Arabic Image Captioning: The Effect of Text Pre-processing on the Attention Weights and the BLEU-N Scores

Moaz T. Lasheen, Nahla H. Barakat
Faculty of Informatics and Computer Science, The British University in Egypt, Cairo, Egypt

***Abstract*—Image captioning using deep neural networks has recently gained increasing attention, mostly for English langue, with only few studies in other languages. Good image captioning model is required to automatically generate sensible, syntactically and semantically correct captions, which in turn requires good models for both computer vision and natural language processing. The process is more challenging in case of data scarcity, and languages with complex morphological structures like the Arabic language. This was the reason why only limited number of studies have been published for Arabic image captioning, compared to those of English language. In this paper, an efficient deep learning model for Arabic image captioning has been proposed. In addition, the effect of using different text pre-processing methods on the obtained BLEU-N scores and the quality of generated images, as well as the attention mechanism behavior were investigated. Furthermore, the "THUMB" framework to assess the quality of the generated captions is used -for the first time- for Arabic captions' evaluation. As shown in the results, a BLEU-4 score of 27.12, has been achieved, which is the highest obtained results so far, for Arabic image captioning. In addition, the best THUMB scores were obtained, compared to previously published results on common images.**

***Keywords*—Arabic image captioning; computer vision; deep learning; image captioning; natural language processing**

I. Introduction

A. Overview

Recently, automatic image captioning became a hot topic, building on the success of deep neural networks in the areas of computer vision and Natural Language Processing (NLP) tasks. Image captioning models require two main components; the first is to extract the image's features, detect its objects, and describe their relationships; while the second is the language model that converts those features to a meaningful word sequence [1, 2]. These models are initially trained on a data set of images, along with their corresponding captions [3].

Studies for caption generation is largely in English due to the availability of data sets and other pre-trained image and language models. The most commonly used architectures are Encoder-Decoder, with or without additional, optional layers, like different attention mechanisms and different embedding models [3-5]. The Encoder–Decoder architecture mainly uses several variations of a Convolutional Neural Network (CNN) as encoders, where high-level feature are extracted from the input images, which are then passed to the decoder (language) model; where Recurrent Neural Networks (RNN) have been widely used. Recently, transformers, as well as Generative Adversarial Neural Networks (GANS) models have been used. Encoder architectures included AlexNet, VGG-16 Net, RESNet, GoogleNet and DenseNet [6]. However, RESNet showed better performance, and had fewer training parameters compared to other common encoders like VGG variants [3]. For language models (the decoders), Long Short-Term Memory (LSTM), RNN, Gated Recurrent Units (GRU) have been adopted [6]. However, the LSTM is the most widely used decoder, for its ability to remember long term dependencies in the generated word sequence [6]. Several attention models have also been proposed; including hard or soft, top-down, bottom-up, semantic, and other attention methods [6, 7]. Attention methods are also used with GANs and Reinforcement Learning, which have shown excellent performance [8]. For more details on English image captioning, please refer to the reviews in [3-9]. The situation is different for Arabic image captioning, as only few models have been proposed with less satisfying results. This can be attributed to the complex morphological structure of the Arabic language, and the scarcity of data sets of images with Arabic captions. Image captioning has many valuable applications; like image indexing and retrieval, assisting visually impaired people, robot vision systems, medical image description, analysis of traffic data, and other industrial applications [10, 11].

In this paper, an efficient model for Arabic image captioning is proposed, utilizing an encoder-decoder architecture, with soft attention mechanism, and beam search to generate best captions. The paper attempts to answer the following research questions: 1) what is the effect of different Arabic text pre-processing methods on the BLEU-N scores, the behavior of the attention mechanism, and quality of the generated captions? 2) Dose beam search improve BLEU-N scores?

As the result section shows, the proposed model achieved the highest BLEU-4 score so far; for Arabic image captioning. In addition, the quality of the generated captions compares favorably to the related work, as measured by THUMB score, which is used for the first time for Arabic captions evaluations, as well as ratings of four Arabic native speakers.

The rest of the paper is organized as follows: Section B summarizes this study's contributions, followed by a review and analysis of the related work in Sections II. The

experimental methodology, results and discussion, are presented in Sections III and IV respectively. Section V discusses the effect of text pre-processing methods on the attention visualization, and the paper conclusion is presented in Section VI.

B. The Paper's Contributions

- The proposed model in this paper for Arabic image captioning achieved the highest BLEU-4 Score so far,
- For the first time, the paper investigates the effect of Arabic text pre-processing on the attention mechanism, as well as the BLEU-N scores,
- For the first time, images' captions are qualitatively evaluated using THUMB scores,
- The paper presents the most comprehensive literature review for Arabic image captioning.

II. Related Work

The first published study on Arabic image captioning was in 2018 by [12]. Since then, the majority of the published studies used Encoder-Decoder architectures, with or without attention mechanisms; and recently, transformers have been used. The following Sections summarize the work in this area.

A. Encoder – Decoder based Models

The model which obtained the best results as measured by the BLEU-4 score is [12]. This model is different from all other published ones, as it uses Region Convolutional Neural Network (RCNN) to map the image objects to Arabic root words, where a transducer based algorithm has been used for this purpose. The output root words are then passed to an LSTM to generate the standard Arabic caption, and a dependency tree constraints algorithm has been used to ensure that the generated caption is grammatically correct. The authors reported the BLEU-N scores on Middle Eastern newspapers & Flickr8 data sets [12]. In [13] , a CNN was used as the encoder, and LSTM as decoder, on a part of Flickr8 data set, and used BLEU-N for evaluation, with two additional measures. A different study by [14] also used the VGG OxfordNet as encoder and RNN-LSTM as decoder. Arabic Flickr8 plus sample of MS COCO data set with Arabized captions have been used. The authors in [2], translated the captions of Arabic Flickr8 data and made it available online. The authors [2] also proposed a model using VGG16 CNN and LSTM as encoder and decoder respectively. They also proposed a base model, which generate English captions, which are then translated to Arabic.

