RESEARCH ARTICLE OPEN ACCESS

Development of an Arabic Image Description System

Rasha Mualla [1], Jafar Alkheir [2]
Department of Computer and Control Engineering,
University of Tishreen
Latakia - Syria

ABSTRACT
Image description models are one of the most important and trend topics in the machine-learning field. Recently, many researches develop systems for image classification and description in many languages especially English. Arabic language had not taken any attention in this field. In this research, two image description models will be introduced; the first one is English-based model while the other is the Arabic-based model. In our study, CNN deep learning networks are used for image feature extraction. In the training stage, LSTM networks are chosen for their ability of memorizing previous words in image description sentences. LSTM are fed with two inputs; the first one is the image features and the other is the image description file. A new JSON image description file for Arabic description model is built, and the research uses a subset of flickr8k dataset consisting of 1500 training images, 250 validation images and 250 test ones. Performance evaluation is computed via Bleu-n and many other metrics for comparison between Arabic and English models.
Keywords :— Machine Learning, Deep Learning, Image Description, CNN, LSTM, Arabic Description, JSON.

I. INTRODUCTION

Due to the increment of layers needed in traditional networks, some real problems are emerging. One of these problems is the additional training time occurring in case of using gradient descent training algorithms like backpropagation. In such algorithms, the gradient decreases significantly with the increasing of network layers (depth), and this causes to slow down the training process [1],[2],[3],[4]. Another problem is the over-fitting in which the training process speeds and terminates early, so the network responds well for the training samples but worse for the test ones. This problem is due to the increment of neurons in each layer [33],[40],[46],[47].
In addition to all those problems, the redundant neurons and layers results in additional weights and biases which should be learned. The full connectivity of neurons in each layer with the neurons with other layers increases the complexity of these networks.
Those problems are solved by using deep learning via different situations like minimizing the interconnections between neurons form layer to another. Another way to reduce the training time is using fixed weights and biases through the network. Due to increasing the number of layers, the deep neural networks use a pooling layer which reduces the dimension of data. This will reduce the training time and extract the most important information from input data [32],[33],[34],[40].
Another deep learning characteristic is the recursive ability. It makes the networks memorize some of previous data which will be benefit for other layers. This memory makes the network useful for application in which the output depends on some input data (i.e. language producing and translation) [18].

At the last ten years, many studies have achieved which use different types of deep learning neural networks in field of image description such as convolutional neural networks (CNN) for image classification, recurrent neural networks (RNN) for image description and Long-short term memory (LSTM) for long memory dependency image description systems.

The remaining paper is organized as follows: in Section II, previous and related studies are introduced. Then, the materials and methods used in this study are described including the proposed image description model in section III. Section IV deals with the results discussion. The paper ends with a conclusion.

II. RELATED WORK

There are many studies in the field of image classification and description. While some of them use symbolic datasets [14],[19],[28],[37], others depend on natural datasets [31],[40]

The first studies at this field focus only on the image component detection. Some studies like [22] detect the components by rounding a box for each component, others [7],[10] define the location of humans in the images. Studies like [23] define objects depending on faces and bodies. Another study [10] uses Caltech dataset which consists of 35000 images. It contains also information about objects and components in image without the need of detection process. Another research [12] uses Pascal VOC dataset which includes 20 classes taken from 1100 images.

Sun dataset [13] consists of 908 different classes and 3819 sub classes. It consists of 7971 images of disk class, 20213 images of wall class and 16080 images of window. It contains less number for other classes such as boat, plane, ground and light.

The famous MS COCO dataset [40] defines the properties of image components from side view of object detection and labelling as well as relationships between images components.

By using 200 classes corresponding to 400000 images form ImageNet dataset, the study [40] detects location of 350000 components using bounding box method.

Studies [5]،[24]،[34],[35],[43] use CNN to generate a brief description of image, while other studies [33],[41],[44] use RNN to generate a full description.

Many recent studies [2],[42] focus on building a description of image components based on the matching between image components and described words.

In other studies [1],[4],[11],[21],[27],[39],[47], the researchers generate only comments (not sentences) of images. They used fixed templates (Cat, House, Dog, Boat, ….etc.) which corresponded to the most common objects in images. Similar study [8] uses annotations but with multi-level form. Another study [3] uses different method to achieve the process of component detection which deals with Semantic object classification.

Several studies [17],[20],[29],[30] deal with the problem of holistic scene understanding which infers the scene type, objects and their locations in the image. All those studies focus on labelling the image objects using a fixed set of categories.

Studies [16],[38] focus on discovering, annotating, and recognizing scene attributes in order to generate description.

Kiros et al. [26] develop a log bilinear model to generate full-length sentences. This study uses a fixed window for generating sentences.

Karpathy et al. [47] use a method depended on RNN for generation of full sentences describing images. They use the log probability to produce the next token based on previous words in context. The research uses the Flickr8k and Flick30k datasets. The same idea found in other studies [6],[9],[15],[25],[32]. The difference between the two studies is that kiros' model uses a fixed window context, while the another study uses Recurrent Neural Network (RNN) model conditions the probability distribution over the next word in a sentence on all previously generated words.

