Automatic Arabic Image Captioning using RNN-LSTM-Based Language Model and CNN

Huda A. Al-muzaini, Tasniem N. Al-yahya, Hafida Benhidour
Dept. of Computer Science
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

***Abstract*—The automatic generation of correct syntaxial and semantical image captions is an essential problem in Artificial Intelligence. The existence of large image caption copra such as Flickr and MS COCO have contributed to the advance of image captioning in English. However, it is still behind for Arabic given the scarcity of image caption corpus for the Arabic language. In this work, an Arabic version that is a part of the Flickr and MS COCO caption dataset is built. Moreover, a generative merge model for Arabic image captioning based on a deep RNN-LSTM and CNN model is developed. The results of the experiments are promising and suggest that the merge model can achieve excellent results for Arabic image captioning if a larger corpus is used.**

Keywords*—AI; image caption; natural language processing; neural network; deep learning convolutional neural network; recurrent neural network; long short-term memory

I. Introduction

Automatic generation of captions for images by describing the content of an image using natural language sentences has become a fundamental task in Artificial Intelligence and has recently attracted the attention of the research community [1]. Given the huge number of images that are available online, image captioning has become nowadays central to image retrieval tasks such as the one carried by search engines or newspaper companies. More specific applications, like describing images for blind persons or teaching children concepts, can also be given as examples on the importance of captioning images.

Image captioning has been identified as a cross-modal task which grounds and relates the visual and the natural language model. Despite the challenging nature of this task, several image caption generation models, one can cite [2]–[6] as examples, have achieved promising results due to the advances in training neural networks [7] and the large image datasets that are now available [8].

The sparsity of annotated resources other than English is an issue in morphological complex language such as Arabic. Thus, there is a need for corpora sufficiently large for image captioning in other languages.

The aim of this work is to take a step towards the goal of developing an image caption generation model for describing images in Arabic language (see Fig. 1). The model is inspired by the merge model proposed in [10] and [11]. It consists of two sub-networks: a deep recurrent neural network (RNN) for sentences and a deep convolutional neural network (CNN) for images. These two sub-networks interact with each other in a merge layer to predicate and generate the caption. Moreover, the first public Arabic image caption corpus is presented. This Arabic version is a subset of the Flickr [11] and MS COCO [12] caption data sets. The remainder of the paper is organized as follows. Influential work as well as state-of-the-art models for image caption generation for English as well as other languages are presented in Section II. Detailed description of the image caption generation model for Arabic language is given in Section III. A description of the image dataset with Arabic captions is presented in Section IV. The process of building the Arabic image caption corpus through crowdsourcing is presented in Section V. The experiment evaluation and results are described in Section VI. Finally, the conclusion with some directions for future work is given in Section VII.

II. Related Work

This section covers recent advances in the development of image caption generation models for different languages including: English, Arabic, Chinese, Japanese, and German.

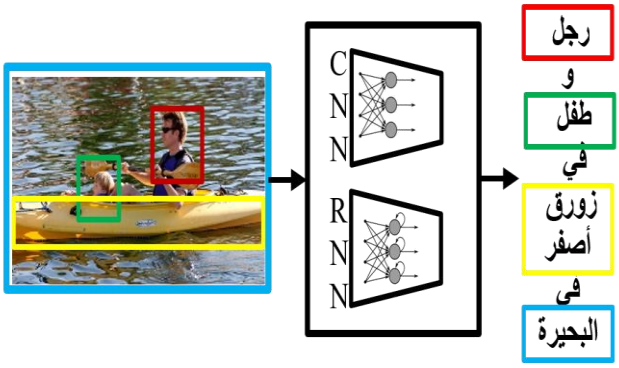

Fig. 1. Example of an Arabic caption generated for an image (caption translation in English: a man and a child in a yellow canoe in the lake).

A. Image Captioning for English Language

The different approaches for image caption generation can be either based on retrievable or constructive approaches as pointed out in [9], [10], [13], [14]. This taxonomy is clearly depicted in Fig. 2. An image caption generator based on a retrievable approach models the problem as a retrieval task. A

database based on image features and captions is constructed. Given an image, the most suitable annotation is retrieved. This approach however lacks the ability of generating novel sentences, does not scale to describe raw images, and the caption generation is limited to the features and the size of the database. Thus, this approach is not suitable for today's demand. Example of work based on this approach includes [1], [15]–[17].

