Tursunov Javlon Jurakul ugli
1st year Master’s student
Tashkent University of Information Technologies
named after Muhammad al-Khwarizmi, Republic of Uzbekistan

Memonova Gulrukh Nazrikulovna
Assistant professor
Karshi branch of Tashkent University of Information Technologies
named after Muhammad al-Khwarizmim, Republic of Uzbekistan

Memonova Gulnoza Nazrikulovna
School teacher
General secondary school 80th under the public education department
of Kasan district, Republic of Uzbekistan

CUSTOM OBJECT DETECTION USING YOLO

Abstract. *These days, deep learning and convolutional neural networks (CNN) are gaining popularity in terms performing far better compared to traditional methods when it comes to object detection, localization and classification, making the job easier for humankind. However, deploying them successfully to perform some certain tasks still remains to be an issue. For instance, in object detection and localization in a video or still image, since existing datasets include only limited number of class objects, sometimes object that needs to be detected and localized may not exist in the dataset, making it impossible to classify the object. Thus, to overcome this isssue, a new aproach has been proposed which makes use of the CNN-based YOLO algorithm. In this work, a new model has been developed which is able to classify and localize objects which don’t exist in the dataset. With the help of this model, just like any other objects, unknown objects can be identified as well after sufficient training of the model. This trained model achieved high accuracy in terms of predicting objects right, making it reliable for other projects which involves detecting objects which are specific to new classes.*

Keywords: *Kaggle dataset, Google colaboratory, supervised learning, unsupervised learning, YOLO.*

Introduction

There is a wide variety of methods to detect, localize and classify different objects in an image or in the video which is obviously consisted of sequence of frames. One of the popular approaches is to use deep learning based on convolutional neural networks like YOLO v2 and R-CNN which can detect objects in images with unsupervised learning. To train any kind of machine learning model, there are mainly two types of trainings: supervised learning and unsupervised learning. Supervised learning is defined whether labelling is used or not along with the dataset when training the algorithms so that they can learn to make classification of data accurately. In supervised learning, labelling or annotation is fed into the network while in unsupervised learning, model is trained without the labelling and in this case, model learns how to classify the objects by using specific characteristics that are typical for a particular object. Supervised learning performs better in terms of accuracy since it is provided with the clear label as a reference [1].

As it was mentioned above, there are various methods and algorithms including YOLO which are used for object detection. Having said that, YOLO is commonly used for object detection that uses the COCO dataset, meaning that classes of objects which are included in the this dataset can be identified and classified using YOLO. COCO dataset includes 91 different classes, with 80 of them being used commonly and this dataset includes most commonly used objects in everyday life [3]. Since it is based on the convolutional neural network(CNN) and neural networks require training before putting them into work, YOLO should be specifically trained to detect objects which are not included in this dataset. Sometimes, there are situations where the number of classes defined in the COCO dataset is not enough and one may want to detect or localize the object which is not typical object and not included in commonly used datasets. In this kind cases, that’s where the training the YOLO comes in to make it identify objects which are included in the commonly used datasets. For example, YOLO can identify cow but cannot identify rabbit and in this case, custom object detection method comes in handy and primary focus of this work is to build model which is capable of detecting particular object which, in this case, is kangaroo. To get a little bit idea of what the YOLO is and how it works, brief information is given below. YOLO is based on CNN, an acronym standing for “You Only Look Once” and has high accuracy, allowing users for the detection of objects whether in image or frames.

YOLO differs completely from other algorithms by using different approach. To make prediction and calculate probabilities of objects, it uses a neural network to process the whole image instead of picking different regions. The convolutional neural network used in YOLO has one depth that divides the image into grids, with each grid having a vector which includes the information like class probability, coordinates of central point and width and height of bounding box. The model gets better progressively, improving the accuracy in terms of identifying the objects right and drawing the bounding boxes [2]. Since YOLO is faster and more accurate compared to other existing algorithms, it is preferred more in computer vision tasks.

Related work

There has been a significant improvement in computer vision over the years, with researchers contributing to the development in object detection, classification and localization.

Rahul Dutt Sharma [4] has made proposition of a technique which works by subtracting the background dynamically designed to track and detect the objects on the move. That proposed technique was discussed widely and after discussion, they came to the conclusion that it performs better compared to other existing works by extracting the foreground effectively.

