Journal of Engineering Sciences and Information Technology
Volume (6), Issue (1): 30 Mar 2022
P: 150 - 164

ISSN: 2522- 3321

مجلة العلوم الهندسية وتكنولوجيا المعلومات
المجلد (6)، العدد (1): 30 مارس 2022م
ص: 150 - 164

Improving The Performance of the Image Captioning Systems Using a Pre- Classification Stage

Rasha Mohammed Mualla
Mechanical & Electrical Engineering Faculty || Tishreen University || Syria
Jafar Alkheir
Artificial Intelligence, Computer Science || Tishreen University || Syria
Samer Sulaiman
Faculty Engineering || Al- Manara University || Syria

Abstract: In this research, we introduce a novel image classification and captioning system by adding a classification layer before the image captioning models. The suggested approach consists of three main steps and inspired by the state- of- art that generating image captioning inside small sub- classes categories is better than the unclassified large dataset. In the first one, we have collected a dataset of two international datasets (MS- COCO and Flickr2k) including 10778 images in which 80% is used for training and 20% for validation. In the next step, dataset images have been classified into 11 classes (10 classes of indoor and outdoor categories and one class of "Null" category) and fed into a deep learning classifier. The classifier is re- trained again using our classes and learned to classify each image to the corresponding category. At the final step, each classified image is used as input of 11 pre- trained classified image captioning models, and the final captioning sentence is generated. The experiments show that adding the pre- classification step before the image captioning stage improves the performance significantly by (8.15% and 8.44%) and (12.7407% and 16.7048%) for Top- 1 and Top- 5 of English and Arabic systems respectively. The classification step achieves a true classification rate of 71.32% and 73.09% for English and Arabic systems respectively.

Keywords: Deep Learning, Natural Language Processing, Arabic Language Image Captioning, English Language Image Captioning, Image Classification, Image captioning.

تحسين أداء أنظمة وصف الصور باستخدام مرحلة التصنيف المسبق للصور

رشا محمد معلا
كلية الهندسة الميكانيكية والكهربائية || جامعة تشرين || سوريا
جعفر الخير
كلية الهندسة المعلوماتية || جامعة تشرين || سوريا
سامر سليمان
كلية الهندسة || جامعة المنارة || سوريا

المستخلص: في هذا البحث قدمنا نظامًا جديدًا لتصنيف الصور ووصفها عن طريق إضافة طبقة تصنيف قبل نماذج وصف الصور بالتسميات التوضيحية. يتكون النهج المقترح من ثلاث خطوات رئيسية ومستوحاة من أحدث التقنيات التي تعتبر توليد تسميات

توضيحية للصور ضمن مجموعات الصور الفرعية الصغيرة المصنفة أفضل منه في مجموعة البيانات الكبيرة غير المصنفة. في الخطوة الأولى، قمنا بتجميع مجموعة بيانات صور من مجموعتي بيانات معياريتين (MS- COCO وFlickr2k) وتتضمن هذه المجموعة 10778 صورة تم تقسيمها إلى 80٪ للتدريب و20٪ للتحقق من الصحة. في الخطوة التالية، تم تصنيف صور مجموعة البيانات إلى 11 فئة (10 فئات من الفئات الداخلية والخارجية وفئة واحدة من فئة Null) وتم إدخالها في مصنف يعتمد مبدأ التعلم العميق. يتم إعادة تدريب المصنف مرة أخرى باستخدام أصناف الصور العشرة التي تم تشكيلها، بحيث يتعلم تصنيف كل صورة إلى الفئة المقابلة لها. في الخطوة الأخيرة، يتم استخدام كل صورة مصنفة كمدخلات لـ 11 نموذج مصنف ومدرب مسبقاً لوصف الصور، بحيث يتم إنشاء جملة التسمية التوضيحية النهائية للصورة. أظهرت التجارب أن إضافة خطوة التصنيف المسبق قبل مرحلة توليد التسميات التوضيحية للصورة تحسن الأداء بشكل ملحوظ بنسبة (8.15٪ و8.44٪) و (12.7407٪ و16.7048٪) من حيث معياري Top- 1 وTop- 5 لأنظمة الوصف التي تعتمد اللغة الإنجليزية والعربية على التوالي. حققت خطوة التصنيف معدلات تصنيف حقيقية بلغت 71.32٪ و73.09٪ للأنظمة الإنجليزية والعربية على التوالي.

الكلمات المفتاحية: التعلم العميق، معالجة اللغات الطبيعية، وصف الصور باللغة العربية، وصف الصور باللغة الإنكليزية، تصنيف الصور، وصف الصور بتسميات توضيحية.

Introduction.

Deep learning DL had been recently used in many computer science applications. One of those essential applications is image captioning in which a description sentence, in a natural language, of each input image, is generated (Zakir, Sohel, Shiratuddin, & Laga, 2019) (Oluwasammi et al., 2021) based on the extracted features and component relationships of that image (Katiyar & Borgohain, 2021). Image captioning, as one of computer vision and natural language processing branches (Simao, Kappeler, Boakye, & Soares, 2019), has many applications such as image retrieval and indexing besides many other usages like transferring image objects into words (Simao, Kappeler, Boakye, & Soares, 2019) minimizing size for transferring and storing.