B. Encoder – Decoder Models with Attention Mechanism

In [15], the authors proposed three different encoders, utilizing CNNs for feature extraction and single, and/or multiple objects detection. A final hybrid model was proposed with attention mechanism, which is used for detected objects prioritization. They used LSTM with soft attention and beam search were used for the decoder. The data set used was MS COCO, and Flicker30. Unlike other studies, the authors assessed the quality of the generated captions by measuring the semantic similarity between the generated captions and ground truth captions. Another method that used attention is [16], as shown in the next section.

C. Encoder – Decoder Architectures with Transformers

Two studies reported their results [16] and [17]. In [16] three models were proposed. The first uses MobileNetV2 network as encoder, LSTM with attention as the decoder. The second was MobileNet V2 (GRU) as encoder, and GRU with attention as decoder. Finally, a transformer-based model was also proposed, where EffeceintNet is used as the encoder, and a transformer based architecture as the decoder. FARASA segmenter has been used for text pre-processing, and BLEU-N scores were reported. Different transformer based models were proposed in [17] which were initialized with AraBERT and GigaBERT pre-trained transformers, then fine-tuned by detecting object tags in images using OSCAR method. Flickr8 and part of MS COCO data sets have been used, and BLEU-N scores are reported.

D. Analysis of the Related Work

A comparison of the BLEU-N scores for the studies reviewed in this section can be found in Section IV, Table IV. From that table, it can be seen that the best reported BLEU-N scores are by [12], who used root words to generate captions, followed by [14], then the transformers based models in [16]. It was also noted that transformers achieved minor improvements on the BLEU-N scores. However, the comparison of BLEU-N scores does not provide a concrete conclusion which model is better; as most of the studies did not use a common train/ validation/ test splits. For example, [16] used 90%, 10 % for training, and testing respectively. Also, in [17], the MS COCO images used only for training, Flickr8 test set have been used for testing. Similarly, in [13], 1500, 250, and 250 images have been used for training, validation and testing respectively. Excluding the results by [12], the BLEU-4 results are close, which does not pinpoint the value of a specific architecture over the others. Furthermore, none of the studies reviewed here investigated the effect of pre-processing methods on the BLEU-N scores or the quality of the generated captions.

III. Experimental Methodology

A. The Dataset

The data set used in this paper is the Arabic-Flickr8 [2], which is a translated version from the original English Flickr8, using Google Translate. The best 3 translated captions are kept and further edited by native Arabic speakers. For the purpose of training, validation, and testing, Karpathy's data splits [18]; 6000, 1000, 1000 images for training, validation, testing respectively are used. Unlike its English version which has 5 captions for each image, the Arabic Flickr 8 has only 3 captions.

B. The Model Architecture

The model used in this paper follows the Encoder- Decoder architecture, with attention mechanism, teacher forcing, and beam search. The selection of the Encoder and Decoder networks is based on their prior excellent performance in computer vision and NLP problems, details as follows:

1) The Encoder: The RESNet-101 [19], is a CNN that has 101 layers, and was chosen -for the first time in Arabic image captioning- as the encoder; due to its proven ability to extracts very rich feature set from an image. RESNet stands for Residual Neural Network architecture [19], which is able to overcome the vanishing gradients problem using skip connections. The output of the encoder is passed to the next part of the architecture with its same dimensions.

2) The attention mechanism: The objective of using attention mechanisms [7] in image captioning is to allow focusing on a specific part of the image, while generating the captions. It calculates the weights of different pixels of the encoded images, which are then used by the Decoder. In this paper, soft attention has been used, which is trained in an End-to-End manner using Back-propagation. The soft attention weights are determined by image features and the LSTM previous output.

3) The decoder: An LSTM network is used as the decoder. The LSTM is a variation of the RNNs with additional gates, and is able to overcome the vanishing gradients problem encountered when processing long sequences. Those gates make the RNN decide which tokens should be retained in the memory and which to forget [3]. In the context of image captioning, the decoder looks at different parts of the image; while producing different parts of the output sequence, by weighting different pixels of the output of the encoder. The LSTM cell and hidden states are initialized using the encoded image at the first step; and the encoded image attention weights alongside the decoder weights at each step are computed. The attention weights with the embedding of the token from the previous step, are then concatenated and the LSTM produces the new states.

C. Pre-processing

In this section, the pre-processing steps used are described:

1) Captions' Pre-processing and Tokenization: In this study, Pyarabic [20], which splits image captions into tokens using spaces, and The FARASA segmenter [21] have been used. FARASA [21] is an Arabic word segmenter, which breaks Arabic words into their constituent clitics. For example, the word "wkatabna" (وكتبنا) meaning: "and we wrote" is composed of three clitics "w+katab+na", namely the conjunction article "w" meaning "and" as prefix, the stem "katab" (كتب), possessive pronoun "na" (نا) as suffix. Another example which is very common is the "AL - ال", which corresponds to "the" in English. In the context of image captioning, using FARASA segmenter results in a smaller unique vocabulary size because all different forms of a word are treated as one, but the number of total number of tokens increases, as different suffixes are separated and counted.

2) Image pre-processing: The image pre-processing is kept to the minimum; where all images are resized to 256 (smaller edge) pixels. The center 224x244 pixels was cropped, before transforming them to Pytorch tensors; and normalize using the mean and standard deviation of the IMAGENET dataset.

