The New Dataset MITWPU-1K for Object Recognition and Image Captioning Tasks

Madhuri Bhalekar
School of Computer Engineering and Technology
Dr. Vishwanath Karad MIT World Peace University
Pune, India
madhuri.bhalekar@mitwpu.edu.in

Mangesh Bedekar
School of Computer Engineering and Technology
Dr. Vishwanath Karad MIT World Peace University
Pune, India
mangesh.bedekar@mitwpu.edu.in

Received: 5 May 2022 | Revised: 19 May 2022 | Accepted: 20 May 2022

Abstract-In the domain of image captioning, many pre-trained datasets are available. Using these datasets, models can be trained to automatically generate image descriptions regarding the contents of an image. Researchers usually do not spend much time in creating and training the new dataset before using it for a specific application, instead, they simply use existing pre-trained datasets. MS COCO, Flicker, and Pascal VOC, are well-known datasets that are widely used in the task of generating image captions. In most available image captioning datasets, image textual information, which can play a vital role in generating more precise image descriptions, is missing. This paper presents the process of creating a new dataset that consists of images along with text and captions. Images of the nearby vicinity of the campus of MIT World Peace University-MITWPU, India, were taken for the new dataset named MITWPU-1K. This dataset can be used in object detection and caption generation of images. The objective of this paper is to highlight the steps required for creating a new dataset. This necessitated a review of the existing dataset models prior to creating the new dataset. A sequential convolutional model for detecting objects on a new dataset is also presented. The process of creating a new image captioning dataset and the gained insights are described.

Keywords-convolutional model; dataset; image captioning; image labelling; object detection

I. INTRODUCTION

Many image datasets (e.g. MS COCO, Flicker, PASCAL) are available and can be used in image captioning systems. We wanted to explore the background and preprocessing required while creating such a dataset. For this purpose, we started to create a new dataset, called MITWPU-1K, of our university, MIT World Peace University, campus. Currently, the dataset consists of around 1500 images which are object labeled and captioned manually. These labels are classes of objects present in the image. Our dataset model is trained using the CNN (Convolutional Neural Network) architecture to detect objects present in an image. The validation while consideration of image selection is also presented. Currently, the MITWPU-1K dataset is having 4500 image descriptions, i.e. for each image, an average of three image annotations is provided. This dataset has a total of 68 objects and we are still working on its expansion. Before creating a new dataset, we studied some of the prominent existing datasets used in image captioning.

II. RELATED WORKS

Many architectures have been proposed for generating captions for a given image. The reviewed research papers belong to the domains of Machine Learning and Neural Networks. Most of the datasets which are used in object recognition tasks are developed by considering the task of object classification, detection, and labeling. In the image caption domain, the MS-COCO [1] dataset is considered a benchmark, in which object recognition is enhanced by advancing scene understanding and by piping the annotation via picture labeling, instance spotting, and instance segmentation pipelines. In the MS-COCO dataset, there are 91 frequent object types with 25,00,000 labeled instances in 3,28,000 images. The training set consists of 82,783 photos, whereas the validation set contains 40,504 images, and the testing set 81,434 images. The Pascal VOC [2] dataset consists of 11,000 pictures divided into 20 object groups, having 2,501 images in the training set, 2,510 images in the validation set, and 4,952 images in the testing set. Imagenet [3] contains 14,197,122 images and almost 21,000 object classes. ImageNet is built upon the hierarchical structure provided by WordNet. CIFAR-10 [4] dataset and contains 100 categories with almost 60,000 images. They have performed object classification on tiny images of size 32×32. The authors explain how they have trained a two-layer convolutional Deep Belief Network (DBN) on a 1.6 million tiny images dataset.

