Arabic Image Captioning using Deep Learning

Obeida ElJundi[1], Mohamad Dhaybi[1], Kotaiba Mokadam[4], Ali Yehya[3], Daniel Asmar[2] and Hazem Hajj[1]

[1]Department of Electrical & Computer Engineering, American University of Beirut
[2]Department of Mechanical Engineering, American University of Beirut
[3]Department of Computer Science, American University of Beirut
[4]Department of Civil Engineering, American University of Beirut

***Abstract*— The art of describing the content of an image by computers is a well-known problem in CV and NLP and know as image captioning. Image captioning can ease the life of children and impaired individuals. Although remarkable work has been accomplished recently for English, and due to the lack of large and publicly available dataset, the progress on Arabic Image Captioning is still lagging. In this paper, a robust Arabic Image Captioning dataset is developed based on one of the most popular English datasets; namely Flickr8K. Moreover, a deep learning model was built and benchmarked on the developed dataset.**

I. Introduction

The internet and social media have facilitated the way we communicate and visualize the world. Since the internet appeared, the online visual data generated by users has been growing exponentially. For instance, each day, around 300 million photos are upload to Facebook [1]. Although understanding the content of image appears to be a simple task, even for children, yet it is quite challenging for computers. Image captioning refers to the ability of automatically generating a syntactically plausible and semantically meaningful sentence that describes the content of an image. Enabling machines to describe the visual world would result in many advantages, such as improved information retrieval, early childhood education, for visually impaired persons, for social media, and so on [2].

Image captioning necessitates skills from the fields of Computer Vision (CV) and Natural Language Processing (NLP). Image captioning not only requires extracting meaningful information from an image, but also needs to express the extracted information in a human-readable sentence. Image captioning goes beyond and above object detection, since it also requires understanding the interactions between the detected objects in the scene. Moreover, the grammatical structure of a sentence generated by image captioning model provides more informative description of an image than a bag of unordered words produced by object detection algorithms.

Despite the difficult nature of image captioning, tremendous achievements have been accomplished recently, thanks to deep neural networks. Inspired by recent advances in neural machine translation [3] [4] [5], the encoder-decoder approach was adopted in several proposed image captioning methods [6] [7] [8]. The intuition is that image captioning can be thought of as translation, where the image is translated to a natural sentence. In machine translation, both the encoder and the decoder are Recurrent Neural Networks (RNN), or one of its variations (e.g., LSTM or GRU). As illustrated in Fig. 1, instead of RNN, the encoder in image captioning is a Convolution Neural Networks (CNN) [9], which is considered the State-of-the-art when dealing with images. Modern deep learning models attempt to apply the visual attention mechanism of humans inspired by cognitive psychology [10] [11] to look at the most important part of the image (see Fig. 2) [12] [13] [14].

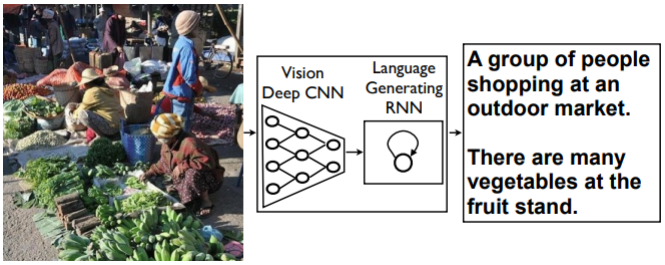

Fig. 1. Sequence-to-Sequence image captioning model (encoder: CNN, decoder: RNN)

