Generating Image Captions in Arabic using Root-Word Based Recurrent Neural Networks and Deep Neural Networks

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

Abstract

Image caption generation has gathered widespread interest in the artificial intelligence community. Automatic generation of an image description requires both computer vision and natural language processing techniques. While, there has been advanced research in English caption generation, research on generating Arabic descriptions of an image is extremely limited. Semitic languages like Arabic are heavily influenced by root-words. We leverage this critical dependency of Arabic to generate captions of an image directly in Arabic using root-word based Recurrent Neural Network and Deep Neural Networks. Experimental results on datasets from various Middle Eastern newspaper websites allow us to report the first BLEU score for direct Arabic caption generation. We also compare the results of our approach with BLEU score captions generated in English and translated into Arabic. Experimental results confirm that generating image captions using root-words directly in Arabic significantly outperforms the English-Arabic translated captions using state-of-the-art methods.

1 Introduction

With the increase in the number of devices with cameras, there is a widespread interest in generating automatic captions from images and videos. Automatic generation of image descriptions is a widely researched problem. However, this problem is significantly more challenging that the image classification or image recognition tasks which gained popularity with ImageNet recognition challenge (Russakovsky et al., 2015). Automatic generation of image captions have a huge impact in the fields of information retrieval, accessibility for the vision impaired, categorization of images etc. Additionally, the automatic generation of the descriptions of images can be used as a frame by frame approach to describe videos and explain their context.

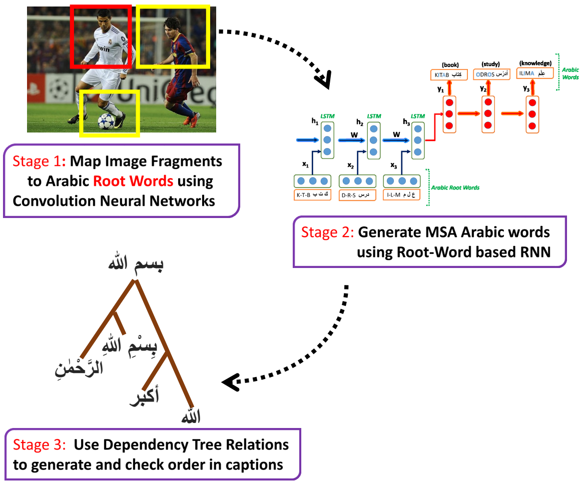
(book)
KITAB كِتاب
(study)
ODROS أدرُس
(knowledge)
ILIMA عِلْمَ
Arabic Words
y1
y2
y3
h1
h2
h3
LSTM
LSTM
LSTM
W
W
x1
x2
x3
K-T-B ك ت ب
D-R-S درس
I-L-M ع ل م
Arabic Root Words
Stage 1: Map Image Fragments to Arabic Root Words using Convolution Neural Networks
Stage 2: Generate MSA Arabic words using Root-Word based RNN
بسم الله
بِسْمِ الله
الرَّحْمٰنِ
أكبر
الله
Stage 3: Use Dependency Tree Relations to generate and check order in captions

Figure 1: Overview of Our Approach

Recent works which utilize large image datasets and deep neural networks have obtained strong results in the field of image recognition (Krizhevsky et al., 2012; Russakovsky et al., 2015). To generate more natural descriptive sentences in English, (Karpathy and Fei-Fei, 2015) introduced a model that generates natural language descriptions of image regions based on weak labels in form of a dataset of images and sentences.

However, most visual recognition models and approaches in the image caption generation community 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 two major reasons: i) lack of existing image corpora in languages other than English ii) the significant di-

Proceedings of NAACL-HLT 2018: Student Research Workshop, pages 144–151
New Orleans, Louisiana, June 2 - 4, 2018. ©2017 Association for Computational Linguistics

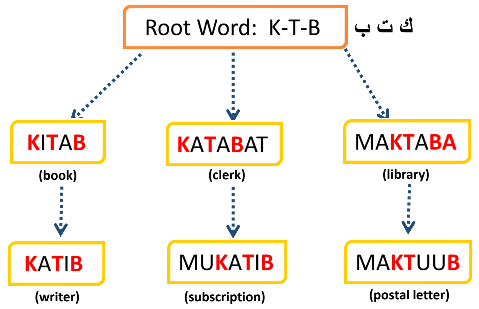

Figure 2: Leveraging Root-Word in Arabic, Constants and Vowels are filled using Recurrent Neural Network