D. Pre-trained Word Embedding

The Aravec; a pre-trained word embedding have been used to provide richer representations for the image captions. In this study, the skip gram model trained with Wikipedia data have been used [22].

E. The Beam Search

For training and validation, teacher forcing has been used [3], as it makes the model learn the context in more efficient way. At the testing stage, beam search has been used to generate the best captions. Beam search [3] works by finding the top-k words with the highest decoder scores at each step, calculate the additive scores for each of the pairs from current and previous steps and get the best combinations, in each decoding step. In this way, beam search outputs the completed sequences with highest scores. The beam size that resulted in captions with best BLEU-4 score is chosen.

F. Modeling

Two models have been designed to generate image captions; and investigate the effect of the following settings on the quality of the generated captions:

- The use of different text pre-processing in particular, FARASA word segmenter and PyArabic tokenization,
- The use of different Beam Sizes.

1) Model 1: It was decided to start with a base model, as a reference for comparison. So, PyArabic tokenizer, with AraVec pre-trained embedding was used, to compensate for the smaller number of captions for the Arabic Flicr8 data set, compared to its English version.

2) Model 2: In this model, FARASA Segmenter has been used to pre-process the captions, and similar to model 1, AraVec embedding model has been utilized.

For both models, beam search has been used to generate best captions, and they were evaluated on Flickr8 test set, as well as 200 images randomly selected from MS COCO data set, to further test the robustness of our models.

G. Evaluation Methods

1) The BLEU–N Score: The BLEU [23] stands for Bilingual Evaluation Understudy, which is a metric originally proposed to evaluate the quality of machine translation models. As image captioning can be thought of as translation from image features to text describing that image, it has been widely used to evaluate image captioning models. BLEU-1, 2, 3, 4, measures the fraction of n-grams that appear in both the generated and ground truth captions, where n takes the values, 1, 2, 3 and 4.

2) The THUMB 1.0: As human evaluation of the generated captions is still considered the gold standard, the THUMB framework for caption's evaluation is used. THUMB stands for "Transparent Human Benchmark" [24]. THUMB is based on two major scores; namely, precision and recall,

which are measured on a scale from 1 to 5. This in addition to penalty scores which are deducted from the average of precision and recall scores, to penalize any problems in the fluency and /or conciseness of the generated captions, as well as any issues concerning the use of inclusive language. The following sections briefly introduce the THUMB 1.0 [24] framework.

a) Precision: As the measure entails, the Precision (P) assess how precise the image is described by the generated caption, which is mainly intended to detect the common failures of the language model part. Precision is measured on a scale from 1 to 5, where 0.1 point is deducted in situations like, minor difference in colors, counts, the caption is not accurate, but do not mainly contradict with image's contents, in addition to other attributes like occasions, locations, etc. [24].

b) Recall: Recall (R) evaluates how good (complete) does the caption describe the image contents; including main objects, their relationships and colors. Therefore, it penalizes the generic, short captions that are usually generated by the majority of image captioning methods. If the image description (caption) is too generic, where different diverse images can be imagined based on that caption, then the recall score tends to be low [24].

c) Penalties: Penalties are given to penalize fluency problems; which assesses the text structure, regardless of the image contents. As most of automatically generated captions do not suffer fluency problems, points are deducted from the average of the precision and recall scores. A penalty of 0.1 is given for grammatical or spellings mistakes, as they are easily corrected. For other more serious problems like duplication, broken sentences, a minimum of 0.5 points are deducted. The Conciseness is also evaluated in this framework, where penalties are given for unnecessarily long, detailed captions, where 0.5 points are deducted. However, as the majority of automatically generated captions tend to be short, this penalty is not very common [24]. The final type of penalties is given on describing humans with terms that deviate from inclusive language**,** which ranges from 0.5 for subjective comments, to 2.0 for more severe problems. A final rule in this framework, is that double penalties should be avoided. If a problem is penalized using precision, it should not be penalized again by recall [24].

IV. Results and Discussion

A. Model 1 Results

Results of model 1 are shown in Table I. From this table, it can be seen that the model achieved BLEU-4 score of 8.29, which is superior to 9 of previously published BLEU-4 scores. This can be attributed to the strong encoder architecture used as well as the use of the attention mechanisms and beam search.

B. Model 2 Results

Results of model 2 are shown in Table II. From this table, it can be seen that the BLEU-N scores have significantly increased for all beam sizes. The best BLEU-4 results were obtained with beam size of 5. The results show that the use of FARASA segmenter significantly improved all the scores. These results are consistent with [12], where root words have been used, which is a similar approach to FARASA pre-processing. This model achieved *the highest BLEU-4 scores* obtained on the Arabic Flickr8 data set so far. Fig. 1 shows the BLEU-4 scores obtained at different beam sizes, for PyArabic, compared to FARASA pre-processing. From this figure, it can be seen that the best results were obtained with beam size 5. This can be attributed to the improvements of the completed sequence of words with beam size of 5, compared to the ground truth captions for the test set.

C. Results on 200 Images from MS COCO

As an additional test of the models quality 200 randomly selected images from MS COCO data set with Arabized captions were used. The results are shown in Fig. 2. From this figure, it can be seen that better results were obtained; again by the model with FARASA segmenter, and beam size of 3, which is the same situation as Flickr8 test set. Fig. 3 compares the performance of the two models on MS COCO data set.