Best recent study [46] fuses the benefits of CNN and RNN together to generate image description. The CNN produces the image features while the RNN takes these features and image captions to produces a full-length description.

In this study, we aim to built an Arabic description model (ADM) and compare it with previous English-based models (EBM). We develop an existing image description model of Karpathy' study [47] and modify it to generate Arabic description. ADM uses CNN for feature extraction of images, LSTM networks to generate an Arabic description sentences, and Bleu evaluation metrics for comparing both models ADM and EDM. We apply our model on a dataset consisting of 2000 images based on flickr8k dataset. Our ADM model is also compared with a translation of EDM sentences to define the best method.

III. MATERIALS AND METHODS

All studies in the literature are used the English language. However, our suggested model, which generates an Arabic description sentence of image, is described in figure (4). We aim to generate a full-length image description given two main inputs which are the image features obtained from CNN and the Arabic notation file (JSON) file. The Arabic description model will be inferred using a special type of RNN which is the LSTM network.

A. Convolutional Neural Networks (CNN)

CNNs are types of deep learning networks which consist of a convolution and pooling layers. The convolution layer makes the convolution process which extracts different features of inputs. The pooling layer reduces the input dimension. Before the pooling process, a non-linear function is applied to remove the outlier points. The process of convolution-pooling is applied several times to minimize the data to a vector size. The last layer of CNNs is the fully connected layer from which the classification results are obtained [9].

CNN components:

The architecture of CNN consists of convolutional, pooling, non-linear and the fully connected layers.

The convolutional layer applies a number of filters to the input image in order to extract specific features like (lines, edges, corners….). Those filters are generated randomly and applied using the sum of product rules, and each filtering process constitutes one result. The number of neurons of this layer is (M*N*K) which M and N are the size of input image and k is the number of filters.

The next layer of CNN is the pooling layer in which the M*N*K feature matrix is minimized into half size (M/2)*(N/2)*K. This operation is called the subsampling process. The pooling stage is done via different ways such as max pooling or average pooling (as illustrated in figure 1) [48],[49].

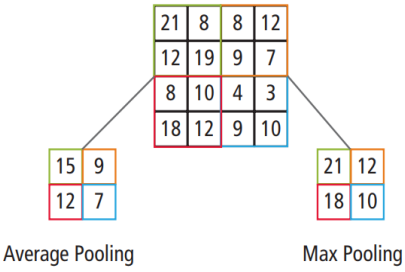


Fig.1 The Pooling Methods

There are two common types of CNN networks relating to number of convolutional layers, these types are 2conv1maxpool and 3conv2maxpool types.

Non-linear layer is the third layer in CNN. In this layer, a non-linear function is applied with equation of y=max(y,0) to remove the outlier points as shown in figure 2.

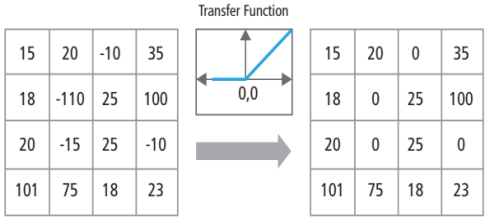


Fig.2 the Non-Linear Function Calculations

The non-linear function reduces the vanishing gradient problem which appears in the case of increasing the number of layers. Most of deep learning networks use the backpropagation algorithm. When adding more layers the

gradient will minimize significantly until it vanishes. However, with non-linear function, the gradient propagated effectively through the network eliminating the vanishing problem. The simplicity and minimal calculations during that non-linear layer make these functions very suitable for CNNs especially for time saving [48],[49].

The last layer in CNNs is the fully connected layer (FC). The neurons from previous layer are fully connected to all neurons in this layer. The values of previous neurons are multiplied by weights of connections of the FC layer, then all multiplications results are summed together to produce one value as output of FC layer. These calculations are repeated until finishing all calculations of FC outputs [48],[49]. Figure 3 shows some of FC calculations.

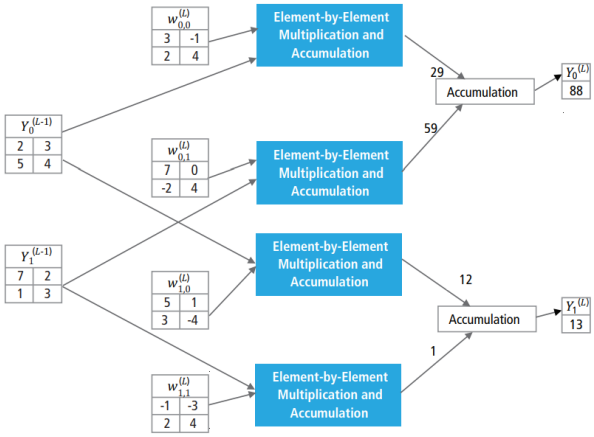


Fig.3 The calculations of FC layer

B. The CNN 4096 features:

The input image is resized into 224*224*3 dimensions and used as input of CNN. The feature vector of CNN is fed into the RNN network in addition to the caption (Image Description File) in order to generate the final representation model.

