استخدام التصنيف المسبق للصور لتحسين دقة أنظمة وصف الصور

م. رشا معلا [1]*

د. جعفر الخير [2]

د. سامر سليمان [3]

الملخص:

يعد التعلم العميق الذي يبنى على هدف وصف الصور وإظهار تسميات توضيحية لها أحد أهم التطبيقات الواعدة في مجال علوم الحاسب. حيث يتكون من جزئيين رئيسيين هما (نموذج وصف الصورة والنموذج النصي). قمنا في أبحاث سابقة بدراسة تأثير استخدام اللغات واختلاف مجموعة بيانات مختلفة على نماذج لوصف الصور، ولكن في هذه الورقة البحثية سندرس تأثير تصنيف مجموعات بيانات الصورة على دقة نماذج الوصف السابقة. تم انشاء مجموعة بيانات مؤلفة من 12000 صورة ومجمعة من مجموعتين للبيانات (Flickr2k و MS-COCO)، حيث أن هذه النماذج المصممة تدعم اللغتين العربية والإنكليزية، أما بالنسبة للجزء المختص بوصف الصورة تم استخدام سيناريوين مختلفين. في السيناريو الأول استخدمنا نماذج الصور مع شبكة CNN والنص LSTM ,بينما في السيناريو الثاني تم استخدام ResNet50 و FastText كنماذج وصف الصورة والنص على التوالي. تم تطبيق عملية التدريب لكل من الأصناف الداخلية والخارجية ثم تم تطبيق سيناريو الاختبار بعد انتهاء عملية التدريب وفقاً لحالتين وبأربع طرق مختلفة وهاتان الحالتان هما نماذج الوصف كلمة -كلمة ونموذج الوصف جملة – جملة. أثبتت نتائج تحليل الأداء أن الفئات المصنفة تتمتع بأداء أعلى من غير المصنفة في حالة مجموعة البيانات المستندة إلى التكرار وغير المتكررة في جميع السيناريوهات.

الكلمات المفتاحية:

تأثير عملية LSTM، شبكة CNN، شبكة ResNet50، شبكة FastTextالتعلم العميق، أنظمة وصف الصور، نموذج التصنيف، مجموعات البيانات المصنفة.

[1] طالبة دكتوراه – قسم هندسة الحاسبات والتحكم الآلي – كلية الهمك – جامعة تشرين. Rasha.mualla@tishreen.edu.sy

[2] أستاذ – قسم الذكاء الصنعي – كلية المعلوماتية – جامعة تشرين. Dr.jafar.alkheir@tishreen.edu.sy

[3] مدرس – قسم الميكاترونيكس – كلية الهندسة– جامعة المنارة. samer.sulaiman@manara.edu.sy

Using Image Pre-classification to improve the accuracy of the image captioning systems

(1)Eng. Rasha Mualla

(2) Jafar Alkheir

(3)Samer Sulaiman

Abstract

Deep learning for the purpose of image description and captioning has been one of the most promised computer science application recently. It consists of two parts; the image and the text description models. In previous researches, we studied the effect of using different languages and datasets on the image description models. In this paper, we study the classification effect of the image dataset on those models. So, a new combined 12000-images dataset consisting of two international datasets (Flickr2k and MS-COCO) is built. The designed models support Arabic and English languages. For the description part, we used two different scenarios. In the first scenario, we used CNN and LSTM models. While for the second one, ResNet50 and FastText are used as image and text models respectively. The training is applied for both indoor and outdoor classes. Tests scenarios are applied in two cases and four ways which are the word-by-word and the sentence-by-sentence models. The performance analysis proves that classified classes have a higher performance than the unclassified ones in case of repeating-based and non-repeating-based datasets in all scenarios.

Keywords:
Deep Learning, Image Captioning Systems, FastText, ResNet50, CNN, LSTM, Classification effects, Classified datasets.

[1]PhD Student in the Department of Computer Engineering of Mechanical and Electrical Engineering of Tishreen University, Rasha.mualla@tishreen.edu.sy

[2] Professor in the Department of Artificial Intelligence of Computer Science of Tishreen University, Dr.jafar.alkheir@tishreen.edu.sy

[3] Teacher in the department of Mechatronics of Faculty Engineering Al-Manara University, samer.sulaiman@manara.edu.sy

I. Introduction

Nowadays, there are a lot of information distributed among social networks and internet sites. Some of this information are images or even videos. Companies like Google, Facebook and Twitter manipulate all this information in order to make sure that customers show the topics that it concerns. To achieve that, very big data is processed using different deep learning algorithms and models for analyzing, captioning and indexing images used in their services (Sebastián et al.., 2018).

Image captioning is the process of generating a description of an image (Zakir et al., 2019) It requires defining the important objects within the image, relationships between them and their attributes. It is used in many applications like image indexing, image retrieval, image recognition and many other ones (Sebastián et al.., 2018), (Zakir et al., 2019).

Image captioning process consists of two basic models; the first one is the image model, while the second model is the text model. The problem is that at every minute, the datasets become bigger and bigger. Deep learning networks are able to handle this big size of datasets. For example, convolutional neural networks (CNNs) (Yan et al., 1998) are used in deep learning as image feature extractor. It can be used as a classifier by using "Softmax" as a final classification layer (Zakir et al., 2019). LSTM (Long Short-Term Memory), on the other side, is a language-based model used to build the coding of the sentence describing the image (Sepp et al., 1997).