Many deep learning neural networks were used for the image captioning task to build the image model, including the Convolutional Neural Networks (CNN) (Kalchbrenner, Grefenstette, & Blunsom, 2014) (Aneja, Deshpande, & Schwing, 2018), AlexNet (Alex, Sutskever, & Hinton, 2012), VGG (Karen & Zisserman, 2014), ResNet (Kaiming, Zhang, Ren, & Sun, 2016), GoogleNet (Szegedy, et al.), etc. On the other hand, many text models (language model) were used like Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM) (Wonmin, Breuel, Raue, & Liwicki, 2015), GloVe and FastText (Mualla, Alkheir, & Sulieman, 2020), etc.

Most of these image and text models are built based on specific datasets like Flickr and MS-COCO. Microsoft Common Objects in Context (COCO) dataset (Tsung- Yi, et al., 2014) is one of the most common image datasets including more than 82783 images for training and 40504 images for validation. It consists of 80 classes including indoor and outdoor scenes. On the other hand, the Flickr2k dataset is a standard sentence- based partial dataset of the entire Flickr8k dataset (Micah, Young, & Hockenmaier, 2013), used for image description consisting of more than 8000 images besides the image- description file.

Study Problem:

The research problem can be described as follows:

1- Image captioning is a big challenge, especially in the case of a global unclassified dataset.
2- Image captioning has many limitations like the computational time required and the dependency on the used natural Languages and the class of images as well.
3- Global unclassified dataset increases the captioning time and decreases the accuracy.

Study Objectives: can be summarized as follows:

1- Enhancing the performance of the Image captioning system using a pre- classification stage.
2- Improve the performance using classified sub- datasets instead of a global unclassified dataset.

Study Importance:

The scientific importance of the study arises from the idea that pre- classification step inserted before the image captioning system can improve the performance significantly. Image captioning systems can be used for many applications like virtual assistant, image to voice applications, image indexing, assistant apps for impaired people, etc.

Study Methodology.

A. Analysis methodology: we suggested using the descriptive statistical analysis depending on the data summary of tables, charts and figures.
B. Data sources: We relied on two different international datasets (MS- COCO and Flickr2K).
C. Study Boarders: We used a pre- classification layer before 10 pre- trained captioning models. The input of our system is the image to be described while the output is the full description of that image in Arabic and English languages.

Study Architecture:

This study had been divided into two topics; the first one deals with the theoretical part of the study and the previous work. While the second one discusses our suggested image classification and captioning methodology besides the results and discussion of the developed pre- classification image captioning models.

The first topic- the theoretical framework and previous studies.

Theoretical framework:

Three basic captioning methods have been introduced in the literature; the first one is template-based image captioning in which specific templates (objects, attributes, actions, etc.) with a number of

blank spaces between them are detected. Le et al. (Li, Kulkarni, Berg, Berg, & Choi, 2011), for example, detected parts of sentences related to the extracted objects and their relationships in the scene. Although this method can generate syntactically correct captions, the predefined templates cannot produce variable- length captions. The second captioning method is the retrieval- based in which captions are generated using a set of existing captions (Hossain, Sohel, Shiratuddin, & Laga, 2019). The entire idea is based on training the image captioning model to detect the visually similar images and their corresponding candidate captions of a specific dataset. However, this method generates grammatically correct captions, but not semantically correct ones.

The most accurate image captioning method is Novel- based captioning. In this method, the image model is separated from the language model so that the visual features of the image are extracted first (for example, by CNN). Then the captions are generated using a language model (for example, RNN) (Sargar & Kinger, 2021).

There are different factors affect the image captioning process, such as the sentence captioning generation method, dataset size and type, image categories and description language (Dao, Nguyen, & Bressan, 2016).

All previous studies introduced the problem of scene classification and image captioning independently or as separated models. Only one previous study dealt with the problem of classification of image datasets but in a manual way (Mualla, Alkheir, & Sulaiman, 2021).

Classification of dataset is an essential process which can be done before the image captioning so that the captioning process can be done in pre- classified sub- datasets instead of the entire dataset (Mualla, Alkheir, & Sulaiman, 2021). So, dealing with many small sub- classes datasets to generate image captioning is much more effective than generating captioning through an entire huge dataset (including all those sub- classes together).

The following sections will be organized as follows. First, we will list the related works, then the materials and methods used in this paper will be explained. After that, the results will be introduced and discussed in detail. Finally, we will conclude our work.

Second- Related work:

Yoon et al. (Yoon, Park, Park, & Lim, 2019) introduced an image description model for the consideration of CAM- based disagreement loss. They use multimodal embedding representing the textual and image features as well as an intermediate layer for joint learning. CNN is used for the image model while the LSTM is used as a text model. Two symbolic datasets are used which are the Oxford Flowers 102 and the Caltech UCSD Birds 200–2011. They evaluate the performance using BLEU, METEOR and CIDErn metrics. The results indicate that the intermediate model achieved enhancement by 12.43% and 3.33% in the CIDEr for the Birds and Flower datasets respectively.

