Deep Learning for Arabic Image Captioning: A Comparative Study of Main Factors and Preprocessing Recommendations

Hani Hejazi[0000-1111-2222-3333], Khaled Shaalan[0000-0003-0823-8390]
Faculty of Engineering and IT
The British University in Dubai UAE

Abstract—**Captioning of images has been a major concern for the last decade, with most of the efforts aimed at English captioning. Due to the lack of work done for Arabic, relying on translation as an alternative to creating Arabic captions will lead to accumulating errors during translation and caption prediction. When working with Arabic datasets, preprocessing is crucial, and handling Arabic morphological features such as Nunation requires additional steps. We tested 32 different variables combinations that affect caption generation, including preprocessing, deep learning techniques (LSTM and GRU), dropout, and features extraction (Inception V3, VGG16). Moreover, our results on the only publicly avail-able Arabic Dataset outperform the best result with BLEU-1=36.5, BLEU-2=21.4, BLEU-3=12 and BLEU4=6.6. As a result of this study, we demonstrated that using Arabic preprocessing and VGG16 image features extraction enhanced Arabic caption quality, but we saw no measurable difference when using Dropout or LSTM instead of GRU.**

Keywords—Deep learning; NLP; Arabic image captioning; Arabic text preprocessing; LSTM; VGG16; INCEPTION V3

I. Introduction

Social media has increased the number of images uploaded to the web. In June, 2019 Facebook received 300 million photos a day, while Instagram received 95 million [1]. Additionally, the advent of smart devices and cameras in public places has created a challenge for automatic captioning of images to search for images by content or by human language, as well as for video context descriptions.

Image Captioning (IC) involves a lot of work since it starts with detecting and identifying objects, then it relates these detected objects, and finally it translates them into human understandable text by using their language syntax and semantics. A lot of efforts were done to overcome these challenges and a good result was achieved using deep learning techniques.

Most of the work was based on western languages. As a result, language translation was applied to benefit from these models in different languages, but the results were not as good as the original language model. For example [2] and [3] show that building an image captioning model that generates Arabic captions outperforms an English based model with the aid of Arabic translation.

Many factors were studied to understand the effect of which on captioning, like Preprocessing method, Deep learning technique, Dropout usage, and image classifier.

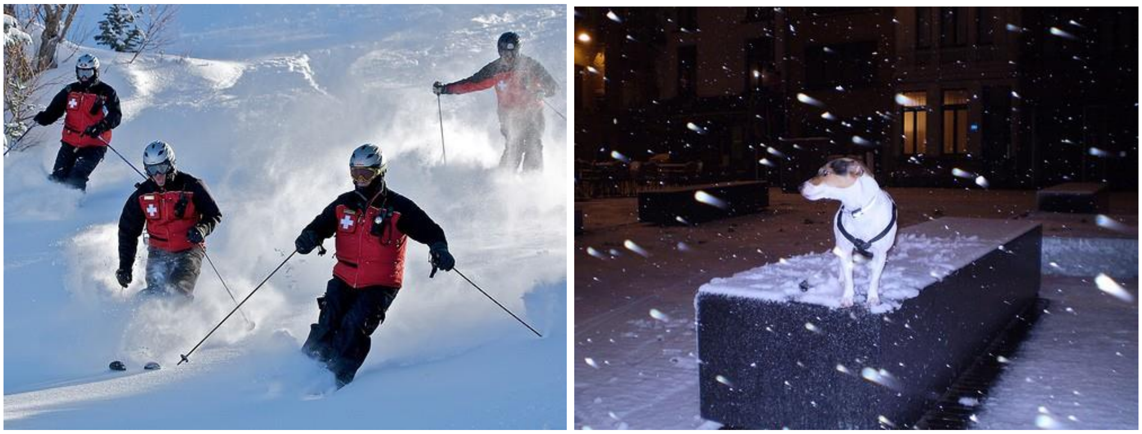
1) Dataset: One public Dataset was found for this task [2] based on Flickr8K but with just three Arabic captions for each image. However, the original Flickr8K has five captions per image. Fig. 1 illustrates two samples of this Dataset.

2) Image Features Extraction: Building CNN is common for this task, but it requires a big dataset and high processing power. An alternative way is to use a pre-trained model, as an example [2] used VGG16 [4] as a features extractor. Our work also utilized Inception V3 [5], which provides a well-optimized trained model that can be utilized even without pre-processing and training.

3) Arabic Text Preprocessing: Arabic is obviously different from English and needs preprocessing. It might have diacritic signs which affect the word’s meaning and use, but it is commonly ignored [6]. Moreover, we noticed that the conjunction Waw "(و)"in the Arabic Dataset is attached to the next word like "ويقول" (and-he-says). As per our preprocessing rule, if the letter Waw "و" (and) appears separately, it is removed as we remove all single character occurrences. Due to this, we decided to fix the typo.

4) Models: Experiments were conducted with two deep learning algorithms (GRU and LSTM), two image classifiers, and four preprocessing methods, resulting in 32 models. They were compared based on their performance.