Recent progress in automatic image captioning is based on a constructive approach. A constructive approach gradually constructs a novel caption for each image. This can be further divided into computer vision and natural language generation methods (CV/NLG) or CNN/RNN methods. For the first category, image attributes are extracted from images using computer vision techniques which are fed to natural language generation models to generate a syntactically correct caption. This approach is the base of the work in [18]. The CNN/RNN approaches have proven to be the most successful ones. They model the caption generation process in two phases, the first phase is image features learning phase and the second is the sentence generation phase. Depending on whether the image is injected to the language model or left out and then later merged with the output of the language model using a feedforward layer, one can distinguish two models, the inject and the merge model. A complete empirical study of these two models can be found in [10] and [11]. In the inject class (see Fig. 3), the language model, such as the RNN, is the primary generation component where an image is directly injected to the model during training time. The output of the RNN is a mixed vector that is handled in a subsequent feedforward layer to predict the next word in the caption. Works under this class includes [2] and [19]. In [2], the Neural Image Caption (NIC) model is presented. This model is based on an end-to-end neural network that works by first pre-training it for an image classification task using a CNN and then using the last hidden layer as an input to the RNN that generates sentences. Experiments on several datasets including Flickr [11], MS COCO [12], Pascal VOC [17], and SBU [16] using different metrics: BLEU-{1,2,3,4} [20], CIDER [21], and METEOR [22] reported an accuracy comparabale to state-of-the-art approaches; for instance, on the Pascal dataset, NIC yielded a BLEU score of 59, to be compared to the current state-of-the-art of 25, while human performance reaches 69. On Flickr30k, an improvement was achived from 56 to 66, and on SBU, from 19 to 28. In [19], the process starts by decomposing the input image by detecting objects and other regions of interest to produce a vector representation richly expressing the image semantics. This feature vector is taken as input by a hierarchical RNN. The hierarchical RNN is composed of two levels: a sentence RNN and a word RNN. The sentence RNN receives image features, decides how many sentences to generate in the resulting paragraph, and produces an input topic vector for each sentence. Given this topic vector, the word RNN generates the words of a single sentence. The model was experimented on a novel dataset of paragraph annotations, comprising of 19,551 MS COCO [12] and Visual Genome [23] images, and evaluated across six language metrics: BLEU-{1,2,3,4} [20], CIDER [21], and METEOR [22]. The scores show the superior advantages of this method over traditional image captioning methods and was close to human performance.

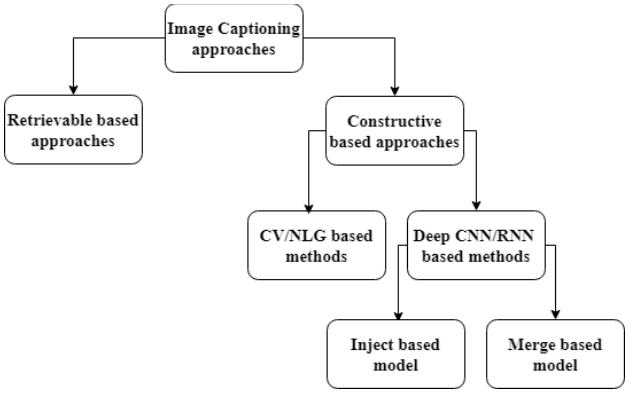

Fig. 2. Taxonomy for English image captioning approaches.

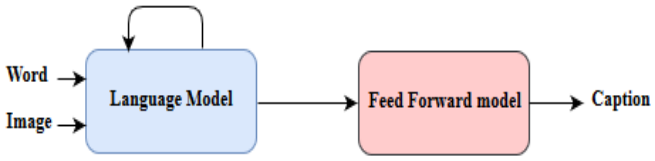

Fig. 3. Inject model.