Eljundi et al. (ElJundi, Dhaybi, Mokadam, Haj, & Asmar., 2020) in 2019 designed two models for the captioning of the components of the image in Arabic (translated from English, and a full description model in Arabic), in which they relied on the VGG- 16 model to extract the features of the image, and on the RNN- LSTM network for the text model. The researchers used the Flickr8k dataset to implement the model, as the training process took 125 seconds for every 5 iterations, while the system achieved a score of 0.5 on the L- gram Bleu benchmark. Also, in 2019, the researcher Zakraoui et al. (Zakraoui, Elloumi, Alja'am, & Yahia, 2019) developed a model for image captioning in which they linked Arabic texts with images based on machine learning and deep learning techniques, for this part, they used CNN for the image captioning model and LSTM for the text model. For the training and testing process, a special data set of images was used in addition to images from the GoogleNet dataset. The research reached an accuracy of 56% for the image captioning process.

For indoor- outdoor scene classification, a lot of researches have been introduced. In 2016, Shahriari et al (Shahriari & Bergevin, 2016) proposed a hierarchical classification approach based on two stages. In the first stage, the image features are generated, while in the second one, the outdoor classes are categorized in several categories against only one indoor class. The experiments were applied on two datasets (15- Scene category and SUN397) and the result accuracies were (97.84% and 55.1%) and (93.09% and 92.4%) for outdoor and indoor classes in the case of 15- Scene and SUN397 datasets respectively.

Kumari et al. (Kumari, Jha, Bhavsar, & Nigam, 2020) used ResNet as an image model to classify indoor- outdoor scenes. They evaluated their approach using different datasets including indoor and outdoor scenes, their results indicated an ideal performance in many cases.

Another study by Neduchal et al (Neduchal, Gruber, & Železný, 2020) introduced the task of indoor and outdoor classification in the field of mobile robotic. They applied many traditional machine learning algorithms (SVM, K- NN, decision trees and Naïve Bayes) and different color and texture feature extraction approaches. They concluded that the best algorithm achieved 96.17% accuracy.

In 2018, Mualla et al (Mualla & Alkheir, 2018) introduced a word- by- word captioning system based on CNN as an image model and LSTM as a text model. In this research, they study the effect of different languages on the performance of the image captioning systems. They built both Arabic and English captioning systems of images from the Flickr2k dataset including 1500 images for training, 250 for validation and 250 for test. They got 51.5 and 34.4 as a BLEU- 1 evaluation metrics for the English and Arabic systems respectively. They continued their work in 2019 (Mualla & Alkheir, 2019) and studied the effect of using different image captioning models on the performance of captioning systems. They used different captioning models like VGG16, VGG19 and ResNet50. They concluded that the ResNet50 model exceeded both models VGG16 and VGG19 in terms of accuracy. In recent research of Mualla et al. (Mualla, Alkheir, & Sulaiman, 2020), they study the effect of using different text representation models on

the performance of captioning systems. They used different text models including Glove and FastText in case of sentence- by- sentence captioning systems. They used 10000 images of MS- COCO datasets. Their results indicate a superiority of systems based on FastText models against the corresponding GloVe models. A performance comparison between two types of image captioning systems using different languages had been introduced by Mualla et al. (Mualla, Alkheir, & Sulieman, 2020). The first model depends on generating a captioning of images (word- word) using CNN and LSTM, while the other generates the captioning in a way (sentence- sentence) based on ResNet50 and FastText, based on Flickr and MS- COCO datasets. The results proved that the English captioning 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.

Recently in 2021, Mualla et al. (Mualla, Alkheir, & Sulaiman, 2021) continued their work by studying the effect of the image's classes on the performance of image captioning models. They combined two datasets to configure a unified 12000 image dataset, for both English and Arabic languages. They used two scenarios. In the first one, they used CNN and LSTM while for the second, they applied ResNet50 and FastText for the image and text model respectively. Experiments were applied in two cases and four ways (word by word, sentence by sentence, repeated and non- repeated dataset) and proved that the classified models had a better performance against the unclassified ones in all cases.

Second topic- The pre- classification based image captioning models.

Materials and methods:

The current study is the first one that uses a pre- classification model before the image captioning in order to improve the accuracy of captioning systems. To achieve that, we develop the model used in (Mualla, Alkheir, & Sulaiman, 2021) by implementing a fully automated image classification as a preprocessing stage in the captioning system. scene classification is firstly done by training an image classifier to identify the appropriate class of a given image (Zhang, 2019). The basic idea of our proposed model is to find out the suitable pre- trained image captioning model automatically by using an image classifier. Thereafter, the input image will be described based only on the predicted class of the input image. That mean, each input image will be described by a classified captioning model (Mualla, Alkheir, & Sulaiman, 2021). based on the predicted class.

