Resources and End-to-End Neural Network Models for Arabic Image Captioning

Obeida ElJundi[1], Mohamad Dhaybi[1], Kotaiba Mokadam[2], Hazem Hajj[1] and Daniel Asmar[3]
[1]*American University of Beirut, Electrical and Computer Engineering Department, Lebanon*
[2]*American University of Beirut, Civil and Environmental Engineering Department, Lebanon*
[3]*American University of Beirut, Mechanical Engineering Department, Lebanon*

Keywords: Deep Learning, Computer Vision, Natural Language Processing, Image Captioning, Arabic.

Abstract: Image Captioning (IC) is the process of automatically augmenting an image with semantically-laden descriptive text. While English IC has made remarkable strides forward in the past decade, very little work exists on IC for other languages. One possible solution to this problem is to boostrap off of existing English IC systems for image understanding, and then translate the outcome to the required language. Unfortunately, as this paper will show, translated IC is lacking due to the error accumulation of the two tasks; IC and translation. In this paper, we address the problem of image captioning in Arabic. We propose an end-to-end model that directly transcribes images into Arabic text. Due to the lack of Arabic resources, we develop an annotated dataset for Arabic image captioning (AIC). We also develop a base model for AIC that relies on text translation from English image captions. The two models are evaluated with the new dataset, and the results show the superiority of our end-to-end model.

1 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 uploaded to Facebook (Inc., 2018). 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, an aid for visually impaired persons, for social media, and so on (Bai and An, 2018).

Image captioning requires extracting meaningful information about the content of the image and expresses the extracted information in a human-readable sentence. As a result, image captioning models need to achieve several objectives including object detection, extraction of relationships among objects, inference of the pragmatic information within the image, transcribing the information into coherent textual forms with correct syntactical and semantic structures.

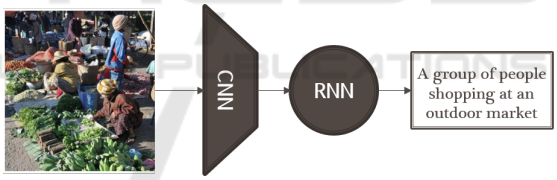

Figure 1: Sequence-to-Sequence image captioning model (encoder: CNN, decoder: RNN).

Despite these significant challenges with image captioning, tremendous achievements have been accomplished recently with deep neural networks. Inspired by recent advances in neural machine translation (Cho et al., 2014) (Kalchbrenner and Blunsom, 2013) (Sutskever et al., 2014), the encoder-decoder approach has been adopted in several proposed image captioning methods (Vinyals et al., 2015) (Jia et al., 2015) (Donahue et al., 2015). The intuition is that image captioning can be thought of as translation from a set of image pixels to a natural sentence. In machine translation, typically, both the encoder and the decoder include 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 Network

ElJundi, O., Dhaybi, M., Mokadam, K., Hajj, H. and Asmar, D.
Resources and End-to-End Neural Network Models for Arabic Image Captioning.
DOI: 10.5220/0008881202330241
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 233-241
ISBN: 978-989-758-402-2; ISSN: 2184-4321

(CNN) (LeCun et al., 1998), which is considered the ideal solution when dealing with unstructured, spatial data such as images. Modern deep learning models attempt to apply the visual attention mechanism of humans inspired by cognitive psychology (Rensink, 2000) (Spratling and Johnson, 2004) to look at the most important part of the image (Xu et al., 2015) (You et al., 2016) (Yang et al., 2016).

In addition to the image captioning challenges mentioned above, the models for text transcription of image content needs to comprehend the semantics of the target language in the text. As a result, while there has been significant advances in models for image captioning for English, those models can not be applied directly to other languages such as Arabic. As a consequence, the progress in Arabic Image Captioning (AIC) is still lagging. Consideration for the Arabic language is important since it is spoken by more than 422 million people in the Arab world, and it is the native language in 22 countries. Arabic is also 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 (Boudad et al., 2017).