While analyzing these different dataset models, we come across the Yolo architecture [5] and Caffe framework [6]. Yolo architecture is used for object detection. Classification is done using CNNs along with localization using regression. Caffe convolutional architecture is used in Region-Based CNNs (RCNNs) for quick feature embedding framework. Caffe has already been employed in several academic research projects. Along with these papers, we also find some latest review papers on object detection methods used in deep learning [7, 8] which provide the metrics used for object detection along with the datasets used. The detailed review of many object detection survey papers is summarized in [7]. The authors observed that

Corresponding author: Madhuri Bhalekar

for object detection, deep learning methods provide a prominent approach. The authors also highlight the future work that can be done in visual object detection like multi-domain object detection, silent object detection, and unsupervised object detection using a deep learning approach. In [8], authors provide the categorization of existing image captioning systems and commonly used datasets for image captioning. They provided the detailed statistics of the following datasets: MS COCO, Flickr30K, Flickr8K, Visual Genome, IAPR TC-12, Stock3M, and MIT-Adobe FiveK dataset. Table I presents the details of some of the existing datasets mostly used in image captioning.

TABLE I. SUMMARY OF PROMINENT EXISTING DATASETS USED IN IMAGE CAPTIONING

Dataset	Images	Objects / classes	Training set	Validation set	Testing set
MS-COCO [1]	3,28,000	80 categories	82,783	40,504	81,434
Pascal VOC [2]	11,000	20	2,501	2,510	4,952
ImageNet [3]	14,197,122	21,000 classes	1.2 million	150,000	-
CIFAR-10 [4]	60,000	100	-	-	-

III. METHODOLOGY USED IN DATASET CREATION

After analyzing the existing dataset models, we have started creating our new dataset which consists of images of the nearby vicinity of the MIT World Peace University campus. We are creating a new dataset which can be further used in object recognition with labeled data and captions for the image. The flow of the creation of the dataset is shown in Figure 1 and can be summarized as:

- Image collection and validation
- Interface to provide image description/annotations
- Labeling objects

These steps are explained in the following sections.

A. Image Collection

For creating the new dataset, we collected images having diversity. The main objective of collecting diverse images was to proper train the dataset to avoid the problem of over fitting or under fitting during training. So, we tried to collect diverse images, like images with different brightness levels, having different foreground and background, containing multiple objects, etc. We took images that include different activities and events carried out on the MITWPU campus. While building this dataset we considered the images of posters and banners of various events on campus which contain some text information. The main goal was to use these textual data from the image to provide more detailed image descriptions during the caption generation task. Initially, we collected the images with the use of a smartphone and a professional camera, due to which the gathered images have different resolution, size, and orientation. After collecting several images, we found that all images could not be considered for the new dataset, so the rectification of these collected images are carried out as mentioned below.

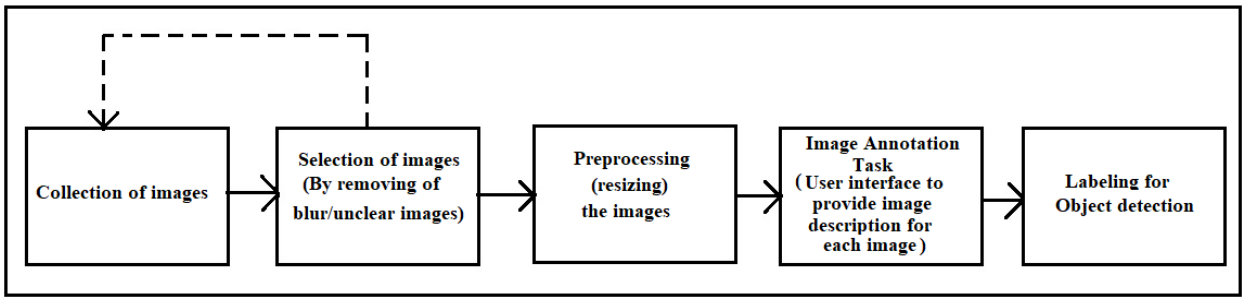


Fig. 1. System flow of dataset creation.

B. Validation

Images from different sources were collected (via a professional camera, smart phones, etc.) from which many had low resolution, poor brightness, whereas some images were redundant, i.e. captured with multiple clicks on the same view. While preparing the dataset we observed that images should provide a good representation of the classes which can further help in the classification process. To achieve this, these images were removed using manual filtration. After validating the selected images at present, we had 1500 images in the dataset. Further, to speed up the process, the images were resized. Some sample images are shown in Figure 2.

We also considered images containing text, like banners and posters, so that in the future we can develop an image captioning system for generating the description of an image along with text extraction as shown in Figure 12.