Two steps are applied for the pre- classification and captioning system. In the first step, the input image is classified into the appropriate class within 10 different categories using a pre- trained object classifier based on deep learning (Zhang, 2019). While in the second step, the image captioning process is performed using pre- trained models of a previous study (Mualla, Alkheir, & Sulaiman, 2021). Figure 1 illustrates the general system diagram.

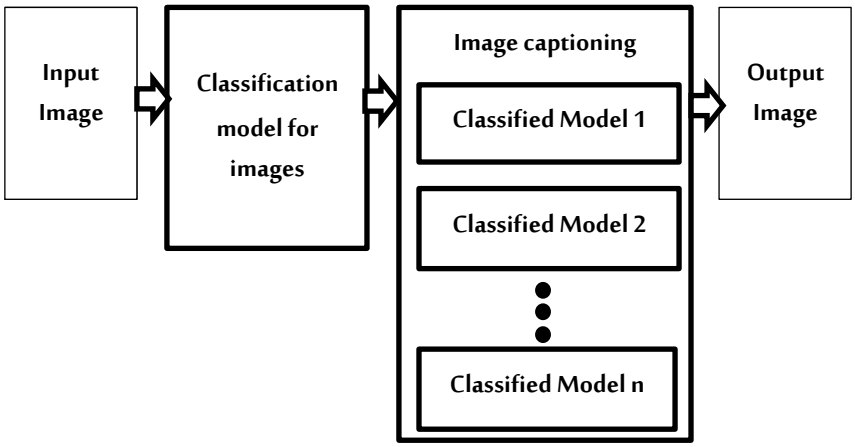


Fig (1) General System Stages

The following Hardware and software tools are used in order to perform the classification and captioning processes: computer with windows 10- 64 bit, 8 GB of RAM, 1 TB of HDD, Nvidia GeForce GTX 1050 and Core- i7 processor; Python 3.6, tensorflow 2.0.0, jupyter notebook, and the Anaconda platform. The research has been done within six months, from 10/1/2021 until 10/6/2021, in the laboratories of the University.

Dataset:

Two different datasets (MS- COCO Val 2014 and Flickr2k) available at (Flickr, n.d.) (Microsoft, n.d.) are used to build our dataset which consists of 10778 images. We try to choose natural images, which unlike symbolic ones composed of several elements. Because of that, it was possible to have some images in more than one category. A detailed explanation of the training, validation and test datasets of the indoor and outdoor categories are listed in tables 1 and 2 respectively.

The indoor category consists of 3843 images including 3072 for training and 771 for validation. On the other hand, the outdoor category has 6935 images including 5547 for training and 1388 for validation. This results in 80% of the entire collected dataset used for training while 20% used for validation.

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

Class	Total count	Train Count	Val count	Test count
Appliance	585	468	117	117
Electronic	606	484	122	121
Food	700	560	140	140
Furniture	1326	1060	266	265
Kitchen	626	500	126	125

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

Class	Total count	Train Count	Val count	Test count
Animal	2480	1984	496	496
Outdoor	715	572	143	143
Person	914	731	183	182
Sports	822	657	165	164
Vehicles	2004	1603	401	400

Classification model:

For the classification stage, we used the deep learning model proposed in (Zhang, 2019). We re-train it using our newly collected dataset. The proposed model consists of three layers, which are the MobileNet model layer, the dropout layer and the dense layer.

MobileNet V2 model uses both inverted residual connections and bottleneck convolutional layers. It consists of 32 filter- based convolutional layers and 19 residual- bottleneck layers; moreover, the training stage uses RELU as a nonlinear activation function, 3x3 kernel size and batch normalization process (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018). On the other hand, the dropout layer is used to avoid the network's overfitting, while the dense layer will reduce the size of the features using the softmax activation function for the last layer of a classification network.

The classification model uses MobileNet V2 to classify images based on the following steps:

- Define image dataset that will be used for the training process.
- Configuring the training pipeline using the transfer learning principle which is based on taking a model trained using different dataset for some different purpose and then modify it to fit our dataset and our task (in our case: we used a trained model for classifying objects captured by mobile camera then transfer the learning into our goal which is classifying the classes of indoor/outdoor scenes).
- Training the model.

We obtained and re- trained this classifier model based on these 11 different categories (classes) which are "Appliance", "Electronic", "Food", "Furniture", "Kitchen", "Animal", "Outdoor", "Person", "Sports", and "Vehicles". Another class is called "Null" representing the case when the classifier fails to find out the suitable class of the input images. As a result, we got a new classifier of our collected dataset so that for any input test image, the classifier will match it to one of the ten classes, otherwise to "Null".