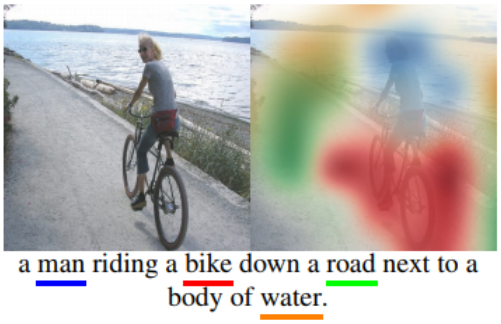

Fig. 2. Image captioning with attention

While English image captioning has attracted significant attention in research recently, the progress in Arabic Image Captioning (AIC) is still lagging in spite of the very large number of people who speak it. The Arabic language is spoken by more than 422 million people in the Arab world, and it is the native language in 22 countries. Arabic is ranked the fourth mostly used language on the web. Moreover, during the last five years, it is the fastest growing language

with a growth rate of 6091.9% in the number of Internet users [15].

In this work, we tackled this problem by developing our own AIC dataset and making it available for public. We translated an English benchmark dataset and developed a deep learning model to test the robustness of our dataset. Dataset and code can be accessed from: `https://github.com/ObeidaElJundi/Arabic-Image-Captioning`.

II. Literature Review

Here, we review the recent progress on both English and Arabic image captioning.

A. Deep Learning for English Image Captioning

Unlike traditional approaches, recent deep learning model tackle the task as end-to-end problem where parameters for both image understanding and language generation are learnt jointly.

Neural Machine Translation (NMT) has achieved significant progress recently, thanks to the Sequence-to-Sequence encoder-decoder framework [3] [4] [5]. Inspired by that, image captioning can be formulated as a translation task where image is translated into natural language. Several methods adopted the encoder-decoder framework for image captioning [6] [7] [8]. Google Neural Image Captioning (NIC) [6] model, *"show and Tell"*, is based on a CNN encoder and an RNN decoder. The CNN will extract relevant features form the image and encode them into a vector, where the RNN aims to decode the encoded vector into sentence. Jia et al. [7] proposed Guided LSTM (gLSTM) to help the decoder not to drift away or lose track of the original image content.

The encoder-decoder approach lacks interpretability since the language model (e.g., LSTM) is fed with an encoded representation of the whole scene and does not account for the spatial aspects of the image that is relevant to the parts of the image captions. Inspired by the human cognitive visual system [10] [11], the attention mechanism was adopted to focus on salient regions of the image while generating words [12] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25]. Xu et al. [12] was the first to adopt attention by proposing a model called *Show, Attend and Tell* based on the seminal Google NIC model, *Show and Tell* [6]. Lu et al. [19] proposed adaptive attention to help the decoder predict when to attend to the image (and if so, to which regions). The adaptive attention mechanism improved the overall performance by allowing the decoder to ignore looking at the image while generating non-visual words such as *"of"*, *"the'*, and *"a"*.

One description might not be enough to completely describe the entire visual scene. Dense captioning models generate several captions to describe many regions of an image [26] [27]. Johnson et al. [26] proposed DenseCap that can localize salient regions inside an image using CNN and generate descriptions for those regions. More advanced challenges, such as the target region overlapping, were addressed by [27].

Compositional approaches were proposed where, in contrast to the end-to-end framework, independent building blocks are combined to generate captions [28] [29] [30] [31] [32] [33]. The first block extract semantic visual concepts (e.g., attributes) using CNN, where the seconds block utilizes extracted concepts to generate captions using a language model (e.g., LSTM). Generated captioned are re-ranked based on similarity methods.

Generative adversarial networks (GAN) are recent architectures well-known by their ability of learning deep features from unlabeled data [34]. Dai et al. [35] and Shetty et al. [36] utilized Conditional Generative Adversarial Networks (CGAN) to improve the naturalness and diversity of the generated captions and achieved remarkable results.

In Reinforcement Learning (RL), instead of learning from labeled data, agents learn by receiving rewards based on actions they perform. Ren et al. [37] proposed a novel RL based image captioning that is consisted of two networks; the policy network predicts the next word based on the current state, where the value network guides the policy network by evaluating its reward. Rennie et al. [38] developed a new optimization approach named self-critical sequence training (SCST) and achieved remarkable results.