Technique developed by Syed Mazhar Abbas [5], which is region-based to detect and classify objects by making use of R-CNN and Faster R-CNN, was trained by custom dataset which included still images. This technique proved itself very effective in terms of object detection. Most importantly, this system was able to operate with low GPU requirement of 3.0 or higher.

Priynka Malhotra [6] came up with the technique for object detection for which, R-CNN , Fast R-CNN and YOLO were indispensable and combined. Despite the fact that R-CNN and Fast R-CNN are comparatively slower than YOLO, they are good in terms of detection of tiny objects while YOLO performs well in regression than classification. YOLO outdoes the R-CNN and Fast R-CNN with high accuracy and speed in terms of classification in real time.

Ajeet Ram Pathak [7] suggested using the Deep Learning applications to detect objects and he made the role of deep learning clear to use some of the techniques which are based on the convolutional neural network (CNN). In his proposal, available frames of deep learning were discussed for which, pascal voc dataset is important.

A new system and a chip were designed by Chun Ju Huangv [8] to detect objects in real time. Developed model was based on Is-r-yolo network which included an algorithm which uses distributed features in the image.

Han.C, Liu et al [9] came up with the brand new approach to make identification of clustered image shape. In image transformation, the transforming features were derived from the shape. Pairs were identified between shape of local image and queried shape by using PAD(Pyramid of arc length descriptor). Those proposed methods are used to measure the transforming shape into domains by taking the wavelet transform and Fourier transform into account. Afterwards, myriad shape descriptors were put forward to make measurement of similarities between shapes. Triangle shapes were used as a reference. But, algorithms, which exist and are used today, match shape descriptors, with their main task being matching shapes.

Kamate.S et al [10] carried out a research to detect and track the objects which are on the move. This proposed model was advised to be used from the vehicle which were totally unmanned to collect information about migrants crossing the borders illegally in the USA. These days, UAV(unmanned aerial vehicle) is considered vital in many industries where abilities of humans are limited or they need extra edge. Due its daunting nature, tracking the moving objects can require many methods to be used like background subtraction and histogram oriented gradients. Primary aim of this research to give humans helping hands to carry out surveillance and tracking suspicious objects. In this work, different methods were under investigation to improve the performance of the model.

Li.C et al [11] did a research on object detector with deep learning and what he accomplished was able to detect tiny objects with the help of the deep learning. Algorithms proposed in this work were designed for sampling generator to switch generators in various scenes. However, since less samples were used during the training process, networks cannot make significant changes for the various tasks.

Erhan.D et al [12] made a proposition of research about detecting the objects by using the neural networks. This work consisted of two distinctive stages. First stage was a part which directly used a method to detect objects using deep convolutional neural network while second part involved using convolutional neural network which works based on the region and this second part is in charge of improving the network accuracy while lowering the time spent on the task. The dataset which was used in this work contained 201 labels which can recognize an image, thus helping to solve the detection of image. What was proposed by him can be seen being used in image retrieval systems and parking systems with automation. In this kind of systems, the main role of object detection is the image classification.

A research was conducted by He.K.Zhang et al [13] into deep learning as classical learning methods in detection of objects. This emerging work related to object detection is mainly geared towards accurate and real-time detection. The method used in subtraction is divided into three different processes which are: background modeling, object detecting and background update. The process which involves subtracting the background is the same as the difference between frames which can occur in time.

Methodology

For this work, the Kaggle dataset was used which is a dataset containing different images used for object identification, classification and localization. Besides, Google colaboratory has been used to deal with computational part. Google colaboratory is a programming environment developed by the Google to carry out heavy-going programs which cannot be compiled by the local machines. With the help of Google colaboratory, free GPU can be used in heavy computational tasks. This work basically follows the work flow described in Fig 1.