Proposed Image Captioning Model:

As mentioned above, our proposed model is composed of two stages. The input image will be first classified via the classifier, in order to determine the suitable classified image captioning model. Each

classified pre- trained model consists of an image representation model based on ResNet50 and a text representation model based on FastText. Its captioning method is based on a sentence- by- sentence model (Mualla, Alkheir, & Sulieman, 2020). ResNet50 network is an image representation model consists of 50 convolutional layers and residual units perform as identity connections. Besides that, there is a dense layer used to reduce the image feature vector into 256 elements, in order to be fused with the 256 text vector. FastText is a pretrained text model which consists of word representation layer. It selects the corresponding weights of words used in the captioning file and produces a 16- embedding vector. These vectors will be used as input for gated recurrent units producing 256 word- vectors (the text representation of image). The image and text models are fused using the dot product operation that trains the model to match a set of true captioning sentences with their corresponding image vectors, as well as to define a set of false captioning sentences along with the same image vectors. So that this captioning model will be trained to match the correct image and text captioning, and at the same time reject the match between the false captioning and image vectors. These pretrained models are selected based on two factors; the first one is the performance, and the second one is the dataset size. Mualla et al (Mualla, Alkheir, & Sulieman, 2020) concluded that the best model combination that performs the best in case of large dataset size (MS- COCO) is the ResNet50 as an image model and FastText as a text model.

Performance Evaluation Metrics

To evaluate our proposed model, we used the following performance metrics, the first four are for the captioning evaluation while the final one is for the classification evaluation.

Top- 1 Similarity criterion, which expresses the similarity of the best resulting description of the image with the original actual description. To calculate this criterion, the arithmetic means value of all similarity values is taken for all the entered test samples (Shankar, et al., 2020).

Top- 5 Similarity criterion, which represents the similarity of the original sentence describing the image with each of the five resulting sentences of best predictions and taking a higher value similarity (Shankar, et al., 2020).

Distribution of Best Depth Description (DBDD) is used to define the ratio of being each of the five generating description sentences the best description based on the similarity with the original one.

Bilingual Evaluation understudy (BLEU), which is an evaluation of the accuracy of the resulting description, so that if the description resulting from the captioning process is very close to the description used in the training, the BLEU standard will give a high value; otherwise, it will be low (Kishore, Roukos, Ward, & Zhu, 2002).

True classification rate (%) which is the percentage that represents the number of images correctly classified out of the total number of test images.

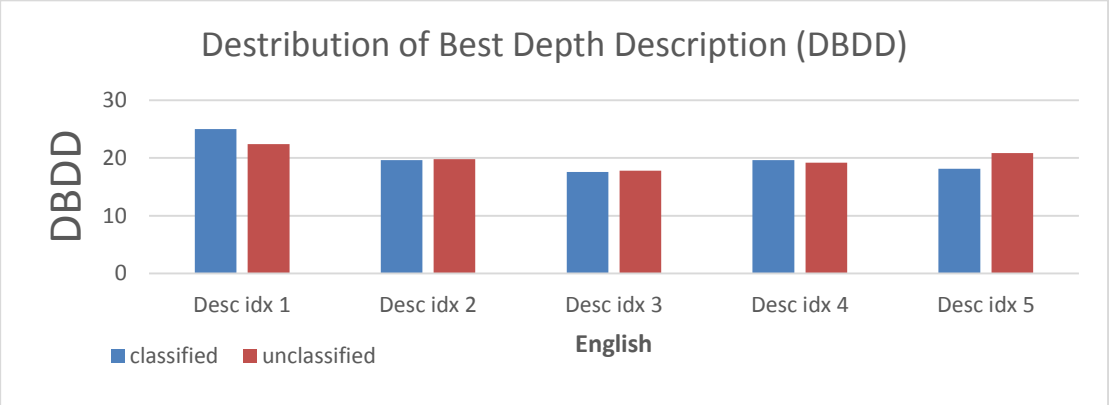
Results and Discussion.

We have selected a random subset of the images for the test stage. The test sub- dataset consists of 20% of the entire dataset used in the training stage. The total number of test images is 1904 and 1925 for English and Arabic, respectively. We performed the classification stage as a preprocessing stage in order to direct the input image to the suitable pre- trained captioning model. Therefore, the captioning output will be based on the classification output. In the end, a performance evaluation is done using metrics such as (Top1 and Top5 similarity, Depth of best similarity, BLEU and True Classification Rate%).

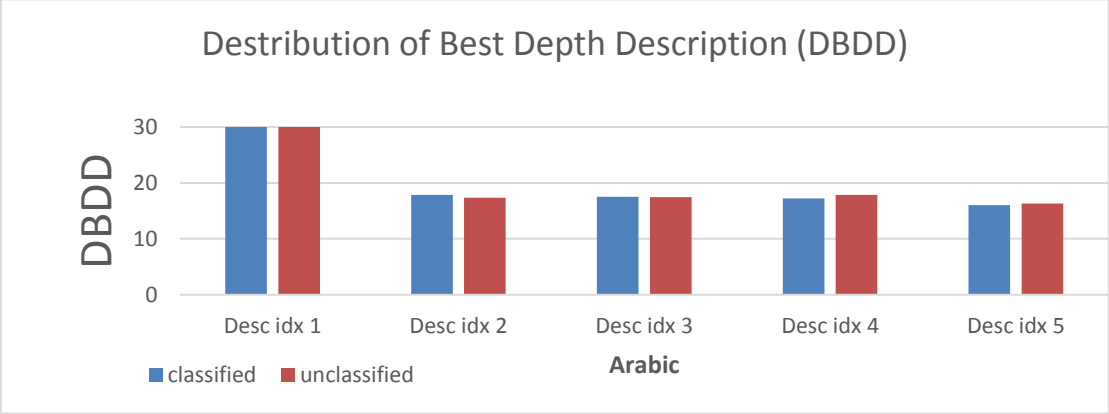
Tests Scenarios:

To evaluate our classification and captioning models, we apply the test image dataset into two scenarios. In the first one, we apply the test to the proposed classified captioning system using the proposed evaluation parameters. In the second scenario. We also computed the same metrics in the second scenario for the model without the classification stage.