TABLE I. Scores of Model 1 on Arabic Flickr8 with Beam Search

Model	BLEU-1	BLEU-2	BLEU-3	BLEU- 4
Model 1 + Beam size 1	39.01	24.45	13.01	7.27
Model 1 + Beam size 3	40.10	25.58	14.28	7.89
Model 1 + Beam size 5	39.10	25.13	13.96	**8.29**

TABLE II. Comparison of BLEU-4 Score for FARASA pre-processing, and Beam Sizes

Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4
FARASA + Beam size 1	57.45	43.79	31.86	22.81
FARASA + Beam size 3	59.90	47.40	36.13	26.89
FARASA + Beam size 5	58.71	46.52	35.71	**27.12**

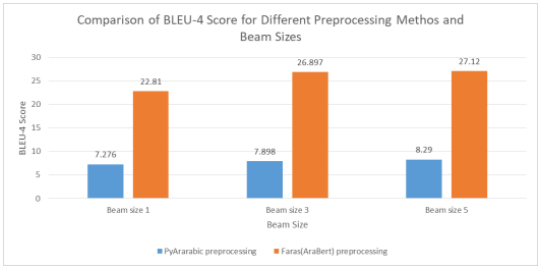


Fig. 1. Comparison of BLEU-4 Score for different Preprocessing Methods and Beam Sizes.

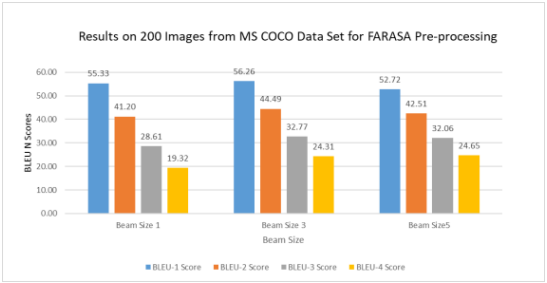


Fig. 2. Comparison of BLEU-4 Score for different Beam Sizes.

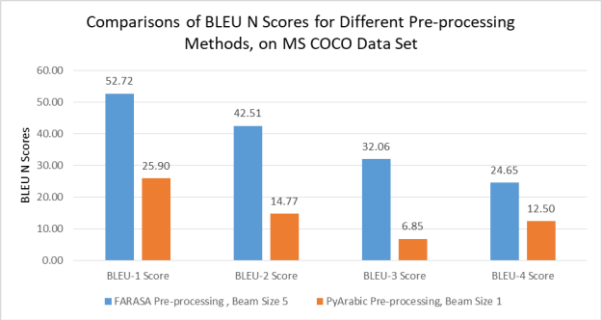


Fig. 3. Comparison of BLEU-N Scores, for Models 1 and 2, on 200 Images from MS COCO.

D. Results' Comparison with Related Work

Fig. 4 shows our BLEU-4 results, compared to previously published work, and Table III shows our BLEU-N scores as compared to previously published results on Arabic Flickr8. From this table, it can be seen that the highest BLEU-4 score are achieved by the model trained with FARASA segmenter and beam size of 5. Furthermore the PyArabic model also achieved high BLEU- N scores, which are superior to 9 of related work results. As noted in the related work Section II, most of those studies did not report the data splits they used, others used different splits like [14] and [17], who used parts of both Flickr8 and MS COCO data sets, and [16], who used 90/10 for training and testing, while [17] used combined data set for training, but the testing was only done on part of Flickr8 data.

E. Qualitative Evaluation of our Results

As a complementary measure to the obtained BLEU–N scores, it was important to seek qualitative evaluations, to validate the BLEU-N scores results, and show that our models predict accurate and meaningful description for the images. Table IV shows a sample of the captions generated by our models, compared to others previously published captions for same images. As human evaluations are still considered the gold standards to evaluate the quality of machine generated captions, four native Arabic speakers were asked to rank ours, and others captions for each image. Based on the average ranking for the evaluators, they reported that our model's captions have better quality in 16 out the 29 (55%) captions listed in Table IV. In particular, those which are generated with models used PyArabic pre-processing. Other related work methods are better in 8 out of 29 (28%), and both have the same quality in 5 out of the 29 (17%) images.

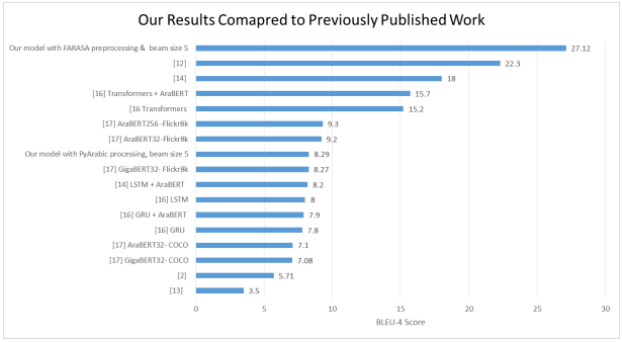


Fig. 4. Our Results Compared to those of Previously Published Methods.

TABLE III. OUR RESULTS AS COMPARED TO PREVIOUSLY PUBLISHED METHODS

Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4
[2]	33.18	19.26	10.49	5.71
[16] Transformers + ARABERT	44.30	N/A	N/A	15.70
[16] Transformers	42.70	N/A	N/A	15.20
[16] LSTM + ARABERT	38.30	N/A	N/A	8.20
[16] LSTM	35.10	N/A	N/A	8.0
[16] GRU + ARABERT	37.6	N/A	N/A	7.9
[16] GRU	35.3	N/A	N/A	7.8
[14]	52.00	46.00	34.00	18.0
[12]	65.8	55.9	40.4	22.30
[13]	34.40	15.40	7.60	3.50
[17] AraBERT32-Flickr8k	39.10	24.6	15.0	9.2
[17] AraBERT32- COCO	36.5	22.1	12.9	7.1
[17] AraBERT256 -Flickr8k	38.7	24.4	15.1	9.3
[17] GigaBERT32- Flickr8k	38.6	24.1	14.4	8.27
[17] GigaBERT32- COCO	36.0	21.5	12.4	7.08
Our model with PyArabic & beam size of 3	39.108	25.131	13.962	8.29
Our model with FARASA & beam size 5	58.708	46.523	35.712	**27.12**