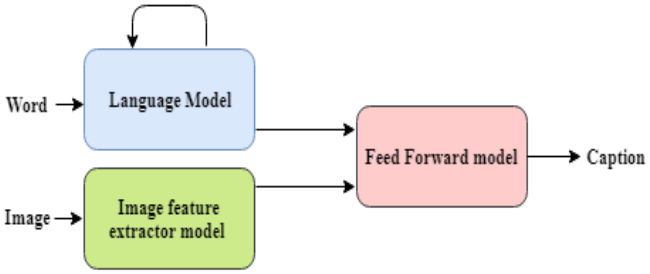

Fig. 4. Merge model.

The second class is the merge model in which the image features and linguistic models are learned independently and then merged in a feed forward model in which the prediction takes place (see Fig. 4). The work of [24] was the first to prpose a merge model for image captioning and shortly after their work was published, several papers appeared with promising results including [4], [25], [26]. This demonstrates the effectiveness of this model. In [24], Mao et al. proposed the merge model then refined it in [4] and [26]. Their image representation is learned independently by a CNN model then inputted to the RNN-LSTM model along with every word in the sentence description. The approach uses the capacity of the RNN-LSTM more efficiently. The RNN-LSTM model incorporates a two-layer word embedding system which learns the word representation more efficiently than the single-layer word embedding. These two models interact with each other in a multimodal layer. The effectiveness of their model was validated on four benchmark datasets IAPR TC-12, Flickr 8K, Flickr30K, and MS COCO. Experimental results based on BLEU-{1,2,3,4} [20], CIDER [21], METEOR [22], and ROUGE [27] showed the outstanding performance of their

model for almost all evaluation metrics. In [25], the Deep Compositional Captioner (DCC) is introduced. DCC builds on recent deep captioning models which combines a CNN and RNN networks for visual and language generation, respectively. Then, both models are combined into a deep caption model which is trained jointly on paired image-sentence data. However, unlike previous models which can only describe objects that are present in paired image-sentence data, DCC is able to generate sentences that describe objects presented in unpaired image/data but not present in paired image/sentence data. To accomplish this task, the training is preformed into three stages: 1) CNN and RNN are trained with unpaired data, then 2) both models are combined into a caption model which is trained on paired image-sentence data, and finally, 3) the knowledge is transferred from words that appear in paired image-sentence data to words that do not appear in paired image-sentence data. DCC performance was empirically evaluated by studying results on a training split of the MS COCO [12] dataset by deliberately excluding certain objects. Moreover, DCC performance to describe objects in the ImageNet7k dataset which are not present in the caption datasets was assisted. DCC scored 69.36 and 23.98 on the BLEU and METEOR metrics respectively. In addition, the F1-score was reported, which indicates that DCC can integrate new vocabulary in captions.

The literature on English caption generation although new, is rich of models that have proven their efficiency. However, few explicit comparison between the performance of the inject and merge architectures has been investigated. In [26], the authors compared the inject and merge architectures based only on the BLEU metric and concluded that merge is superior. The first work that studies extensively and systematically the difference between the inject and merge architecture is presented in [10] and [11]. Experimental evaluation concluded the following: 1) inject architectures tends to be slightly better on standard corpus-based metrics such as CIDER [21], 2) merge architectures produce sentences that are rich in vocabulary; that is inject models tends to re-generate captions wholesale from the training data, 3) inject models tend towards more generic and less image specific captions, especially for longer captions; a problem that merge models is not susceptible of, and 4) from an engineering perspective, merge architectures make better use of their RNN memory and avoids overfitting.

B. Image Caption for Arabic language

Automatic image captioning in Arabic was addressed only by the work of [28] by using root-word based RNN and Deep Belief Network (DBN). The approach adopted can be summarized in three stages. In the first stage, a Region CNN (RCNN) [29] is used to map image objects to Arabic root words by the aid of a transducer based algorithm for Arabic root extraction [30]. After that, stage two uses a word based RNN with LSTM memory cell to generate the most appropriate words for an image in Modern Standard Arabic (MSA). Finally, the caption sentences are generated by using dependency tree relations; specifically the Prague Arabic Dependency Treebank (PADT) [31]. For evaluation, two datasets were created. The first consists of annotating 10,000 images from the ImageNet dataset with Arabic captions and the second 100,000 images from Al-Jazeera news website. Experiments show a promising result considering BLEU-1 score with value 34.8 for Arabic caption generation.