Table I illustrates some of the seminal work and their results on three benchmark datasets: Flickr8k [39], Flickr30k [40], and MS COCO [41].

B. Deep Learning for Arabic Image Captioning

Mualla and Alkheir [42] built an Arabic Description Model (ADM) which generates a full image description in Arabic by taking as input image features obtained from a CNN and a JSON file containing image descriptions in English. The English JSON description file is translated to Arabic and fed to an LSTM network along with the feature vector generated by the CNN. Authors reported that it is a bad idea to just translate the generated captions of English image captioning models to Arabic since it suffers from bad construction of Arabic sentence.

Jindal [43] leveraged the heavy influence of root words in Arabic by generating root words from images instead of captions. The contribution is divided into three stages: 1- using CNN, root words are extracted instead of actual sentence to map fragments in image. 2- Root-word based RNN then converts extracted roots to more appropriate morphological inflections to describe an image. 3- The ordering of words in sentences was checked by dependency tree relations of obtained words. The results show that generating Arabic captions is better than generating English captions and translating them to Arabic.

Al-Muzaini et al. [44] built an Arabic dataset based on two English benchmark dataset, Flickr8k [39] and MS COCO [41], using Crowd-Flower Crowdsourcing [45]. Moreover, a merge model was developed basaed on LSTM and CNN to achieve excellent results.

Previous work on AIC used small dataset for training, which may result in poor performance. Al-Muzaini et al. [44], for example, trained their deep learning model on 2400

		Encoder-Decoder		With Attention	
		Google NIC [6]	gLSTM [7]	Attend and Tell [12]	Adaptive [19]
Flickr8k	B1	63	64.7	67	-
	B2	41	45.9	45.7	-
	B3	27	31.8	31.4	-
	B4	-	21.6	21.3	-
	M	-	20.6	20.3	-
Flickr30k	B1	66.3	64.6	66.9	67.7
	B2	42.3	44.6	43.9	49.4
	B3	27.7	30.5	29.6	35.4
	B4	18.3	20.6	19.9	25.1
	M	-	18.6	18.5	20.4
MS COCO	B1	66.6	67	71.8	74.2
	B2	46.1	49.1	50.4	58
	B3	32.9	35.8	35.7	43.9
	B4	24.6	26.4	25	33.2
	M	-	23.3	23.9	26.6

TABLE I

SEMINAL WORK RESULTS ON THREE BENCHMARK DATASETS. B1: BLEU 1, B2: BLEU 2, B3: BLEU 3, B4: BLEU 4, M: METEOR

sample only. Moreover, to the best of our knowledge, no Arabic dataset is avaibale for public yet. Therefore, we aim to develop and publish the first AIC dataset. Our Arabic dataset will be similar to one of the English benchmark datasets; namely Flickr8K [39].

III. METHODOLOGY

in section III-A, we illustrate the sequence-to-sequence encoder-decoder framework for Neural Machine Translation (NMT). Next, inspired by NMT, we describe our proposed Arabic image captioning model in section III-B. In section III-C transfer learning is introduced to improve training and generalization. In section III-D, we talk about why and how we are going to prepare our own dataset.

A. Encoder-Decoder for Neural Machine Translation

The Recurrent Neural Network (RNN) [46] is a special type of neural networks specialized for sequence data, such as time series and text. RNN processes a sequence of inputs $(x_1, ..., x_T)$ and produces a sequence of outputs $(y_1, ..., y_T)$ by iterating the following equation:

$$h_t = f(W_h x_t + U_h h_{t-1} + b_h)$$

$$y_t = f(W_y h_t + b_y)$$

where x_t is the current input vector, h_t is the current hidden state vector, y_t is the current output vector, f is a non-linear activation function such as sigmoid σ, and W, U, and b are parameters to be learnt.

