Active Learning Based Framework for Image Captioning Corpus Creation

Moustapha Cheikh[1,3](✉) and Mounir Zrigui[2,3]

[1] Faculty of Economics Management, University of Sfax, Sfax, Tunisia
moustapha.ml.cheikh@gmail.com

[2] Faculty of sciences of Monastir, University of Monastir, Monastir, Tunisia
mounir.zrigui@fsm.rnu.tn

[3] Research Laboratory in Algebra, Numbers Theory and Intelligent Systems,
University of Monastir, Monastir, Tunisia

Abstract. Image captioning aims at analyzing the content of an image in order to subsequently generate a textual description through verbally expressing the important aspects of it. In spite of the fact that the task of automatic image description is not bound to the English language, yet, the recent advances mostly focus on English descriptions. Collecting captions for images is an expensive process that requires time and labor cost. In this paper, we introduce a novel active learning framework with human in the loop for image captioning corpus creation, using a translated version of existing datasets. We implemented this framework to create a new dataset called ArabicFlickr1K. This dataset has 1095 images, each is associated with three to five descriptions. We also propose a neural network architecture to automatically generate Arabic captions for images. This architecture relies on an encoder-decoder framework. Our model scored 47% on BLUE-1.

Keywords: Image captioning · Computer vision · Natural language processing

1 Introduction

Overs the last few years has been renewed interest in tasks that require a combination of linguistic and visual information [1]. This interest is largely motivated by the amount of available data on the internet and the recent advances in computer vision.

Image captioning [10] has become a key task with the interest of both natural language processing (NLP) and computer vision communities. This task consists of analyzing the content of an image in order to subsequently generate a textual description of it by verbally expressing the important aspects of that image.

Image captioning may play an important role in many applications. For instance the generated captions can be used in text based information retrieval [12], video indexing [44] and several other NLP applications.

I. S. Kotsireas and P. M. Pardalos (Eds.): LION 14 2020, LNCS 12096, pp. 128–142, 2020.
https://doi.org/10.1007/978-3-030-53552-0_14

The description could be difficult because it could in principle, taking into consideration any visual aspect of the image, include the description of the objects and their properties, as well as the way in which people and objects of the image are interacting. Nevertheless, image captioning is a complex task since it requires not only a complete understanding of the image, but also a sophisticated generation of natural language.

A brief look at an image is enough for a human being to point out and describe an important amount of details about the visual scene. Our visual system can't recognize a lot of gray shares compared to hundreds of thousands of different color shades and intensities. The images considered in this work are color images, and this is due to the immense deal of information that can be found in color images.

In an image captioning system, we have as an input an RGB image I and we are required to generate a sequence of words $= (s_1, s_2, ..., s_N)$. The possible words $s_i \in V$ at time-step i are subsumed in a discrete set V of options. The number of possible options $|V|$ easily reaches several thousands. There are special tokens in the set of option V that mark any word that is not in the set, the start of a sequence and its end. In practice, those tokens are used to identify whether a word exists in the set of options V or it is either the start or the end of a sequence.

Given a training set $D = \{(I, S^*)\}$, which contains pairs (I, S^*) of image input I and corresponding ground-truth caption $S^* = (s_1^*, s_2^*, ..., s_N^*)$, consisting of words $s_i^* \in V, I \in \{1, 2, .., N\}$, we maximize, with respect to parameters W, a probability model $P_W(s_1, s_2, ..., s_N|I)$.

Collecting captions for images or videos [26] is an expensive process. This is not unique to image caption, however, it's much easier to create a corpus, compared to other tasks in NLP, that suffers from lack of resources [25].

The main contribution of this paper is the focus on solving the lack of resource for Arabic image captioning. Our contributions are as follows:

- We proposed a novel active learning framework with human in the loop for image captioning corpus creation, using a translated version of existing datasets.
  Human annotators help refine the translation of the automatic translation model and identify the correct one. As annotators label quality of the translations, our system ranks the rest of the translated sentences and propose new instances that have the highest probability of being correct for human verification. The idea behind this is to reduce the time that would be spent to find the correct translation in the translated version.
- We proposed a new dataset of Arabic image captions named ArabicFlickr1k. This dataset contains 1095 images, every image is associated with at least three captions.
- We introduced a deep learning model based on Encoder-Decoder architecture for Arabic image captioning.