In this work, we address the challenge of image captioning in Arabic including the lack of Arabic resources. We develop a new AIC dataset and propose two separate models for the evaluation of AIC. The first model uses English image captioning models then translates the English text to Arabic. The second model is an end-to-end model that directly transcribes Arabic text from images. The models are compared using our AIC dataset, and the results show the superiority of our end-to-end AIC model. The dataset and code are made publicly available: https://github.com/aub-mind/Arabic-Image-Captioning.

The rest of the paper is organized as follows. Section 2 summarizes the literature of English and Arabic image captioning. We describe our dataset and end-to-end system in section 3. Our experiments and results are demonstrated in section 4. Finally, section 5 concludes the paper.

2 RELATED WORK

In this section, we review the recent progress on both English and Arabic image captioning.

2.1 Deep Learning for English Image Captioning

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

Neural Machine Translation (NMT) has achieved significant progress recently, thanks to the Sequence-to-Sequence encoder-decoder framework (Cho et al., 2014) (Kalchbrenner and Blunsom, 2013) (Sutskever et al., 2014). Inspired by that, image captioning can be formulated as a translation task, where an image is translated into natural language. Several methods adopted the encoder-decoder framework for image captioning (Vinyals et al., 2015) (Jia et al., 2015) (Donahue et al., 2015). Google's Neural Image Captioning (NIC) (Vinyals et al., 2015) model, *"show and Tell"*, is based on a CNN encoder and an RNN decoder. The CNN extracts relevant features form an image and encodes them into a vector; on the other hand, the RNN aims to decode the encoded vector into a sentence. Jia et al. (Jia et al., 2015) proposed Guided LSTM (gLSTM) to help the decoder not 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 entire 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 (Rensink, 2000) (Spratling and Johnson, 2004), the attention mechanism was adopted to focus on salient regions of the image while generating words (Xu et al., 2015) (Wu and Cohen, 2016) (Jin et al., 2015) (Pedersoli et al., 2016) (Lu et al., 2017) (Liu et al., 2017) (Chen et al., 2017) (Tavakoliy et al., 2017) (Anderson et al., 2017) (Chunseong Park et al., 2017) (Sugano and Bulling, 2016). Xu et al. (Xu et al., 2015) 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* (Vinyals et al., 2015). Lu et al. (Lu et al., 2017) 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 (Johnson et al., 2016) (Yang et al., 2017). (Johnson et al., 2016) proposed DenseCap, which can localize salient regions inside an image using a CNN, and generate descriptions for those regions. More advanced challenges, such as the target region overlapping, were addressed by (Yang et al., 2017).

Table 1: Previous work results on 3 benchmark datasets. B1:BLEU 1, B2:BLEU 2, B3:BLEU 3, B4:BLEU 4, M:METEOR.

		Encoder-Decoder		With Attention	
		Google NIC (Vinyals et al., 2015)	gLSTM (Jia et al., 2015)	Attend and Tell (Xu et al., 2015)	Adaptive (Lu et al., 2017)
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

Compositional approaches were proposed where, in contrast to the end-to-end framework, independent building blocks are combined to generate captions (Fang et al., 2015) (Tran et al., 2016) (Fu et al., 2017) (Ma and Han, 2016) (Oruganti et al., 2016) (Wang et al., 2016). The first block extracts semantic visual concepts (e.g., attributes) using a CNN, and the second 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 to learn deep features from unlabeled data (Goodfellow et al., 2014). Dai et al. (Dai et al., 2017) and Shetty el al. (Shetty et al., 2017) 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. (Ren et al., 2017) proposed a novel RL based image captioning that consists of two networks: the policy network predicts the next word based on the current state, and the value network guides the policy network by evaluating its reward. Rennie et al. (Rennie et al., 2017) developed a new optimization approach named self-critical sequence training (SCST) and achieved good results.

Table 1 illustrates some of the previous work and their results in terms of BLEU and METEOR on three benchmark datasets: Flickr8k (Hodosh et al., 2013), Flickr30k (Young et al., 2014), and MS COCO (Chen et al., 2015).

