks to Mustansiriyah University.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

[1] M. T. Lasheen and N. H. Barakat, "Arabic image captioning: The effect of text pre-processing on the attention weights and the bleu-n scores," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, pp. 413–423, 2022. doi: 10.14569/IJACSA.2022.0130751.

[2] M. Al-Tai, B. M. Nema, and A. . Al-Sherbaz, "Deep learning for fake news detection: Literature review," *Al-Mustansiriyah Journal of Science*, vol. 34, no. 2, pp. 70–81, Jun. 2023. doi: 10.23851/mjs.v34i2.1292.

[3] Z. A. Ramadhan and D. Alzubaydi, "Text detection in natural image by connected component labeling," *Al-Mustansiriyah Journal of Science*, vol. 30, no. 1, p. 111, 2019. doi: 10.23851/mjs.v30i1.531.

[4] N. M. Khassaf and S. H. Shaker, "Image retrieval based convolutional neural network," *Al-Mustansiriyah Journal of Science*, vol. 31, no. 4, pp. 43–54, Dec. 2020. doi: 10.23851/mjs.v31i4.897.

[5] A. Salaberria, G. Azkune, O. L. de Lacalle, A. Soroa, and E. Agirre, "Image captioning for effective use of language models in knowledge-based visual question answering," *Expert Systems with Applications*, vol. 212, p. 118 669, 2023. doi: 10.1016/j.eswa.2022.118669.

[6] S. M. Sabri, "Arabic image captioning using deep learning with attention," M.S. thesis, Institute for Artificial Intelligence, University of Georgia, 2021.

[7] A. Attai and A. Elnagar, "A survey on arabic image captioning systems using deep learning models," in *Proceedings of the 14th International Conference on Innovations in Information Technology (IIT)*, 2020, pp. 114–119. doi: 10.1109/IIT50501.2020.9299027.

[8] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep learning approaches on image captioning: A review," *arXiv*, 2022. doi: 10.1145/3617592.

[9] O. ElJundi, M. Dhaybi, K. Mokadam, H. Hajj, and D. Asmar, "Resources and end-to-end neural network models for arabic image captioning," in *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 5, 2020, pp. 233–241. doi: 10.5220/0008881202330241.

[10] H. Hejazi and K. Shaalan, "Deep learning for arabic image captioning: A comparative study of main factors and preprocessing recommendations," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, pp. 37–44, 2021. doi: 10.14569/IJACSA.2021.0121105.

[11] H. D. Hejazi, "Arabic image captioning (aic): Utilizing deep learning and main factors comparison and prioritization," M.S. thesis, The British University in Dubai (BUiD), 2022.

[12] H. A. Al-muzaini, T. N. Al-yahya, and H. Benhidour, "Automatic arabic image captioning using rnn-lstm-based language model and cnn," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 6, pp. 67–73, 2018. doi: 10.14569/IJACSA.2018.090610.

[13] R. Mualla and J. Alkheir, "Development of an arabic image description system," *International Journal of Computer Science Trends and Technology (IJCST)*, vol. 6, no. 3, pp. 205–213, 2018.

[14] M. T. Lasheen and N. H. Barakat, "Arabic image captioning: The effect of text pre-processing on the attention weights and the bleu-n scores," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, pp. 413–423, 2022. doi: 10.14569/IJACSA.2022.0130751.

[15] J. Emami, P. Nugues, A. Elnagar, and I. Afyouni, "Arabic image captioning using pre-training of deep bidirectional transformers," in *Proceedings of the 15th International Conference on Natural Language Generation*, 2022, pp. 40–51. doi: 10.18653/v1/2022.inlg-main.4.

[16] I. Afyouni, I. Azhara, and A. Elnagar, "Aracap: A hybrid deep learning architecture for arabic image captioning," *Procedia Computer Science*, vol. 189, pp. 382–389, Jan. 2021. doi: 10.1016/j.procs.2021.05.108.

[17] H. Siraj and D. N. Mezaal, *Arabic-image-captioning-haneen-siraj-and-dr-narjis-mezaal*, https://github.com/Haneensiraj/Arabic-image-captioning-Haneen-Siraj-and-Dr-Narjis-Mezaal.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017. doi: 10.1609/aaai.v31i1.11231.

[20] H. Mubarak, "Build fast and accurate lemmatization for arabic," *Language Resources and Evaluation*, pp. 1128–1132, May 2018.

[21] K. Darwish and H. Mubarak, "Farasa: A new fast and accurate arabic word segmenter," *Language Resources and Evaluation*, pp. 1070–1074, Jan. 2016.

[22] M. I. Jordan, "Serial order: A parallel distributed processing approach," in *Neural-Network Models of Cognition - Biobehavioral Foundations*, 1997, pp. 471–495. doi: 10.1016/S0166-4115(97)80111-2.