The remainder of the paper is organized in sections as follows: Sect. 2 aims at presenting a detailed review of existing datasets and approaches for automatic image captioning in the literature. Section 3 is about providing a description of each component of the proposed active learning framework for image captioning corpus creation. In Sect. 4, we are describing end to end architecture for Arabic image captioning. Last but not least, the experimental evaluation and the results are provided. Finally, in Sect. 6 conclusion and future research directions are presented.

2 Literature Review

The availability of datasets, containing images mapped to their descriptions, has contributed to the advance of image captioning research. Image captioning model benefits from the quality and the size of this datasets. Serious progress has been made in the English language. However, other languages are behind, given the scarcity of image captions corpora [2]. The following datasets are the most commonly used of the literature.

2.1 Datsets

Flickr8k [15] is a collection of 8092 images taken from the Flickr website and made public by the University of Illinois. The images contain no person or famous place, so that the entire image can be described according to all the different objects of the image. Each image contains five different captions for reference with an average length of 11.8 words written by humans. They used Amazon Mechanical Turk crowdsourcing service to collect this descriptions. They asked people on the platform to describe the objects, scenes and activities in the images without providing them any information about the images. Only With the information that can be found in the image they were able to collect conceptual descriptions that describe the images.

Flickr30k [42] is an extension of Flickr8k. It contains 31,783 images of people involved in everyday activities and events from the Flickr website. Each image is associated with five descriptions in English, which were collected from Amazon Mechanical Turk. These descriptions are required to accurately describe the objects, scenes, and activities displayed in the image. The dataset contains 158,915 descriptions. Usually, 1000 images are selected as validation data, 1000 images as test data, and the remaining images are used as train data.

Microsoft COCO [23] is a large scale dataset, containing 123,287 images. Most images contain multiple objects and meaningful contextual information we encounter in everyday scenes, and each image is accompanied by five English descriptions annotated by humans. Microsoft COCO is widely used for various computer vision tasks.

STAIR Captions [40] is a Japanese version of Microsoft COCO, it consists of 820,310 Japanese captions for 164,062 images. The authors proposed a model combining the English and Japanese captions. The resulting bilingual model has better performance when compared with the monolingual model that uses only the Japanese caption corpus.

Multi30K [9] is the German version of Flickr30K. The authors extended the Flick30k dataset by collecting five descriptions in German for the 31014 images. They used the Crowdflower platform to hire 185 people for 31 days to describe each image. They collected five independent descriptions for each image. They also translated 31,000 descriptions (about 6200 images) of the English version, translated by professional translators without seeing the images.

Recently, a growing number of research focused on the task of associating images and sentences from both the computer vision and the NLP researchers. In The literature, there are two traditional well-studied directions. The first approaches are known as the language model-based approaches or generation based approaches. They start by converting images into words describing a fixed number of scenes, objects, their attributes and their spatial relations. After that, they formulate new coherent sentences from those words. The second approaches are known as retrieving based approaches. They produce the description by transferring existing description from other images. In the remaining of this section, we will see works done on those approaches and other approaches based on neural networks.

2.2 Generation Based Approaches

Generation based approaches differ in the way they represent images and the technique they use to generate the descriptions.

We mention in this category [22]. Their approach comprises a first step that uses Image Recognition models to extract visual information from the image [11]. They extract a fixed number of objects, including things like birds and cars, they also extract stuff like grass and water. For each object extracted from the image, they also extract their attributes like color. Finally, they extract the special relationships between those objects (near, under). This information is next used for composing sentences to describe the image. The generation step relies on Web-scale N-grams [5]. They did not take actions into consideration in the extraction step.