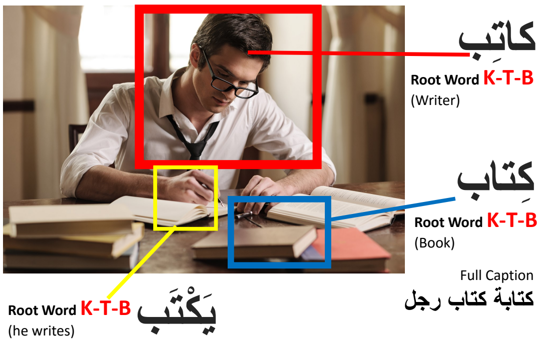

Figure 3: State-of-the-Art: Man studying with books *Ours (translated from Arabic for reader's readability): Writer writing on notebook*

alects of Arabic and the challenges in translating images to natural sounding sentences. Translation of English generated captions to Arabic captions may not always be efficient due to the various Arabic morphologies, dialects and phonologies which results in losing the descriptive nature of the generated captions. A cross-lingual image caption generation approach in Japanese concluded that a bilingual comparable corpus has better performance than a monolingual corpus in image caption generation (Miyazaki and Shimizu, 2016).

Arabic is ranked as the fifth most widely spoken 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 for Arabic caption generation.

1.1 Novel Contributions

Semitic languages like Arabic are significantly influenced by their original root-word. Figure 2 explains how simple root-words can form new words with similar context. We leverage this critical aspect of Arabic to formalize a three stage approach integrating root-word based Deep Neural Networks and root-word based Recurrent Neural Network and dependency relations between these root words to generate Arabic captions. Fragments of images are extracted using pre-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. *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 the most appropriate words for an image by using a root-word based Recurrent Neural Network.
- Finally, using dependency tree relations of these obtained words to check order in sentences in Arabic

To the best of our knowledge, this is the first work that leverage root words to generate captions in Arabic (Jindal, 2017). We also report the first BLEU scores for Arabic caption generation. *Additionally, this opens a new field of research to use root-words to generate captions from images in Semitic languages.* For the purpose of clarity, we use the term ”root-words” throughout this paper to represent the roots of an Arabic word.

2 Background

2.1 Previous Works

The adoption of deep neural networks (Krizhevsky et al., 2012; Jia et al., 2014; Sharif Razavian et al., 2014) has tremendously improved both image recognition and natural language processing tasks. Furthermore, machine translation using recurrent neural networks have gained attention with sequence-to-sequence training approaches (Cho et al., 2014; Bahdanau et al., 2014; Kalchbrenner and Blunsom, 2013).

Recently, many researchers, have started to combine both a convolutional neural network and a recurrent neural network. Vinyals et al. used a convolutional neural network (CNN) with inception modules for visual recognition and long short-term memory (LSTM) for language modeling (Vinyals et al., 2015).

However, to the best of our knowledge most caption generation approaches were performed on English. Recently, authors in (Miyazaki and Shimizu, 2016) presented results on the first cross-lingual image caption generation on the Japanese language. (Peng and Li, 2016) generated Chinese captions on the Flickr30 dataset. There has been no single work addressing to the generation of captions in Semitic languages like Arabic. Furthermore, all previously proposed approaches map image fragments to actual words/phrases. We rather propose to leverage the significance of root-words in Semitic languages and map image fragments to root-words and use these root-words in a root-word based recurrent neural network.

2.2 Arabic Morphology and Challenges

Arabic belongs to the family of Semitic languages and has significant morphological, syntactical and semantical differences from other languages. It consists of 28 letters and can be extended to 90 by adding shapes, marks, and vowels. Arabic is written from right to left and letters have different styles based on the position in the word. 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. Verbs, nouns and adjectives inflect for both gender and number.