C. Interface to Provide Image Description

As the dataset that can be utilized in an image captioning system was developed, we added descriptions for each image. To accomplish this task, a user interface was created in python as shown in Figure 3. Using this interface manually we provided captions for every image, which were stored in the form of csv and json files, and can be further used to train our dataset for image captioning (Figure 4). While providing the

manual captions we took care to provide different captions for the same image considering all visual elements of the image. For this, we provided a minimum of three captions for each images. Finally, along with each new image, an image description file containing the image annotations was provided.

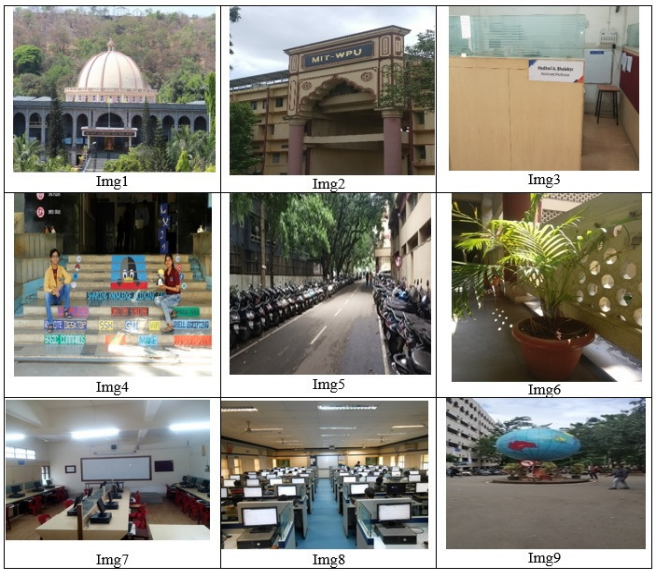


Fig. 2. Sample mages from the dataset.

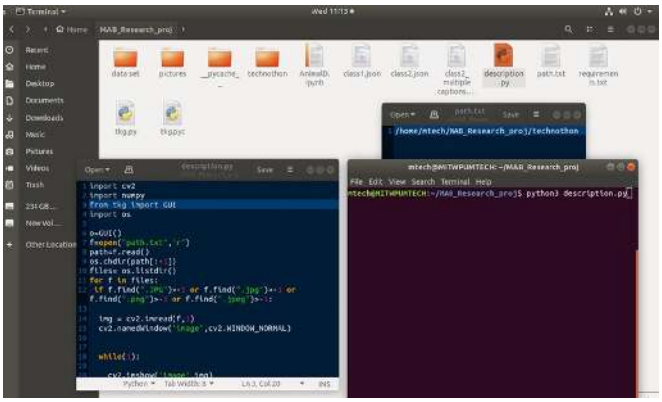


Fig. 3. Dataset creation module.

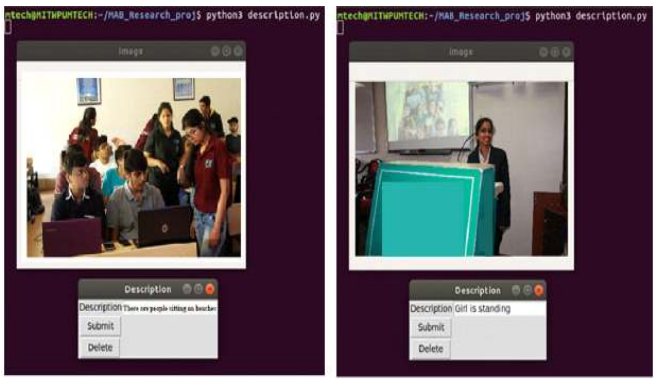


Fig. 4. Interface providing image description.

D. Image Description Format

With the image identification number we keep track of its associated descriptions or annotations. We maintained track of image file as {"file_name": "IMG_100.jpg","id": 100}, and the included annotations for each image were in the format mentioned below in either csv or json file format:

- {"image_id": 100, "id": 1,"caption": "Description 1"},
- {"image_id": 100,"id": 2,"caption": " Description 2"},
- {"image_id": 100,"id": 3,"caption": " Description 3"},

This way, in the new dataset two folders were maintained, one which contains all the images and another that includes the manually assigned annotations for each image.