Another work that considers actions, by relying on an external corpus to predict the relationships between objects, is [39]. In which they fill in a sentence template by predicting the likely objects, verbs, scenes and prepositions that make up the core sentence structure. This is based on a Hidden Markov Model (HHM). They started by detecting objects and scenes from the image using the state of art image recognition models at that time. After this step, they used a language model trained on the English Gigaword corpus to predict the verbs given the objects detected in the image. Using the predicted actions, they estimated the probability that a preposition co-locates with a scene using an existing data. They used a HMM model to find the likely sentence structure given the predicted objects, verbs, scenes and propositions. The last step is the generation, using the results from the previous steps they fill in a fixed sentence template. They limited the number of objects, Verbs, Scenes and Prepositions to cover only what is commonly encountered in images. In addition, the sentence generated for each image is limited to at most two objects occurring in a unique scene.

In this context too, [20] generate descriptions from using pre-trained object detectors and a fixed template based method for description generation. Their system use object recognition models to detect objects (bird, car person, grass, trees). For each detected object they pass it to an attribute classifier and store the detected attributes. Same for every object and region, they predict prepositional relationships. They combine the output of the above in a CRF to produce input for language generation methods and generate the description using a fixed template.

[28] used computer vision models to predict the bounding boxes of objects in the image. For each detected object in the image, they extract attributes such as shape and texture. They also associate detected actions from the image to objects. Finally, they use preposition functions to predict a set of spatial relations that is held between each pair of objects based on their bounding box. A step before the description generation filters detected attributes that are unlikely and place objects into an ordered syntactic structure. Finally, they generate a large set of syntactically well-formed sentence fragments and then recombine these using a tree-substitution grammar.

2.3 Retrieval Based Approaches

Retrieval based approaches in the literature can be branded into two main categories. The first one uses a visual space to extract similar image for a given query image. The other category combines textual and visual information in one space.

[21] work falls into this category. For a given test image, their system retrieves visually similar images from the training data. From those images, they extract segments of their corresponding descriptions that are potentially useful. Then they selectively use those text segments to produce a new description. In order to compose the description, they proposed a new stochastic composing algorithm. A downside of their system is that the produced sentences rely on how correctly the retrieved text segments can describe the given image.

A collection of one million images associated with visually relevant descriptions was introduced in [30]. These descriptions are written by people on the Flicker website. The authors also proposed two methods for automatic descriptions generation. The first method uses two global image descriptors to retrieve similar images. The second method integrates global descriptors and specific content estimators. The specific content estimators extract objects, actions, staff, attributes and scenes from the image. Relaying on the large parallel corpus that they collected, they used both methods to produce relevant image descriptions. Since the descriptions associated with images were written by humans, this corpus enabled the proposed methods to yield descriptions that have a high linguistic quality.

The second category of retrieval-based approaches produces a co-embedding of images and descriptions in the same space. Among the works that have opted for this approach, we find [13]. Where they proposed Stacked Auxiliary Embedding (SAE); an approach based on weakly annotated images data. They were able to improve the performance of description retrieval using SAE, to transfer knowledge from a large-scale data of weakly annotated images. Even with large amount of dataset, retrieval-based approaches do not have the ability to generate new description for unseen image with new combination of objects.

2.4 Retrieval Based Approaches

Recently research in image automatic description has been limited by the existing techniques in image recognition systems and their efficacy. However, this systems begins to improve with the advances of neural network approaches [19].

[18] is the first to use only neural networks for automatic image description in the same period as [35] where they proposed a representation that map images and sentences into the same space using recursive neural networks. After that, they can map a given image into this space, rank all the sentences and chose the first one as description. However, unlike Socher, Kiros proposed a multi-layer perceptron (MLP) that uses a group of word representation vectors biased by features from the image. This means the image features condition the language model output. The image features are extracted from a convolutional neural network.

Advances in machine translation and computer vision enabled [37] to produce a new model based on deep neural network for image description. Their model consists of a convolutional neural network that represents the image in a context vector which then is passed to a language model based on LSTM. The joined model takes an image as input and is trained to maximize the probability of a description associated with a the given image. The model is fully trainable using Stochastic gradient descent and has the state of the art performance at that time on MS COCO. Similar work at the same period has been done by [17] where they used VGGNet [34] to represent the images and obtained the state of the art performance on Flickr8K, Flickr30K and MSCOCO.