5) Evaluation: Bilingual evaluation understudy (BLEU) metric is used to evaluate between different language translation and image captioning accuracy. For the purpose of comparing the effects of each understudy factor, we have used BLEU-1, BLEU-2, BLUE-3, and BLUE-4.



1 الناس يتزلجون على تلة مغطاة بالثلوج.
2 المتزلجين في الزي الموحد يتقدمون نزولا على منحدر ثلجي.
3 هناك أربعة متزلجين على الثلج يتزلجون على جانب التل.

1 كلب يقف على مقعد على الثلج.
2 كلب يقف على مقعد بينما الثلج يتساقط.
3 الكلب يقف على شيء بينما الثلج يسقط حوله.

Fig. 1. Sample Images with Three Captions from [2] Dataset.

The contribution of this paper is to:

- Build 32 models using different parameters: 2 Deep learning methods (LSTM, GRU) X 2 With/Without Dropout X 4 Preprocessing techniques X 2 image classifiers (VGG16, INCEPTION V3), and compare the results to show the most significant factors.
- Compare the four Arabic language preprocessing techniques and compare their effects to illustrate the importance of preprocessing for Arabic versus English, where all reviewed articles do not preprocess the text.
- Develop an Arabic Image Captioning model that outperforms the best results on the publicly available dataset and use the latest Arabic Image Captioning (AIC) dataset as input to the model. Analyze the results from the perspective of Arabic preprocessing and the model's performance.

In the next section we review the related work done for both Arabic and English IC. In Section III, Methodology, experiment, and Dataset are described, then the results are discussed and comparisons were illustrated to show the enhancement achieved by each experimented factor in Section IV, at the last section we give some concluding remarks.

II. Related Work

Recent work on Image Captioning is reviewed for both Arabic and English. We noticed that there is a lack of Arabic image captioning datasets available for tackling this task in Arabic compared to English.

A. *English Image Captioning*

The author in [7] introduced a convolution framework for image captioning consisting of four parts that begin with embedding layer for the input text, embedding for the input image, and then convolution model at the end embedding for output generation. A comparison is made against the LSTM model on the challenging MSCOCO dataset. Another experiment was done based on feed forward network that can operate over all words in parallel, and the results outperformed the baseline LSTM model.

The author in [8] introduced a novel method for image captioning by using visual regions relationships, graph neural network and context aware attention mechanism for caption generation, memorizing previous visual content was the competitive edge in the model. The model is trained and tested on MSCOCO and Flickr30K Dataset, the reported results showed that this model can outperform the state-of-the-art attention-based methods as per the authors.

The author in [9] proposed new Visual Question Answering (VQA) model based on Cascading Top-Down attention (CTDA) captioning where each keyword in question is mapped to a region in the image. A good performance was demonstrated with VQA V2.0 and V1.0 datasets.

The author in [10] applied reinforcement learning with self-critical sequence training (SCST) with CIDEr metric as a reward. It is applied on MSCOCO dataset and the result was promising in its time.

The author in [11] introduced Bottom-up attention CNN by dividing the image into regions and features vector. The model was built on MSCOCO Dataset and showed a promising result.

The author in [12] built a model for captioning images, which was then applied to question answering based on MSCOCO datasets.

B. *Arabic Image Captioning*

The author in [2] have built end to end model for Arabic Image Captioning (AIC) based on image features extractor VGG16 and LSTM for language model. Also introduced a new public dataset for AIC. They found that directly generating captions from an Arabic dataset yielded better results than translating captions from English datasets based on models generated from those datasets.

The author in [3] has used a subset of Fliket8K that consists of 2000 images and their Arabic caption in Jason file. A CNN was used for image features extraction for captions using LSTM. Two models for English and Arabic captions were introduced and the results showed that Arabic based captioning from genuine Arabic dataset has better results than those derived from English-to-Arabic translation dataset.

While the author in [13], explored generating the text based on the Arabic root using CNN ImageNet and mapping each root to an image region. Then finding the best word to describe the image using root words trained on RNN. The caption is generated through a dependency tree representing the generated words and their relations. 405,000 images from newspapers with their captions as well as those provided by Fliker8K were translated by professional translators. Unfortunately, this dataset was not yet made public.

The author in [14] also used two datasets: one with 5358 captions for 1176 images translated by human and the second has 150 images along with 750 captions. RNN was used. The evaluation showed promising results for a larger dataset.

The objective of this section is to provide a review of the various methods used for Image Captioning and to compare them with AIC research so we can identify any gaps that need to be addressed.

It is obvious that applying machine learning approach to AIC requires big data. Our study indicates that there is less research performed in AIC and this can be due to a lack of publicly available dataset for this task. Moreover, no results yet outperformed English captioning performance.

The majority of work is focused on reapplying the deep learning method used in English image captioning without considering the Arabic language and differences. As a result, we decided to examine the factors that influence Arabic image captioning. In addition, we found one public Arabic image captioning dataset that we can use for our experiments. Using this dataset, we will choose different factors that affect the task. The purpose is to identify factors that can outperform these studies' results.