In the recent years, many image models had been introduced like AlexNet (Alex et al., 2012), GoogLeNet (Christian et al., 2015), VGG (Karen et al., 2014) and ResNet (Kaiming et al., 2016) which all are enhanced CNN-based architecture. Expanding the depth and width of the CNN architecture in VGG (Karen et al., 2014) and GoogLeNet (Christian et al., 2015) increased the performance specially after using the inception modules. ResNets (Kaiming et al., 2016) developed the residual learning block by using a shortcut connection of identity mapping. This connection forced the network to split over the obstacle of lots of layers coping their values to the next layers. This results in enhancing the performance significantly (they got with 96.4% accuracy) (Kaiming et al., 2016).

Form "dataset" point of view, the researchers used many types of image datasets like MS-COCO, Flickr and many other ones. Microsoft Common Objects in Context (COCO) (Tsung et al., 2014) is a large image captioning dataset, which consists of 80 classes, and provides more than 82783 images for train, 40504 for validation, and 8000 images for test sets. COCO dataset al.so contains description file including one description sentence for each image. On the other hand, Flickr2k dataset is a partial dataset of flickr8k (Micah et al., 2013), which is a standard sentence-based image description dataset. It consists of more than 8000 images and their corresponding description sentences (5 sentence per image).

Our paper will focus on studying the impact of different types of natural images on the performance of image captioning systems. A modified architecture of dataset is used. We classify dataset into indoor and outdoor classes. After that, each class will be also classified into other different five sub-classes in order to achieve more efficient classification and enhance the image captioning process. This paper will also evaluate the individual sub-classes accuracy under different languages. The designed models are built twice, once in Arabic and the other in English.

The remainder of this paper is structured as follows: the next section will discuss the Related work, materials and methods of our model. Then, a detailed training and test scenario will be described. After that, experimental results and discussion will be viewed.

II. Related Work

Byeon et al.. (Wonmin et al., 2015) developed a 2D LSTM (Long-Short Term Memory) based deep learning system. The input image of the proposed model is subdivided into non-overlapping windows, and then are fed into four separate LSTMs memory blocks. The designed model reduced the total computations which results in reducing the complexity on a single-core CPU in addition to its simplicity. The model got 78.56% accuracy on Stanford Background dataset and 70.11 accuracy on SiftFlow English dataset.

Hayat, et al.. (Munawar et al., 2016) extracted spatial layout and scale invariant features from images using different deep learning network. He applied an intermediate level of information via extracting mid-level patches. Then, pyramidal image representation was applied to insure getting the scale invariance. This provides multiple distinctive features for "indoor" scenes. These features utilized information in making the final decision. The research used MIT-67, Scene-15, Sports-8, Graz-02, and NYU data sets. They got 74.4% accuracy for using VGG net on MIT-67 dataset. They denoted that the models achieved 93.1% accuracy when using 15 category scene dataset, 98% accuracy for Graz-02 dataset (using 3 categories Cars, People and Bikes), 81.2 % accuracy for NYU indoor scene dataset. The results showed that when using classified categories, the performance had increased.

Narang, et al.. (Neeru et al., 2017) introduced a deep learning system for tri-level hierarchical classification of mobile phone face datasets in heterogeneous environment. They suggested using CNN to automatically categorize face data captured scenes under different levels. In the Level 1, face images are classified based on phone type. While in the Level 2, face images are further classified into indoor and outdoor images. While in the third level, images are classified into close and far ones. The final results showed that classification accuracy was improved from 95 to 98% and from 90% to 99% for levels 1 and 2 respectively.

(Sebastián et al., 2018) combined textual and image representation for multimodal author profiling using CNN for the image model and fastText for textual model. They designed three different language description models: English, Arabic and Spanich. They obtained 0.80, 0.74 and 0.81 as accuracy for the multimodal scenario for English, Spanish and Arabic models respectively.

(Mualla et al., 2018) developed an Arabic-English image description system using CNN as image model and LSTM as text model. They built a new Arabic dataset containing Arabic description files for the Flickr2k dataset. They used 1500 images, 250 images and 250 images for training, validation and test respectively. Their model got 51.5 as a BLEU-1 metrics for the English-based model while it got only 34.4 for the Arabic one.

In 2019, (Mualla et al., 2019) used ResNet50, VGG16 and VGG19 description models. A subset of MS-COCO dataset was used with 10,000 images (9,000 of which were taken for training and 1000 for validation). The results show that the ResNet50 model outperforms both models VGG16 and VGG19 in terms of the accuracy. They continued their research by studying the effect of using different text models on the performance of image captioning systems (Mualla et al., 2020). They applied two different language models (FastText and GloVe). They used a subset of MS-COCO dataset and found that FastText models had a better performance than GloVe ones.

In a recent research of (Mualla et al., 2020), they introduced a performance comparison between two types of image captioning systems using different languages. The first one depends on generating a description of images (word - word), while the other generates the description in a way (sentence - sentence). They used VGG for the first model and ResNet50 for the second one. For the text models, they used LSTM and FastText. They build two different models; the first one used the Flickr2k dataset and its description file. The second model used a subset of MS-COCO dataset (10000 images). Both datasets are modified to be compatible with both models. So that for the Flickr dataset, only one description sentence is chosen for each image, while the subset MS-COCO dataset is modified to contain 5 sentences for both English and Arabic languages. The results proved that the English description systems had better performance than Arabic ones. The (CNN + LSTM) models with small dataset sizes and the (ResNet50 + FastText) with large dataset sizes achieved the best performance.

(Sumanth et al., 2020) presented a new deep learning system for image retrieval. They obtained the representation of images in a higher dimension of the MS-COCO dataset. They designed the system as a baseline score by fusion of captioning feature vector and image feature vector. For the image model, ResNet network is used. While FastText model is used for the text model. For the captioning model, the proposed system used Gated Recurrent Unit (GRU) to reduce the vanishing gradient problem. The system also used LSTM model to retain the relevant information for the sequence model. Another layer (self-attention layer) was added to the text and image model as an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence. The results showed that the cross-attention fusion applied between text and image improved the model performance significantly.