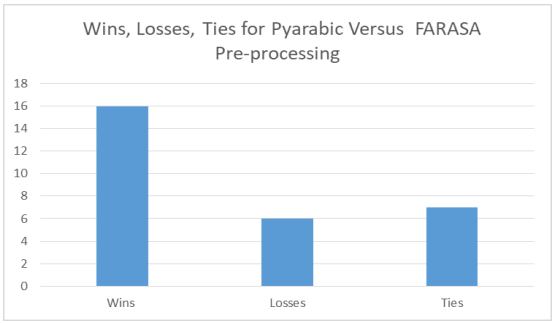


Fig. 5. Summary of THUMB Score Comparison between PyArabic, and FARASA based Models.

To further evaluate and quantify the differences in the captions' quality, THUMB 1.0 [24] was employed, which is utilized for Arabic image captioning for the first time in this paper. Therefore, three different native Arabic speakers were asked to use the THUMB framework, and evaluate the captions in Table IV. Based on the average of ratings by the three evaluators; Fig. 5 compares the THUMB scores of our two models. From this figure, it can be seen that PyArabic model has better quality captions, where it obtained 16 wins, 6 losses and 7 ties; compared to FARASA based model. Similarly, Fig. 6 summarizes the comparison results between our models and the related work, showing our models' wins, losses and ties. Again, and similar to the first four evaluators, our models obtained higher THUMB scores, where 15 and 17 wins; 5 and 7 losses; and 9 and 5 ties were obtained by our PyArabic and FARASA based models respectively. Fig. 7 shows detailed comparison between our model's Precision, Recall, and Penalties, compared to those of related work. From this table that our models obtained better scores, but had more penalties, in particular for FARASA based model. Even though

FARASA based model compares favorably to the related work models; however, it has more penalties, due to token repetitions, like example 2, location issues, like example 6, where the caption "little girl in people" should have been "little girl with people" , in Arabic "فتاة صغيرة في الناس" should have been "فتاة صغيرة مع الناس". Again, in spite of the fact that BLEU-N scores for FARASA based model are much higher than PyArabic based model, it tends to output short, more generic descriptions, compared to PyArabic based models, where the structure and/or the semantics of the captions are better for the latter. The improved results for FARASA based models could be partially attributed to the increase of the number of tokens, like the suffix "AL" "ال", possessive pronoun "na" (نا), "TAA Marboota" "ة", while number of unique stem words decreases. Another reason is that; due to the tendency of producing short, more generic captions, then the overlap of the N-grams with the ground truth captions in the test set increases, hence the higher BLEU-N scores. The important question here is: should FARASA based models be used as they have higher BLEU-N scores? The answer depends mainly on the nature of the image to be captioned. For example, the captions for images 1, 5, 7, 13 are of same quality for both models. However, if the image scene is busier, then PyArabic based models would be better, like the case of images 2, 3, 6, 18, etc. As Fig. 5 shows, PyArabic based model produces better, stronger sentences semantically; however, the difference of the models scores are minor in most of the cases, which favors the use of FARASA based models. It should be noted here that it is hard to conclude that PyArabic based models are qualitatively better than FARASA based models; as only a sample of 29 images were shown here, which are common between all published work, and obtaining the THUMB scores for more images could reverse the situation, which is likely, as the BLEU-4 score for FARASA based model is much higher. One more thing that could give us a hint, and may help in answering the question; is to be able to understand the internal logic of the model during the caption generation process. This can be achieved by visualizing the attention mechanism weights during caption generation, which is explored in Section V.

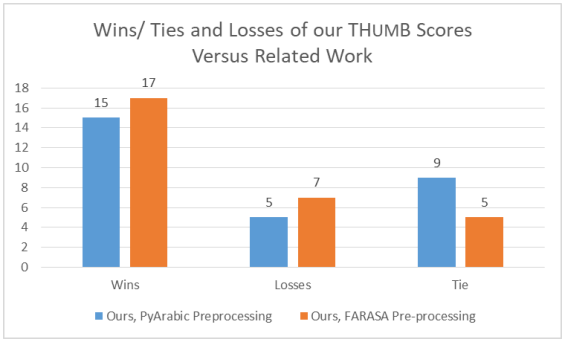


Fig. 6. Summary of THUMB Score for our Two Models, Compared to Related Work Models.

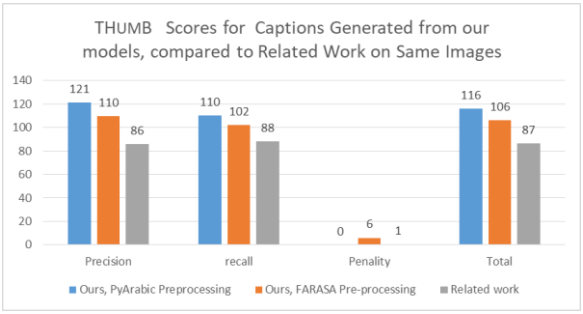
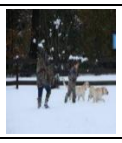


Fig. 7. Detailed THUMB Scores for Our Two Models, Compared to Related Work Models.