III. Methodology

In this section, we describe the characteristics of the AIC dataset. We show how we apply the preprocessing task to produce appropriate training datasets. Nevertheless, we describe Deep learning models that act as image classifiers which we are able to use for extracting features from the images.

A. Dataset

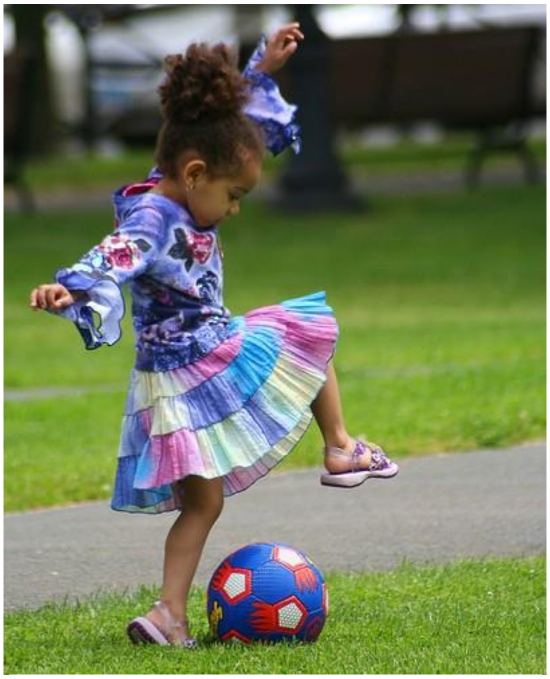
For the Image Captioning (IC) task, finding or creating a dataset is crucial in general for having better prediction results. In English, there are many benchmark IC Datasets. For example, Flickr8K [15] contains 8000 images with 5 English captions per image. Likewise, Flicker30K [16] contains 30,000 images with 150,000 captions.

Flickr30K entities [17] are reusable images which contain the caption text for either a specific entity or region and can be used for searches or retrieval tasks.

The largest dataset is MS COCO [18] that contains more than half million captions, 330,000 images with five independent captions for consistent evaluation.

1) A little girl in a dress playing with a soccer ball.
2) A little girl in a colorful dress is playing with a blue and red soccer ball.
3) Girls in brightly-colored clothes plays with a blue ball.
4) The young girl is kicking a blue and red soccer ball.
5) Young girl in blue dress stepping over a soccer ball.

For Arabic captioning, [2] introduced the first publicly avail-able AIC dataset that is based on Fliker8K, with 8000 images, 6000 for training, 1000 for validation, and the remaining 1000 for testing. Fig. 1 shows a sample of images and captions from this dataset. The author in [2] translated Flickr8K output using Google Translate API and the best three translations is post-edited, if needed by human expert. Since the dataset was generated by machine translation, some low-quality Arabic sentences appear in Fig. 2).



1 الفتاة الصغيرة ترفس كرة القدم الزرقاء والحمراء.
2 فتاة صغيرة في ثوب لعب كرة القدم.
3 تلعب فتاة صغيرة فى ثوب ملون بكرة القدم الزرقاء والحمراء.

Fig. 2. Sample Image with Translated English Caption Result in Inaccurate Arabic Sentences from [2] Dataset.

B. Preprocessing Techniques

We have used four Preprocessing techniques. Each technique generates a different dataset, namely: A, B, C, and D. Below, we provide the detailed description of each of which:

1) Original Text (Method A): To evaluate the effect of text preparation in the experiment, we used the captions as is.

2) Base Preprocessing (Method B): Both [2], [19] used the traditional technique proposed by [6]. In this method, punctuation, diacritics, non-Arabic letters, single letter words were dropped. Also, a lexographic normalization process took place to unify similar letters, including " آأإاا " - "ا" , "ى" - "ي" , "ئ ؤ " - "ء" , "ة" - "ه" , "گ" - "ك" .

3) Removing the Alef with the Nunation (Method C): We have noticed that when removing Tanween diacritic the extra Alef is not removed. So, we removed this extra Alef too, such that the word "قميصًا" (shirt-with extra nunation-) becomes "قميص" (shirt-without nunation-) instead of "قميصا" (shirt-with Alef as partial nunation-). Applying this technique would reduce the total vocabularies because in the previous method each surface form was considered a different vocabulary as illustrated in Fig. 3. Moreover, we separated and removed the Waw conjunction from next word, e.g. "ويقول" (and-he-says) becomes "يقول" (he-says).

4) Full Preprocessing (Method D): We partially followed Method C, but we kept the conjunction Waw. In all previous methods all single letter words was removed including the isolated conjunction Waw, e.g. "ويقول" becomes "يقول" "و." but we think this highly affect syntactic and semantic of the captions. Fig. 3 shows differences in the frequency counts for preprocessing methods B, C, and D.

The final caption is then surrounded by a start and end tags. The length of each caption is set to 25 words; shorter captions are padded with nulls.

Table I shows the output of the four preprocessing methods along with their statistics. Since we dropped words with single appearance we can notice in the third column of the table a big reduction in the repeated vocabularies count. For example, applying Method C to the dataset produces 9,713 unique vocabularies but only 5,344 of them were repeated and the remaining 4,369 should be removed.