2.2 Deep Learning for Arabic Image Captioning

Mualla and Alkheir (Mualla and Alkheir,) built an Arabic Description Model (ADM) that 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 recognized captions in English to Arabic because the poor structure of the resulting Arabic sentences.

Jindal (Jindal, 2018) leveraged the heavy influence of root words in Arabic to generate root words from images instead of from captions. Using a CNN, root words instead of actual sentence are extracted from images; the roots are then translated to morphological inflections, and finally, a dependency tree is used to verify the ordering of words in sentences. The results show that generating Arabic captions directly in one stage, produced superior results to a two stage English caption→Arabic translation process.

Using Crowd-Flower Crowdsourcing (https://www.figure eight.com/,), Al-Muzaini et al. (Al-Muzaini et al., 2018) built an Arabic dataset based on two English benchmark datasets, Flickr8k (Hodosh et al., 2013) and MS COCO (Chen et al., 2015). Moreover, a merge model was developed

Figure 2: Sample from our translated dataset.

based on LSTM and CNN to achieve excellent results.

The contributions of this paper include the following:

- First, although previous Arabic captioning datasets do exist, to the best of our knowledge, none of them is public. In this paper, we provide public access to the first large Arabic captioning dataset.[1]
- Second, most previous work on AIC used small datasets for training, which risk over-fitting. For example, Al-Muzaini et al. (Al-Muzaini et al., 2018) trained their network on only 2400 samples. We are proposing training on a considerably larger dataset, consisting of 8092 images, each having 3 captions.
- Third, the merge model is the most common deep learning model in AIC literature. To the best of our knowledge, we are the first to propose approaching AIC using the sequence-to-sequence encoder-decoder framework.

3 METHODOLOGY

In this section, we first describe the Arabic captioning dataset we are releasing. Next, we review the sequence-to-sequence encoder-decoder framework for Neural Machine Translation (NMT). Finally, inspired by NMT, we describe our proposed Arabic image captioning model.

3.1 Proposed Arabic Captioned Dataset

The proposed Arabic captioning dataset is built by translating the Flickr8K (Hodosh et al., 2013) dataset, containing 8000 images, each captioned five times by humans. Images are extracted from flickr[2] and mainly contain humans and animals. Translation to Arabic is performed in two steps: first, all English captions are translated using the Google Translate API[3]; the process is automated using a python client of the API[4]. During a second stage, all translated captions are edited and validated by a professional Arabic translator. This is necessary because of the many contextual errors Google Translate performs. Fig. 2 shows examples of some of the Arabic captioned images.

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 entire Arabic sentence incoherent. Therefore, translated captions were then verified by choosing the best three translated captions out of five, and modifying some captions if needed.

3.2 End-to-End Model for Arabic Image Captioning

The Recurrent Neural Network (RNN) (Elman, 1990) is specialized for sequence data, such as time series and text. RNNs process a sequence of inputs $(x_1,...,x_T)$ and produce 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.

RNNs suffers from two main issues: first, RNN input and output sequences are expected to have the