(Madhavan et al., 2020) introduced an image to language understanding (captioning) approach, in which two different types of models (Encoder-Decoder and Multi-model) were compared. In the Encoder-Decoder approach, inject and merge architecture were used. The inject architecture was built based on CNN as Encoder and LSTM as decoder, while the merge architecture was built based on CNN as Encoder and LSTM as image decoder (image feature encoder) and LSTM as sentence encoder. Researcher developed a full-fledged application in which the input was the image and the output was the captions. The designed models were trained via Google Conceptual captions dataset that contains 100,000 images. The researchers concluded that Inception LSTM Inject with threshold architecture was the best model with BLEU, METEOR and ROUGE values of 0.13, 0.14, and 0.18, respectively.

This research deals with the problem of different classes of image used and the effect of the data set (images) and image classes (components of the image and their relationship with each other) on the image captioning systems.

In image understanding datasets, classes can be either things or stuff (Holger et al., 2020), (Wenjie et al., 2020) Things are classes with defined shape like car, house, person etc. on the other hand, stuff are classes with undefined shape (background) like sky, grass, road etc. (Panagiotis et al., 2020). Most previous studies focus on the things due to their effective role to describe images (Geremy et al., 2008), (Holger et al., 2018). According to (Munawar et al., 2016) which proved that using sub-classes (indoor/outdoor) instead of using the entire dataset as one package, increased the performance of captioning systems, we will depend on the things (indoor and outdoor) classes, and we will take five sub-classes of each main classes (indoor and outdoor).

III. Materials and Methods

In this paper, we depend on two different image-captioning models. The first one depends on previous model (Mualla et al., 2018) using CNN as the image description model and the LSTM for the text model. The CNN takes image with size 224*224 as input, applies convolution through two convolutional networks, reduces the size of the convolution images by pooling layer and then produces the final feature vector by using a fully connected layer FC. The CNN takes image under size 224*224 (Mualla et al., 2018) and produces a feature vector consisting of 4096 items. The text model is the LSTM network which takes 5 description sentences per image. The model uses flickr2k dataset which is a part of flickr8k dataset, consisting of 2000 images with 5 sentences for each image. It produces description word by word.

On the other side, the second model uses a subset of MS-COCO dataset consisting of 10000 images with one description sentence for each image (Mualla et al., 2020). This model uses ResNet50 model (Kaiming et al., 2016) to get the image feature vector (image representation) and a pre-trained Public FastText model (Mualla et al., 2019), (FastText, 2019) to get the text representation. ResNet50 consists of 50 convolutional layers adding the idea of identity connection that improve network accuracy. The FastText model consists of selected words of the public FastText model counting 13432 words for Arabic and 11693 for English consisting the model vocabulary. The text and image feature vectors have the size of 256 items in order to fuse them together using dot product to make a unified feature vector of both positive and negative pairs. This fusion model produces captioning sentence by sentence.

In this research, we will use a combined dataset consisting of 12000 images (2000 images from flickr2k and 10000 images from MS-COCO (Flickr., 2019), (MS-COCO, 2019) The used images are not symbolic rather they are natural so that many categories can be found at the same scene making the Classification and description more challenging. Since the two datasets differ in the description sentence number, so we adapt both datasets to fit the requirement of each model. The combined dataset was classified into two main categories: indoor and outdoor. Each category consists of different sub classes. Indoor category consists of five sub classes which are Appliance, Electronic, Food, Furniture and Kitchen. On the other hand, outdoor category includes Animal, Outdoor, Person, Sports and Vehicle. Table 1 and 2 include detailed information about the number of images per each class used in our proposed dataset for indoor /outdoor categories.

The final combined dataset is split into 80% for training, 20% for validation, and randomly selected 20% for test datasets for each class.

Table 1: Components of category (indoor) and their sub-classes counts.

Class	Total count	Train Count	Val count	Test count
Appliance	515	412	103	103
Electronic	520	416	104	104
Food	536	428	108	107
Furniture	1287	1029	258	257
Kitchen	550	440	110	110

Table 2: Components of category (outdoor) and their sub-classes counts.

Class	Total count	Train Count	Val count	Test count
Animal	1811	1488	362	362
Outdoor	584	467	117	116
Person	721	576	145	144
Sports	821	656	165	164
Vehicle	1800	1440	360	360

The entire system architecture is illustrated in figure 1.

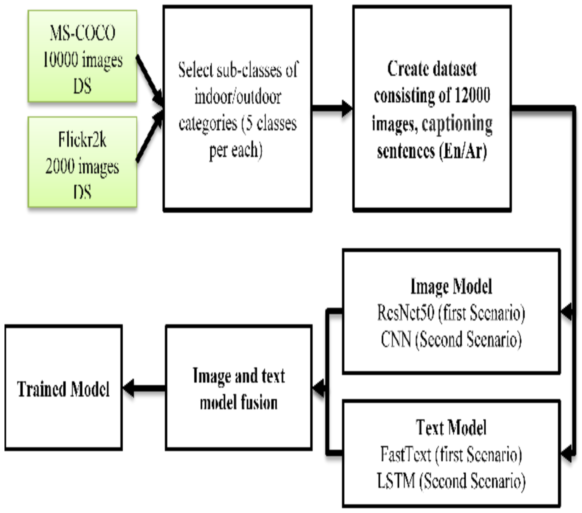

MS-COCO 10000 images DS
Flickr2k 2000 images DS
Select sub-classes of indoor/outdoor categories (5 classes per each)
Create dataset consisting of 12000 images, captioning sentences (En/Ar)
Image Model ResNet50 (first Scenario) CNN (Second Scenario)
Text Model FastText (first Scenario) LSTM (Second Scenario)
Image and text model fusion
Trained Model