C. Image Caption for Other Languages

The limitation of image description corpora in languages other than English is an issue, particularly for morphologically rich languages such Arabic and Japanese. In [32] a Japanese version of MS COCO caption dataset has been created using Yahoo! Crowdsourcing. The authors developed a model for image caption generation for Japanese language using deep learning. They pre-trained the model with the English portion of the corpus to improve the performance then trained it using Japanese captions. The resulting bilingual model has better performance comparing to the monolingual model that uses only the Japanese caption corpus. Cross-lingual image captioning for Chinese language has been developed by applying machine translation [33]. The experiment has been done on Flickr8k-cn and Flickr30-cn datasets. To improve the translated English-Chinese sentences, a fluency-guided learning framework has been proposed using LSTM neural network. The proposed approach improves both the fluency and the relevance without using any manually written caption in Chinese. In [34], an RNN model for generating Chinese captions has been presented. The authors developed two methods, one that takes the list of words from a Chinese sentence as input, and the second takes the list of characters and feed them to the same RNN model. The Chinese caption is obtained by translating Flickr30 dataset from English to Chinese using Google Translation API. They observed that the character level method outperform the word level in this task.

Multi30K, a Germen version of Flickr30K dataset, has been presented in [35]. Each image has a German translation of the English description obtained from Flickr30K dataset and five independent German captions obtained using Crowd flower platform. The translated sentences were collected by professional English-German translators without seeing the image.

III. Methodology

The Arabic image captioning model proposed in this work follows the merge architecture that was previously described in [10], [11]. This architecture is a simplified version of the architecture in [2]. It was chosen for its simplicity whilst still being the best performing system in the 2015 MS COCO [12] image captioning challenge.

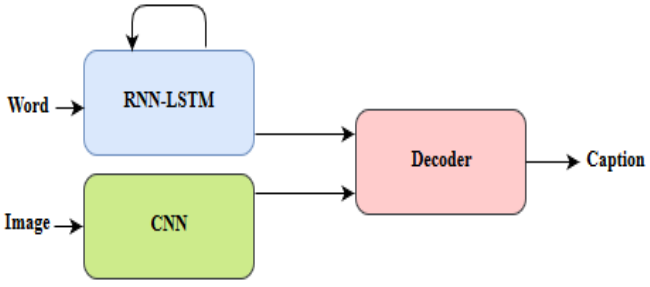

Fig. 5. The proposed model.

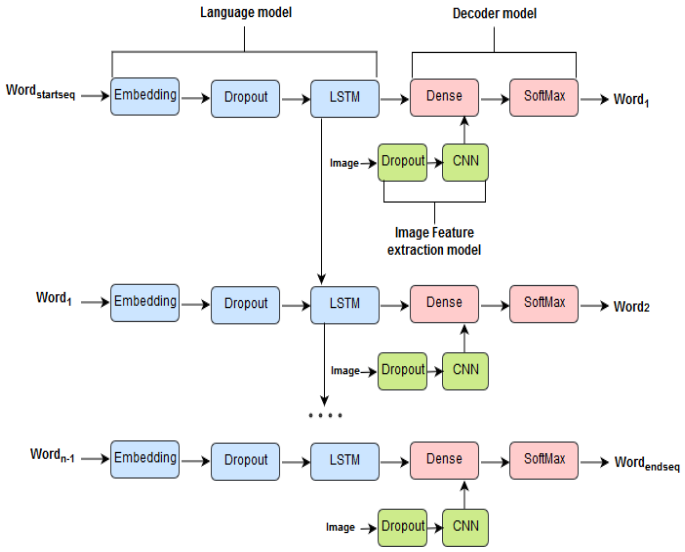

Fig. 6. Image captioning system for Arabic based on merge model.