[1] https://github.com/aub-mind/Arabic-Image-Captioning

[2] https://www.flickr.com/

[3] https://cloud.google.com/translate

[4] https://googleapis.github.io/google-cloud-python/latest/translate

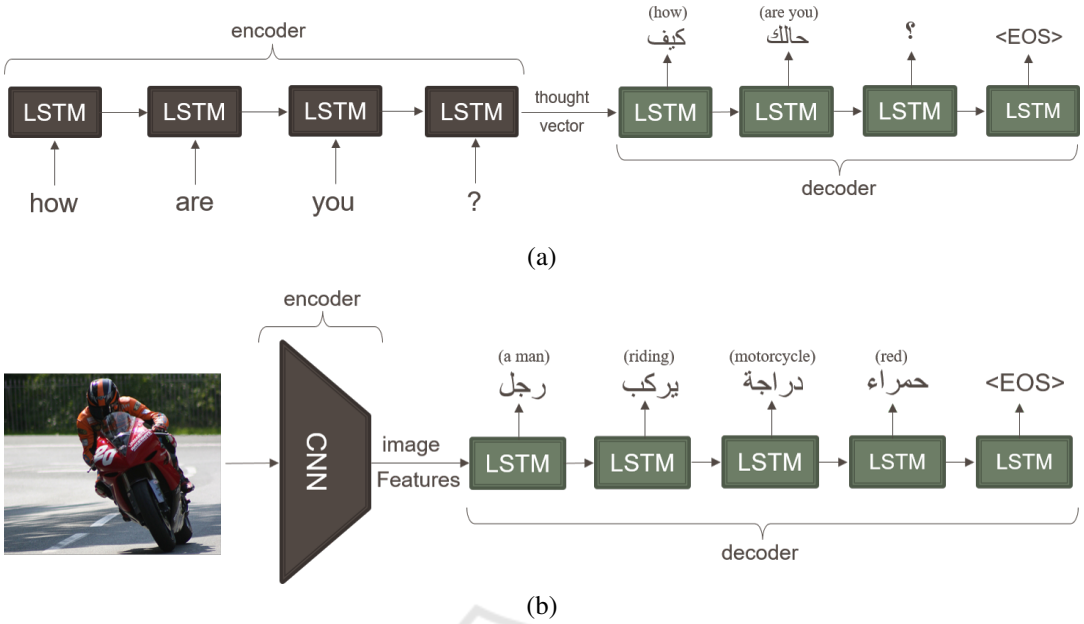

Figure 3: Sequence-to-Sequence Encoder-Decoder framework for NMT (a) and for for Arabic Image Captioning (b).

same length, which is not the case in natural language translation, where the word count for the source and destination languages may differ. A simple strategy around this issue is to encode the input sequence into a fixed-sized vector (usually referred to as thought vector) 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 a *Sequence-to-Sequence Encoder-Decoder* framework.

Second, in a standard RNN the gradient tends to vanish during training due to its inability to handle long-term dependencies (Bengio et al., 1994) (Hochreiter et al., 2001). To mitigate this issue, standard RNN cells are replaced by Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) cells because of their superiority in learning long-range temporal dependencies. The aforementioned *Sequence-to-Sequence Encoder-Decoder* framework can be utilized for image captioning by encoding the input image into a feature vector, and then decoding that feature vector into an Arabic sentence. The only difference between an NMT and the image captioning framework is that a CNN is used for input encoding instead of an RNN, as illustrated in Figure 3.

Given any input image and its corresponding Arabic caption, the Arabic image captioning encoder-decoder model maximizes 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 learned, and $y = y_1,...,y_t$ is the corresponding Arabic caption.

During Arabic captioning, each image is fed to a CNN to generate its visual representation (or image features), represented as $x_{-1} = CNN(I)$, which is the first input of the upcoming LSTM. Subsequent inputs at different time stamps are now represented by word embeddings, which are vectors of numbers that reflect semantics; words with similar meaning have close embeddings. The embedding for each word is calculated as $x_t = W_e S_t \quad for \quad t = 0,...,N$, where W_e is a $300X|V|$ word embedding matrix, meaning each word will be represented by a vector of length 300. $|V|$ denotes the vocabulary length, which is 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 $p_{t+1}LSTM(x_t)$.

Transfer Learning. Instead of initializing our decoder CNN weights randomly and train from scratch, we will use the weights of a pre-trained CNN. This is known as transfer learning, which refers to the situation where what has been learned in one setting (task) is exploited to learn another setting (task). Transfer learning is used a lot in the literature to improve model generalization and speed up training. For our CNN, we use VGG16 (Simonyan and Zisserman, 2014), one of the state-of-the-art models for object detection. VGG16 contains thirteen convolution layers

and three fully connected layers, and is able to detect approximately one thousand different objects.