Figure 2 (A and B) shows the Distribution of Best Depth Description (DBDD) resulted for both Arabic and English captioning models and for both scenarios. The results indicate that the first resulted captioning sentence is the most similar captioning sentence among the five captioning sentences for both Arabic and English- based models. We can also see that the distribution of depth description for Arabic models is more closed compared to the English models, and this is because the Arabic language has a smaller number of words within the sentence (for example, in English we can say "A man" and "The man" which are two- words sentence, one of them is the same" man", while in Arabic we say "رجل" and "الرجل" which are only one- word sentence but different words).



(A)



(B)

Fig (2) Distribution of Best Depth Description (DBDD) of test set for Arabic and English models (A) for English, (B) for Arabic.

From the "Top- 1", "Top- 5" and "BLEU" points of view, the English- based description models achieved a better performance against the Arabic ones as figures 3, 4 and 5 illustrate. This is due to the fact that dealing with grammar and word parts is easier in English than Arabic language. Figure 3 and 4 also show that the confidence interval with a level equals to 95% around the mean value of Top- 1 and Top- 5 is very low which means that the mean values of the Top- 1 and Top- 5 values can be taken as accurate values.

Comparing the classified models with the unclassified ones, we can see that the classification enhances the performance significantly by 8.15% and 8.44% for Top- 1 of English and Arabic systems respectively. Similarly, the Top- 5 is improved by 12.7407% and 16.7048% for English and Arabic systems respectively. Moreover, the BLEU accuracy increased by 35.07 and 32.79 of English and Arabic systems respectively. That means, the adding of the classification stage before the captioning stage enhances the performance significantly.