The model is composed of three parts as shown in Fig. 5: 1) a language model based on RNN-LSTM [36] to encode linguistic sequences of varying length, 2) an image feature extractor model based on CNN [7] to extract image features in the form of a fixed-length vector, and 3) a decoder model that takes as input the outputted fixed vectors from the previous models and makes a final prediction.

A detailed illustration of the three parts of the proposed system is shown in Fig. 6. First, the language model inputs sequences with a pre-defined max length -the maximum words in the longest caption- which are fed then into an embedding layer that uses a mask to ignore padded values. Further, a 50% dropout is performed in a form of a regularization then the output is forwarded to the LSTM layer with 256 memory units. Independently, the second stage is the image feature extractor model that expects an input photo features to be a vector of 4,096 elements. A 50% dropout is also done before the image being processed by a CNN layer to produce a 256-element representation of the image. The final stage is the Decoder model that merges the 256-output of both models to an output Softmax layer that makes the final prediction over the entire output vocabulary for the next word in the caption.

IV. Dataset

The model has been trained and tested with images, from MS COCO dataset [37] and Flickr8K datasets [38]. The MS COCO images contain multiple objects in the scene collected by searching for pairs of 80 object categories. This dataset contains 2.5 million captions labelling over 330,000 images. To gather Arabic captions for the images, Crowd-Flower Crowdsourcing service [39] was used. Given 1166 images taken from the training set of the MS COCO dataset, a total of 5358 captions were collected. The images have on average 4.6 captions; the maximum number was 6 and the minimum was 4. Flickr8K dataset contains 8000 images and each comes with 5 English sentences. The images were selected with different locations and scenes from 6 Flickr groups. The first 2261 images from the training set were selected. A professional English-Arabic translator translated the captions of 150 images from Flicker, a total of 750 Arabic captions. The rest of the images (2111) were translated to Arabic using Google translator and then checked by Arabic native speakers. The total of images from both datasets (COCO and Flicker) is 3427, with a vocabulary size of 9854 and the longest caption consisting of 27 words. Since the dataset consists of some images from MS COCO and some from Flicker training sets, all images were divided for the experiments to 2400 for training, 411 for the development, and 616 for testing with a percentage of 70:12:18 respectively (see Table I).

V. Crowdsourcing Procedure

All captions used to build the dataset were human generated using Crowd-Flower Crowdsourcing [39]. A job was posted that asked the contributors to describe an image. In the job page, a user interface was provided with instructions in Arabic and one example. Each task includes only 5 images in each page to prevent contributor's exhaustion. Some of the instructions were translated directly from English instructions that were used in the MS COCO captions [37] and instructions specific to Arabic language were added. The job has the following instructions:

1) Please adhere to the standard Arabic language.

2) Write a useful sentence that ends with a period (.). Do not just type multiple words or phrases.

3) The sentence must contain at least 20 Arabic letters.

4) Use a polite style of speech and correct punctuation marks.

5) Please comment on the image by giving only factual data:

a) Do not write about things that may happen in the future.

b) Do not write about sounds, such as, the child heard the sound of the horn.

c) Do not speculate or imagine. Do not write about something that makes you feel uncertain.

d) Do not write about your feelings regarding the scene in the picture.

e) Do not use excessive poetic style.

1) Do not use demonstrative pronouns such as 'this' or the adverb of place 'here'.

2) Please do not write the names of the persons, places or nationalities; e.g. Washington City, American Flag.

3) Please describe all important parts of the scene; do not describe unimportant details.

TABLE I. Train/Dev/Test Split

Train	Train	2400
	Dev	411
Evaluate	Test	616
Total		3427

The job only appears for Arabic contributors to be sure that non-Arabic workers do not participate in this job. Six captions per image were collected. To guarantee the quality of the captions and that they are well written in Arabic and not using an Arabic dialect, a data-cleaning task was assigned to a professional Arabic language specialist. For some images, he selected the best 4 captions, for others he kept all 6 captions with small modifications.

VI. Experimental Evaluation

In this section the relative importance of different components of the proposed model is assessed, the implementation environment is defined, and finally the obtained results are presented and analyzed.