3.3 English Captions Translation

To demonstrate the necessity of our end-to-end AIC system, we develop and train an English image captioning (EIC) system similar to our Arabic IC described in section 3.2 using the original Flicker8K dataset. We then translate the generated English captions to Arabic using a pre-trained NMT model, namely Google Translate, which is based on the sequence-to-sequence encoder-decoder framework (Sutskever et al., 2014). The translated Arabic captions are evaluated and compared with our end-to-end Arabic IC output. A high level comparison of AIC against translated EIC is illustrated in Figure 4.

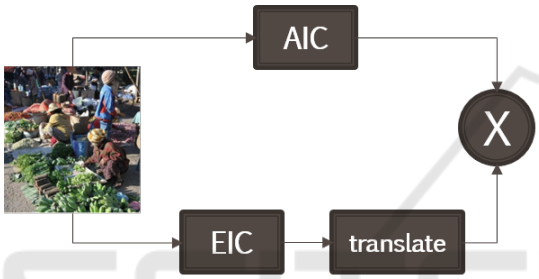

Figure 4: AIC against translated EIC.

4 EXPERIMENTS AND RESULTS

To validate the efficiency of our proposed Arabic image captioning system, we perform experiments on our dataset of 8092 images, the last 800 samples are used for testing, while the rest are utilized for training.

4.1 Preprocessing

Datasets contain 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 (Shoukry and Rafea, 2012): 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.

4.2 Evaluation of AIC End-to-End Model

The complete model was implemented in Python[5] using the latest version of Keras (Chollet et al., 2015), a deep learning framework built on top of TensorFlow (Abadi et al., 2015). 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, in order to ensure generating captions related to a specific image. The loss function was Softmax Cross Entropy. Figure 7 shows the error rate (loss) of our model after each epoch during training. The optimization was done with mini batch Gradient Descent with Adam optimizer and batch size of 1024. The total number of epochs was 5. We consider an epoch as a single pass of the complete training dataset through the training process. Each epoch took around 25 seconds.

Some examples of accurate and inaccurate results of our encoder-decoder model are shown in Figures 5 and 6, respectively.

Following previous works, the model was evaluated on the BLEU-1,2,3,4 (Papineni et al., 2002), which assesses a candidate sentence by measuring the fraction of n-grams that appear in a set of references. BLEU scores for our training and testing are illustrated in Figure 8.

The results obtained for Arabic captioning are inferior to those of English captioning. We attribute this to two reasons: first, the Arabic language is by nature more complex than English; discarding the vocalization of letters during training has negatively impacted the accuracy of the captioning network. Second, the dataset used for English captioning is larger than that of Arabic; increasing the dataset size in the future will most likely increase the performance of the Arabic captioning network.

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

Figure 5: Accurate results generated by our AIC model.

Figure 6: Inaccurate results generated by our AIC model.

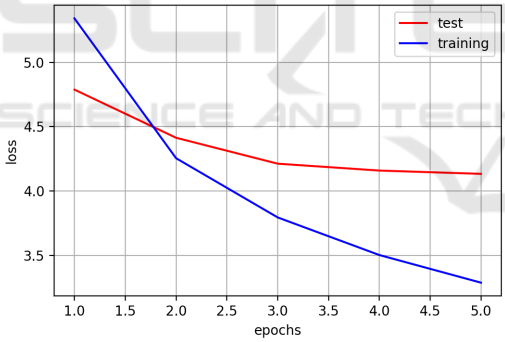

Figure 7: Model loss after each epoch.

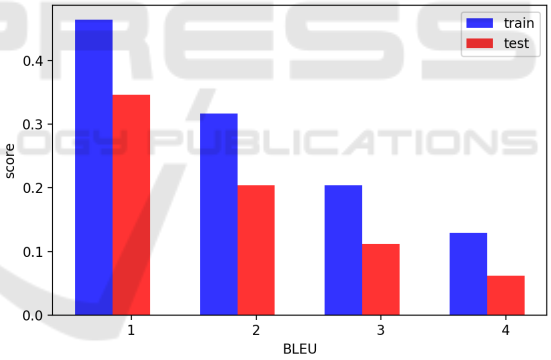

Figure 8: BLEU scores for our AIC model.