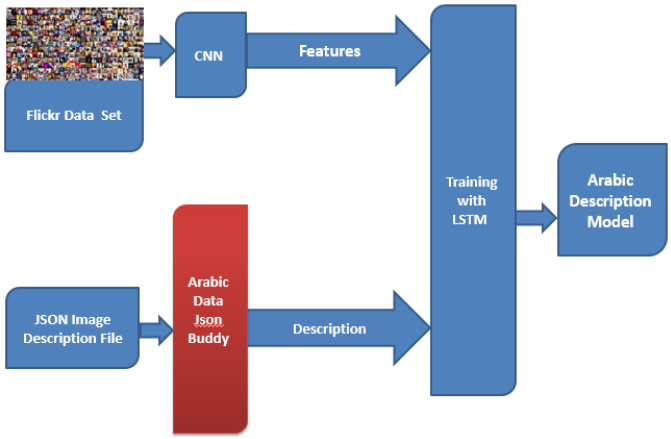


Fig.4 Our Suggested Training Model

In Karpath model [49], CNN consists of four (2-convolutional 1-max pooling) layers in which the input matrix is passed to different convolution filter sizes (64, 128, 256 and 512) convolution. Then the filtered matrix is down subsampled into half size each time (4 times). So, the final feature matrix becames 7*7*512 size. The last process is the sum of production of the last (512*7*7) vector layer with all neurons of the last fully connected layer. The filter of the last layer is also of size 512*7*7. The final feature vector has 4096 samples. Figure (5) shows that calculations clearly.

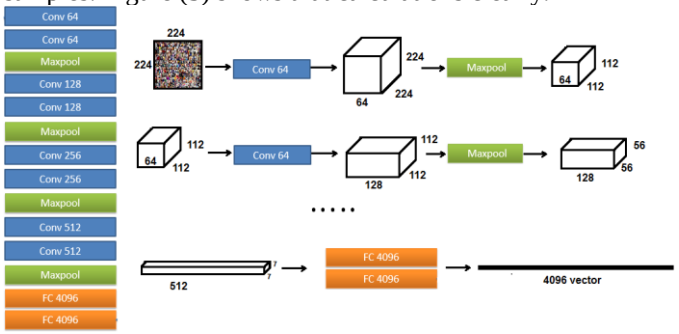


Fig.5 CNN layers and Calculations

C. JSON image description file:

Image description file is a file containing two parts. The first is "images" part and the second is the "dataset" part. The "images" part also contains two sections which are the "imgid" and "sentences". The "imgid" represents the number of each image, while the sentences contain the information of each sentence describing the image. For each image, there are five sentences describing the image. The "sentences" section consists of two parts which are the "tokens" and "raw". While "token" part represents the words individually, the "raw" part is the full-length sentence. The second part of "dataset" represents the name of the dataset. Figure (6) illustrates a simple part of an English JSON description file of Flickr8k dataset.

{"images": [{"sentids": [0, 1, 2, 3, 4], "imgid": 0, "sentences": [{"tokens": ["a", "black", "dog", "is", "running", "after", "a", "white", "dog", "in", "the", "snow"], "raw": "A black dog is running after a white dog in the snow .", "imgid": 0, "sentid": 0}, {"tokens": ["black", "dog", "chasing", "brown", "dog", "through", "snow"], "raw": "Black dog chasing brown dog through snow", "imgid": 0, "sentid": 1}, {"tokens": ["two", "dogs", "chase", "each", "other", "across", "the", "snowy", "ground"], "raw": "Two dogs chase each other across the snowy ground .", "imgid": 0, "sentid": 2}, {"tokens": ["two", "dogs", "play", "together", "in", "the", "snow"], "raw": "Two dogs play together in the snow .", "imgid": 0, "sentid": 3}, {"tokens": ["two", "dogs", "running", "through", "a", "low", "lying", "body", "of", "water"], "raw": "Two dogs running through a low lying body of water .", "imgid": 0, "sentid": 4}], "split": "train", "filename": "2513260012_03d33305cf.jpg"}, {"sentids": [5, 6, 7, 8, 9], "imgid": 1, "sentences": [{"tokens": ["a", "little", "baby", "plays", "croquet"], "raw": "A little baby plays croquet .", "imgid": 1, "sentid": 5}, {"tokens": ["a", "little", "girl", "plays", "croquet", "next", "to", "a", "truck"], "raw": "A little girl plays croquet next to a truck .", "imgid": 1, "sentid": 6}, {"tokens": ["the", "child", "is", "playing", "croquette", "by", "the", "truck"], "raw": "The child is playing croquette by the truck .", "imgid": 1, "sentid": 7}, {"tokens": ["the", "kid", "is", "in", "front", "of", "a", "car",

Fig.6 Part of Flickr8k JSON English Description File

D. JSON Arabic image description file:

To generate an Arabic description of a data set image, we may think two different ways. The first is by training a description model based on English description file, and then translating the output of the model. The other method, which is better and more efficient than the first, based on training a model using an Arabic description file. For this solution, the output of the model is an Arabic description sentences.