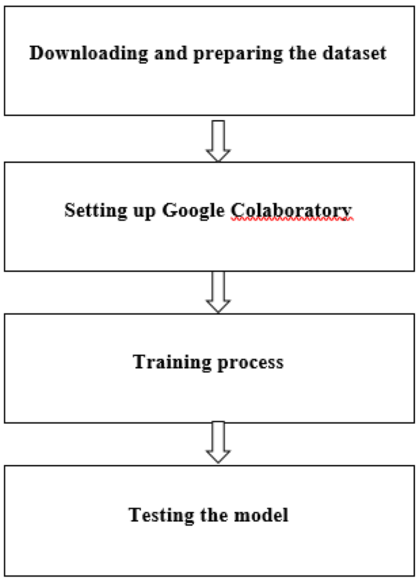
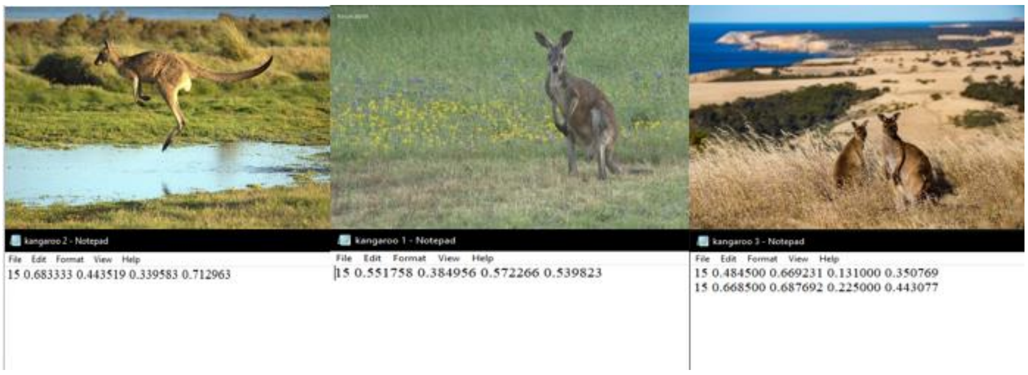

Downloading and preparing the dataset
Setting up Google Colaboratory
Training process
Testing the model

Fig. 1. **Work flow for developing custom-object detection**

A) Downloading and preparing the dataset. As shown in the diagram, dataset firstly should be prepared for the work. Since this works involves developing model to detect custom object which is not included in the COCO dataset, images of that object which cannot be detected by the YOLO algorithm were collected and for the case of this work, custom object is a kangaroo animal which is not identified in the images by the any of the algorithms. By developing this model further, any other wanted custom object can be detected. Since the Kaggle dataset [15] includes the kangaroo images, that dataset was downloaded to make the job easier and below Fig 2 shows some of the images in the dataset.

Fig. 2. **A glance at the dataset**

B) Setting up Google Colaboratory. Once the dataset is ready and since supervised learning is used for this work, YOLO algorithm based on CNN needs annotations for training the network. Annotations and images are fed into the neural network, and it uses the annotations to detect the class and position of the object. To make annotations, the special software was used which can be found through the link in the references section [14]. Annotations are prepared in the simple text files which include five numeric values which describes class of objects, coordinates of the object center and dimensions of the bounding box which is drawn around the object in object detection.


kangaroo 2 - Notepad
File Edit Format View Help
kangaroo 1 - Notepad
File Edit Format View Help
kangaroo 3 - Notepad
File Edit Format View Help

```
15 0.683333 0.443519 0.339583 0.712963
```

```
15 0.551758 0.384956 0.572266 0.539823
```

```
15 0.484500 0.669231 0.131000 0.350769
15 0.668500 0.687692 0.225000 0.443077
```

Fig. 3. **Dataset and annotations**

The number of vectors in the file differ depending on the number of classes of objects in the image. Above Fig 3 shows some random images and annotations used in neural network along with the image dataset. The folder includes all the images and annotations and annotation files have the same names corresponding to the images. Once the pre-processing of dataset is done, Google colaboratory is set up which is mainly involved for the training process.

C) Training process Once the dataset and Google Colaboratory are done, written code was executed to start the training process with YOLO v3 algorithm which is built on the convolutional neural networks. A clear look on the summary of the model shows how many trainable parameters there are. In the process of the training, the number of epochs was 50 which is an indication of how many times all the images included in the dataset should be passed through the network. In this work, the dataset is comprised of more than 330 images which is enough to develop a model that is capable of predicting objects accurately. As the accuracy of the model is considered as one of the important factors, it was under monitor and the more different images are included in the dataset and the more time spent on the training, the higher accuracy model can have. There are some factors used to measure how well the model performs. For instance, there are two metrics used are accuracy and loss. The loss indicates how bad the model predicts. If the loss is higher, model performs badly in terms of prediction and for this reason, model should be trained more so that accuracy of prediction can increase, decreasing the loss simultaneously. In the Fig 4 and Fig 5, it can be seen how accuracy is increasing while loss is decreasing respectively in the training process as the number of epochs is incrementing.