For the image encoder, a fully convolutional network based on Visual Geometry Group (VGG) OxfordNet 16-layer CNN [40] is adopted. Prior to training, all images were vectorized using the activation values which is trained to perform object recognition on a 4096-element vector and returns a 256 vector. For the language model, a single hidden RNN-LSTM layer with 256 memory units is defined. This layer is supported in the Keras [41] API library. The network uses a dropout of 50% on both models.

The complete model was implemented in python using latest version 2.1.6 of Keras [41]. Eexperiments were conducted on the commercial cloud server FloydHub [42]. FloydHub servr uses Nvidia Tesla K80 GPUs (12GB vRAM) and 61GB RAM and supports Keras [41] API.

In the experiment, the maximum length of a description within the data set is 27 words. This value is essential because it defines the input length to RNN-LSTM model. Given the amount of training data, the model was fit for 10 epochs, and the model stabilized after the 6th epoch. At the end of the 6th model, the loss computed was 4.278 on the training dataset and a loss of 4.859 was on the development dataset. The model generates correct descriptions of images (see Fig. 7), the syntax and the semantic of the sentences is accurate. For the middle image in the second row, the obtained caption “Dish has a dish inside”, even though correct, it fails to describe the important part of the image which is the food.

Following previous works, the model was evaluated on the BLEU-{1,2,3,4} [20], which evaluates a candidate sentence by measuring the fraction of n-grams that appear in a set of references. The model scored a value of 46 which is considered excellent in the BLUE scale. All BLEU scores obtained by the proposed model are given in Table II. Moreover, Table II gives a comparison of the proposed model with the BLEU scores of the Arabic captions obtained by translating the English captions derived from the NIC model [2] using Google Translate [28]. This translated model is evaluated on the Flickr8K dataset. As seen from Table II, the Arabic caption based merge model is comparable on the BLEU-{1} score. Also comparing the proposed model with [28], the proposed model results a 10% higher BLEU-{1} score. The obtained results are promising and can be improved with the availability of more data.

TABLE II. BLEU-{1,2,3,4} METRICS FOR THE ARABIC MODEL & NIC [2]/GOOGLE TRANSLATION

Test Dataset	Model	BLEU-			
		{1}	{2}	{3}	{4}
Flickr616	**Arabic caption based on merge model**	46	26	19	8
Flickr8K	**NIC [2]/Google translate**	52	46	34	18