In order to achieve our goal, we need to generate the JSON Arabic caption file. This is done by the following steps:

1- Obtain the English JSON dataset notation.
2- Translate the sentences by Ultra edit (Smart Translation). This program memorizes the previous

words translation, which facilitates the process. The problem with Ultra edit is that the translation may contain some lacks such as brackets. In other cases, the brackets could exist in false places.

3- The third step is the debug step that corrects the output of Ultra edit translation. JSON Buddy does the debug process.

{"images": [{ "sentids": [0, 1, 2, 3, 4], "imgid": 0, "sentences": [{ "imgid": 0,"sentid": 0, "tokens": ["يركض ", "كلب", "اسود", "خلف", "كلب", "ابيض", "في", "الثلج", "."],"raw": "يركض كلب اسود خلف كلب ابيض في الثلج."}, {"imgid": 0, "sentid": 1, "tokens": ["كلب", "اسود", "يطارد", "كلب", "بني", "عبر", "الثلج", "."], "raw": " كلب اسود يطارد كلب بني عبر الثلج."}, { "imgid": 0, "sentid": 2, "tokens": ["الارض", "عبر", "البعض", "بعضهما", "كلبان", "يطارد", "."], "raw": "يطارد كلبان بعضهما البعض عبر الارض الثلجية."}, { "imgid": 0, "sentid": 3, "tokens": ["يلعب", "كلبان", "معا", "في", "الثلج", "."], "raw": "يلعب كلبان معا في الثلج."}, { "imgid": 0, "sentid": 4, "tokens": ["كلبان", "يركضان", "عبر", "جسم", "مائي", "منخفض", "."],"raw": "كلبان يركضان عبر جسم مائي منخفض."}], "split": "train", "filename": "2513260012_03d33305cf.jpg" }, { "sentids": [5, 6, 7, 8, 9], "imgid": 1,"sentences": [

Fig.7 Part of Flickr8k JSON English Description File

By revising the original JSON file and the Arabic one (figure 7), we find, for example, that the sentence

"**A black dog is running after a white dog in the snow**" becomes

"**يركض كلب اسود خلف كلب ابيض في الثلج**"

which is 100% correct in Arabic parsing.

E. Training stage:

The input of this stage is the Arabic JSON description file and the features extracted by the CNN network. There are many choices for training stage. Recurrent neural network (RNN) is a very common choice for generate image description because of its ability to produce output based on given the previous input. The main problem of RNN is that its memory is very short. Therefore, with the more network depth, the more complexity will face us.

Long-short term memory (LSTM) neural networks is another type of memory deep learning NN which is an improved type of RNN in which we can produce an output based on given different previous time intervals of inputs.

So, for generating a natural language description of an image, the LSTM will be more efficient. The architecture of our LSTM training model is described in figure (8) below.

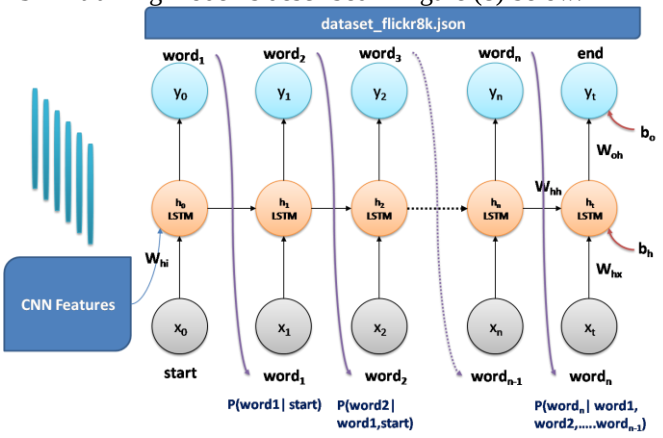


Fig.8 The proposed LSTM model for generating image description

The training process is concluded by the following equations:

For t=1:t

bv = Whi[CNNθc (I)]

ht = f(Whxxt + Whhht−1 + bh + 1(t = 1) Ɵ bv)

yt = sof tmax(Wohht + bo).