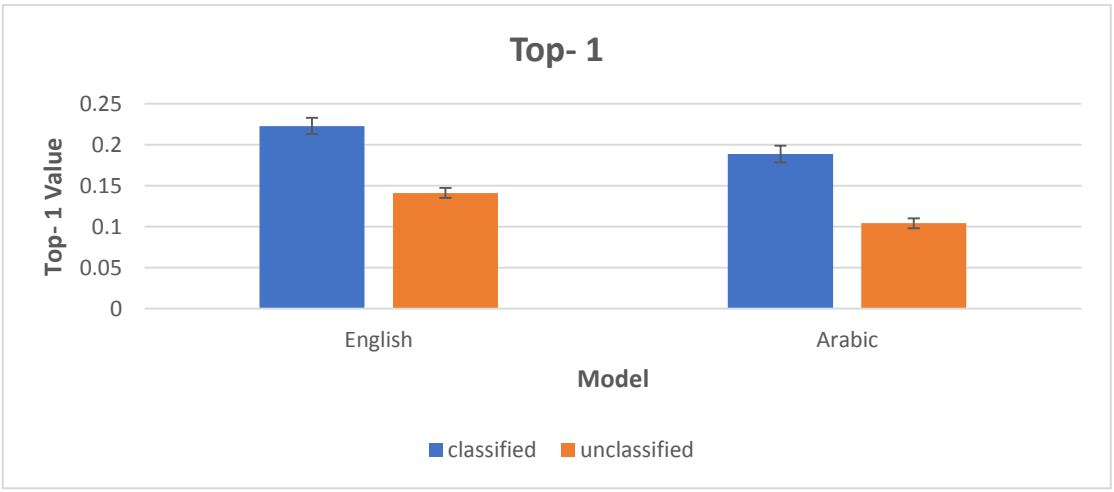


Fig (3) Top- 1 of test set for Arabic and English models

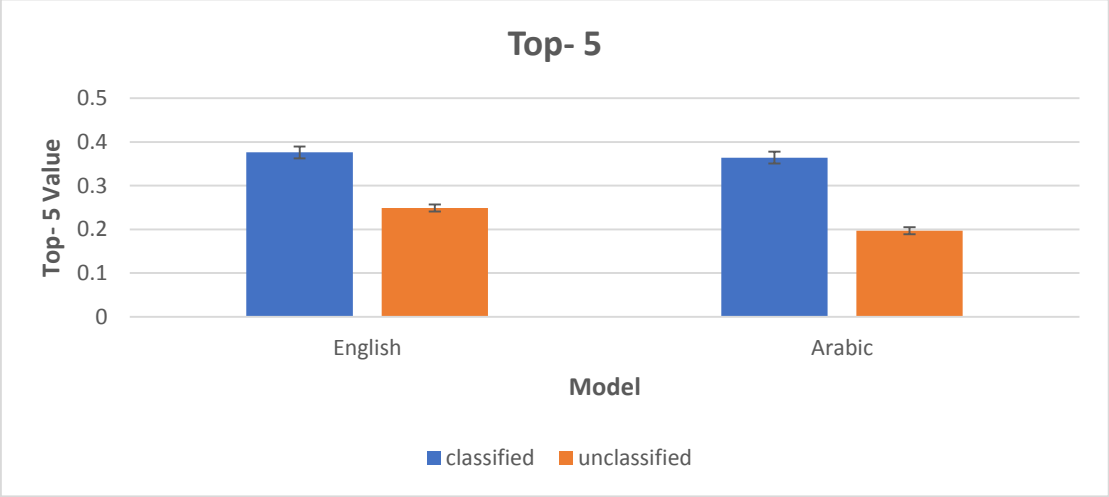


Fig (4) Top- 5 of test set for Arabic and English models

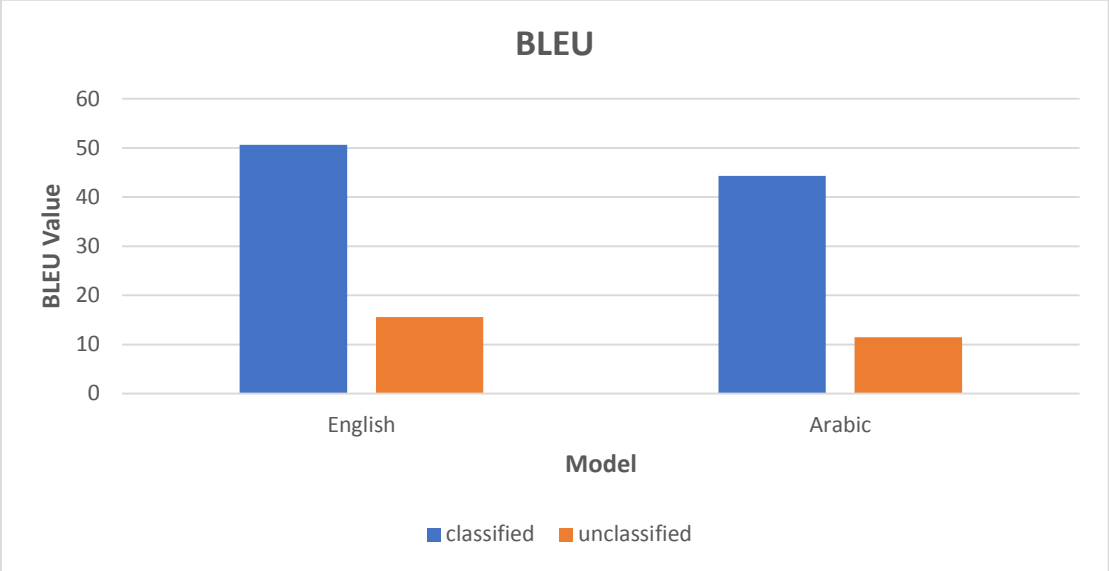


Fig (5) BLEU of test set for Arabic and English models

Testing the classification stage:

For the classification part, the true classification rates are 73.09% and 71.32% for Arabic and English datasets respectively. This classification rate affects the image description accuracy so that some test images will be incorrectly described because of the false classification. Although this limitation of the classification step, the performance of the description model has increased significantly in case of using a pre- classification stage. Therefore, the improvement of the used classifier will minimize this type of error and improve the performance.

Conclusion.

In the current research, we have proposed a novel method for the image captioning systems in Arabic and English languages. The new approach depends on adding a new classification step before the image captioning step aiming to pre- classification of the input image into the best corresponding class

including indoor and outdoor categories. The class will be used only to generate a captioning using a specific pre- trained image captioning model (which is trained previously on a specific category). This will make the image captioning based on sub- classes datasets instead of the unclassified dataset minimizing the errors and increasing the overall accuracy. The experiments are applied using a subset of MS- COCO and Flickr2k datasets. They prove the fact that the adding of pre- classification stage increases the performance significantly comparing to the same image captioning models without classification stage as in (Mualla, Alkheir, & Sulieman, 2020). We can conclude that the pre- classification step increased the performance significantly in terms of BLEU, Top- 1 and Top- 5 metrics.

Recommendation.

The current research recommends the following

1. Study the effect of the classification step on another scenario of the image description model (word- by- word).
2. Compare its results with the result of the current (sentence- by- sentence) scenario.
3. Enhancement of the classification step in order to improve the performance.

References.