For creating the MS COCO dataset [1], huge crowdsourcing was involved in the annotation task and by using the Precision-Recall metric the quality of the annotation task was measured. We measured the quality of annotations of our dataset by applying the same concept. The quality of annotations was analyzed by a group of people including the authors of the paper. In some cases, we observed low precision and recall value, so we tried to search the false positive and false negative patterns and again perform the annotation correction task. For validation purposes, after assigning descriptions to all the images, we manually checked for redundancy the descriptions and verified them. Presently, the new dataset contains 1500 images with 4500 descriptions/annotations. Some image samples with the assigned captions are shown in Figure 5.

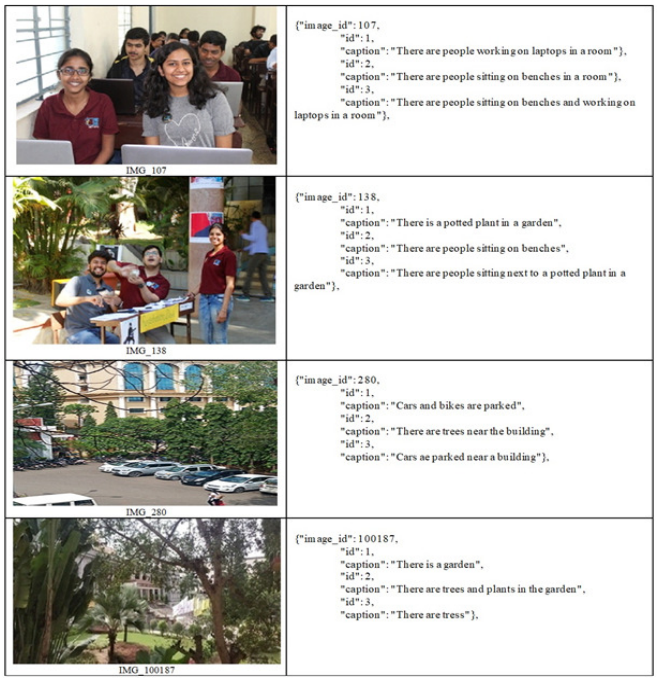


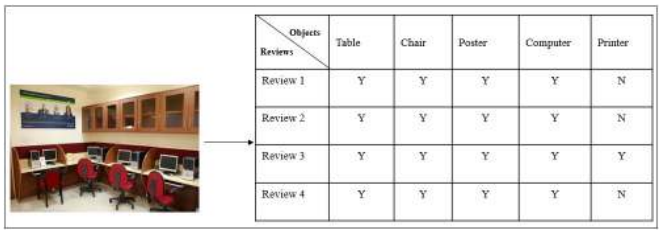
Fig. 5. Sample images with assigned captions from the created dataset.

E. Object Labelling

After acquiring images and their associated descriptions, the next task was to perform object detection, but before that,

we performed labelling on the images as shown in Figure 6. When multiple appearances of the same object occur in one image, only one label will be considered. Before using the new dataset, we performed object detection. Identifying objects from an image becomes complex when the image contains multiple objects. For this, we used the TensorFlow object detector [9] and done the review process used during the ImageNet dataset creation for detecting objects present in the image (Figure 7) [3]. The process involves visually checking the objects assigned with the images in order to check the classification accuracy of the new dataset.

Fig. 6. Labelling the objects of an image.



Reviews \ Objects	Table	Chair	Poster	Computer	Printer
Review 1	Y	Y	Y	Y	N
Review 2	Y	Y	Y	Y	N
Review 3	Y	Y	Y	Y	Y
Review 4	Y	Y	Y	Y	N

Fig. 7. Review method of identifying objects present in the image.

After performing the review method thoroughly, a complete list of classes in the new dataset was acquired. At present, our dataset contains 68 different object classes, such as person, chair, room, garden, plant, car, road, door, window, building, globe, printer, pen, notebook, bag, cupboard, book, bench, table, staircase, etc.