RNN suffers from two main issues. RNN input and output sequences should have the same length. This is not the case with natural language translation, where the source language and the destination language words count may differ. Therefore, a simple strategy is to encode the input sequence into a fixed-sized vector, usually called input representation or thought context, using one RNN, and then to decode the vector to the target sequence with another RNN, as shown in figure 3. Thus, the decoder RNN can keep generating outputs until it ends with a special end-of-sentence symbol *"<EOS>"*. This well-known framework is called *Sequence-to-Sequence Encoder-Decoder* framework, and it is shown in figure 3. The other issue with standard RNN is the vanishing gradient during training due to its disability of handling long term dependencies [47] [48]. Therefore, standard RNN cells are replaced by Long Short-Term Memory (LSTM) [49] cells in most settings due to their superiority of learning long range temporal dependencies.

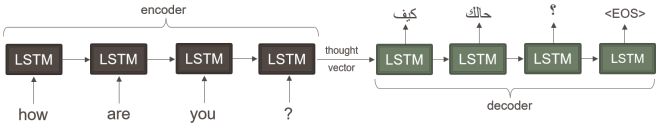

Fig. 3. Sequence-to-Sequence Encoder-Decoder framework for NMT

B. Encoder-Decoder for Image Captioning

The aforementioned *Sequence-to-Sequence Encoder-Decoder* framework can be utilized for image captioning by encoding the input image into features vector and decoding that features vector to Arabic sentence. The only difference between NMT framework and image captioning framework is that CNN is used for input encoding instead or RNN, as illustrated in figure 4.

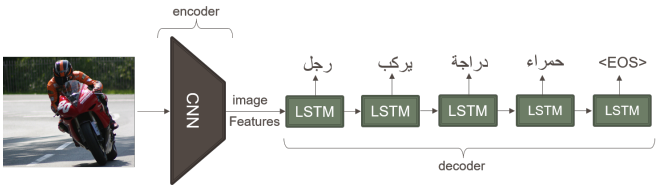

Fig. 4. Sequence-to-Sequence Encoder-Decoder framework for Arabic Image Captioning

Given any input image and its corresponding Arabic caption, the Arabic image captioning encoder-decoder model

should maximize the following loss function:

$$\arg\max_{\theta} \sum_{(I,y)} log\ p(y|I;\theta)$$

where I is the input image, θ are parameters to be learnt, and $y = y_1, ..., y_t$ is the corresponding Arabic caption.

The flow is the following. The image is fed to a CNN to generate its visual representation (or image features), represented as x_{-1} which is the first input of the upcoming LSTM.

$$x_{-1} = CNN(I)$$

Subsequent inputs at different time states are word embeddings; vector of numbers that reflect semantics where words with similar meaning have close embeddings. The embedding for each word is calculated as follows:

$$x_t = W_e S_t \quad for \quad t = 0, ..., N$$

where W_e is $300X|V|$ word embedding matrix, meaning each word will be represented by a vector of length 300. $|V|$ denotes the vocabulary length; the number of unique words in our dataset. S_t is a $|V|X1$ one hot vector representing word i. Each hidden state of the LSTM emits a prediction for the next word in the sentence, denoted by p_{t+1}, as follows:

$$p_{t+1} = LSTM(x_t)$$

C. Transfer Learning

Instead of initializing our decoder CNN weights randomly and train from scratch, we will use the weights of a pre-trained CNNs. This is known as transfer learning, which refers to the situation where what has been learned in one setting (task) is exploited to learn other setting (task). Transfer learning is used a lot in the literature to improve model generalization and fasten training. For our CNN, we will use VGG16 [50], one of the state-of-the-art models for object detection. VGG16 contains 13 convolution layers and 3 fully connected layers and is able to detect around 1000 different objects. VGG16 is available online for public[1], and its architecture is illustrated in figure 5.