The reason of having these words sparse in the caption dataset might be due to misspelled words or the use of rare words. If size of the dataset is small, it might make the caption not a good representative for the Arabic Language model, since many words rarely appear or do not appear at all. This raises the need for a big enough dataset for AIC.

Low frequency words affect the prediction process, so they have to be treated at the preprocessing stage since often they are typos. Fig. 3 shows how the proposed methods C and D reduced the occurrences of words with just one appearance from 4963 (Method B) to 4369 a decrease of about 12%. As per (Fig. 3), the number of low frequency words reduced in most cases, but we can observe an increase of the number of words with 12 and 13 frequency; this might be due to the matching between words with low frequencies after applying the preprocessing task.

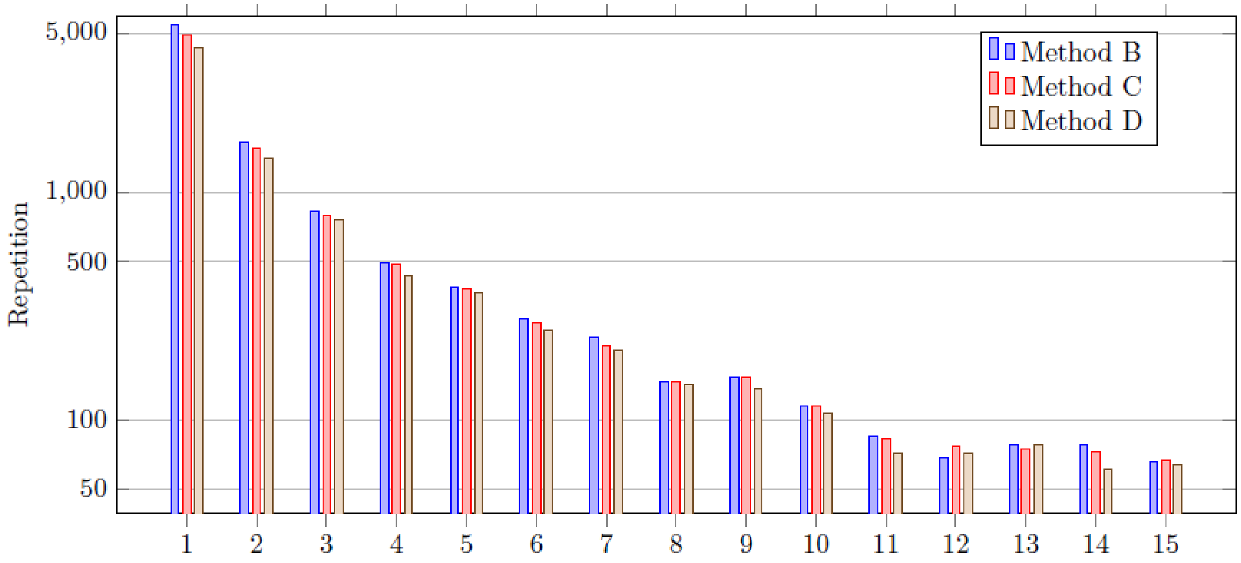


Fig. 3. Variation in the Frequency Counts for Each Preprocessing Method (Rare Counts).

TABLE I. PREPROCESSING METHODS USED, WITH SAMPLE CAPTIONS AND NUMBER OF DETECTED VOCABULARIES

Preprocessing method	**Sample Caption**	**Total Vocabularies**	**Unique Vocabularies**	**Unique Repeated Vocabularies**
Method A	صبي يرتدي نظارات و قميصًا أحمر	179,532	11,386	5,893
Method B	صبي يرتدي نظارات قميصا احمر	178,176	10,692	5,729
Method C	صبي يرتدي نظارات قميص احمر	178,175	9,713	5,344
Method D	صبي يرتدي نظارات و قميص احمر	183,342	9,714	5,345

C. *Models*

Recurrent Neural Networks (RNN) is best used for time series data, but it suffers from the short term memory problem or the vanishing gradient where the earlier inputs effect starts to be exponentially smaller when we move more steps forward in the prediction. We can resolve this by using one of the following variations: Gated Recurrent Unit (GRU) or Long Short Term Memory (LSTM) where a gates are used to control the older sequence information by saving in memory unit and propagate to next units.

Since text is considered a Time Series prediction we propose to use GRU and LSTM network in our experiment and compare their performance and effect on the results.

D. *Experiment*

Experiments were designed to test the impact of our independent variable on the quality and accuracy of Arabic captions. We have conducted experiments that involved 32 variable combinations: 4 Datasets, 2 image classifiers, 2 dropout usage, and 2 Deep Learning methods.

Fig. 4 shows the experiment design where we have indicated four labels to highlight the variant stages of the experiments. In the first stage (1) images are passed to one of two features extractors (Inception V3, VGG16). Next, a vector that contains image features is produced, captions are preprocessed using the four methods then tokenized, and then passed to embedding layer.