Furthermore, Arabic is widely categorized as a diglossia (Ferguson, 1959). A diglossia refers to a language where the formal usage of speech in written communication is significantly different in grammatical properties from the informal usage in verbal day to day communication. Arabic morphology consists of a bare root verb form that is trilateral, quadrilateral, or pentalateral. 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. In contrast, in most European languages words are formed by concatenating morphemes.

Stem pattern are often difficult to parse in Arabic as they interlock with root consonants (Al Barrag, 2014). Arabic is also influenced by infixes which may be consonants and vowels and can be misinterpreted as root-words. One of the major problem is the use of a consonant, hamza. Hamza is not always pronounced and can be a vowel. This creates a severe orthographic problem as words may have differently positioned hamzas making them different strings yet having similar meaning.

Furthermore, diacritics are critical in Arabic. For example, two words formed from ”zhb” meaning ”to go” and ”gold” differ by just one diacritic. The two words can only be distinguished using diacritics. The word ”go” may appear in a variety of images involving movement while ”gold” is more likely to appear in images containing jewelry.

3 Methodology

Our methodology is divided into three main stages. Figure 1 gives an overview of our approach. In Stage 1, we map image fragments onto root words in Arabic. Then, in Stage 2, we used root word based Recurrent Neural Networks with LSTM memory cell to generate the most appropriate words for an image in Modern Standard Arabic (MSA). Finally, in Stage 3, we use dependency tree relations of these obtained words to check the word order of the RNN formed sentences in Arabic. Each step is described in detail in following subsections.

3.1 Image Fragments to Root-Words using DNN

We extract fragments from images using the state-of-the-art deep neural networks. According to (Kulkarni et al., 2011; Karpathy et al., 2014), objects and their attributes are critical in generating sentence descriptions. Therefore, it is important to efficiently detect as many objects as possible in the image.

We apply the approach given in (Jia et al., 2014; Girshick et al., 2014) to detect objects in every image with a Region Convolutional Neural Network (RCNN). The CNN is pre-trained on ImageNet (Deng et al., 2009) and fine-tuned on the 200 classes of the ImageNet Detection Challenge. We also use the top 19 detected locations as given by Karpathy et al in addition to the whole image and compute the representations based on the pixels inside each bounding box as suggested in (Karpathy and Fei-Fei, 2015). It should be noted that the output of the convolutional neural network are Arabic root-words. To achieve this, at any given time when English labels of objects were

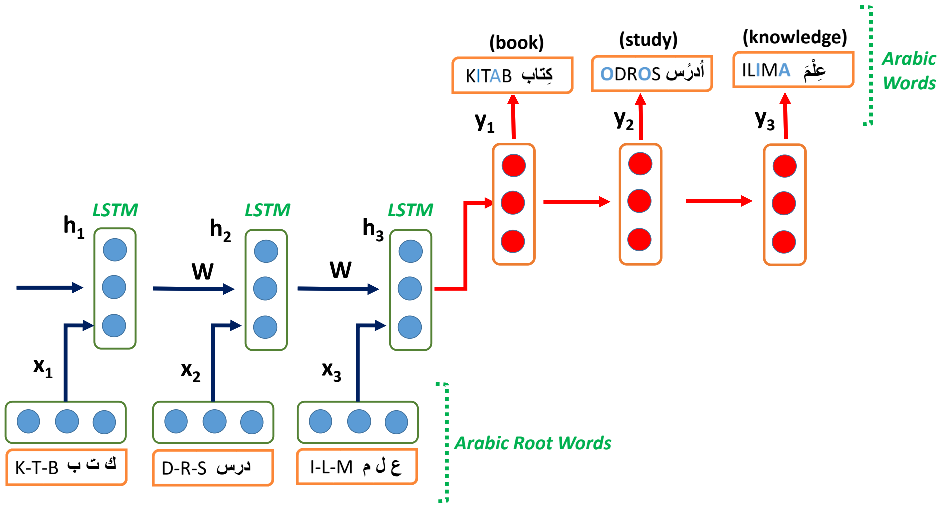

Figure 4: Our Root-Word based Recurrent Neural Network

used in training of the convolution neural network, **Arabic root-words of the object were also given as input** in the training phase. (Yaseen and Hmeidi, 2014; Yousef et al., 2014) proposed the well-known transducer based algorithm for Arabic root extraction which is used to extract root-words from an Arabic word in the training stage. Given the Arabic influence on root-words and the limited 4 verb prefixes, 12 noun prefixes and 20 common suffixes, the approach is optimized for initial training. Briefly, the algorithm has following steps given the morphology of Arabic.
1. Construct all noun/verb transducer
2. Construct all noun/verb patterns transducer
3. Construct all noun/verb suffixes transducer
4. Concatenate noun transducers/verb transducers obtained in steps 1, 2 and 3.
5. Sum the two transducers obtained in step 4.