TABLE IV. OUR GENERATED CAPTIONS, COMPARED TO PREVIOUS STUDIES FOR THE SAME IMAGES

	Image	*PyArabic based model Captions*	*FARASA based model captions*	*Previous work captions*
1		مجموعة من الرجال يلعبون كرة القدم Group of men playing football	مجموعة من الناس يلعبون كرة القدم Group of people playing football	**[1]** لاعب كرة قدم يرتدى قميص احمر اللون فى الملعب A soccer player wears a red shirt on the field
2		شخص يركب دراجة نارية Person riding a motorcycle	راكب الدراجة الدراجة الدراجة Cyclist bike bike	**[1]** رجل يركب دراجة ترابية Man riding a dirt bike
3		مجموعة من الناس يقفون في الشارع A group of people standing in the street	مجموعة من الناس في الشارع Group of people in the street	**[1]** مجموعة من الناس فى الخارج فى مدينة مزدحمة A group of people outside in a crowded city
4		صبي صغير يرتدي سترة حمراء Little boy wearing a red jacket	رجل يرتدي قبعة صغيرة man wearing beanie	**[16]** صبي صغير يرتدى زى القراصنة يرفع علم القراصنة A little boy in a pirate costume raises a pirate flag

5		صبي صغير يقفز في الهواء Little boy jumping in the air	صبي صغير يقفز في الهواء little boy jumping in the air	**[16]** صبي صغير يقفز في الهواء little boy jumping in the air
6		مجموعة من الناس يلعبون في الهواء A group of people playing in the air	فتاة صغيرة في الناس little girl in people	**[16]** صبي صغير فى قميص ازرق و جينز ازرق Little boy in a blue shirt and blue jeans
7		كلب بني يركض عبر العشب Brown dog running through the grass	كلب بني يركض في العشب Brown dog running in the grass	[16] كلب بني يركض في حقل Brown dog running in a field
8		فتاة صغيرة في قميص أزرق Little girl in blue shirt	امرأة صغيرة في الهواء little woman in the air	**[16]** امرأة فى ثوب السباحة تمشى فى بركة A woman in a bathing suit walking in a pool
9		رجل في قميص أزرق يلعب كرة السلة A man in a blue shirt playing basketball	رجل يرتدي سترة السلة السلة A man wearing a basketball jacket the basket the basket	**[16]** صبي صغير يلعب كرة السلة فى ملعب رياضى Little boy playing basketball in the sports field
10		صبي صغير يقفز في الشارع Little boy jumping the street	امرأة صغيرة في الشارع little woman in the street	**[16]** رجل يعزف على الجيتار فى الشارع A man playing guitar in the street
11		كلب أسود يركض على العشب Black dog running on the grass	اثنين من الكلاب يلعبون في العشب Two dogs playing in the grass	**[2]** كلب بنى يقف فى الماء Brown dog standing in the water
12		كلب أسود يركض في الثلج Black dog running in the snow	اثنين من الكلاب يلعبون في الثلج Two dogs playing in the snow	**[14]** كلب أسود وأبيض يقفز على سجادة A black and white dog jumps on a carpet
13		رجل يركب الأمواج Man surfing	راكب الأمواج في الماء Surfer in the water	**[14].** رجل يمارس رياضة ركوب الأمواج Man surfing
14		رجل يرتدي سترة حمراء على لوح التزلج على الجليد A man wearing a red jacket on a snowboard	المتزلجان في الثلج Snow skaters	**[14]** رجل يرتدي خوذة حمراء قيف على تلة ثلجية A man wearing a red helmet stood on a snow hill
15		صبي صغير يقفز في الثلج Little boy jumping in the snow	المتزلج في الثلج snow skater	**[2]** صبى فى سترة حمراء يلعب فى الماء A boy in a red jacket playing in the water
16		كلب أسود يقفز في الهواء Black dog jumping in the air	كلب أسود يقفز في الهواء Black dog jumping in the air	**[2]** كلب اسود يقفز فى الهواء Black dog jumping in the air
17		فتاة صغيرة في الماء Little girl in the water	فتاة صغيرة في الماء على الشاطئ little girl in the water on the beach	**[2]** صبي فى ثوب سباحة يلعب فى الماء Boy in a bathing suit playing in the water

18		فتاة صغيرة في قميص أحمر في حقل Little girl in a red shirt in a field	فتاة صغيرة في العشب little girl in the grass	**[13]** فتاة صغيرة ترتدي فستان ملون تحمل كوب bear cup little girl wearing a colorful dress
19		كلب أبيض يركض في الثلج A white dog running on the snow	كلب أبيض يركض في الثلج A white dog running on the snow	**[2]** كلب أبيض يركض في الثلج White dog running in the snow
20		رجل في قميص أزرق Man in blue shirt	امرأة ترتدي سترة وامرأة Woman wearing a jacket and woman	**[1]** تحمل الزهور من باقة ا مرأة فى سترة حمراء Woman in red jacket carrying flowers from a bouquet
21		رجل يجلس على لوح التزلج على الشاطئ A man sitting on a surfboard on the beach	رجل يقفز في الماء Man jumping in water	**[16]** صبيان يستعدان للقفز من رصيف يقع على جسم كبير فى الماء Boys preparing to jump from a pier that falls on a large body in the water
22		رجل يقف على مقعد في الشارع Man standing on bench in the street	امرأ ة من الناس في الشارع Woman of people in the street	**[2]** مجموعة من الناس يحملون المشروبات و يشيرون الى الكاميرا A group of people carrying drinks and pointing at the camera
23		شخص يركب دراجة على الجليد person riding a bicycle on ice	شخصان في الثلج Two people in the snow	**[13]** رجل يرتدي خوذة زرقاء اللون يقفز فوق لافتة تقول Man wearing blue helmet color jumps over a sign that says
24		رجل يركب دراجة نارية Man riding a motorcycle	مجموعة من الدراجة النارية set of motorcycle	**[13]** رجل يرتدي خوذة حمراء يقود دراجة نارية في الهواء Man wearing red helmet driving a motorcycle in the air
25		رجل في قميص أحمر يقفز في الهواء A man in a red shirt jumps in the air	رجل يقفز على الهواء man jumping on air	**[15]** امرأة تحمل مضربا فوق ملعب تنس Woman holding a racket on a tennis court
26		كلب بني يركض على العشب Brown dog running on the grass	الكلب البني والبني Brown dog brown dog	**[13]** كلب بني و اسود اللون يقفز فوق سياج ابيض و ابيض Brown and black dog jumping Over a white and white fence
27		اثنين من الكلاب يلعبون في الثلج Two dogs playing in the snow	ثلاثة من الكلاب في الثلج Three dogs in the snow	**[13]** كلب اسود و كلب بني اللون في حقل عشبى Black dog and brown dog in a grassy field
28		كلب أسود يركض عبر العشب Black dog running across the grass	اثنين من الكلاب يلعبون في العشب Two dogs playing in the grass	[13] كلب اسود و كلب بني اللون في حقل عشبي. Black dog and brown dog in grass field
29		اثنين من الناس في الهواء Two people in the air	مجموعة من الناس في الشارع group of people on the street	[16] امرأة فى سترة برتقالية تتحدث على هاتفها المحمول A woman in an orange jacket talking on her mobile phone