Afterwards, a dropout layer is used, if required by experiment, and the results are passed to either LSTM or GRU. At the end a Dense layer is used for prediction. Each model is saved, and test images are passed to it for caption prediction. All predicted captions are recorded and compared with the actual ones. BLEU- 1/2/3/4 scores are calculated and stored per each experiment. Table II shows the recorded results which we analyze and discuss in the next sections. In each experiment one path is chosen at a time until all combinations are covered. Many experiments were repeated with lower epoch when Overfitting is detected.

The configuration of the hardware used is: Intel(R) Core(TM) i7 10th generation (6 core, 12 logical processors) with NVIDIA GeForce GTX1 1650 (4GB) for processing, 16 GB RAM Memory, total accumulated training time for latest models about 7 hours.

The collected experiment data was analyzed to find the effect of each factor. Also, a t-test is applied to find the significance of each variable.

E. *Overfitting*

Since the size of Dataset is small training and testing (validation) loss value is monitored after each epoch, if the testing loss increases or stays the same while the training loss decreased, this means an overfitting is detected and we observe a lower prediction accuracy from that model.

Then lower number of epochs are made to reach the lower testing loss value and a better model accuracy (BLEU measure).

F. *Evaluation*

To evaluate each experiment result, BLEU-1/2/3/4 are used. BLEU is a precision-based metric that ranges between zero (lowest) and one (best). The number of n-grams that appears in the candidate text is compared to total n-grams in the reference text. This metric is used by [2], which we use to compare our results with their results.

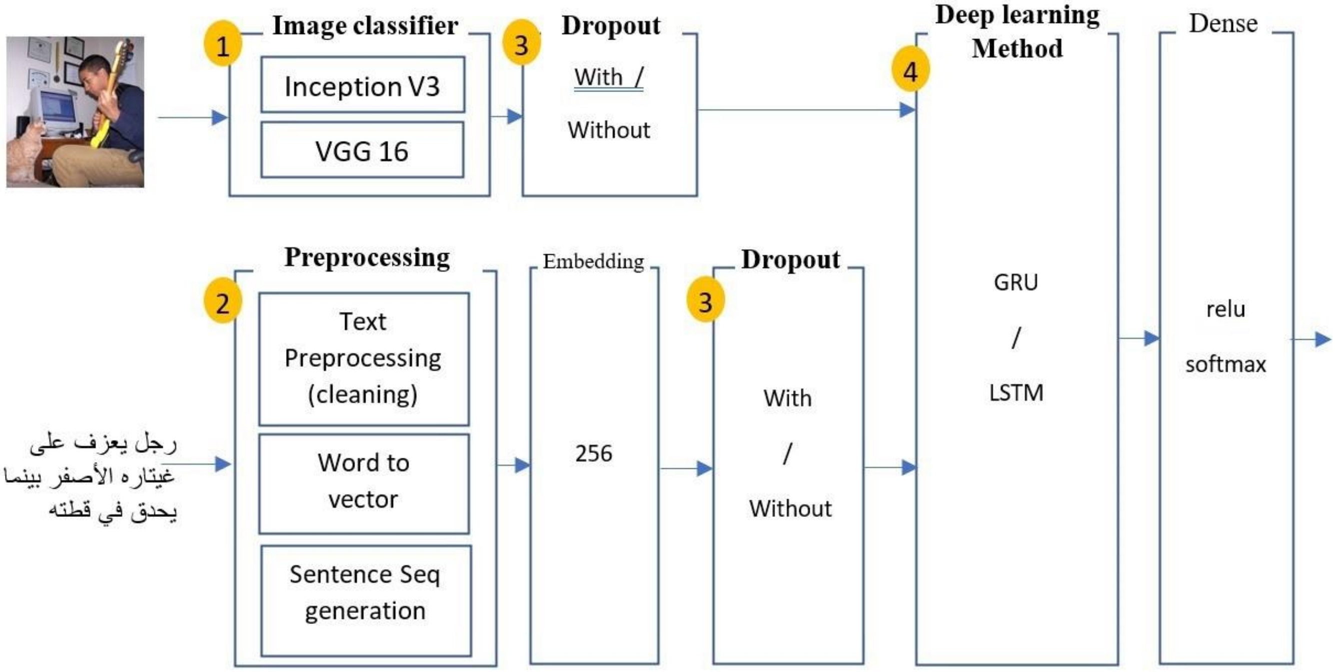


Fig. 4. Experiment Flow that Yields a Total of 32 Experiments: (1) Two Image Classifiers, (2) 4 Preprocessing Methods, (3) Dropout, (4) Two Deeplearning Techniques.

