A Deep Learning Approach for Arabic Caption Generation Using Roots-Words

Vasu Jindal
University of Texas at Dallas, Texas, USA
vasu.jindal@utdallas.edu

Abstract

Automatic caption generation is a key research field in the machine learning community. However, most of the current research is performed on English caption generation ignoring other languages like Arabic and Persian. In this paper, we propose a novel technique leveraging the heavy influence of root words in Arabic to automatically generate captions in Arabic. Fragments of the images are associated with root words and deep belief network pre-trained using Restricted Boltzmann Machines are used to extract words associated with image. Finally, dependency tree relations are used to generate sentence-captions by using the dependency on root words. Our approach is robust and attains BLEU-1 score of 34.8.

With the increase in number of devices with cameras, there is a widespread interest in generating automatic captions from images and videos. Generating image description have huge impact in the fields of information retrieval, accessibility for the vision impaired, categorization of images etc. Additionally, generation automatic descriptions of images can be used as a frame by frame approach to describe videos and explain their context. Automatic generation of image descriptions is a widely researched problem. However, most visual recognition models and approaches in this fields are focused on Western languages, ignoring Semitic and Middle-Eastern languages like Arabic, Hebrew, Urdu and Persian. As discussed further in related works, almost all major caption generation models have validated their approaches using English. This is primarily due to the significant dialects between different forms of Arabic and the challenges in translating images to natural sounding sentences.

Arabic is ranked as the fifth most native language among the population. Furthermore, Arabic has tremendous impact on the social and political aspects in the current community and is listed as one of the six official languages of the United Nations. Given the high influence of Arabic, it is necessary for a robust approach to generate captions of images in Arabic.

In this paper, we propose a three-stage root word based for generation of captions in Arabic for images. Briefly, we first create fragments of images using a previously trained deep neural network on ImageNet. However, unlike other published approaches for English caption generation (Socher et al. 2014), (Karpathy and Fei-Fei 2015), (Vinyals et al. 2015), we map these fragments to a set of root words in Arabic rather than actual words or sentences in English. We used deep belief networks pre-trained by Restricted Boltzmann Machines to select different root words associated with the image fragments and extract the most appropriate words for the image. A rank based approach is used to select the best image-sentence pairing from other false image sentences pairings. Finally, we use dependency tree relations to create sentence captions from the obtained words. Our main contribution in this paper is three-fold:

- Mapping of image fragments onto root words in Arabic rather than actual sentences or words/fragments of sentences as suggested in previously proposed approaches.
- Finding most appropriate words for an image by choosing set of vowels required to be added to the root words using Deep Learning
- Using dependency tree relations of these obtained words to finally form sentences in Arabic

To the best of our knowledge, this is the first work that leverage root words dependency relation to generate captions in Arabic. We rely on previously published approaches specifically ImageNet and Caffe on object detection to extract features from the images. For the purpose of clarity, we use the term root-words throughout this paper to represent the roots of an Arabic word.

Arabic Morphology

Arabic belongs to the family of Semitic languages and has significant morphological, syntactical and semantical differences from other languages. The base words of Arabic inflect to express eight main features. Verbs inflect for aspect, mood, person and voice. Nouns and adjectives inflect for case and state. Arabic morphology consists from a bare root verb form that is trilateral, quadrilateral, or pentaliteral. The derivational morphology can be lexeme = Root + Pattern or inflection morphology (word = Lexeme + Features) where features are noun specific, verb specific or single letter conjunctions. We leverage this critical aspect of Arabic morphology to generate the captions of image. For example, the root meaning write has the form k-t-b. More words can be formed using vowels and additional consonants in the root

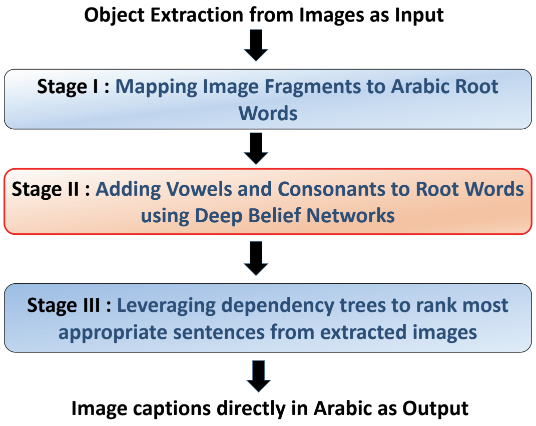
Object Extraction from Images as Input
Stage I : Mapping Image Fragments to Arabic Root Words
Stage II : Adding Vowels and Consonants to Root Words using Deep Belief Networks
Stage III : Leveraging dependency trees to rank most appropriate sentences from extracted images
Image captions directly in Arabic as Output

Figure 1: General view of our proposed method

word. Some of the words that can be formed using k-t-b are kitab meaning ”book”, kutub meaning ”books”, katib meaning ”writer”, yaktubu meaning ”he writes” etc.