V. Attention Mechanism and Pre-processing Methods

Motivated by the results obtained in Section IV, where the use of PyArabic and FARASA segmenter lead to different BLEU-N; as well as THUMB scores, we thought it would be interesting to visualize the sequence of the tokens generated with images' attended parts by each model. Therefore, attention scores visualization is used to investigating the effect of different text pre-processing methods on the generated captions. Put it another way; it is required to investigate whether the attention mechanism attends the image salient features the same way during word sequence generation; for different pre-processing methods.

Fig. 8 to 15 show the attention visualization for four different images and their captions, while the alignment of tokens is shown with the images' attended parts during sequence generation. Interestingly, it is noticed from those figures, that the attention mechanism behavior is somehow different in FARAS based model compared to PyArabic based model. Take for example Fig. 8 and 9, which visualize the attention by FARASA and PyArabic models respectively. Even though the length of the sequence is shorter in Fig. 8 (FARASA based model), it can be noticed that the model perfectly attends the token "فتاة","Girl" , which spans images 2,3,4, then "in" "فى"in 5 , then " the water" "ماء+ال"6,7,8, which made the attended image features very prominent. This is not the case in Fig. 9, where an additional word "water""ماء" is not clearly visualized like Fig. 8. Looking at Fig. 10 and 11, it can be seen that the attention mechanism better attends the image's salient features in case of FARASA based model, where the attention shows the dogs and the snow clearly, which is not the case in Fig. 11 for PyArabic. One explanation is that FARASA pre-processing outputs larger number of tokens for the same image region, therefore, higher attention scores.

In addition to the above, attention visualization is also useful in detecting a model's failure, even if it is not very clear in the caption. Consider for example Fig. 12 and 13, where the generated caption is " child in the air", while the attention is attending the cat face, which is the same for PyArabic based model, "Man is sitting on a chair", still the attention is on the cat's face. This is a clear failure of both models, which may not be noticed in case of PyArabic model, as there is a man sitting there in the image. From those examples, it can be concluded that FARASA based model better attends the image's salient features during caption generation, compared to PyArabic model. A third, and more interesting example is shown in Fig. 14 and 15, where an images composed of two individual images, for two different scenes are stacked together. For Fig. 14, the generated caption is "three people in the water", and the attention shows that the dog is counted as one of the three persons in the image, and the water is clearly attended as well. This is not the case for Fig. 15, with the caption "Man Surfing", where one person is attended, while the surfing token is aligned with the dog, which is a clear failure of the model.

From those examples, it can be concluded that FARASA segmenter improves the attention mechanism behavior, which explains the high BLEU-N scores of its model. However, more examples needs to be looked at, to confirm such conclusion.

Fig. 8. Attention Weights Visualization for FARASA based Model: The Caption is "Girl in the Water".

Fig. 9. Attention Weights Visualization for PyArabic based Model. The Caption is "Little Girl in the Water."

Fig. 10. Attention Visualization for FARASA based Model. The Caption is: "Three Dogs in the Snow".

Fig. 11. Attention Visualization for PyArabic based Model. The Caption is: "Two Dogs Playing in the Snow".

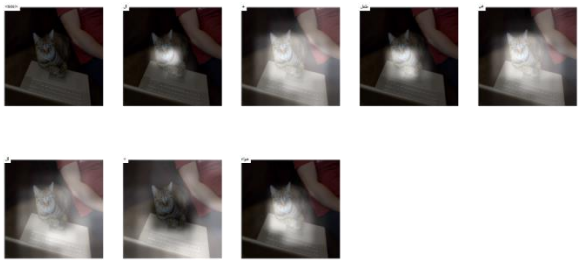


Fig. 12. Attention Visualization for FARASA based Model. The Caption is: "Child in the Air".

Fig. 13. Attention Visualization for PyArabic based Model. The Caption is: "Man Sitting on Chair".

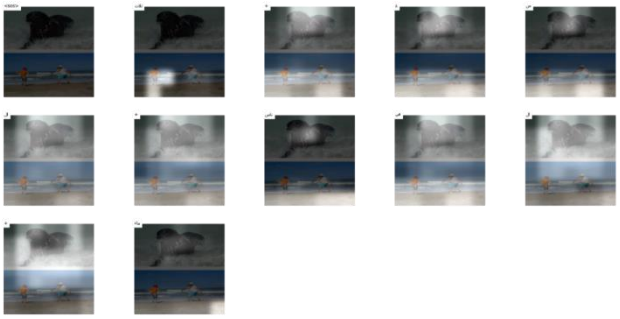


Fig. 14. Attention Visualization with FARASA based Model. The Caption is: "Three People in Water".