Similar to (Vinyals et al., 2015), we used a discriminative model to maximize the probability of the correct description given the image. Formally, this can be represented using:

$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1}; \theta) \tag{1}$$

where θ are the parameters of our model, I is an image, and S its correct transcription and N is a particular length of a caption. This $p(S_t|I, S_0, \ldots, S_{t-1}; \theta)$ is modeled using a root-word based Recurrent Neural Network (RNN).

3.2 Root-Word Based Recurrent Neural Network and Dependency Relations

We propose a root-word based recurrent neural network (rwRNN). The model takes different root-words extracted from text, and predicts the most appropriate words for captions in Arabic, essentially also learning the context and environment of the image. The structure of the rwRNN is based on a standard many-to-many recurrent neural network, where current input (x) and previous information is connected through hidden states (h) by applying a certain (e.g. sigmoid) function (g) with linear transformation parameters (W) at each time step (t). Each hidden state is a long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) cell to solve the vanishing gradient issue of vanilla recurrent neural networks and inefficiency in learning long distance dependencies.

While a standard input vector for RNN derives from either a word or a character, the input vector in rwRNN consists of a root-word specified with 3 letters (r_{1n}, r_{2n}, r_{3n}) that correspond to the characters in root-words' position. Most root-words in Arabic are trilateral very few being quadilateral or pentalateral. If a particular root-word is quadilateral (pentalateral) then the r_{2n} represents the middle three (four) letters of the root-word. Formally:

$$x_n = \begin{bmatrix} r_{1n} \\ r_{2n} \\ r_{3n} \end{bmatrix} \tag{2}$$

The final output (i.e. the predicted actual Arabic word y_n), the hidden state vector (h_n) of the

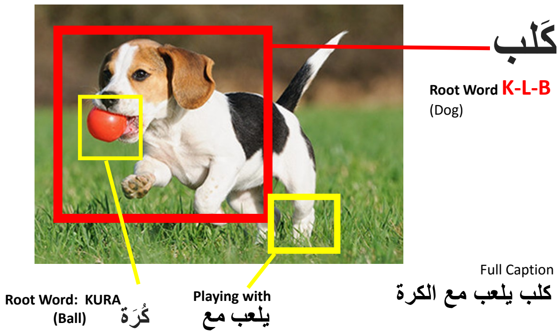

Figure 5: Arabic-English: Dog playing with ball
Ours: Dog plays with a ball

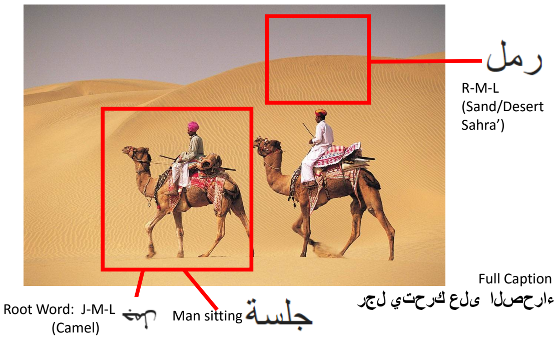

Figure 6: Arabic-English: Man sitting on camel
Ours: Man moving on camel in desert

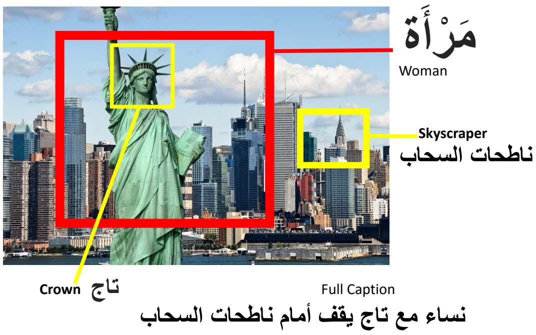

Figure 7: Arabic-English: Woman standing in city
Ours: Woman with crown standing in front of skysrapers