Karpathy and Vinyals models provide the image features vector as an input only at the first step of the generation. In [8], the proposed model uses the image features vector at each time step. They represent the images using a pretrained CNN model (VGGNet [34] and CaffeNet [16]) on ImageNet. They explored three variants of their image description architecture, and evaluated the effect of depth in the LSTM language model. Their work also covered video description and Activity Recognition.

Work cited above that uses neural networks do not pay attention to a particular area or objects in the image when generating the description. The concept of attention was first introduced by [38] for image description. They proposed two variations of spatial attention and demonstrate that their models are able to focus on specific region in the image while generating the description. This can be used to gain insight on how their models work.

The success of spatial attention proposed by [38] was followed by the semantic attention in [41]. Spatial attention enables the generation component to focus on relevant places and regions in the image to compose more accurate image description. While, the semantic attention helps the generation step to incorporate semantically relevant concepts like actions and objects detected from the image. First, they map visual concepts (regions, objects, attributes, etc.) detected in the image to words. After that, they use a pretrained convolutional neural network to extract visual features. Then, the model learns through the semantic attention to selectively fuse this words with the visual features into the hidden states and outputs of recurrent neural networks that generates the image description. This words are represented with word embeddings, witch means that they can use external resource, not only for the image representation, but also for text representation.

While [41] extracted visual concepts and used them to help the visual attention, another text based type of attention was proposed by [29], in which they used the description associated with an image to guide the visual attention. During the training; they guide the visual attention with the description to help the model focus only on relevant visual objects in the image. This description is retrieved from visually similar images in the training dataset. They showed that this approach yield better performance on MS-coco at that time.

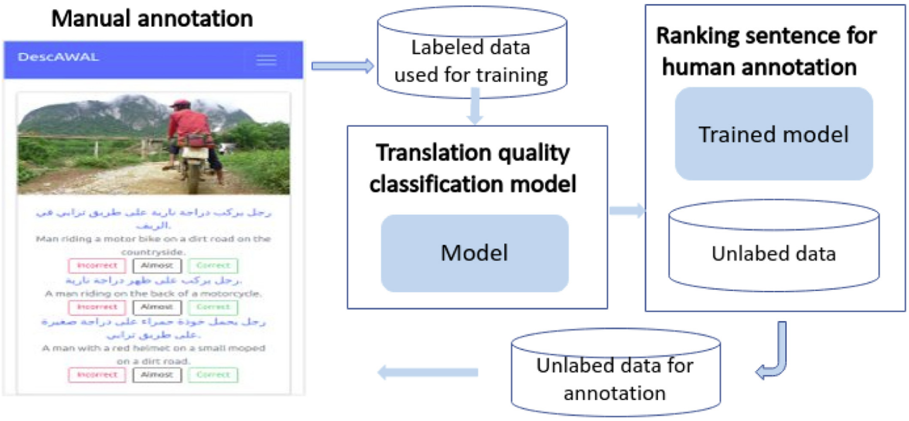
Neural networks based approaches before that [24] typically provide the language model with the image features at every step of the description generation. The authors argue that the language model does not need visual features to generate every word in the description. They introduced an adaptive attention encoder-decoder model. This model can automatically choose to ignore visual features when generating the next word of the description and to only use the language model. The adaptive attention decides when the language model should look at the image and also where it should look. This is done by a new extension of LSTM that relies on a new spacial attention that they introduced. They reported a significant performance over previous methods on MS COCO and Flickr30.

3 Corpus Creation Framework

There is a wide variety of resources on the internet like Facebook, Flickr and other websites from which we can collect images with captions. The only problem with those captions is that they do not describe the image specifically, but rather they give information about what cannot be seen in the image. In [15], the authors suggested that the description should focus on conceptual information that refer to objects, attributes, events and other literal content of the image.

While the task of automatic image description is not bound to the English language, yet, the recent advances have been mostly focusing on English descriptions. It is clear that the creation of resources like [9] costs tens of thousands of dollars and is a time-consuming task. However, The creation of new resources for Arabic image captioning will have a great impact on future research.

In the following subsections we explain in details the different components of our active learning based framework for image captioning corpus creation.