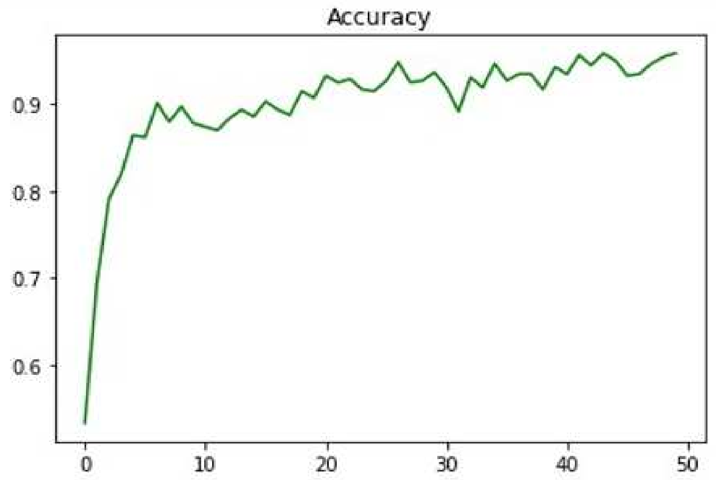


Fig. 4. **Increasing accuracy**

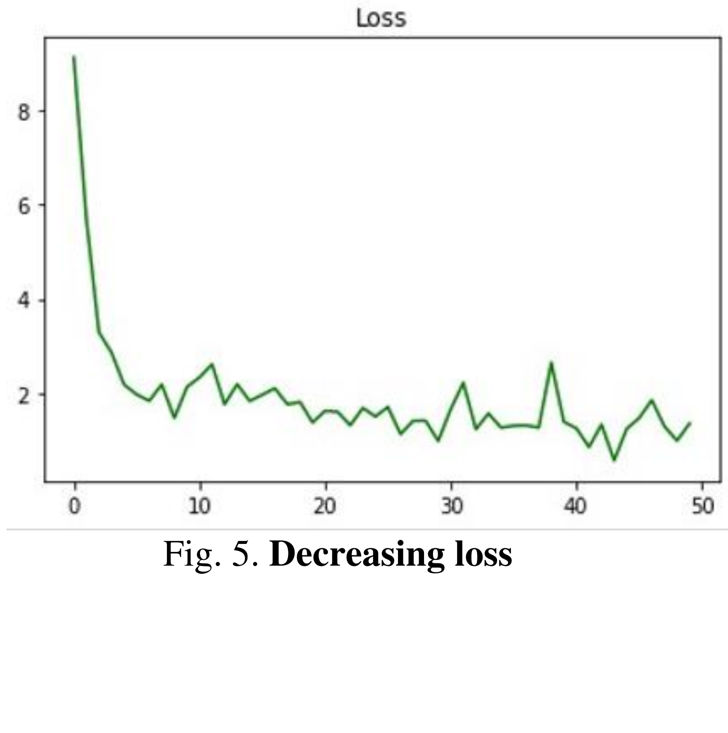


Fig. 5. **Decreasing loss**

As it can be seen in above graph, the accuracy can reach more than 90 percent which is quite reliable in terms of prediction. After YOLO v3 training has been completed, the best model is saved in Google Colaboratory which was used in the python programming environment.

D) Testing the model. To make sure if model is doing it is job right without errors, images, which are not incorporated in the dataset used in the training process, were used when checking the model. The below Fig 6 shows how model is carrying out what is meant to do to detect and localize objects in the image.