Figure 1: The System Stages

V. Experimental Test and Discussion

Training Scenarios

We build two different scenarios for training phase. The first one is based on (Word-by-Word) while the other one is based on (sentence-by-sentence) models. Figure 2 illustrates the general description of both scenarios. Each scenario is performed for both Arabic and English description models. The used combined dataset has two types one for Arabic and the other one for English. Each dataset is divided into five indoor and five outdoor classes. Indoor classes include Appliance, Electronic, Food, Furniture and kitchen which are all classes inside our homes and places. On the other hand, the outdoor classes contain classes like Animal, Outdoor, Person, Sports, and Vehicle. All those classes are combined together to configure the unclassified dataset.

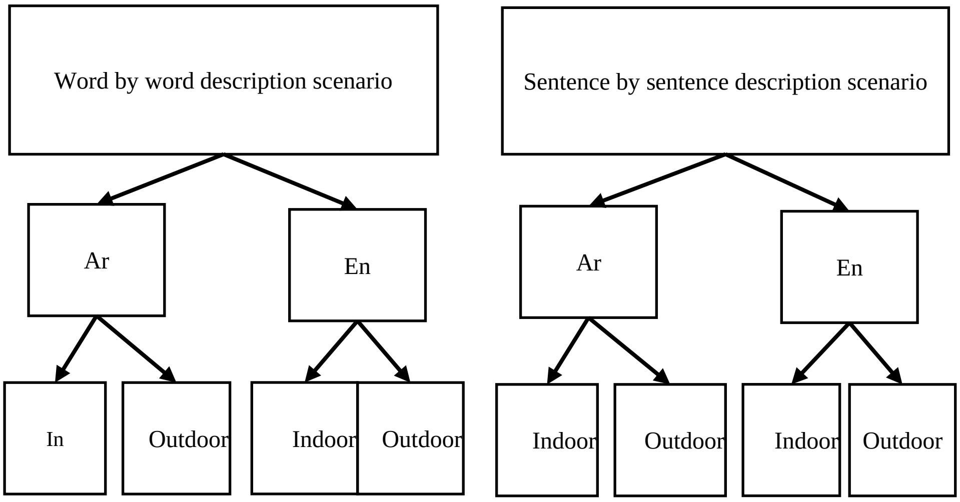

Word by word description scenario
Sentence by sentence description scenario
Ar
En
Ar
En
In
Outdoor
Indoor
Outdoor
Indoor
Outdoor
Indoor
Outdoor

Figure 2: The Training Scenarios

Test Scenarios:

Arabic and English description sentences and the test datasets have been unified in order to justify the test process for both systems. Two main different test methods are used. The first method uses the entire classes of the combined dataset taking in account repeating images among classes. While the second one uses the unique classes after omitting the repeating images.

The tests are separated into two different categories; the first one is the unclassified category in which the images will be tested according to the global dataset (12000 images). While for the second category, the images will be tested according to subclassed datasets (Animal, Food, Sports, etc.).

In terms of previous perceptions, we divided the test process into 4 parts which will be applied for both word-by-word and sentence-by-sentence models.

- Unclassified dataset experiments in terms of repeating classes' dataset for Arabic captioning system.

- Classified dataset experiments in terms of repeating classes' dataset for Arabic captioning system.
- Unclassified dataset experiments in terms of non-repeating classes' dataset for Arabic captioning system.
- Classified dataset experiments in terms of non-repeating classes' dataset for Arabic captioning system.
- Unclassified dataset experiments in terms of repeating classes' dataset for English captioning system.
- Classified dataset experiments in terms of repeating classes' dataset for English captioning system.
- Unclassified dataset experiments in terms of non-repeating classes' dataset for English captioning system.

Classified dataset experiments in terms of non-repeating classes' dataset for English captioning system.

Results:

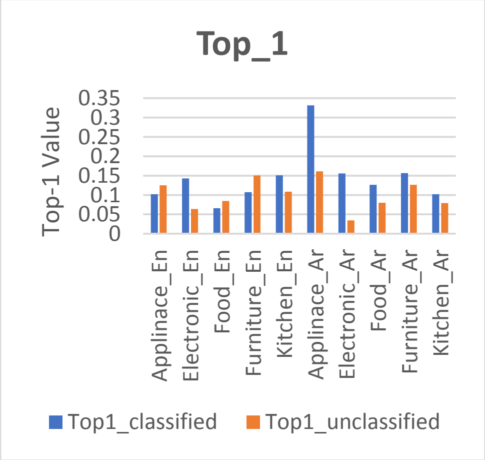
In order to evaluate our models, the following performance metrics are used, which are the evaluation parameters. They are the most important and most widely used in this field

1- Top-1 Similarity criterion, which expresses the degree of match of the best resulting image description against the original description.

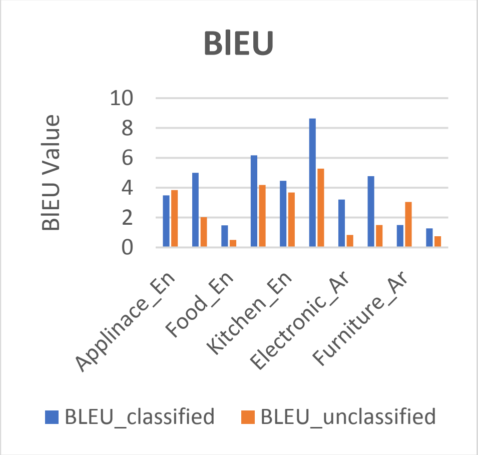
2- Bilingual Evaluation understudy (BLEU), which is an evaluation of the accuracy of the resulting description, so that if the description resulting from the testing process is very close to the description used in the training, the BLEU standard will give a high value, otherwise it will be low.