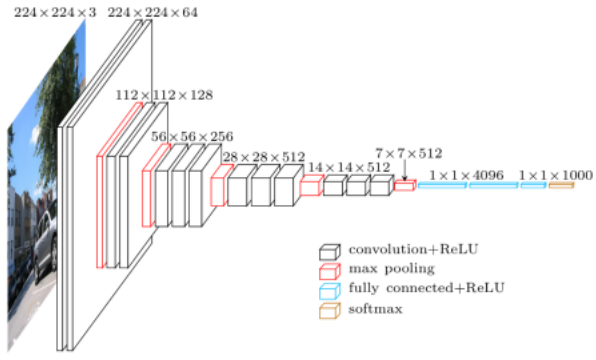

Fig. 5. VGG16 model for object detection

[1]http://www.robots.ox.ac.uk/~vgg/research/very_deep/

D. Data Preparation

Data is the most essential factor to train any Machine or Deep Learning algorithm. Work on Arabic Image Captioning is still lagging, specially when it comes to datasets. Previous work on Arabic Image Captioning developed their own datasets by translating English Image Captioning datasets, such as MS COCO [41] or Flickr8K [39]. Translation was done either in lab by experts or by crowdsourcing using Crowd-Flower [45]. Nevertheless, authors did not make their Arabic datasets available online for public yet.

Therefore, we aim to develop and share our own dataset by translating one of the English benchmark datasets, Flickr8K [39]. Flickr8K contains 8000 images, each has 5 English captions annotated by humans. Images are extracted from flickr[2] and mainly contain humans and animals. 6000 images are reserved for training, 1000 for validation, and 1000 for testing. We'll be translating all the English captions to Arabic using Google Translate API[3]. The translation process will be automated using a python client of the API[4]. Google Translate API performance suffers when it comes to Arabic translations, so translated captions will be validated by us to fix any Arabic-specific translation issues.

IV. Experiments & Results

A. Dataset

Deep learning models are known to be data hungry algorithms; meaning they require much data to learn the input-output mapping. AIC suffers form the lack of big and robust dataset. Previous work on AIC not only translated small portions of benchmark English datasets, such as [44] who trained on 2400 samples only, but also did not make their translated dataset available for public. To address this issue, we developed our own AIC dataset and ran preliminary experiments on it to ensure its quality. We translated Flickr8K [39], one of the most popular datasets for English Image Captioning. Translation was automated using Google Translate API. Google uses Neural Machine Translation, described in section III-A, as its main translation engine. Translation from or to Arabic has not yet achieved desired results and still suffers from Arabic specific issues. For example, few translated captions contain a word that was translated literally and out of context, which makes the whole Arabic sentence incoherent. Therefore, translated captions were then verified by choosing the best three translated captions out of 5 and modifying some captions if needed. Figure 6 illustrates three examples from our translated dataset. Up until the moment of writing this paper, the dataset contains 4102 translated and verified captions, and the rest will be added soon. 500 samples will be used for testing, while the rest will be utilized for training. Our dataset is available for public, and it can be found here: https://github.com/ObeidaElJundi/Arabic-Image-Captioning.

[2]https://www.flickr.com/
[3]https://cloud.google.com/translate
[4]https://googleapis.github.io/google-cloud-python/latest/translate

Fig. 6. sample from our translated dataset

B. Preprocessing

Dataset contains raw text, which may include useless textual information. It is crucial to clean and preprocess our data before feeding it to any model because *"garbage in, garbage out"*. We followed Arabic preprocessing techniques recommended by [51]: Diacritics were removed, the *"hamza"* on characters was normalized in addition to normalizing some word ending characters such as the *"t marbouta"* and *"ya' maqsoura"*. Moreover, we got rid of punctuation as well as non Arabic letters. Finally, a special start and end token were added at the beginning and the end of each caption to mark the starting and the ending point of each caption. Short captions were padded with a special padding token to ensure having captions of the same length.

C. Training & Results

The complete model was implemented in python[5] using the latest version of Keras [52], a deep learning framework built on top of Google's famous framework: TensorFlow [53]. Training was done on a local PC utilizing NVIDIA GTX 1080 GPU (8GB vRAM) and 32GB RAM.