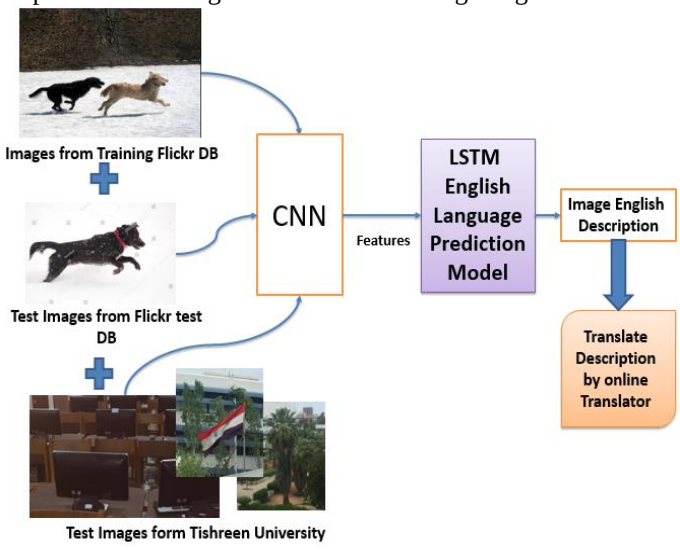
where t is the number of iterations, CNNθc (I) is the 4096 features of CNN, Whi is the weight of the connection between input feature vector and the first hidden layer, Whx is the weight of the connections between the input layer neurons and the hidden layers ones, xt is the input at period time t, Whh is the weight of the connections between the hidden layer neurons, bh is the bias of the hidden layer neurons, Woh is the weight of the connections between the hidden layer neurons and the output layers ones, bo is the bias of the ouput layer neurons, and softmax is the activation function of the output layer.

As shown in figure (8), each word is predicted in presence of all the previous predicted words in the sentence. So, the last word probability is given as: P(wordn| word1, word2,…..wordn-1).

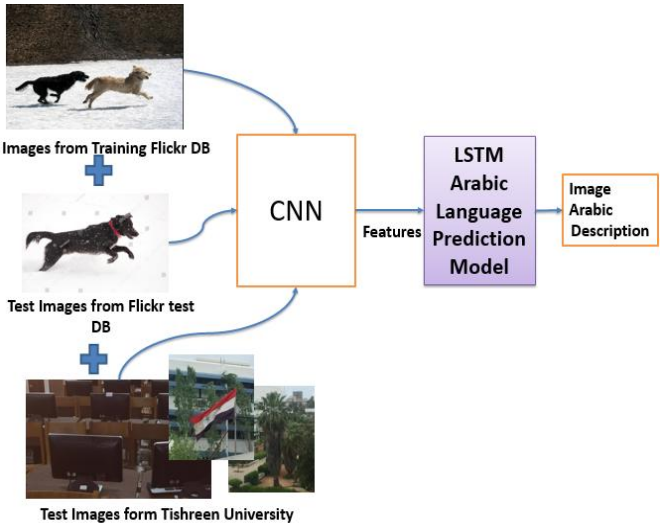
F. Test Scenarios:

For the test stage, we suggested two different scenarios. The first one is to translate the output of the English description model of images using common translator such as Google. This scenario suffers from real problem which is the bad construction of Arabic sentence. The second scenario solves this problem by generating the Arabic description using the trained LSTM Arabic language prediction model. It gets Arabic sentences as output of the system describing the test images with a correctly Arabic construction. Figure (9) A and B illustrates the two different scenarios.

For the test scenario, we collected three different test sets. First, we examine our model by simulating a number of trained images from filckr8k dataset. Second, test images (different from trained images) are chosen from flickr8k test dataset for using in our trained model. Last, we build a newly test dataset from our college in Tishreen University for more confidence of the experiments and use it as input for the model. In the three cases, the English and Arabic models gave outputs as a full length sentences describing images.



(A)



(B)

Fig.9 Our test scenarios: (A) English model, (B) Arabic Model

G. Performance Evaluation Metrics:

G.1 Bleu:

Bleu is a method for automatic evaluation of machine translation or image description systems. It calculates degree of match between two descriptions, the first is the reference one (JSON description sentences of each image) and the other is the result description sentence (output of our systems). Bleu is given as follows [50]:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (1)$$

Where **BP** is the **Brevity penalty** which gives information about the match degree between the reference and result description. BP is computed as the following equation [50]:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases} \quad (2)$$

W_n is weight which is 1/N (N is the number of generated words in the description sentence).

log(Pn) is the logarithm of n-gram value which is the number of matched between reference and generated sentence.

Ratio is the rate between length of translation and length of reference sentence.

G.2 Log Probability (LP)

LP is another performance metric which declares the probability of similarity between description and components of described image, and is calculated as follows:

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t) \quad (3)$$

Where log pt(st) is the logarithm of n-gram value which is the number of matched between reference and generated sentence. The more LP values are, the more accurate results obtain.

IV. RESULTS DISCUSSION

For the experiments stage, we use a subset from Flickr8k dataset in both language English and Arabic. (four different dataset, MSCoco, Flickr30k, Flickr8k and Flickr2k, and extra simple Arabic one.) . We use 1500 images for train stage, 250 images for val stage and 250 images for tests stage.

A. Train samples test results:

Some of test samples with their corresponding reference and result description sentences are introduced as following. The results are in Arabic and English.

Reference Image Description

1 A black and a brown dog are running toward the camera .
2 A black and a yellow dog .
3 A black dog and a brown dog with a ball .
4 A large black dog and a large yellow dog appear to be playing with a soccer ball .
5 The black dog is by a soccer ball and the brown dog is running .