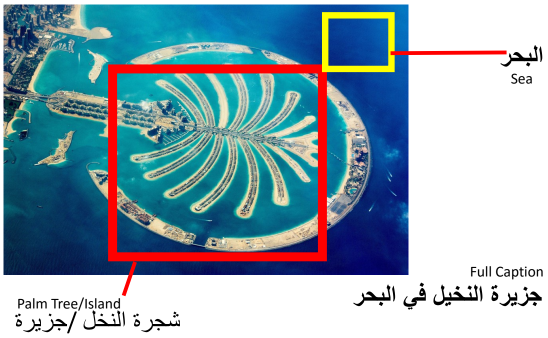

Figure 8: Arabic-English: Island in Sea
Ours: Palm Shaped Islands in the Sea

LSTM is taken as input to the following softmax function layer with a fixed vocabulary size (v).

$$y_n = \frac{\exp\left(W_h \cdot h_n\right)}{\sum_v \exp\left(W_h \cdot h_n\right)} \tag{3}$$

Cross-entropy training criterion is applied to the output layer to make the model learn the weight matrices (W) to maximize the likelihood of the training data. Figure 4 gives an overview of our root-word based Recurrent Neural Network. The dependency tree relations are used to check if the order of Recurrent Neural Network is correct.

Dependency tree constraints (Kuznetsova et al., 2013) checks the caption generated from RNN to be grammatically valid in Modern Standard Arabic and robust to different diacritics in Arabic. The model also ensures that the relations between image-text pair and verbs generated from RNN are still maintained. Formally, the following objective function is maximized:

$$\begin{aligned} &\text{Maximize} \quad F(y;x) = \Phi(y;x,v) + \Psi(y;x) \\ &\text{subject to} \quad \Omega(y;x,v) \end{aligned} \tag{4}$$

where $x = x_i$ is the input caption from RNN (a sentence), v is the accompanying image, $y = y_i$ is the output sentence, $\Phi(y;x,v)$ is the content selection score, $\Phi(y;x)$ is the linguistic fluency score, and $\Omega(y;x,v)$ is the set of hard dependency tree constraints. The most popular Prague Arabic Dependency Treebank (PADT) consisting of multi-level linguistic annotations over Modern Standard Arabic is used for the dependency tree constraints (Hajic et al., 2004).

4 Experimental Results

Figure 3, 5-8 gives a sample of our approach in action. *For the convenience of our readers who are not familiar with Arabic, Figure 5, 6, 7, 8 have the English caption generated using (Xu et al., 2015) denoted as "Arabic-English" and "Ours" denote a professional English translation of the Arabic caption generated from our approach.* We evaluate our technique using two datasets: Flickr8k dataset with manually written captions in Arabic by professional Arabic translators and 405,000 im-

Table 1: BLEU-1,2,3,4/METEOR metrics compared to other methods, (—) indicates an unknown metric

		BLEU				
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	BRNN (Karpathy and Fei-Fei, 2015)	48.2	45.1	29.4	15.5	—
	Google (Vinyals et al., 2015)	52.4	46.1	34.8	18.6	—
	Visual Attention (Xu et al., 2015)	54.2	48.4	36.2	19.4	—
	Ours	**65.8**	**55.9**	**40.4**	**22.3**	**20.09**
Middle Eastern News Websites	MS Research (Fang et al., 2015)	44.1	33.9	26.8	13.7	9.25
	BRNN (Karpathy and Fei-Fei, 2015)	45.4	34.8	27.6	13.9	12.11
	Google (Vinyals et al., 2015)	46.2	36.4	28.5	14.3	15.18
	Visual Attention (Xu et al., 2015)	48.5	38.1	30.9	15.4	16.82
	Ours	**55.6**	**43.3**	**34.5**	**18.9**	**18.01**

ages with captions from various Middle Eastern countries' newspapers. All these newspapers publish articles with images and their captions in both Arabic and English.