Word-By-Word Scenario test results: repeating vs non-repeating indoor classes Arabic and English description results:

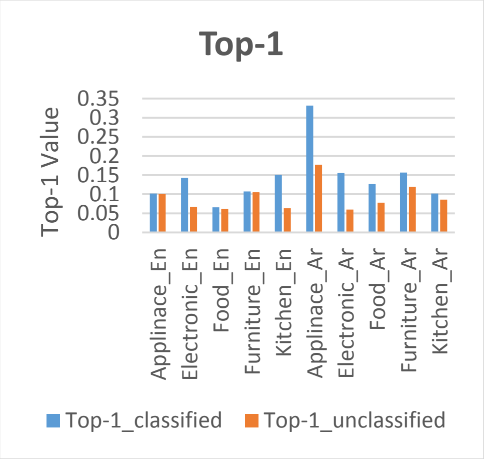
Figure 3 illustrates comparisons between Arabic and English description systems using indoor classes in terms of Top-1 and BLEU performance metrics respectively, in case of repeating classes. It also describes the performance comparisons between Arabic and English description systems using indoor classes in terms of Top-1 and BLEU performance metrics respectively, in case of non-repeating classes.



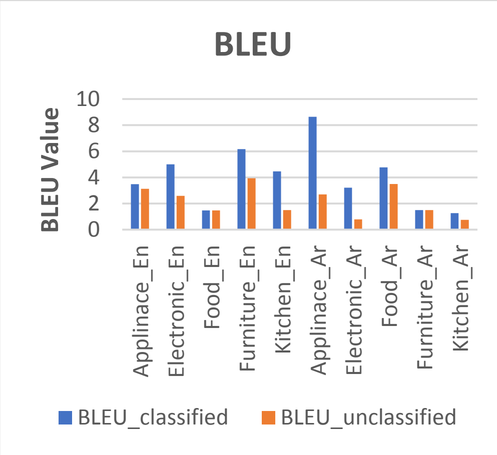
Top-1 of classified and unclassified cases of indoor repeating classes



BLEU of classified and unclassified cases of indoor repeating classes



Top-1 of classified and unclassified cases of indoor non-repeating classes

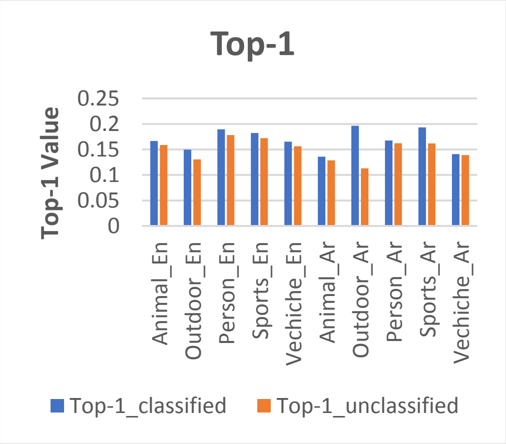


BLEU of classified and unclassified cases of indoor non-repeating classes

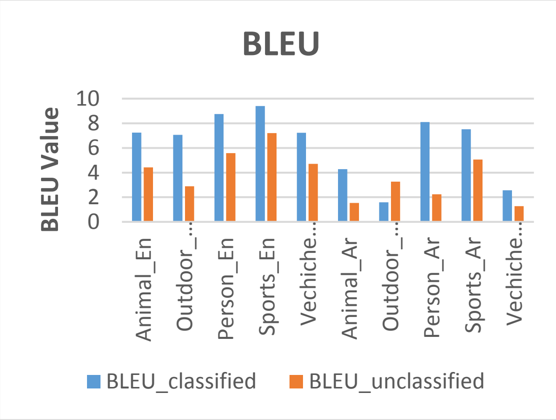
Figure 3. Performance evaluation of English/Arabic Repeating Non-Repeating indoor classes of the first scenario (word-by- word)

Repeating vs non-repeating outdoor classes Arabic and English description results:

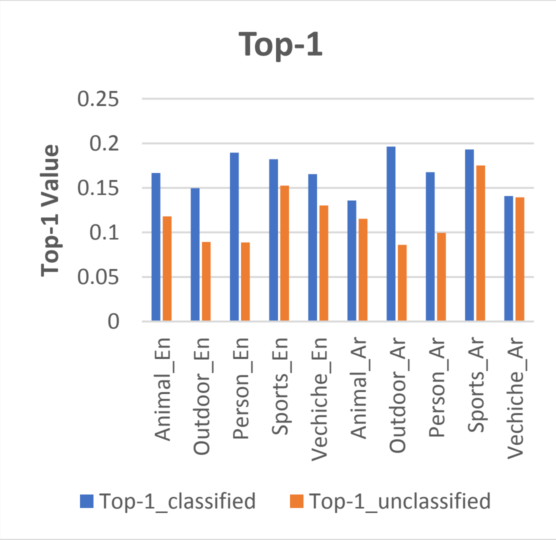
Again, same experiments of outdoor classes have been applied and the results are shown in figure 4 for repeating and non-repeating classes' datasets.