Result Description (English):
a dog is running through a field

Result Description (Arabic):
. كلب اسود و كلب بني اللون في حقل عشبي.

Reference Image Description

1 . يركض كلب اسود و بني اللون نحو الكاميرا
2 . كلب اسود و بني اللون
3 . كلب اسود و كلب بني مع كرة
4 . يظهر كلب اسود كبير الحجم و كلب بني كبير الحجم ليلعبوا بكرة القدم
5 . يوجد كلب اسود بجانب كرة القدم و يركض كلب بني

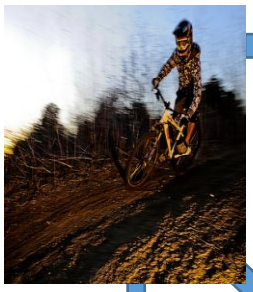
1A brown and white dog is running through woodland.
2 A large brown , white , and grey dog is running through a forest covered in leaves.
3 A tan long-haired dog is running in the forest.
4 Brown and white dog running through fall leaves on the ground .
5 The brown dog bounds through the leaves and the underbrush .

Result Description (English):
a dog runs through a field

Result Description (Arabic):
. كلب بني و اسود اللون يقفز فوق سياج ابيض و ابيض

Reference Image Description

1 . يركض كلب بني و اسود اللون عبر الغابة
2 . يركض كلب بني و ابيض و رمادي اللون حجمه كبير عبر غابة مغطاة بالاوراق
3 . يركض كلب بني ذو شعر طويل في الغابة
4 . يركض كلب بني و ابيض اللون عبر اوراق متساقطة في الارض
5 . يقفز الكلب البني عبر الاوراق و الشجيرات النامية



Reference Image Description
1A biker rides on a dirt road .
2 A person on a bmx bike .
3 A person on a BMX bike is riding on an outdoor course .
4 A person rides his mountain bike down a dirt road very fast .
5 Dirt bike rider jumping down the hill .

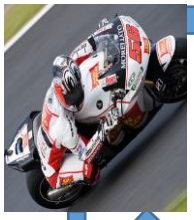
Result Description (English):
a man is standing on a rock looking at the view of the mountains .
Result Description (Arabic):
. رجل يركب دراجة نارية في الهواء

Reference Image Description
1. يقود دراج في طريق ترابي
2 . دراج على دراجة بي ام اكس
3 . يقود دراج على دراجة بي ام اكس في الهواء الطلق
4 . يقود دراج دراجته الجبلية بسرعة كبيرة في طريق ترابي
5 . يقفز سائق دراجة ترابية اسفل الهضبة

Reference Image Description
1A girl in a dress holds a tennis racket .
2 A little girl in a red dress and cape holds a large tennis racquet .
3 A little girl in a red flowered dress with a tennis racket .
4 A young girl , in an orange dress , holding a tennis racket .
5 The little girl in the red flowered dress and pink cape is holding a racket .

Result Description (English):
a girl in a pink shirt is playing with a ball
Result Description (Arabic):
. فتاة صغيرة ترتدي فستان ملون تحمل كوب

Reference Image Description
1 . تحمل فتاة ترتدي فستان مضرب التنس
2 . تحمل فتاة ترتدي فستان برتقالي و قبعة مضرب تنس كبير
3 . فتاة صغيرة بفستان برتقالي مزهر مع مضرب تنس.
4 . تحمل فتاة بفستان برتقالي مضرب التنس
5 . تحمل فتاة بفستان برتقالي مزهر و قبعة زهرية اللون مضرب التنس



Reference Image Description
1 A black and white motorcycle is being ridden around a racetrack .
2 A man on a motorcycle is racing on a track .
3 A motorcycle racer speeds down a track .
4 A orange , white , and black motorcyclist racing .
5 A professional motor racer in motion .

Result Description (English):
a man in a red shirt is riding a bike on a dirt bike
Result Description (Arabic):
. رجل يرتدي خوذة حمراء يقود دراجة نارية في الهواء

Reference Image Description
1 . تمت قيادة دراجة سوداء و بيضاء حول مسار سباق
2 . يتسابق دراج يقود دراجة في مسار السباق
3 . يسرع متسابق على مسار سباق
4 . تتسابق دراجة نارية برتقالية و بيضاء و سوداء
5 . سائق دراجة محترف يتحرك

Reference Image Description
1A boy , an adult and two dogs running through the snow .
2 A man and a child walk through the snow with two dogs .
3 Family running with dogs in the snow .
4 Two people are walking two golden dogs in the snow .
5 Two people wearing camouflage clothing in the snow with two dogs .