We also compare the results of our approach with generating English captions using previously proposed approaches and translating them to Arabic using Google translate. To evaluate the performance, automatic metrics are computed using human generated ground-truth captions. All our images in the dataset were translated using professional Arabic translations as ground-truth. The most commonly used metric to compare generated image captions is BLEU score (Papineni et al., 2002). BLEU is the precision of word n-grams between generated and reference sentences. Additionally, scores like METEOR (Vedantam et al., 2015) which capture perplexity of models for a given transcription have gained widespread attention. Perplexity is the geometric mean of the inverse probability for each predicted word. We report both the BLEU and METEOR score for Arabic captions using root-words. Additionally, this opens a new field of research to use root-words to generate captions from images in Semitic languages and may also be applied to English for words originating from Latin. To the best of our knowledge, our scores are the first reported score for Arabic captions. *Furthermore, the results also show that generating captions directly in Arabic attains a much better BLEU scores compared to generating captions in English and translating them to Arabic.* All results shown in Table 1 are captions generated using the corresponding approaches in English and translating them to Arabic using Google Translate. According to Table 1, we can see that our root-word based approach outperforms all current English based approaches and translated to Arabic using Google Translate. An interesting observation is in Figure 8. While all current approaches fail to describe the actual "Palm Jumeriah Island" which is a man-made island in shape of Palm tree in Dubai, our approach learns the context of "sea", "island" and "palm" and produces the correct result. Most inefficient cases in our algorithm are due to random outliers like some recent words which are not influenced by root-words. This can be further improved by using a larger dataset and using new dialectal captions in the training phase.

5 Conclusion and Future Work

This paper presents a novel three-stage technique for automatic image caption generation using a combination of root-word based recurrent neural network and root-word based deep convolution neural network. This is the first reported BLEU score for Arabic caption generation and the experimental results show a promising performance. We propose to directly generate captions in Arabic as opposed to generating in English and translating to a target language. However, our research proves, using the BLEU metric, that generating captions directly in Arabic has much better results rather than generating captions in English and translating them to Arabic. Our technique is robust against different diacritics, many dialects and complex morphology of Arabic. Furthermore, this procedure can be extended other Semitic languages like Hebrew which intensively depend on root-words. Future work includes exploring other Arabic morphologies like lemmas used in Arabic dependency parsing (Marton et al., 2010; Haralambous et al., 2014). We also plan to apply this approach to other Semitic languages and release appropriate datasets for the new Semitic languages.

References

Thamir Al Barrag. 2014. Noun phrases in urban hijazi arabic: A distributed morphology approach.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, 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.

Charles A Ferguson. 1959. Diglossia. *word*, 15(2):325–340.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Jan Hajic, Otakar Smrz, Petr Zemánek, Jan Šnaidauf, and Emanuel Beška. 2004. Prague arabic dependency treebank: Development in data and tools. In *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, pages 110–117.

Yannis Haralambous, Yassir Elidrissi, and Philippe Lenca. 2014. Arabic language text classification using dependency syntax-based feature selection. *arXiv preprint arXiv:1410.4863*.

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

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM.

Vasu Jindal. 2017. A deep learning approach for arabic caption generation using roots-words. In *AAAI*, pages 4941–4942.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP*, 39, page 413.

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

Andrej Karpathy, Armand Joulin, and Fei Fei F Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*.

Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. 2013. Generalizing image captions for image-text parallel corpus. In *ACL (2)*, pages 790–796.

Yuval Marton, Nizar Habash, and Owen Rambow. 2010. Improving arabic dependency parsing with lexical and inflectional morphological features. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 13–21. Association for Computational Linguistics.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1780–1790.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Hao Peng and Nianhen Li. 2016. Generating chinese captions for flickr30k images.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81.

Qussai Yaseen and Ismail Hmeidi. 2014. Extracting the roots of arabic words without removing affixes. *Journal of Information Science*, 40(3):376–385.

Nidal Yousef, Aymen Abu-Errub, Ashraf Odeh, and Hayel Khafajeh. 2014. An improved arabic word's roots extraction method using n-gram technique. *Journal of Computer Science*, 10(4):716.