Manual annotation
DescAWAL
Incorrect
Almost
Correct
Labeled data used for training
Translation quality classification model
Model
Ranking sentence for human annotation
Trained model
Unlabed data
Unlabed data for annotation

رجل يركب دراجة نارية على طريق ترابي في الريف.

Man riding a motor bike on a dirt road on the countryside.

رجل يركب على ظهر دراجة نارية.

A man riding on the back of a motorcycle.

رجل يحمل خوذة حمراء على دراجة صغيرة على طريق ترابي.

A man with a red helmet on a small moped on a dirt road.

Fig. 1. Active learning based framework for image captioning corpus creation.

The first component is a manual annotation tool created with Django and Vue js to speed the process for annotators. This tool is used by Arabic native speakers to classify the quality of the translation presented for them. We use a translation quality classification model to rank and decide what images be passed to the manual verification.

3.1 Manual Annotation

This step consists of an interface that present an image with the associated descriptions to the annotators. Every description comes with an Arabic translation. The human annotators are instructed to verify the quality of the translation on a scale of incorrect, almost correct and incorrect. Almost correct is for caption that needs one or two word refinements. The next component is responsible for choosing which image to be present for annotation to minimize the human effort of finding good translations.

Initially, we picked a random set of descriptions and presented them to human annotators. We used their annotations as initial training data for the next component of the active learning framework.

3.2 Translation Quality Classification Model

The translation quality classification model is an essential component. We use a text classification [43] model. After training, we use this model to classify a small set of the unlabeled data and rank them by the model confidence. We then choose the first batch and pass it to human annotators. The idea is to get more correct translations in a batch compared with random selection. The model is a combination of different layers. The details of each layer are presented next.

Embedding Layer. Word embedding is a technique for representing words by fixed-size vectors, so that words that have a similar meaning also have close vectors (that is to say, vectors whose Euclidean distance is small). In other words, the representation implies the semantic meaning of words [3]. In the Embedding layer, each word is mapped to a dense vector of dimension d.

These vectors are initialized based on the word embeddings model that was proposed in [27]. With a simple and efficient neural network structure, their model made it possible to train on a huge amounts of textual data in a short period of time. The authors introduced two models called Continuous Bag of Words (CBOW) and Skip-Gram. The Skip-Gram architecture tries to predict the context of a given word. Both of these models have become very popular in recent years, showing several improvements in the field of NLP.

Bidirectional Gated Recurrent Units Layer. The units of this layer are composed of the GRU architecture proposed in [6]. GRUs are a more recent variation of LSTM networks. GRU first calculates an update gate based on the current input vector and the hidden state.

$$z^{(t)} = \sigma(W_z x^{(t)} + R_z h^{(t-1)} + b_z) \tag{1}$$

Then, it calculate the reset gate in a similar way but with different weights using a new memory content. If the reset gate is 0, then it skips the previous memory and stores only the new information. The final memory at the time step combines the current and previous time steps.

$$\bar{h}^{(t)} = \sigma(W_h x^{(t)} + r^t \odot R_h h^{(t-1)} + b_h) \tag{2}$$

$$h^{(t)} = z^{(t)} \odot h^{(t-1)} + (1 - z^{(t)}) \odot \bar{h}^{(t)} \tag{3}$$

Bidirectional GRUs process data in both directions with forward and backward hidden layers. Compared with the unidirectional, the number of parameters doubles. Bidirectional GRU returns a vector for each direction. The average of the outputs is taken, giving a vector with the dimension equals to the number of GRU units in the layer. It has been shown that GRUs work better than regular LSTMs and are faster thanks to a simpler architecture [7].

Attention Layer. Taking the representation sequence h, outputted by the BiGRU layer as input, the attention layer produces a new representation vector c with the dimension equal to time steps. This attention is proposed in [32].

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{4}$$

Where

$$e_t = a\left(h_t\right), \alpha_t = \frac{\exp\left(e_t\right)}{\sum_{k=1}^{T} \exp\left(e_k\right)}, \tag{5}$$

$$a\left(h_t\right) = \tanh\left(Wh_t + b\right) \tag{6}$$

W, b are learned with the model.