Top-1 of classified and unclassified cases of outdoor repeating classes

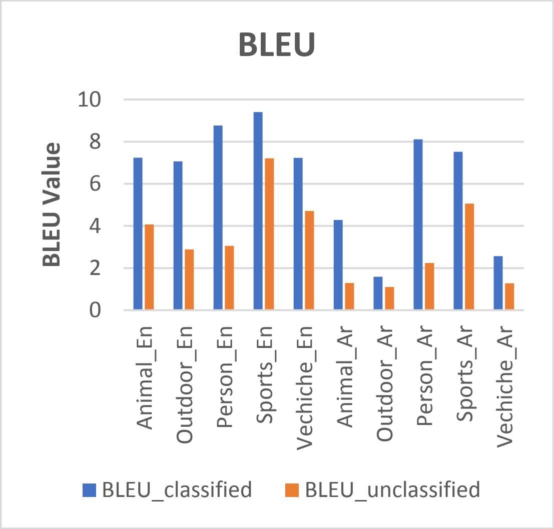


BLEU of classified and unclassified cases of outdoor repeating classes



Top-1 of classified and unclassified cases

of outdoor non-repeating classes



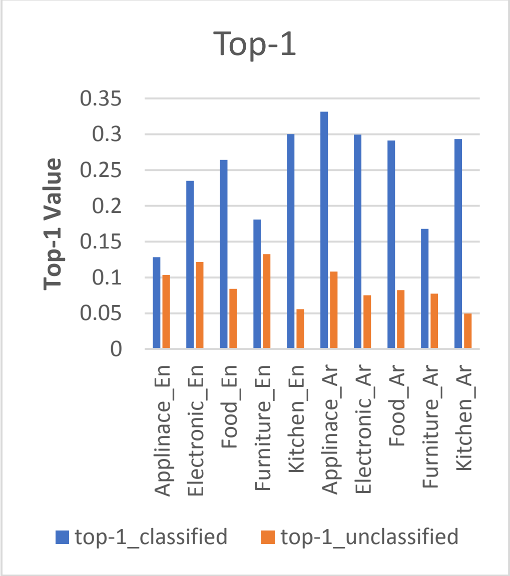
BLEU of classified and unclassified cases of outdoor non-repeating classes

Figure 4. Performance evaluation of English/Arabic Repeating Non-Repeating outdoor classes of the first scenario (word-by- word)

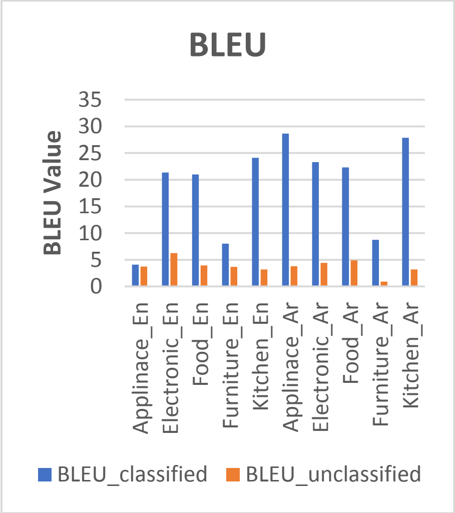
Sentence-By- Sentence Scenario test results:

Repeating vs non-repeating indoor classes Arabic and English description results:

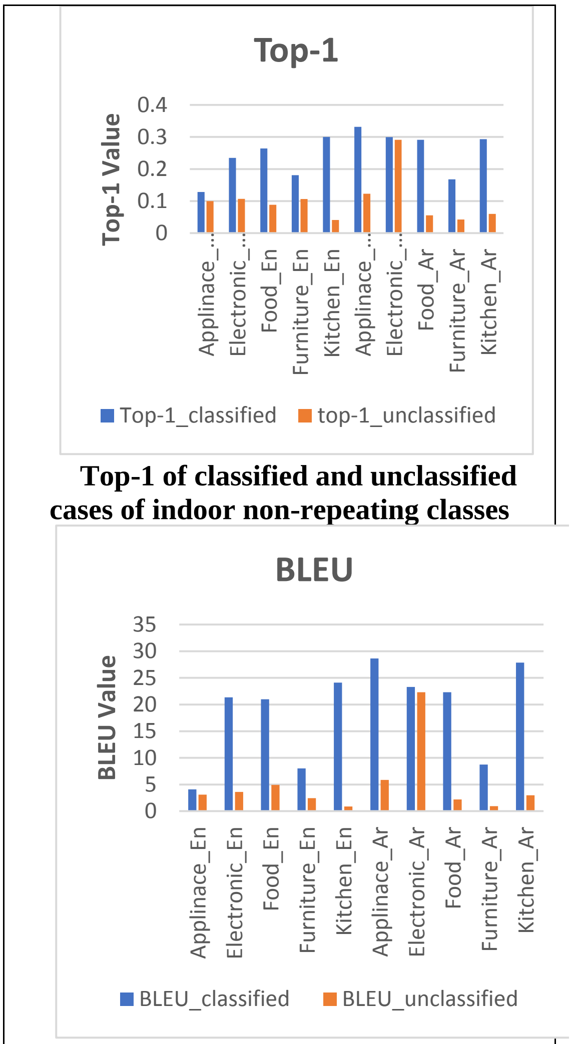
Figure 5 and 6 include the same comparisons of previous (word-by-word) model, but here for the second model (sentence-by-sentence).