- Alex, K., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25(2012), 1097- 1105.
- Aneja, J., Deshpande, A., & Schwing, A. (2018). Convolutional image captioning. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5561- 5570). Salt Lake, Utah: IEEE.
- Dao, D.- C., Nguyen, T.- O., & Bressan, S. (2016). Factors Influencing The Performance of Image Captioning Model: An Evaluation. Proceedings of the 14th International Conference on Advances in Mobile Computing and Multi Media, (pp. 235- 243).
- ElJundi, O., Dhaybi, M., Mokadam, K., Haj, H., & Asmar., D. (2020). Resources and End- to- End Neural Network Models for Arabic Image Captioning. VISIGRAPP, 233- 241.
- Flickr. (n.d.). Flickr dataset. (Flickr) Retrieved 2020, from https: //www.Flickr.com/photos/ tags/dataset/
- Hossain, Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR), 51(6), 1- 36.
- Kaiming, H., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770- 778). San Juan, PR, USA: IEEE.
- Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). convolutional neural network for modelling sentences. arXiv preprint arXiv: 1404.2188.

- Karen, S., & Zisserman, A. (2014). Very deep convolutional networks for large- scale image recognition. arXiv preprint arXiv: 1409.1556.
- Katiyar, S., & Borgohain, S. K. (2021). Image Captioning using Deep Stacked LSTMs, Contextual Word Embeddings and Data Augmentation. arXiv preprint arXiv: 2102.11237.
- Kishore, P., Roukos, S., Ward, T., & Zhu, W.- J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. 40th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 311- 318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.
- Kumari, S., Jha, R. R., Bhavsar, A., & Nigam, A. (2020). Indoor–outdoor scene classification with residual convolutional neural network. In Springer (Ed.), Proceedings of 3rd International Conference on Computer Vision and Image Processing, (pp. 325- 337). Singapore.
- Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., & Choi, Y. (2011). Composing simple image descriptions using web- scale n- grams. Proceedings of the Fifteenth Conference on Computational Natural Language, (pp. 220–228).
- Micah, H., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47(2013), 853- 899.
- Microsoft. (n.d.). COCOdatset. (Microsoft) Retrieved 2020, from https: //COCOdatset.org
- Mualla, R., & Alkheir, J. (2018). Development of an Arabic Image Description System. International Journal of Computer Science Trends and Technology, 6(3), 205- 213.
- Mualla, R., & 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, 41(2).
- Mualla, R., Alkheir, J., & Sulaiman, S. (2020). Performance Evaluation on the Effect of Different Text Representation Models on the Image Captioning Systems. Tishreen University Journal.
- Mualla, R., Alkheir, J., & Sulaiman, S. (2021). Image Pre- classification to improve the accuracy of the image captioning systems. Damascus university journal.
- Mualla, R., Alkheir, J., & Sulieman, 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, 157.
- Neduchal, P., Gruber, I., & Železný, M. (2020). Indoor vs. Outdoor Scene Classification for Mobile Robots. In Springer (Ed.), International Conference on Interactive Collaborative Robotics (pp. 243- 252). Petersburg: Springer.
- Oluwasammi, A., Aftab, M. U., Qin, Z., Ngo, S. T., Doan, T. V., Nguyen, S. B., Nguyen, H. S. & Nguyen, G. H. (2021). Features to Text: A Comprehensive Survey of Deep Learning on Semantic Segmentation and Image Captioning. Complexity, 2021.

- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.- C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510- 4520). IEEE.
- Sargar, O., & Kinger, S. (2021). Image Captioning Methods and Metrics. 2021 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 522- 526). IEEE.
- Shahriari, M., & Bergevin, R. (2016). A two- stage outdoor – indoor scene classification. In IEEE (Ed.), 2016 International Conference on Digital Image Computing: Techniques and Applications (pp. 1- 8). Gold Coast, QLD, Australia: IEEE.
- Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., & Schmidt, L. (2020). Evaluating Machine Accuracy on ImageNet. 37th International Conference on Machine. Vienna, Austria.
- Simao, H., Kappeler, A., Boakye, K., & Soares, J. (2019). Image captioning: Transforming objects into words. arXiv preprint arXiv: 1906.05963, 1- 11.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D.,... Rabinovich, A. (n.d.).
- Tsung- Yi, L., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D.,... Zitnick, L. (2014). Microsoft coco: Common objects in context. European conference on computer vision (pp. 740- 755). Switzerland: Springer.
- Wonmin, B., Breuel, T., Raue, F., & Liwicki, M. (2015). Scene labeling with lstm recurrent neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3547- 3555). Boston, MA, USA: IEEE.
- Yoon, Y., Park, S., Park, S., & Lim, H. (2019). Image classification and captioning model considering a CAM‐based disagreement loss. ETRI Journal, 42(1), 67–77.
- Zakir, H., Sohel, F., Shiratuddin, M., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR), 51(6), 1- 36.
- Zakraoui, J., Elloumi, S., Alja'am, J. M., & Yahia, S. B. (2019). Improving Arabic text to image mapping using a robust machine learning technique. IEEE access, 7, 18772- 18782.
- Zhang, C. (2019, 2 12). How to train an object detection model easy for free. Retrieved 1 1, 2021, from https: //medium.com/swlh/how- to- train- an- object- detection- model- easy- for- free- f388ff3663e