Output Layer. This layer takes as an input the output of the attention layer. The input is fed to a feed forward neural network, with output going through the Softmax function to give the predictions.

3.3 Ranking Sentence for Human Annotation

Initially, we choose random batch of descriptions for annotation, this is done in the first few batches. We use those descriptions as a training data for the translation quality classification model.

After the model is trained, we predict the classes of a random set of descriptions. The model outputs the probability of every class (Correct, Incorrect, and mostly correct) for a given translation instance. We then sort the images based on the number of correct translations and the degree of confidence given by the model. After that, we chose set from the top and send them to human for annotation. After this step, we retrain the model again and repeat.

4 Arabic Image Captioning Model

The encoder-decoder architecture was introduced for the first time in [36], Since then, it has become the standard Neural Machine Translation (NMT) approach. This architecture especially if given large amount of data, outperforms classical Machine translation (MT) methods [4].

Our model was inspired from this architecture. In image captioning, the core idea is to utilize a Convolutional Neural Network (CNN) as an encoder to extract visual features and a Recurrent Neural Networks (RNN) as a decoder to generate the caption.

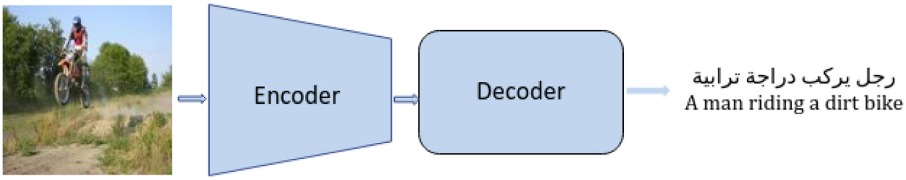


Fig. 2. Arabic image captioning system based on encoder-decoder architecture.

4.1 Encoder

For many years, training deep Neural network was difficult because of a problem known as vanishing gradient. The gradient of the loss function shrinks to zero when the chain rule is applied several times. This prevented the network weights from getting updated so the learning is not performed. ResNets [14] solve this problem where the gradient flow backwards using the skip connections.

To extract hierarchical visual information from the image, we used a pre-trained Resnet101 on Imagenet. In our model Fig. 2, the encoder is first applied to extract both global and regional visual information from the input image. Then we pass those features to the encoder to generate a description. The encoder can be fine-tuned during the training phase.

4.2 Decoder

The main idea behind the decoder is that of the conditional language model. A language model calculates the probability of a sentence by the following equation, where x_i is the next word and $x_1, x_2, ..., x_{i-1}$ represent the context.

$$P(X) = \prod_{i=1}^{n} P(x_i | x_1, x_2, ..., x_{i-1}) \tag{7}$$

To model the image caption generation problem, we use a conditional language model with the image I. s_j is the next word in the image description and $I, x_1, x_2, ..., x_{i-1}$ represent the context used to generate s_j.

$$P(S|I) = \prod_{j=1}^{n} P(s_j | I, s_1, s_2, ..., s_{j-1}) \tag{8}$$

We used an embedding layer, followed by an LSTM layer and a feed forward network layer. The model is trained end to end using cross-entropy loss. At each step, the decoder produces a probability distribution over possible next works. The embedding layer is initialized with pre-trained word2vec model on Arabic Wikipedia.

5 Experimental Evaluation

The proposed framework for Arabic images captioning corpus creation is based on the translation of existing dataset. We translated the Flick30K [42] dataset using Google Translation API. To evaluate this framework, we used Flick30K but the same steps are valid for a much bigger dataset like MS coco.