TABLE II. BLEU-1/2/3/4 RESULT OF THE EXPERIMENT PER VARIABLES COMBINATIONS

Image Classifier	Model	Dataset	Dropout				No Dropout			
				BLEU%				BLEU%		
			1	2	3	4	1	2	3	4
Inception V3	GRU	A	26.6	13.4	6.8	3.6	29.5	14.9	7.8	4.2
Inception V3	GRU	B	28.3	14.7	7.4	3.0	28.3	13.5	6.7	3.0
Inception V3	GRU	C	30.1	15.8	7.9	3.9	29.9	15.7	8.3	4.6
Inception V3	GRU	D	34.1	17.7	9.5	5.3	29.9	16.6	9.4	5.1
Inception V3	LSTM	A	24.4	10.7	4.8	1.8	22.6	10.7	5.1	2.0
Inception V3	LSTM	B	27.6	11.7	4.7	2.0	24.1	11.4	5.1	2.1
Inception V3	LSTM	C	27.8	13.5	6.5	3.0	26.3	11.1	4.5	2.1
Inception V3	LSTM	D	31.8	15.3	8.0	4.6	27.1	12.2	5.7	2.9
VGG16	GRU	A	24.6	13.3	7.2	4.0	24.0	12.9	6.4	3.1
VGG16	GRU	B	23.5	13.2	7.1	3.6	28.2	15.1	8.3	4.6
VGG16	GRU	C	31.1	17.5	9.0	4.1	30.8	16.8	8.8	4.4
VGG16	GRU	D	26.5	15.1	8.8	5.1	36.5	21.4	12.0	6.6
VGG16	LSTM	A	33.6	20.1	11.2	6.4	32.3	18.5	9.8	5.3
VGG16	LSTM	B	33.9	19.5	10.5	5.7	31.2	17.9	9.7	5.5
VGG16	LSTM	C	35.1	20.9	11.5	6.3	33.1	18.9	10.1	5.2
VGG16	LSTM	D	30.7	18.2	10.1	5.4	34.2	19.9	10.8	6.1

IV. RESULT

In this section, we present results from 32 experiments. Table II shows the BLEU results of each experiment. Fig. 5 illustrates these results.

A. BLEU

BLEU-1/2/3/4 was used to measure accuracy of each model prediction. Table II shows the results of these experiments.

We can notice that the best BLEU scores achieved from using VGG16 with GRU on the Dataset generated using the method D, and without dropout, are BLEU-1=36.5, BLEU-2=21.4, BLEU-3=12, and BLEU-4=6.6.

B. Preprocessing Methods Comparison (Datasets)

Each Dataset is produced using a different Preprocessing method, we compared the three Datasets (B,C,D) to show the effect of Preprocessing on the results accuracy. Fig. 6 illustrates the BLEU-1's result.

We can notice that the proposed new Preprocessing methods give higher BLEU measure. The reason might be due to less infrequent words that arise from consistent typo, such as concatenating Waw with the next word, or keeping the Alef of nunation, which produces a vocabulary that is irrelevant to the original word.

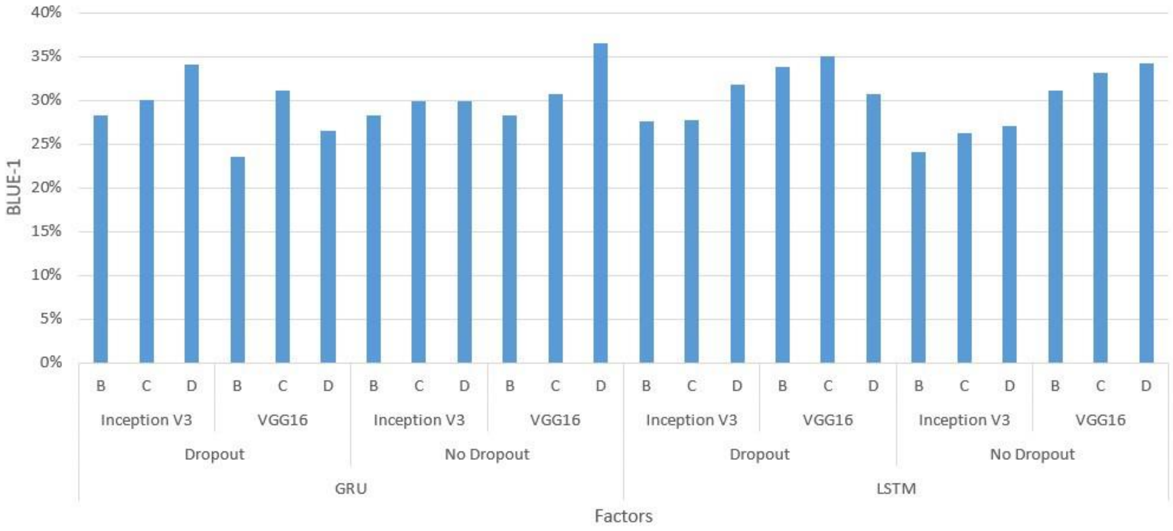


Fig. 5. Experiments Results for BLEU-1 upon Different Parameters.

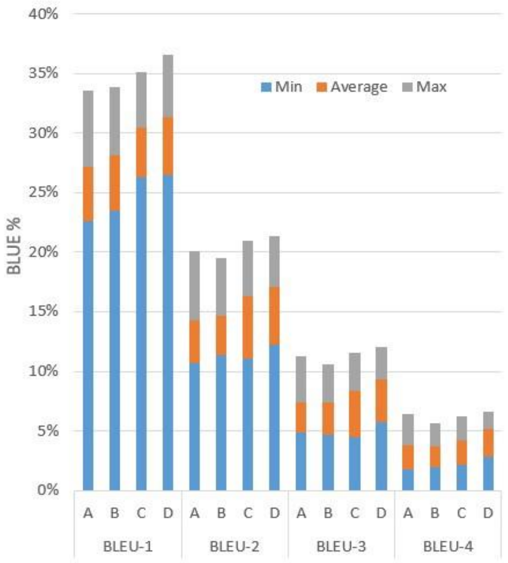


Fig. 6. Average, Minimum, and Maximum Value of BLEU-1/2/3/4 achieved per each Preprocessing Method.