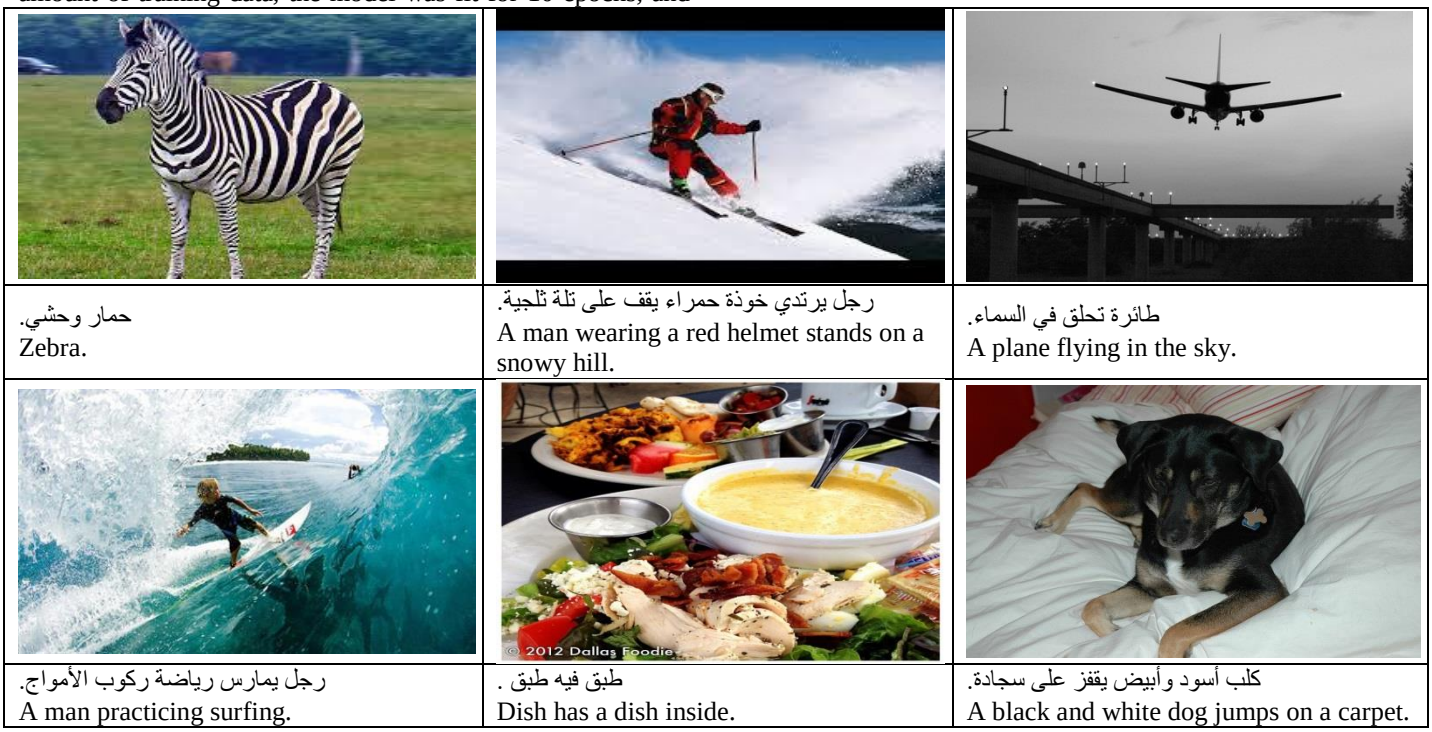

Fig. 7. Examples of image captions generated using the proposed model.

VII. Conclusion

A novel corpus of image captions in Arabic is built by collecting 5358 captions for 1176 images using a Crowd-Flower Crowdsourcing service, and 750 captions for 150 images were obtained from a human translator. The rest of image captions were translated from English to Arabic using Google translator. The RNN model trained by Arabic captions works well for image caption generation even with the small dataset that has been used for training and validating the model. Till now, no other RNN models were proposed for image caption generation for Arabic language except the paper of Jindal [28] that used a different methodology based on Deep Belief Network. The performance of the proposed model on the test set gave a promising result of 46.2 for the BLEU-1 score, which is 10% higher than the Jindal result.

The proposed model can give better performance with larger dataset. Therefore, for future research the image dataset with Arabic captions will be expanded and made publicly available. Further experiments will be conducted with the expanded corpus.

Acknowledgment

The authors would like to express their deep gratitude and sincere appreciation to their families that collaborated and were one hand to build the Arabic image captioning pilot corpus. Especial thanks to Mrs. Hend Al-yahya, -an English linguistic researcher- for her ongoing support and skills in auditing part of this dataset.

References

[1] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, "Automatic image captioning," in IEEE International Conference on Multimedia and Expo (ICME), 2004, vol. 3, no. 22, pp. 0–3.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, vol. 07–12–June, pp. 3156–3164.

[3] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, "Learning like a child: Fast novel visual concept learning from sentence descriptions of images," in IEEE International Conference on Computer Vision (ICCV), 2015, vol. 2015 Inter, pp. 2533–2541.

[5] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 2625–2634, 2017.

[6] A. Karpathy and F. F. Li, "Deep visual-semantic alignments for generating image descriptions," IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 3128–3137, 2015.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in 25th International Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.

[8] O. Russakovsky et al., "Imagenet large scale visual recognition challenge," arXiv Prepr. arXiv1409.0575, 2014.

[9] M. Tanti, A. Gatt, and K. P. Camilleri, "What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?," Aug. 2017.

[10] M. Tanti, A. Gatt, and K. P. Camilleri, "Where to put the Image in an Image Caption Generator," Mar. 2017.

[11] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting Image Annotations Using Amazon's Mechanical Turk," in NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010, no. June, pp. 139–147.