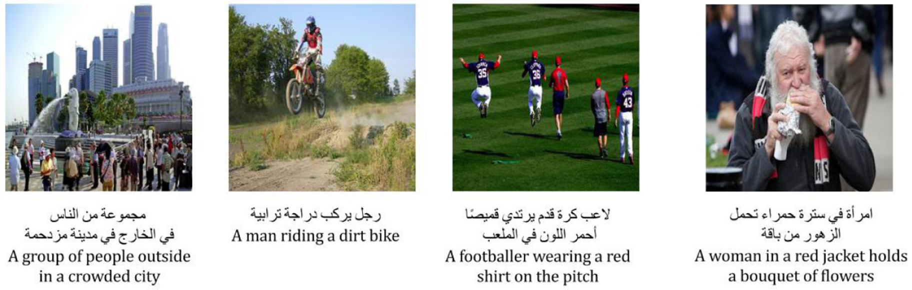
First, the annotators were given a set of 3430 descriptions. They were asked to classify them into three classes. Correct if the translation corresponding to the original English version is correct, almost correct if the translation needs one or two word editing and incorrect otherwise. On average, we found about 6% incorrect translation, 29% almost correct and the rest 65% is correct.

Then we used the result from the above step to train the translation quality classification model to classify the quality based on two classes (correct and incorrect). We could use the three classes, but we focused only on the correct translations. The embedding layer is initialized with word2vec weights trained on Arabic Wikipedia articles using Gensim [33].

We chose a random batch of 2000 images and classified their translated descriptions using the translation quality classification model and passed the top 885 images with the correct translations ranked by the model confidence to the annotators. We found 73% correct descriptions, that is an improvement on the random selection strategy. Finally, we ended up with 1095 images, each image has at least three correct descriptions validated by humans.

The caption model was implemented in Pytorch with the help of Scikit-learn and Tensorflow. All the experiments were done on an Ubuntu system. We used one NVIDIA 1080Ti and 32 GB RAM. We split our data set to 895 images for training, 100 for validation and 100 for the test. We applied some transformation to the images before feeding them to the encoder. All images are scaled to 3 * 224 * 224 and normalized. We prepossessed all captions. We started by tokenizing and then removing words that occur less than two times and then added tokens to mark the start and the end of each caption. In the encoder layer we used an LSTM layer with 512 units. We used 300 for the embedding layer size.

All metrics use for language evaluation output a score indicating a similarity between the candidate sentence and the reference sentences. A popular metric used for automatic image captioning evaluation is BLEU.


مجموعة من الناس
في الخارج في مدينة مزدحمة
A group of people outside
in a crowded city

رجل يركب دراجة ترابية
A man riding a dirt bike

لاعب كرة قدم يرتدي قميصًا
أحمر اللون في الملعب
A footballer wearing a red
shirt on the pitch

امرأة في سترة حمراء تحمل
الزهور من باقة
A woman in a red jacket holds
a bouquet of flowers

Fig. 3. Arabic image descriptions generated using the proposed model with their translation in English.

BLEU (Bilingual evaluation understudy) [31] computes the geometric mean of n-gram precision scores multiplied by a brevity penalty in order to avoid overly short sentences. It is a metric that can be used to measure the quality of machine generated text in tasks like text summarization, Speech recognition and automatic image captioning. This metric was first introduced for machine translation as a reasonable correlation with human judgments of quality.

The caption model is trained end to end with the cross entropy loss. The performance of the proposed model on the test set gave a promising result of 47 for the BLEU-1, 24 for the BLEU-2, 20 for the BLEU-3 and 11 for the BLEU-4.

6 Conclusions

In this paper, we proposed a novel active learning based framework for Arabic image captioning corpus creation. This framework relies on the translations of existing datasets. We also proposed a new corpus for Arabic image captioning (ArabicFlickr1K). We did a detailed review of the literature and the existing resources. We introduced a deep learning model based on the Encoder-Decoder architecture for Arabic image captioning. Our model scored 47% on BLUE-1. Future research directions will go towards leveraging unsupervised data, using more complex language models in the Decoder and more supervised fine-tuning in the training phase.