For the image model, a pre-trained VGG16, excluding the last later, was used to map images to embeddings, a vector of length 4096. The image embeddings vector was then mapped to a vector of 256 by a fully connected layer with tanh activation function to force the output values to be between -1 and 1. For the language model, a single hidden LSTM layer with 256 memory units was defined. The initial state of the LSTM was set to be the image embeddings to ensure generating captions related to a specific image.

The loss function was Softmax Cross Entropy, and the optimization was done with mini batch Gradient Descent with Adam optimizer and batch size of 128. The total number of epochs was 10. We consider an epoch as a single pass of the complete training dataset through the training process. Each epoch took around 12 seconds.

Following previous works, the model was evaluated on the BLEU-1,2,3,4 [54], which evaluates a candidate sentence by measuring the fraction of n-grams that appear in a set of references. BLEU scores of our training and testing are illustrated in figure 7. Finally, some accurate and inaccurate results of our encoder-decoder model are demonstrated in figures 8 and 9, respectively.

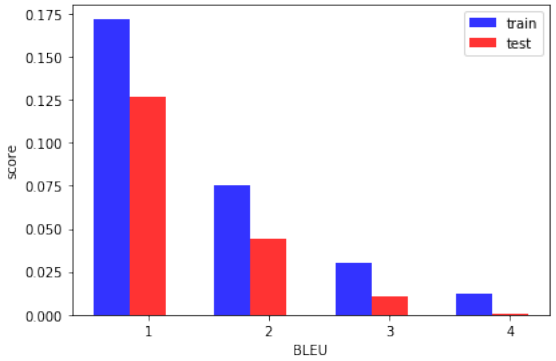

Fig. 7. BLEU scores for our AIC model

Our results are not as good as the work done for English. This is due to the fact that the Arabic language is by nature more complicated than English. Furthermore, increasing the dataset size will improve our performance noticeably.

V. Conclusion & Future Work

In this work, we developed a benchmark Arabic dataset for image captioning by translating a popular English dataset. Verifying translated captions was accomplished by experts to ensure robustness and clarity. Arabic captioned were then preprocessed to clean any useless and confusing textual information. Finally, a sequence-to-sequence encoder-decoder model was trained on our dataset, and results were reported. Model was shown to perform good on some images and poorly on others.

We believe that AIC performance can be improved tremendously by training on more samples. Therefore, out future work aims to increase the size of our dataset by translating and verifying more benchmark English datasets, such as MS COCO [41]. Furthermore, we plan to apply attention mechanisms for our encoder-decoder model, thus, improving performance and interpretability.

[5] https://www.python.org/

كلب أسود يقفز في الهواء

كلب أبيض يركض في الثلج

صبي في ثوب سباحة يلعب في الماء

Fig. 8. accurate results generated by our AIC model

صبي في سترة حمراء يلعب في الماء

كلب بني يقف في الماء

مجموعة من الناس يحملون المشروبات ويشيرون إلى الكاميرا

Fig. 9. inaccurate results generated by our AIC model

References

[1] *Top 20 facebook statistics - updated september 2018*, Oct. 2018. [Online]. Available: https://zephoria.com/top-15-valuable-facebook-statistics/.

[2] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, 2018.

[3] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[4] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1700–1709.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

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

[7] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2407–2415.

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

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[10] R. A. Rensink, "The dynamic representation of scenes," *Visual cognition*, vol. 7, no. 1-3, pp. 17–42, 2000.

[11] M. W. Spratling and M. H. Johnson, "A feedback model of visual attention," *Journal of cognitive neuroscience*, vol. 16, no. 2, pp. 219–237, 2004.

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

[13] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[14] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," in *Advances in Neural Information Processing Systems*, 2016, pp. 2361–2369.