[12] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," 2014, pp. 740–755.

[13] J. Devlin et al., "Language Models for Image Captioning: The Quirks and What Works," arXiv Prepr. arXiv1505.01809, 2015.

[14] S. Pratap, S. Gurjar, S. Gupta, and R. Srivastava, "Automatic image annotation model using LSTM approach," Signal Image Process., vol. 8, no. 4, 2017.

[15] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick, "Exploring Nearest Neighbor Approaches for Image Captioning," arXiv:1505.04467, 2015.

[16] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: Describing Images Using 1 Million Captioned Photographs," NIPS, pp. 1143–1151, 2011.

[17] A. Farhadi et al., "Every Picture Tells a Story: Generating Sentences from Images," in European Conference on Computer Vision (ECCV), 2010, pp. 15–29.

[18] D. Elliott and F. Keller, "Image Description using Visual Dependency Representations," Emnlp, no. October, pp. 1292–1302, 2013.

[19] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A Hierarchical Approach for Generating Descriptive Image Paragraphs," in Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3337–3345.

[20] K. Papineni, S. Roukos, T. Ward, and W. Z. Ibm, "BLEU: a Method for Automatic Evaluation of Machine Translation," in 40th Annual Meeting on Association for Computational Linguistics (ACL), 2002, no. July, pp. 311–318.

[21] V. Ramakrishna, L. Zitnick, and P. Devi, "Cider: Consensus-based image description evaluation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[22] M. Denkowski and A. Lavie, "Meteor Universal: Language Specific Translation Evaluation for Any Target Language," in 9th Workshop on Statistical Machine Translation, 2014, pp. 376–380.

[23] R. Krishna et al., "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," Int. J. Comput. Vis., vol. 123, no. 1, pp. 32–73, May 2017.

[24] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain Images with Multimodal Recurrent Neural Networks," arXiv:1410.1090, Oct. 2014.

[25] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1–10.

[26] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," in nternational Conference on Learning Representations (ICLR), 2015.

[27] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Workshop on text summarization branches out (WAS 2004), 2004, no. 1, pp. 25–26.

[28] V. Jindal, "A Deep Learning Approach for Arabic Caption Generation Using Roots-Words," AAAI, pp. 4941–4942, 2017.

[29] Y. Jia et al., "Caffe: Convolutional Architecture for Fast Feature Embedding," arXiv:1408.5093, 2014.

[30] Q. Yaseen and I. Hmeidi, "Extracting the roots of Arabic words without removing affixes," J. Inf. Sci., vol. 40, no. 3, pp. 376–385, Jun. 2014.

[31] J. Hajič, O. Smrz, P. Zemánek, J. Šnaidauf, and E. Beška, "Prague Arabic dependency treebank: Development in data and tools," in International Conference on Arabic Language Resources and Tools (NEMLAR), 2004, pp. 110–117.

[32] T. Miyazaki and N. Shimizu, "Cross-Lingual Image Caption Generation," in The Association for Computational Linguistics (ACL), 2016, pp. 1780–1790.

[33] W. Lan, X. Li, and J. Dong, "Fluency-Guided Cross-Lingual Image Captioning," in ACM on Multimedia Conference (ACMM), 2017, pp. 1549–1557.

[34] H. Peng and N. Li, "Generating Chinese Captions for Flickr30K Images," 2016.

[35] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30K: Multilingual English-German Image Descriptions," in 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 70–74.

[36] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[37] X. Chen et al., "Microsoft COCO Captions: Data Collection and Evaluation Server," arXiv:1405.0312, 2015.

[38] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," in 24th International Joint Conference on Artificial Intelligence (IJCAI), 2015.

[39] CrowdFlower, "Machine Learning, Training Data, and Artificial Intelligence Solutions: Figure Eight." [Online]. Available: https://www.figure-eight.com/.

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

[41] F. Chollet, "Keras Documentation," Keras.Io, 2015. [Online]. Available: https://keras.io/. [Accessed: 28-Apr-2018].

[42] "FloydHub Documentation." [Online]. Available: https://docs.floydhub.com/. [Accessed: 28-Apr-2018].