References

1. Agrawal, A., et al.: VQA: visual question answering. Int. J. Comput. Vision **123**(1), 4–31 (2017)
2. Al-Muzaini, H.A., Al-Yahya, T.N., Benhidour, H.: Automatic Arabic image captioning using RNN-LSTM-based language model and CNN. Database **9**(6) (2018)
3. Ayadi, R., Maraoui, M., Zrigui, M.: LDA and LSI as a dimensionality reduction method in Arabic document classification. In: Dregvaite, G., Damasevicius, R. (eds.) ICIST 2015. CCIS, vol. 538, pp. 491–502. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24770-0_42
4. Bacha, K., Zrigui, M.: Machine translation system on the pair of Arabic/English. In: KEOD, pp. 347–351 (2012)
5. Brants, T.: Web 1t 5-gram version 1. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13 (2006)
6. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
7. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
8. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634 (2015)
9. Elliott, D., Frank, S., Sima'an, K., Specia, L.: Multi30k: multilingual English-German image descriptions. arXiv preprint arXiv:1605.00459 (2016)
10. Farhani, N., Terbeh, N., Zrigui, M.: Image to text conversion: state of the art and extended work. In: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), pp. 937–943. IEEE (2017)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)

12. Filipe, J., Fred, A.L.N. (eds.): ICAART 2013 - Proceedings of the 5th International Conference on Agents and Artificial Intelligence, Barcelona, Spain, 15–18 February 2013, vol. 2. SciTePress (2013)
13. Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., Lazebnik, S.: Improving image-sentence embeddings using large weakly annotated photo collections. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 529–545. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_35
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
15. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: data, models and evaluation metrics. J. Artif. Intell. Res. **47**, 853–899 (2013)
16. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
17. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
18. Kiros, R., Salakhutdinov, R., Zemel, R.: Multimodal neural language models. In: International Conference on Machine Learning, pp. 595–603 (2014)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
20. Kulkarni, G., et al.: Baby talk: understanding and generating simple image descriptions. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, pp. 1601–1608 (2011)
21. Kuznetsova, P., Ordonez, V., Berg, T., Choi, Y.: TREETALK: composition and compression of trees for image descriptions. Trans. Assoc. Comput. Linguist. **2**(1), 351–362 (2014)
22. Li, S., Kulkarni, G., Berg, T.L., Berg, A.C., Choi, Y.: Composing simple image descriptions using web-scale N-grams. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp. 220–228. Association for Computational Linguistics (2011)
23. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
24. Lu, J., Xiong, C., Parikh, D., Socher, R.: 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, p. 2 (2017)
25. Mahmoud, A., Zrigui, M.: Artificial method for building monolingual plagiarized Arabic corpus. Computación Sistemas **22**(3), 767–776 (2018)
26. Mansouri, S., Charhad, M., Zrigui, M.: A heuristic approach to detect and localize text in Arabic news video. Computación Sistemas **22**(1), 75–82 (2018)
27. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
28. Mitchell, M., et al.: Midge: generating image descriptions from computer vision detections. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 747–756. Association for Computational Linguistics (2012)

29. Mun, J., Cho, M., Han, B.: Text-guided attention model for image captioning. In: AAAI, pp. 4233–4239 (2017)
30. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2Text: describing images using 1 million captioned photographs. In: Advances in Neural Information Processing Systems, pp. 1143–1151 (2011)
31. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
32. Raffel, C., Ellis, D.P.: Feed-forward networks with attention can solve some long-term memory problems. arXiv preprint arXiv:1512.08756 (2015)
33. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA, Valletta, Malta, May 2010. http://is.muni.cz/publication/884893/en
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
35. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. Trans. Assoc. Comput. Linguis. **2**(1), 207–218 (2014)
36. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)
37. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
38. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
39. Yang, Y., Teo, C.L., Daumé III, H., Aloimonos, Y.: Corpus-guided sentence generation of natural images. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 444–454. Association for Computational Linguistics (2011)
40. Yoshikawa, Y., Shigeto, Y., Takeuchi, A.: Stair captions: constructing a large-scale Japanese image caption dataset. arXiv preprint arXiv:1705.00823 (2017)
41. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4651–4659 (2016)
42. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguist. **2**, 67–78 (2014)
43. Zrigui, M., Ayadi, R., Mars, M., Maraoui, M.: Arabic text classification framework based on latent dirichlet allocation. J. Comput. Inf. Technol. **20**(2), 125–140 (2012)
44. Zrigui, M., Charhad, M., Zouaghi, A.: A framework of indexation and document video retrieval based on the conceptual graphs. J. Comput. Inf. Technol. **18**(3), 245–256 (2010)