Result Description (English):
a dog is jumping over a fallen tree in the forest.
Result Description (Arabic):
. كلب اسود و كلب بني اللون في حقل عشبي

Reference Image Description
1 . يركض صبي و رجل و كلبان عبر الثلج
2 . يسير رجل و صبي عبر الثلج مصطحبان كلبان
3 . تركض عائلة مع الكلاب في الثلج
4 . يقوم شخصان بتسيير كلبان بنيان في الثلج
5 . شخصان يرتديان ملابس عسكرية متواجدان في الثلج مع كلبان

Reference Image Description
1A person on a snowmobile in mid jump .
2 A snowmobiler flies through the air .
3 A snowmobile rider flies through the air on his or her machine in front of tall pine trees .
4 Rider jumps snowmobile high in rural area .
5 The man is up in the air .

Result Description (English):
a man is riding a bike on a dirt bike
Result Description (Arabic):
. رجل يرتدي خوذة زرقاء اللون يقفز فوق لافتة تقول

Reference Image Description
1 . متزلج في منتصف القفزة
2 . متزلج يطير في الهواء
3 . متزلج يطير عاليا بزلاجته امام صف من اشجار الصنوبر العالية
4 . المتزلج يقفز عاليا بالمزلجة في ارض جرداء
5 . المتزلج عاليا في الهواء

Reference Image Description
1A brown dog is chewing on a bone with a stuffed animal underneath it .
2 A dog chews on a bone .
3 Big brown dog chews on bone lying down .
4 Big hairy dog chews on a bone while lying on a furry toy .
5 Brown dog chews on bone while laying on the rug .

Result Description (English):
a dog is running through a field .
Result Description (Arabic):
. كلب بني و اسود اللون يقفز فوق سياج ابيض و ابيض

Reference Image Description
1 . كلب بني يمضغ عظمة مع حيوان محشو تحتها
2 . كلب يمضغ عظمة
3 . كلب كبير بني يمضغ عظمة و هو مستلقي
4 . كلب كبير ذو شعر كثيف يعض عظمة بينما يستلقي على دمية فروية
5 . كلب بني يعض عظمة بينما يستلقي على السجادة

Fig 10. Image Description Results of some Test Samples

To illustrate the difference between ADM and the translation of EDM, we translate the EDM of many images and compared them with the corresponding ADM, the figure 11 shows clearly that our ADM gives better results.

Image	Translated EDM	ADM
	. كلب يمر من خلال حقل	. كلب اسود و كلب بني اللون في حقل عشبي
	. رجل يقف على صخرة ينظر الى منظر الجبال	. رجل يركب دراجة نارية في الهواء
	. كلب يمتد من خلال حقل	. كلب بني و اسود اللون يقفز فوق سياج ابيض

Fig 11. Comparison of translated EDM and ADM of some Test Samples

B. Performance Evaluation:

Table 1 illustrates the Bleu metrics of the EDM and ADM description models. It can be noticed that the English model has better values. Also, we can see that the increment of n value decreases the performance and this is normal because of the fact that description will be more complicated with more words of the sentence.

Table 1 The Bleu-n metrics of EDM and ADM.

Model/ Bleu-n	Bleu-1	Bleu-2	Bleu-3	Bleu-4
English-Based model	51.5	31.4	17.4	9.9
Arabic-Based Model	34.4	15.4	7.6	3.5

Table 2 shows the other Bleu detailed metrics which are the brevity penalty (BP) and the Ratio. BP measures the degree of match between reference and generated descriptions. Ratio represents the ratio between generated sentence length (Hyp_len) and reference length (Ref_len).

Table 2 Other metrics of English and Arabic Models

Model/ Metric	BP	Ratio	Hyp_len	Ref_len
English-Based model	1	1.049	2856	2722
Arabic-Based Model	1	1.095	2686	2454

We also computed the log probabilities (LP) of the test images in order to define the accuracy of the generated description. The EDM model gives LP with values between -6 to -18 while ADM gives LP with values between -7 to -24. Those values indicate that EDM gives more accurate results than ADM.

V. CONCLUSIONS

In this work, we propose a novel Arabic-based image description model by developing an English-based model. We use CNN networks for image feature extraction and use them as an input of LSTM network. An Arabic description file is built in order to provide it into the LSTM for the training stage. We use a subset of Flickr8k dataset (1500 for training, 250 form val and 250 for test). Performance evaluation via Bleu-n, Ratio and BP metrics is calculated and discussed. The results show that English-based model had better performance than Arabic model. However, the proposed Arabic model show better results than the translated model. In the future, we will work on the enhancement of Arabic description model.

ACKNOWLEDGMENT

Portions of the research in this paper use the Flickr8k dataset which had been downloaded for free from link:

http://nlp.cs.illinois.edu/HockenmaierGroup/Framing_Image_Description/KCCA.html

REFERENCES

[1] Barbu A., Bridge A., Burchill Z., Coroian D., Dickinson S.,Fidler S., Michaux A., Mussman S., Narayanaswamy S.,Salvi D.,Video in sentences out. arXiv preprint arXiv:1204.2742, 2012

[2] Barnard K., P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. JMLR, 2003.