Methodology

Figure 1 gives an overview of our approach and Figure 2 describes our deep learning framework. In the first stage, we extract different objects from images using Caffe and ImageNet. For example, in Figure 3, the two fragments extracted are boxed in red. We incorporate Restricted Boltzmann Machines (RBM) to pre-train the Deep Belief Network (DBN)and find good initial weights for training. Contrastive Divergence-1 (CD-1) algorithm is used to obtain samples after 1-step Gibbs sampling. Subsequently, the pre-trained DBN network is fine-tuned by vanilla back-propagation with labeled segments as the input layer. We map every root word using 1-of-k encoding vector w using a dictionary of 100,00 words and map every dependency triplet (R, w_1, w_2) into the embedding space as follows where r_1, r_2 and r_3 are the three most similar root words. W_1, W_2 and W_3 are weights learnt from the deep belief network.

$$s = f(W_n \begin{bmatrix} W_{r_1} W_1 \\ W_{r_2} W_2 \\ W_{r_3} W_3 \end{bmatrix}) \tag{1}$$

Results

We evaluate our technique using two datasets: ImageNet dataset with manually written captions in Arabic by professional Arabic translators (10,000 images) and 100,000 images associated with news from Al-Jazeera news website (http://www.aljazeera.net) which regularly publishes articles in both Arabic and English. We use 80,000 images for training and perform our testing on the other 30,000 images. We also compare our approach results with directly generating English captions and translating them to Arabic using Google translate. Table I shows the BLEU-1 scores of our experiment. Our technique achieves 10-fold cross validation BLEU-1 score of 34.8. This is the first reported BLEU-1 score for Arabic captions. Our experiments show a promising result considering the first reported BLEU-1 score for Arabic caption generation from images. Furthermore, the results also show that generating captions directly in Arabic attains a much better BLEU-1 score rather than generating captions in English and translating them to Arabic.

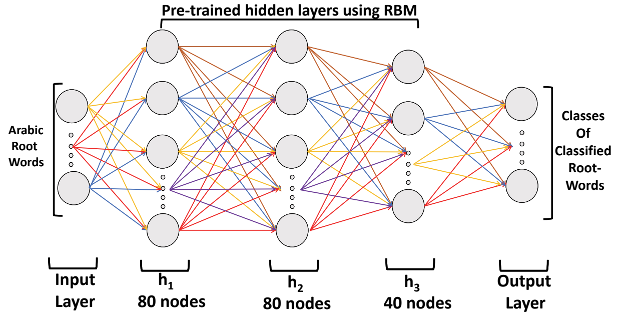

Figure 2: Deep Belief Network to add vowels to root words

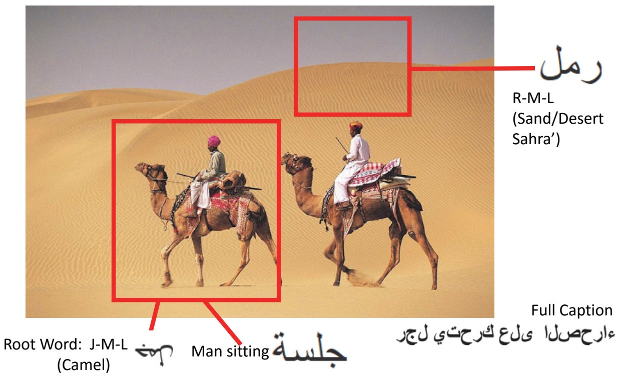

Figure 3: Sample result of sentence generation using our framework *(Man moving on camel in desert)*

Approach	BLEU-1 Score
English	48.4
English-Arabic (Google Translate)	27.2
Our Approach	**34.8**

References

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.

Socher, R.; Karpathy, A.; Le, Q. V.; Manning, C. D.; and Ng, A. Y. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics* 2:207–218.

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*, 3156–3164.