IV. BUILDING AND TRAINING THE CLASSIFICATION MODEL

To perform image classification, there are many existing pre-trained architectures available which are trained on huge size image datasets like MS COCO [1] and ImageNet [3]. Instead of using these available pre-trained architectures, we have come up with a simple convolutional classifier model which was trained on the new dataset. The main reason to come up with a different convolutional classifier was to avoid the over fitting problem because currently our dataset size is limited. We have created a sequential convolutional model, which gives the flexibility to keep adding different layers to the model as per requirements. The basic steps we followed while creating our classifier model which was trained on the new dataset are:

- Build the model
- Compile the model
- Train the model with the new dataset
- Validate the model
- Measure the model's performance in terms of accuracy

The classifier model summary is shown in Figure 8. As already mentioned, in order to detect the 68 different object types present in our dataset, we have used a dense layer size (68). The model is trained to identify more than one object in an image. Some samples of object detection are shown in Figure 9. We initially started with bigger size objects, then proceeded towards smaller size objects. Our trained classification model detected objects like monitor, mouse, and keyboard in the given image shown in Figure 9.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 224, 224, 32)      896
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 112, 112, 32)      0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 110, 110, 64)      18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 55, 55, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 53, 53, 124)       71548
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 27, 27, 124)       0
_________________________________________________________________
flatten (Flatten)            (None, 90396)             0
_________________________________________________________________
dense (Dense)                (None, 68)                6146996
=================================================================
Total params: 6,237,936
Trainable params: 6,237,936
Non-trainable params: 0
_________________________________________________________________
```

Fig. 8. Model summary of the sequential classifier.

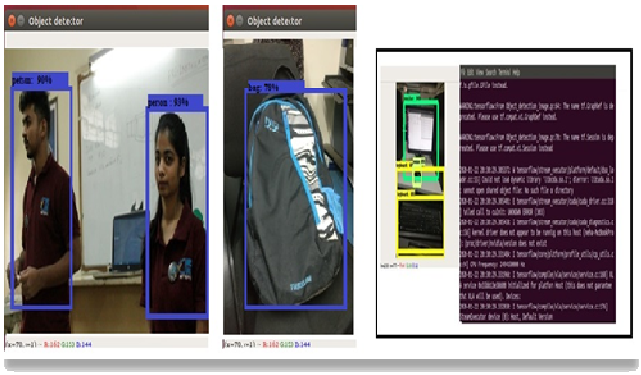


Fig. 9. Results of multiple detected objects.

A. Results

The new dataset was created for image captioning tasks, but it can also be used in object detection. To perform object detection, we performed labelling on the images which are further used during the training phase. Instead of using any pretrained classifier model, we presented the sequential classifier model. To check and compare the accuracy of the

presented sequential model, we performed the review method of [3]. Object detection with the sequential classifier model provided 84% accuracy. The comparative analysis graph plot of the review method and the sequential classifier model is shown in Figure 10.

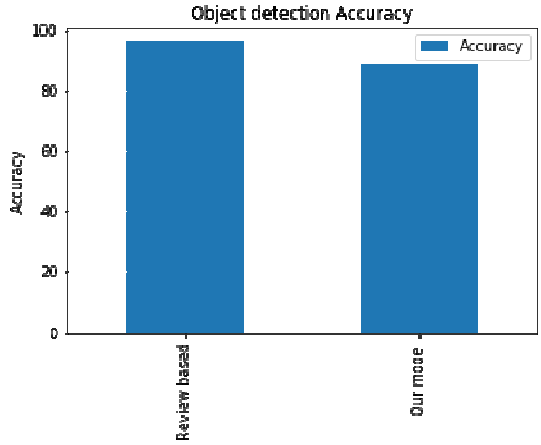


Fig. 10. Object detection accuracy comparison.

TABLE II. MITWPU-1K DATASET SUMMARY

Dataset	Images	Objects	Image descriptions/annotations
MITWPU-1K	1500	68	4500

Fig. 11. Sample images of identified objects in the MITWPU-1K dataset.

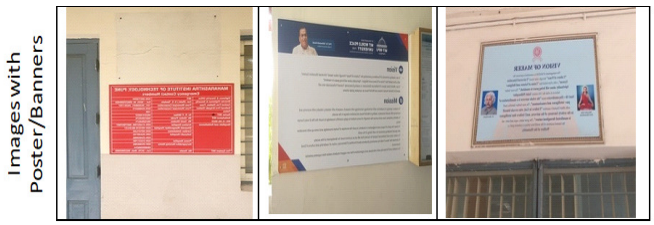


Fig. 12. Sample images containing poster/banner.

Our dataset contains 1500 images with 68 different objects and 4500 image descriptions (Table II). These objects are mostly present in any college or university campus. Some of the samples of the identified objects in the created dataset are shown in Figure 11. The dataset contains some images which contain text in the form of a banner or poster as shown in Figure 12. Using this dataset, we proposed a new deep learning model [10] for performing image captioning and text extraction. The obtained results of image captioning, including textual information present in the image, were satisfactory, giving accuracy up to 83% and performed as well as the state of the art methods.