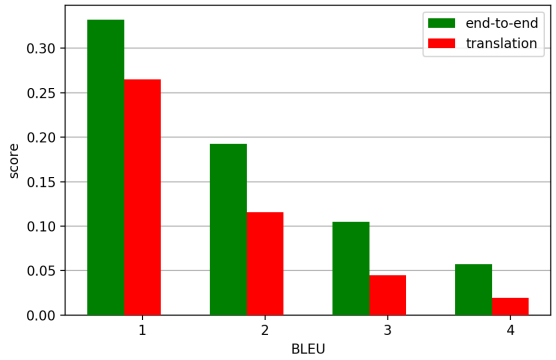

Figure 9: BLEU scores of end-to-end AIC vs translating English captions.

An end-to-end approach of directly generating Arabic captions outperformed translating English generated captions (See Fig. 9). One possible explanation is that using a deep learning model for English captioning followed by a second deep learning model for English-to-Arabic translation may accumulate both models errors and uncertainties.

5 CONCLUSION

This paper investigated the problem of Arabic image captioning. We prepared a large dataset of Arabic-captioned images, and will make it publicly avail-

able. To validate this work, we trained and tested a Encoder-Decoder Network on this dataset, and although the results are not at par with state of-the-art English captioning systems, they are considerably superior to the English-captioning →English-Arabic-translate. Our current work includes adding vocalization to the Arabic training set, as well as increasing the size of the dataset.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Al-Muzaini, H. A., Al-Yahya, T. N., and Benhidour, H. (2018). Automatic arabic image captioning using rnn-lstm-based language model and cnn. *International Journal of Advanced Computer Science and Applications*, 9(6):67–73.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2017). Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2(4):8.

Bai, S. and An, S. (2018). A survey on automatic image caption generation. *Neurocomputing*.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

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

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

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

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

Chollet, F. et al. (2015). Keras. https://keras.io.

Chunseong Park, C., Kim, B., and Kim, G. (2017). Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 895–903.

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

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

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

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

Fu, K., Jin, J., Cui, R., Sha, F., and Zhang, C. (2017). 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*, 39(12):2321–2334.

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

Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., et al. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

https://www.figure eight.com/. Figure eight. High Quality Training Data Platform for ML Models.

Inc., Z. (2018). Top 20 facebook statistics - updated september 2018. https://zephoria.com/top-15-valuable-facebook-statistics/.

Jia, X., Gavves, E., Fernando, B., and Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2407–2415.

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

Jindal, V. (2018). 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*, pages 144–151.

Johnson, J., Karpathy, A., and Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Liu, C., Mao, J., Sha, F., and Yuille, A. L. (2017). Attention correctness in neural image captioning. In *AAAI*, pages 4176–4182.

Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017). 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)*, volume 6, page 2.

Ma, S. and Han, Y. (2016). Describing images by feeding lstm with structural words. In *Multimedia and Expo (ICME), 2016 IEEE International Conference on*, pages 1–6. IEEE.

Mualla, R. and Alkheir, J. Development of an arabic image description system.

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Pedersoli, M., Lucas, T., Schmid, C., and Verbeek, J. (2016). Areas of attention for image captioning. *arXiv preprint arXiv:1612.01033*.

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

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). Self-critical sequence training for image captioning. In *CVPR*, volume 1, page 3.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42.

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

Shoukry, A. and Rafea, A. (2012). Preprocessing egyptian dialect tweets for sentiment mining. In *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages*, page 47.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Spratling, M. W. and Johnson, M. H. (2004). A feedback model of visual attention. *Journal of cognitive neuroscience*, 16(2):219–237.

Sugano, Y. and Bulling, A. (2016). Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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

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

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

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

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

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Yang, L., Tang, K. D., Yang, J., and Li, L.-J. (2017). Dense captioning with joint inference and visual context. In *CVPR*, pages 1978–1987.

Yang, Z., Yuan, Y., Wu, Y., Cohen, W. W., and Salakhutdinov, R. R. (2016). Review networks for caption generation. In *Advances in Neural Information Processing Systems*, pages 2361–2369.

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.