A paired-samples t-test was conducted to compare the Dataset C with the Dataset B. There is a significant difference in the scores from Dataset C (M=0.1482, SD=0.1045) and Dataset B (M=0.1346, SD=0.0978) under the conditions: t(31)=5.0344, p = .0.000019.

These results suggest that removing the Alef of the nunation affect the BLEU results and increases it.

Another paired-samples t-test was conducted to compare Dataset D with Dataset C. There was a significant difference in the scores for Dataset C (M=0.1482, SD=0.1045) and Dataset D (M=0.1571, SD=0.0.1044) under the conditions: t(31)=-2.2136, p = .0.000019 These results suggest that keeping the Waw in the preprocessing phase affect the BLEU results and increases it.

C. Image Features Model Comparison

We involved two image models to extract image features, VGG16 and Inception v3. Fig. 7 illustrates a comparison of BLEU results of both models.

A paired-samples t-test was conducted to compare using VGG16 and Inception V3 as image features extractor.

There is a significant difference in the scores for VGG16 (M=0.1564, SD=0.1011) and Inception V3 (M=0.1294, SD=0.0.0976 under the conditions: t(63)=5.6714, p = .0.00000038 These results suggest that using VGG16 over Inception V3 affect the BLEU results and increases it.

D. DropOut Comparison

We have studied the impact of using the Dropout with Arabic image captioning process. Fig. 8 illustrates the results of experiments with/without Dropout.

A paired-samples t-test was conducted to compare the results with and without Dropout. There was not a significant difference in the scores for using Dropout (M=0.1423, SD=0.1005) and not using dropout (M=0.1436, SD=0.0.1001 conditions; t(63)=-0.46, p = .0.647.

There is no evidence that using Dropout will affect the BLEU results of the generated captions.

E. GRU vs LSTM

Two Deep Learning methods were compared (GRU, LSTM). Fig. 9 illustrates the BLEU results per each method.

The use of GRU or LSTM as a text prediction model was compared using a paired-samples t-test. There is no significant difference in the scores for GRU (M=0.142, SD=0.097) and LSTM (M=0.1438, SD=0.0.1035) under the conditions: t(63)=0.419, p = .0.6766.

These results cannot support that using GRU instead of LSTM may affect the BLEU results of the generated captions.

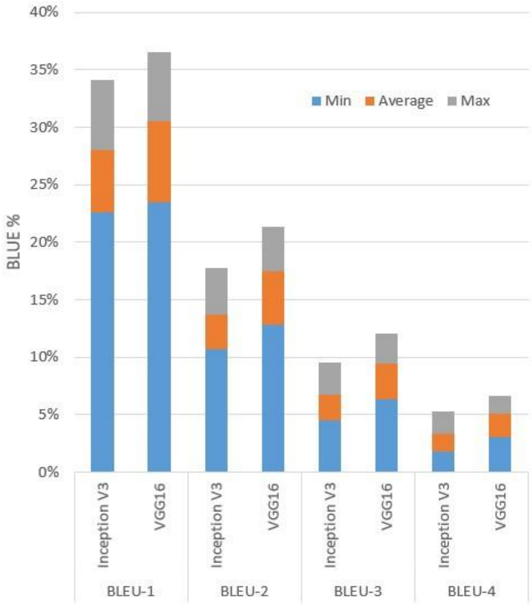


Fig. 7. Average, Minimum, and Maximum Value of BLEU-1/2/3/4 achieved per each Image Features Extraction Model.

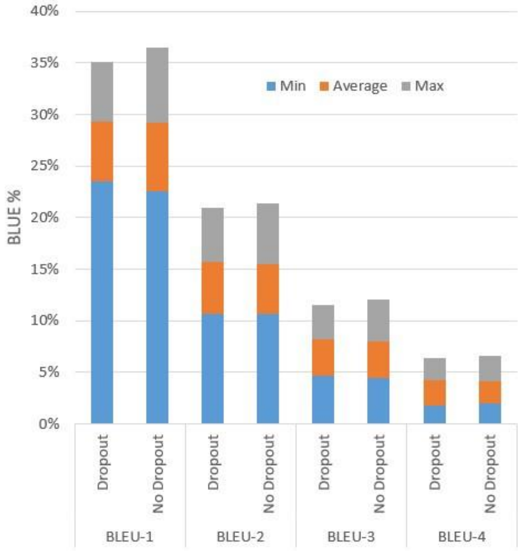


Fig. 8. Average, Minimum, and Maximum Value of BLEU-1/2/3/4 achieved per Dropout usage.