Top-1 of classified and unclassified cases of indoor repeating classes



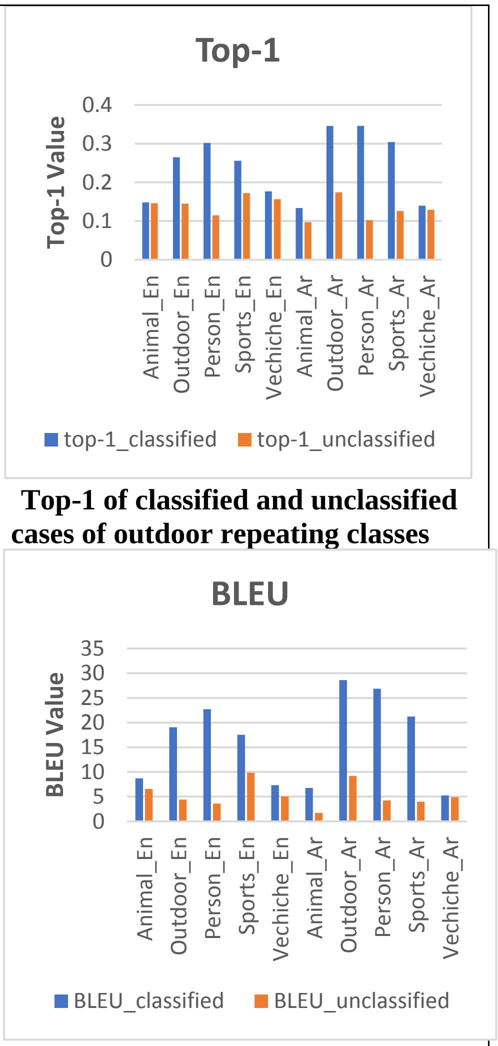
BLEU of classified and unclassified cases of indoor repeating classes



Top-1 of classified and unclassified cases of indoor non-repeating classes

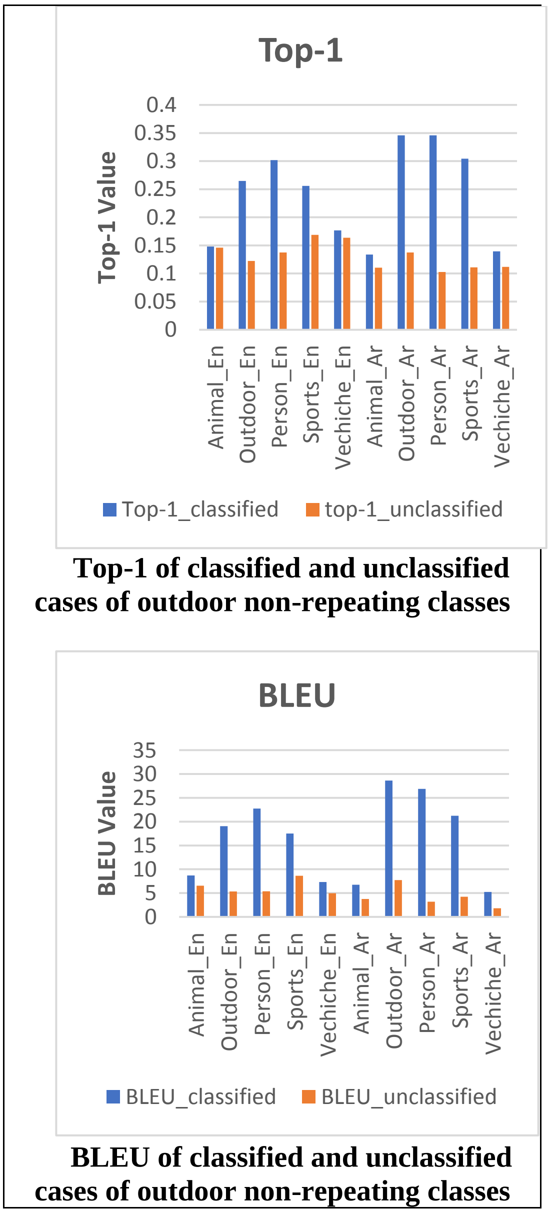
BLEU of classified and unclassified cases of indoor non-repeating classes

Figure 5. Performance evaluation of English/Arabic Repeating Non-Repeating indoor classes of the second scenario (sentence-by- sentence)



Top-1 of classified and unclassified cases of outdoor repeating classes

BLEU of classified and unclassified cases of outdoor repeating classes



Top-1 of classified and unclassified cases of outdoor non-repeating classes

BLEU of classified and unclassified cases of outdoor non-repeating classes

Figure 6. Performance evaluation of English/Arabic Repeating Non-Repeating outdoor classes of the second scenario (sentence-by- sentence)

VI. Discussion

The results show that the performance of classified datasets exceeds the performance of the unclassified ones in terms of Top-1 and BLEU. The results indicate that when the dataset is not classified, the performance was bad, while by classifying images, the performance was better. So, classifying the images as preprocessing phase before using as input data in the image captioning system will increase the performance. This conclusion is right in case of indoor and outdoor classes. The results also show that the best performance of captioning systems is corresponded to the "food" and "person" classes of the English-based indoor and outdoor categories for the sentence-by-sentence and word-by-word models respectively. However, the classes "appliance" and "outdoor" are the best performance of outdoor Arabic-based categories for all scenarios.

By comparing the results of repeating and non-repeating cases, we find a little bit changes in the results of the classified datasets. In some cases, the classified repeating classes have higher performance than the corresponding non-repeating ones, while the opposite is happened in some other cases. This change is because of fact that the repeating effects appears clearly in classified datasets (small sizes) compared to the larger unclassified ones.

The results also indicate that all experiments of the second scenario exceed the performance of the corresponding ones of the first scenario (word-by-word), and this is due to the fact that when we classified the images, the description became similar, so the classified description model will perform better than the unclassified one. So the similarity of the images inside the same class will result in a good description. However, for the (word-by-word) model, the description sentence is generated by the correlation of description words which is depended on the pretrained language model.