The new dataset MITWPU-1K can be used in a wide range of applications such as object detection, caption generation, text detection, etc. Regarding future work, the size of the dataset will be increased by adding some domain-specific images along with captions.

V. CONCLUSIONS

This paper presented the followed process while creating a new image dataset for image captioning generation. The main purpose was to explore the processing behind creating a dataset that can be used in generating image captions. Before and during the creation of this new dataset, existing dataset models were studied in detail. At present, the new dataset consists of 1500 images and its name is MITWPU-1K in line with the current trend to name the dataset in a particular manner. The dataset includes 4500 annotations as we have assigned the minimum of three annotations for each image. The dataset has a subset of images that include text, so along with image description, text extraction can be done to further extend its application in domains like image captioning or other similar tasks.

Using a sequential convolutional model, we performed object detection tasks on the MITWP-1K dataset. The dataset creation model can be extended to include all images of a particular domain such as sports, night vision images, animals, automobiles, computers, etc. In the application of image captioning, all available datasets are mostly generic. The accuracy of such systems can be improved by creating or using domain-specific image datasets.

ACKNOWLEDGMENT

The authors would like to thank the authorities of Dr. Vishwanath Karad MIT World Peace University for their support.

REFERENCES

[1] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, 2014, pp. 740–755, https://doi.org/10.1007/978-3-319-10602-1_48.

[2] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015, https://doi.org/10.1007/s11263-014-0733-5.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, Jun. 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.

[4] R. Doon, T. Kumar Rawat, and S. Gautam, "Cifar-10 Classification using Deep Convolutional Neural Network," in *2018 IEEE Punecon*, Pune, India, Aug. 2018, https://doi.org/10.1109/PUNECON.2018.8745428.

[5] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6517–6525, https://doi.org/10.1109/CVPR.2017.690.

[6] Y. Jia *et al.*, "Caffe: Convolutional Architecture for Fast Feature Embedding," in *MM '14: Proceedings of the 22nd ACM international conference on Multimedia*, New York, NY, USA, Aug. 2014, pp. 675–678, https://doi.org/10.1145/2647868.2654889.

[7] V. Sharma and R. N. Mir, "A comprehensive and systematic look up into deep learning based object detection techniques: A review," *Computer Science Review*, vol. 38, Nov. 2020, Art. no. 100301, https://doi.org/10.1016/j.cosrev.2020.100301.

[8] M. D. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A Comprehensive Survey of Deep Learning for Image Captioning," *ACM Computing Surveys*, vol. 51, no. 6, pp. 118:1-118:36, Oct. 2019, https://doi.org/10.1145/3295748.

[9] G. Tanner, "Creating your own object detector," *Towards Data Science*, Feb. 06, 2019. https://towardsdatascience.com/creating-your-own-object-detector-ad69dda69c85.

[10] M. Bhalekar and M. Bedekar, "D-CNN: A New model for Generating Image Captions with Text Extraction Using Deep Learning for Visually Challenged Individuals," *Engineering, Technology & Applied Science Research*, vol. 12, no. 2, pp. 8366–8373, Apr. 2022, https://doi.org/10.48084/etasr.4772.

[11] B. Ahmed, G. Ali, A. Hussain, A. Baseer, and J. Ahmed, "Analysis of Text Feature Extractors using Deep Learning on Fake News," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 7001–7005, Apr. 2021, https://doi.org/10.48084/etasr.4069.

[12] S. Nuanmeesri, "A Hybrid Deep Learning and Optimized Machine Learning Approach for Rose Leaf Disease Classification," *Engineering, Technology & Applied Science Research*, vol. 11, no. 5, pp. 7678–7683, Oct. 2021, https://doi.org/10.48084/etasr.4455.