Fig. 15. Attention Visualization for PyArabic based Model. The Caption is: "Man Surfing".

VI. Conclusion and Future Work

In this paper, an efficient model for Arabic image captioning is proposed, and the effect using beam search, as well as the pre-processing on the testing BLEU-4 score is investigated. The model with FARASA segmenter achieved the state of the art BLEU-4 score. The results are also consistent with the results of the model which used root words to generate Arabic captions. In addition to the BLEU-N scores, the generated captions were also qualitatively evaluated by two teams of Arabic native speakers, the first team the captions, while the other used the "THUMB" framework for evaluation. The generated captions using two different text pre-processing models achieved the best THUMB scores, where the model with PyArabic pre-processing showed better results on the sample used. Another interesting finding in this paper is that the different text pre-processing methods influence the attention mechanism, where FARASA based model showed better attention visualization for the used samples. The generated captions also compares favorably to all previous related work, quantitatively, and qualitatively. The paper also show that the choice of the right architecture, with the right pre-processing of Arabic text and the use of beam search can significantly improve the quality of the generated captions.

From the work done in the area of Arabic image captioning including this study, it can be concluded that the use of transformers did neither significantly improve the BLEU-N results, nor the use of larger data sets in training.

As a direction of future research, more efficient models can be investigated to improve the obtained results in this area, utilizing Generative Adversarial Networks. Another direction is to propose new text pre-processing methods and additional evaluation methods to cope with morphologically rich languages like the Arabic language.

References

[1] M. Cheikh and M. Zrigui, "Active Learning Based Framework for Image Captioning Corpus Creation," in International Conference on Learning and Intelligent Optimization, Cham, 2020: Springer International Publishing, in Learning and Intelligent Optimization, pp. 128-142.

[2] O. ElJundi, M. Dhaybi, K. Mokadam, H. Hajj, and D. Asmar, "Resources and End-to-End Neural Network Models for Arabic Image Captioning," in 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) 2020, vol. 5, pp. 233-241.

[3] M. Stefanini, M. Cornia, L. Baraldi, Silvia Cascianelli, G. Fiameni, and R. Cucchiara, "From Show to Tell: A Survey on Deep Learning-based Image Captioning," arXiv:2107.06912v3, 2021.

[4] H. Wang, Yue Zhang, and X. Yu, "An Overview of Image Caption Generation Methods," Computational Intelligence and Neuroscience, vol. 2020, p. 13, 2020.

[5] A. Pal, S. Kar, A. Taneja, and V. K. Jadoun, "Image Captioning and Comparison of Different Encoders," Journal of Physics: Conference Series, vol. 1478, no. 012004, 2020.

[6] R. Staniute and D. Šešok, "A Systematic Literature Review on Image Captioning," Applied Sciences, vol. 9, no. 2024, p. 20, 2019.

[7] Z. Zohourianshahzadi and J. K. Kalita, "Neural Attention for Image Captioning: Review of Outstanding Methods," Artificial Intelligence Review, 2021.

[8] A. Elhagry and K. Kadaoui, "A Thorough Review on Recent Deep Learning Methodologies for Image Captioning," arxiv.2107.13114, 2021.

[9] S. Li, Z. Tao, K. Li, and Y. Fu, "Visual to Text: Survey of Image and Video Captioning," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 3, no. 4, pp. 297-312, 2019.

[10] T. Ghandi, H. Pourreza, and H. Mahyar, "DEEP LEARNING APPROACHES ON IMAGE CAPTIONING: A REVIEW " arXiv:2201.12944, 2022.

[11] S. AMIRIAN, K. RASHEED, T. R. TAHA, and H. R. ARABNIA, "Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap," IEEE Access, vol. 8, pp. 218386- 218400, 2020.

[12] V. Jindal, "Generating Image Captions in Arabic Using Root-Word Based Recurrent Neural Networks and Deep Neural Networks," in The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, Louisiana, USA., 2018: Association for the Advancement of Artificial Intelligence, pp. 8093-8094.

[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] 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.

[15] I. Afyouni, I. Azhar, and A. Elnagara, "AraCap: A hybrid deep learning architecture for Arabic Image Captioning," Procedia Computer Science, vol. 189, pp. 382-389, 2021.

[16] S. M. Sabri, "ARABIC IMAGE CAPTIONING USING DEEP LEARNING WITH ATTENTION," Masters, University of GeorgiaProQuest Dissertations, ATHENS, GEORGIA, 2021.

[17] J. Emami, "Arabic Image Captioning using Pre-training of Deep Bidirectional Transformers," Masters, Computer Science, LUND UNIVERSITY, 2022.

[18] 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, pp. 3128-3137.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in IEEE conference on computer vision and pattern recognition, 2016: IEEE, pp. 770-778.

[20] Pyarabic, An Arabic language library for Python. (2010). [Online]. Available: https://pypi.python.org/pypi/pyarabic/.

[21] K. Darwish and H. Mubarak, "Farasa: A New Fast and Accurate ArabicWord Segmenter," in LREC 2016, Tenth International Conference on Language Resources and Evaluation, Slovenia, 2016, pp. 1070-1074.

[22] A. B. Soliman, K. Eissa, and S. R.El-Beltagy, "AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP," Procedia Computer Science, vol. 117, pp. 256-265, 2017.

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

[24] J. Kasai et al., "Transparent Human Evaluation for Image Captioning," arXiv:2111.08940v2, vol. cs.CL, May 2022.