[15] N. Boudad, R. Faizi, R. O. H. Thami, and R. Chiheb, "Sentiment analysis in arabic: A review of the literature," *Ain Shams Engineering Journal*, 2017.

[16] Z. Y. Y. Y. Y. Wu and R. S. W. W. Cohen, "Encode, review, and decode: Reviewer module for caption generation," *arXiv preprint arXiv:1605.07912*, 2016.

[17] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang, "Aligning where to see and what to tell: Image caption with region-based attention and scene factorization," *arXiv preprint arXiv:1506.06272*, 2015.

[18] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," *arXiv preprint arXiv:1612.01033*, 2016.

[19] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 6, 2017, p. 2.

[20] C. Liu, J. Mao, F. Sha, and A. L. Yuille, "Attention correctness in neural image captioning.," in *AAAI*, 2017, pp. 4176–4182.

[21] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, pp. 6298–6306.

[22] H. R. Tavakoliy, R. Shetty, A. Borji, and J. Laaksonen, "Paying attention to descriptions generated by image captioning models," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, IEEE, 2017, pp. 2506–2515.

[23] 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 vqa," *arXiv preprint arXiv:1707.07998*, vol. 2, no. 4, p. 8, 2017.

[24] C. Chunseong Park, B. Kim, and G. Kim, "Attend to you: Personalized image captioning with context sequence memory networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 895–903.

[25] Y. Sugano and A. Bulling, "Seeing with humans: Gaze-assisted neural image captioning," *arXiv preprint arXiv:1608.05203*, 2016.

[26] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4565–4574.

[27] L. Yang, K. D. Tang, J. Yang, and L.-J. Li, "Dense captioning with joint inference and visual context.," in *CVPR*, 2017, pp. 1978–1987.

[28] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, *et al.*, "From captions to visual concepts and back," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1473–1482.

[29] K. Tran, X. He, L. Zhang, J. Sun, C. Carapcea, C. Thrasher, C. Buehler, and C. Sienkiewicz, "Rich image captioning in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 49–56.

[30] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2321–2334, 2017.

[31] S. Ma and Y. Han, "Describing images by feeding lstm with structural words," in *Multimedia and Expo (ICME), 2016 IEEE International Conference on*, IEEE, 2016, pp. 1–6.

[32] R. M. Oruganti, S. Sah, S. Pillai, and R. Ptucha, "Image description through fusion based recurrent multi-modal learning," in *Image Processing (ICIP), 2016 IEEE International Conference on*, IEEE, 2016, pp. 3613–3617.

[33] M. Wang, L. Song, X. Yang, and C. Luo, "A parallel-fusion rnn-lstm architecture for image caption generation," in *Image Processing (ICIP), 2016 IEEE International Conference on*, IEEE, 2016, pp. 4448–4452.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[35] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional gan," *arXiv preprint arXiv:1703.06029*, 2017.

[36] R. Shetty, M. Rohrbach, L. A. Hendricks, M. Fritz, and B. Schiele, "Speaking the same language: Matching machine to human captions by adversarial training," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[37] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," *arXiv preprint arXiv:1704.03899*, 2017.

[38] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *CVPR*, vol. 1, 2017, p. 3.

[39] 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.

[40] 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.

[41] 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.

[42] R. Mualla and J. Alkheir, "Development of an arabic image description system,"

[43] V. Jindal, "Generating image captions in arabic using root-word based recurrent neural networks and deep neural networks," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 2018, pp. 144–151.

[44] 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.

[45] *High quality training data platform for ml models*. [Online]. Available: `https://www.figure-eight.com/`.

[46] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[47] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[48] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber, *et al.*, *Gradient flow in recurrent nets: The difficulty of learning long-term dependencies*, 2001.

[49] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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

[51] 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.

[52] F. Chollet *et al.*, *Keras*, `https://keras.io`, 2015.

[53] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: `https://www.tensorflow.org/`.

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