[3] Brostow, G., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high- de_nition ground truth database. PRL 30(2) (2009),pp: 88-97

[4] Bell, S., Upchurch, P., Snavely, N., Bala, K.: OpenSurfaces: A richly annotated catalog of surface appearance. SIGGRAPH 32(4) (2013).

[5] Chen X., H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325, 2015.

[6] Chen X. and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. CoRR, abs/1411, pp:5654, 2014

[7] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005) .

[8] Deng, J., Russakovsky, O., Krause, J., Bernstein, M., Berg, A., Fei-Fei, L.: Scalable multi-label annotation. In: CHI. (2014).

[9] Donahue J., L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. arXiv preprint arXiv:1411.4389, 2014.

[10] Doll_ar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. PAMI 34 (2012).

[11] Elliott D. and F. Keller. Image description using visual dependency representations. In EMNLP, pages 1292–1302, 2013.

[12]. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. IJCV 88(2) (June 2010), pp: 303-338.

[13] Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: Large-scale scene recognition from abbey to zoo. In: CVPR. (2010).

[14] Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object cate-gories. In: CVPR Workshop of Generative Model Based Vision (WGMBV), (2004).

[15] Fang H., S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Doll´ar, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. arXiv preprint arXiv:1411.4952, 2014.

[16] Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their at-tributes. In: CVPR. (2009).

[17] Fidler S., A. Sharma, and R. Urtasun. A sentence is worth a thousand pixels. In CVPR, 2013.

[18] Frome A., G. S. Corrado, J. Shlens, S. Bengio, J. Dean,T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In NIPS, 2013.

[19] Gri_n, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology (2007).

[20] Gould S., R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In Computer Vision, 2009 IEEE 12th International Conference on, pp:1–8. IEEE, 2009.

[21] Gupta A. and P. Mannem. From image annotation to image description. In Neural information processing. Springer, 2012.

[22] Hjelm_as, E., Low, B.: Face detection: A survey. CVIU 83(3) (2001), pp:236-274.

[23] Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild. Technical Report 07-49, University of Massachusetts, Amherst (October 2007).

[24] JeffreyPennington R. and C. Manning. Glove: Global vectors for word representation. 2014.

[25] Kiros R., R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539, 2014.

[26] Kiros R., R. S. Zemel, and R. Salakhutdinov. Multimodal neural language models. ICML, 2014.

[27] Kulkarni G., V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In CVPR, 2011.

[28] Lecun, Y., Cortes, C.: The MNIST database of handwritten digits (1998).

[29] L.-J. Li and L. Fei-Fei. What, where and who? Classifying events by scene and object recognition. In ICCV, 2007.

[30] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 2036–2043. IEEE, 2009.

[31] Lin T, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Doll P, Lawrence C, Microsoft COCO: Common Objects in Context, 2015.

[32] Mao J., W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090, 2014.

[33] Mikolov T., M. Karafi´at, L. Burget, J. Cernock`y, and S. Khudanpur. Recurrent neural network based language model. In INTERSPEECH, 2010.

[34] Mikolov T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In NIPS, 2013.

[35] Mitchell M., X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daum´e, III. Midge: Generating image descriptions from computer vision detections. In EACL, 2012.

[36] Malik, J ,Bourdev, L.,.: Poselets: Body part detectors trained using 3D human pose annotations. In: ICCV, (2009).

[37] Nene, S.A., Nayar, S.K., Murase, H.: Columbia object image library (coil-20). Technical report, Columbia Universty (1996).

[38]. Patterson, G., Hays, J.: SUN attribute database: Discovering, annotating, and recognizing scene attributes. In: CVPR. (2012).

[39] Russell, B., Torralba, A., Murphy, K., Freeman, W.: LabelMe: a database and web-based tool for image annotation. IJCV 77(1-3) (2008), pp: 157-173.

[40] Russakovsky, O., Deng, J., Huang, Z., Berg, A., Fei-Fei, L.: Detecting avocados to zucchinis: what have we done, and where are we going? In: ICCV, (2013).

[41] Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: ECCV, (2012).

[42] Socher R. and L. Fei-Fei. Connecting modalities: Semisupervised segmentation and annotation of images using unaligned text corpora. In CVPR, 2010.

[43] Shotton, J., Winn, J., Rother, C., Criminisi, A.: TextonBoost for image under-standing: Multi-class object recognition and segmentation by jointly modeling tex-ture, layout, and context. IJCV 81(1) (2009), pp: 2-23.

[44] Sutskever I., J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In ICML, 2011.

[45] Vinyals O., A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. arXiv preprint arXiv:1411.4555, 2014.

[46] Vinyals O, Toshev A, Bengio S, and Erhan D, "Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge", IEEE TRANSACTION ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, sep, 2016.

[47] Karpathy A, Fei-Fei L, "DeepVisual Semantic Alignments for Generating Image Descriptions", Stanford University.

[48] http://ufldl.stanford.edu/tutorial/supervised/ConvolutionalNeuralNetwork/, last access 1-5-2018.

[49] http://cs231n.github.io/convolutional-networks/, last access 1-5-2018.

[50] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.