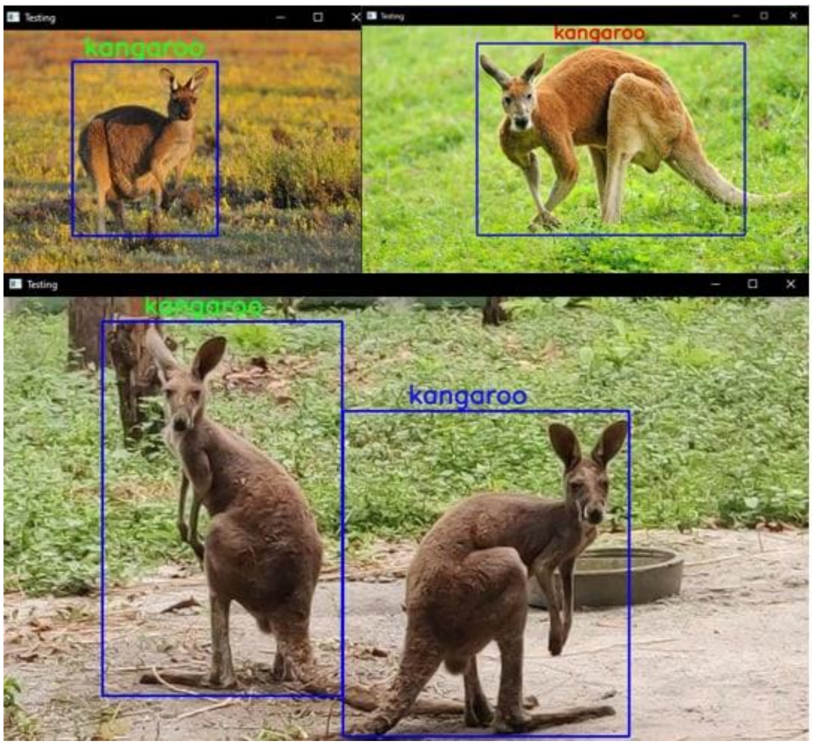


Fig. 6. **Testing the model**

It can be seen that model is doing it is job right, it is detecting objects and localizing them in the image. Now, this model can be used in any image or video to detect custom object. During the testing process, overall 200 images were used to check the accuracy of the model. The above image show some of the custom objects being detected by the model.

Conclusion

In this work, CNN based YOLO v3 algorithm was used which involved supervised learning with labelled data. To deal with the computational part, Google

Colaboratory was used, enabling the use of online GPU and programming environment. Once the training part is finished, the model was put into test to make sure it is working correctly and like expected, it detected objects right and localized them in the still image. This model comes in handy in situations where custom object detection is crucial. For instance, user may want to monitor a certain object via video surveillance but that object to be detected is not included in any dataset, thus creating a huge problem. In this kind of scenarios, the developed model can be used as a reference for further development of models which are capable of detecting multiple objects in the image or video.

References:

1. Ciro Donalek, "Supervised and Unsupervised Learning", Ay/Bi 199 –April 2011
2. Dr. Suwarna Gothane, "A Practice for Object Detection Using YOLO Algorithm", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume 7, Issue 2, 268-272
3. Holger Caesar, Jasper Uijlings and Vittorio Ferrari, "COCO-Stuff thing and Stuff Classes in Context", 2018.
4. Rahul Dutt Sharma, Shubh Lakshmi Agrwal, Sandeep K.Gupta, Anil Prajapti, "Optimized dynamic background subtraction technique for moving object detection and tracking", 2017 2nd International Conference on Telecommunications and Networks.
5. Syed Mazhar Abbas , Dr. Shailendra Narayan Singh, "Region-based object detection and classification using Fast R-CNN", International Conference on "Computational Intelligence and Communication Technology (CICT 2018).
6. Priyanka Malhotra, Ekansh Garg, " Object detection techniques: a comparison" IEEE 7th International Conference on Smart Structures and Systems ICSSS 2020.
7. Ajeet Ram Pathak, Amnjusha Pandey, Siddhahrth Rautaray, "Applications of deep learning for object detection", International Conference on Computational Intelligence and Data Science (ICCIDS 2018).
8. Chun Ju Huang, Ting -Wei Chenm Yu-Chen Fan, "Chip and System Design of Real Time Object Detection Based on LS-R-YOLO Network", 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE).
9. Han, C., Liu, X., Sinn, L. T., & Wong, T. T. (2018). TransHist: Occlusion-robust shape detection in cluttered images. Computational Visual Media, vol4,no. 2,pp. 161-172.
10. Kamate, S., &Yilmazer, N. (2015). Application of object detection and tracking techniques

for unmanned aerial vehicles. Procedia Computer Science, vol61,no.3, pp. 436-441.

11. Li, C., Zhang, Y., &Qu, Y. (2018, March). Object detection based on deep learning of small samples. In Advanced Computational Intelligence (ICACI), 2018 Tenth International Conference on . IEEE,vol2, no.3, pp. 449-454.
12. Erhan, D., Szegedy, C., Toshev, A., &Anguelov, D. (2014). Scalable object detection using deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition vol 3, no 4,pp. 2147-2154.
13. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition vol 2, no.3, pp. 770-778.
14. https://tzutalin.github.io/labelImg
15. https://www.kaggle.com/datasets/hugozanini1/kangaroodataset