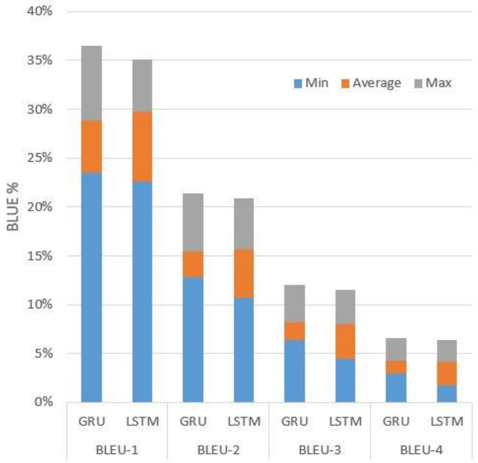


Fig. 9. Average, Minimum, and Maximum Value of BLEU-1/2/3/4 achieved per each Deep Learning Method.

V. Conclusion

Arabic Image Captioning resources are scarce. Fortunately, one public dataset is available. We created an AIC model with tuned factors that outperformed the best results on the publicly available dataset. According to paired t-tests conducted on the results, Arabic text preprocessing and image features extractors have a major role to play in improving the AIC results. For the purpose of comparison, two preprocessing techniques for Arabic captions were proposed and found to yield better results.

A total of 32 experiments were conducted to analyze the effects of four variables. We considered the following variables: preprocessing techniques (original text, normal preprocessing, Alef removal with nunation, and keeping conjunction Waw), Waw typo correction, Deep learning techniques (LSTM, GRU), inclusion and exclusion of Dropout, and two Image features extraction methods (Inception V3, VGG16).

As a result, BLEU1=36.5, BLEU-2=21.4, BLEU-3=12, and BLEU-4=6.6 were the best results we reached. The results were compared using paired t-tests, and the Arabic preprocessing methods exhibited an enhanced level of quality, and VGG16 significantly outperformed Inception V3. Using Dropout or LSTM instead of GRU, however, did not have a major effect.

VI. Limitations and Future Work

The main limitation was the relatively small Dataset size since there was only one publicly available Dataset for AIC. Other Preprocessing and Deeplearning methods could be included in the comparisons but doing that will increase the number of experiments and require more resources, therefore we can consider it in the future work.

As a future work, researchers can benefit from the outcomes of this study by employing it to their future research, particularly, a larger dataset can be created and made public to avail linguistic resources research in this area.

Not to mention, having a big dataset provides several possibilities to tailor the use of extra deep learning techniques and come up with better word representation and features that can significantly improve the performance of the Arabic Image Captioning.

References

[1] D. Stout, Social Media Statistics, 2020 (accessed June 27, 2020).

[2] O. ElJundi, M. Dhaybi, K. Mokadam, H. M. Hajj, and D. C. Asmar, ``Resources and end-to-end neural network models for arabic image captioning." in VISIGRAPP (5: VISAPP), 2020, pp. 233-241.

[3] R. Mualla and J. Alkheir, ``Development of an arabic image description system," International Journal of Computer Science Trends and Technology (IJCST), vol. 8 no. 3, 2018.

[4] K. Simonyan and A. Zisserman, ``Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv: 1409. 1556, .2014.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, ``Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818-2826.

[6] A. Shoukry and A. Rafea, ``Preprocessing Egyptian dialect tweets for sentiment mining," in The Fourth Workshop on Computational Approaches to Arabic Script-based Languages, 2012, p. 47.

[7] J. Aneja, A. Deshpande, and A. G. Schwing, ``Convolutional image captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5561-5570.

[8] J. Wang, W. Wang, L. Wang, Z. Wang, D. D. Feng, and T. Tan, ``Learning visual relationship and context-aware attention for image captioning," Pattern Recognition, vol. 98, p. 107075, 2020.

[9] W. Tian, R. Zhou, and Z. Zhao, ``Cascading top-down attention for visual question answering," in 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020, pp.1-7.

[10] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, ``Selfcritical sequence training for image captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7008-7024.

[11] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, ``Bottom-up and top-down attention for image captioning and visual question answering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6077-6086.

[12] J. Wu, Z. Hu, and R. J. Mooney, ``Generating question relevant captions to aid visual question answering," arXiv preprint arXiv:1906..00513, 2019.

[13] V. Jindal, ``Generating image captions in arabic using root-word based recurrent neural networks and deep neural networks," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

[14] H. A. Al-Muzaini, T. N. Al-Yahya, and H. Benhidour, ``Automatic arabic image captioning using rnn-lst m-based language model and cnn," International Journal of Advanced Computer Science and Applications, vol 9, no.6, 2018.

[15] 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.

[16] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, ``From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," Transactions of the Association for Computational Linguistics, vol. 2, pp. 67-78, 2014.

[17] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2641–2649.

[18] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, “Microsoft coco captions: Data collection and evaluation server,” arXiv preprint arXiv:1504.00325, 2015.

[19] H. D. Hejazi, A. A. Khamees, M. Alshurideh, and S. A. Salloum, “Arabic text generation: Deep learning for poetry synthesis,” in Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2021. Springer International Publishing, 2021, pp. 104–116.