In some categories like animals and electrical devices, the performance of the unclassified and classified models is almost the same and this is because of the fact that these categories don't exist alone in dataset's images (animal could be found together with persons, kitchen, garden or street). On the other hand, the description of images including person category based on other categories will be very difficult so that person can be repeated in more than on class.

Because the word-word model depends on joining of words with each other, and in the case that there are common image between the classes, inappropriate word-joining may be generated for some classes, which leads to generate captioning sentences composed of words different from the captioning sentences included in the test images set. This reduces the evaluation parameter TOP1-similarity

In general, the results showed that despite the small size of the classified data set used in the training process, the performance was better

VII. Conclusion

In this research, we build a global description dataset based on datasets (flickr2k and MS-COCO), consisting of 12000 images with their description files for the English and Arabic languages. The designed description models consist of two different models; the first is the image description model (CNN for the first scenario and ResNet for the second one) while the text description model is (LSTM for the first scenario and FastText for the second one). Therefore, the training scenarios are applied once for the word-by-word model and another once for sentence-by-sentence model. The tests are applied across four different methods for each scenario. First two methods are applied on repeating and non-repeating indoor Unclassified/Classified dataset while the other two ones are applied on the repeating and non-repeating outdoor Unclassified/Classified dataset. Results indicates the transcendence of the classified dataset performance against unclassified one in all scenarios and for all methods. This prove the fact that classifying the images before using as input data in the image captioning model is the best choice. The results also show that the performance of the (sentence-by-sentence) model is higher than the (word-by-word) model.

In the future work, we will add a preprocessing phase to the image captioning system, in which the images will be classified into subclasses in order to direct them to the suitable pretrained classified model.

References

[1] Alex K, Sutskever I, and Hinton G. 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 1097-1105.

[2] Christian S, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, and Rabinovich A. 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9.

[3]FastTextmodel, https://FastText.cc/doc/en/crawl-vectors/html, , last access at 20/12/2019.

[4] Flickr dataset, https://www.flickr.com/photos/tags/dataset/, last access at 1/9/2019.

[5] Geremy H, and Koller D. 2008. Learning spatial context: Using stuff to find things. In European conference on computer vision, pp. 30-43. Springer, Berlin, Heidelberg,.

[6] Holger C, Uijlings J, and Ferrari V. 2018. Coco-stuff: Thing and stuff classes in context In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1209-1218.

[7] Kaiming H, Zhang X, Ren S, and Sun J. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.

[8] Karen S, and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[9] Madhavan S, Srikanth M, and Belov M. 2020. Image to Language Understanding: Captioning approach. arXiv preprint arXiv:2002.09536.

[10] Micah H, Young P, and Hockenmaier J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research 47, 853-899.

[11] MS-COCO Dataset, https://COCOdatset.org , last access at 1/9/2019.

[12] Munawar H, Khan S, Bennamoun M, and An S. 2016. A spatial layout and scale invariant feature representation for indoor scene

classification. IEEE Transactions on Image Processing 25, no. 10 (2016): 4829-4841.

[13] Neeru N, Martin M, Metaxas D, and Bourlai T. 2017. Learning deep features for hierarchical classification of mobile phone face datasets in heterogeneous environments. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 186-193. IEEE.

[14] Panagiotis M, Wen X, Lu C, Geus D, and Dubbelman G. 2020. Cityscapes-Panoptic-Parts and PASCAL-Panoptic-Parts datasets for Scene Understanding. arXiv preprint arXiv:2004.07944.

[15] Rasha M, and Alkheir J. 2018. Development of an Arabic Image Description System, International Journal of Computer Science Trends and Technology (IJCST), 6(3), 205-213.

[16] Rasha M, and Alkheir J. 2019. Performance Evaluation of Image Description Systems Based on Different Deep Learning Models, Tishreen University Journal for scientific studies and researches, Engineering series, Vol.41, No.2.

[17] Rasha M, and Alkheir J, Samer S. 2020. Performance Evaluation on the Effect of Different Text Representation Models on the Image Captioning Systems, Tishreen University Journal for scientific studies and researches, Engineering series, Vol.42, No.4.

[18] Rasha M, and Alkheir J, Samer S. 2020. Evaluating the impact of different languages on the performance of Image Captioning Systems, Aleppo University Journal for scientific studies and researches, Engineering series, No.157.

[19] Sebastián S, Fabio G. 2018. Combining textual and visual representations for multimodal author profiling. Working Notes Papers of the CLEF. 2018; Published in CLEF, Computer Science 2125:219-28.

[20] Sepp H, Jãijrgen S. 1997. Long short-term Memory; Neural Computation 9, 8 (1997), 1735–1780.

[21] Sumanth R, Ikram S, and Ramesh P. 2020. Deep Learning System for Image Retrieval. In 2020 Indo–Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN), pp. 119-124. IEEE.

[22] Tsung-Yi L, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, and Zitnick L. 2014. Microsoft coco: Common objects in context. In European conference on computer vision, pp. 740-755. Springer, Cham.

[23] Wenjie C, Xiong Z, Sun X, Rosin P, Jin L, and Peng X. 2020. Panoptic Segmentation-Based Attention for Image Captioning. Applied Sciences 10, no. 1 (2020): 391.

[24] Wonmin B, Breuel T, Raue F, and Liwicki M. 2015. "Scene labeling with lstm recurrent neural networks." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3547-3555.

[25] Yann L, Bottou L, Bengio Y, and Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, no. 11 (1998): 2278-2324.

[26] Zakir H, Ferdous S, Mohd S, Hamid L. 2019. A comprehensive survey of deep learning for image captioning; ACM Computing Surveys (CSUR); Article No: 118; 51(6):1-36.

Received	2021/6/22	إيداع البحث
Accepted for Publ.	2021/7/14	قبول البحث للنشر