Players vs Referees - object detection

University of Verona
Master Degree in Artificial Intelligence
A.Y. 2022/2023

Computer Vision & Deep Learning
Advanced Programming

Gaetano Alberto Caporusso - VR489760
Sebastiano D'Arconso - VR489066

Contents

1 GitHub

Here you'll find the GitHub page of the project Player-vs-Referee. The most important scripts are the *main.py* file and all the files in the *utils_proj* folder.

2 Motivation and rationale

Our project centers around creating an object detection system tailored specifically for NBA players and referees. By leveraging our self-created dataset, we aim to train a computer vision model capable of detecting and distinguish between these two specific classes. We wanted to combine our passion for NBA with the techincal skills acquired during the course and see if we were able to achieve this goal.

3 State of the Art

Over the past few years, YOLO has emerged as a groundbreaking and highly influential approach in the field of object detection. As of today, YOLOv7 is probably the benchmark for object detection algorithms, followed by EfficientDet, RetinaNet and Faster R-CNN.

4 Objectives

Our goal was to implement an object recognition model that could be able to detect and classify between two specific classes: NBA players and referees. We also wanted to try different models with different approaches in order to have different results to compare and see which model or approach could be the best for our specific task.

5 Methodology

In this section we will explain in detail the dataset selection and the models used for this project.

5.1 Dataset selection

Due to the unavailability of a pre-existing dataset containing NBA players and referees images, along with their respective bounding boxes, it was necessary for us to create our own dataset. This involved the process of sourcing relevant images on the internet and manually labelling each image with the corresponding bounding boxes. By undertaking this task, we ensured that we had a comprehensive and customized dataset specifically tailored to our

experiment's requirements. Each image of the dataset has been labeled using the *labelimg* tool.

5.1.1 Labels

We first labeled our images following the YOLO format, since YOLO was our first choice of model, then we used a python script to convert all the labels in COCO format, since we also wanted to try a Faster R-CNN trained on COCO.

```
# YOLO label format
<object-class> <x-center> <y-center> <width> <height>
```

For YOLO each image has to be associated to the corresponding labels.txt file, if there are more object in the same image each object class and bounding box informations are listed in the same labels.txt file. For COCO format the labels are a bit more complicated, since all the dataset has to be represented by a JSON file, the next structure represent a snippet of our JSON file and in particular the sections for the images and the labels (annotations):

```
# COCO label format
"images": [
     {
         "file_name": "test_10.jpg",
         "height": 256,
         "width": 256,
         "id": 0
     }
]
"annotations": [
     {
         "id": 0,
         "image_id": 0,
         "category_id": 2,   # class
         "bbox": [
             45,
             2,
             85,
             85
         ],
         "area": 7225,       # area of the bbox
         "segmentation": [], # empty because we don't provide a segmentation mask
         "iscrowd": 0        # object is not a crowd or group of instances
     }
]
```

Where under "images" there are all the basic informations about each image like its filename and dimensions, and under "annotations" there are

all the relevant informations of each image. In the COCO format the coordinates of the bounding box differ from the coordinates used in the YOLO format, in fact, in the COCO format the bounding box is represented by the x and y coordinates of the top-left corner and by the width and height of the box.

5.2 Dataset

(a) Referee train

(b) Player train

(c) Test image

Dataset	Number of images
Train	755
Test	157
Validation	54

5.3 Model selection

For our project we decided to try two different models:

- Faster R-CNN with two different backbones:
  - ResNet50
  - MobileNet
- YOLOv5s (s = small)

Model	Backbone	Classes	Params	Trained on
Faster R-CNN	MobileNet	91	19.4M	COCO
Faster R-CNN	ResNet50	91	38.2M	COCO
YOLOv5s	YOLOv5 v6.0	80	7.2M	COCO

6 Differences between the models

We decided to employ these two models for our task as they represent the two primary approaches, we were interested in exploring both methods to

gain a comprehensive understanding and evaluate their performance. Next we will describe the main differences between the two models, more in-depth differences will be discussed during the presentation.

6.1 Faster R-CNN

Faster R-CNN differs from its predecessors by a simple and yet powerful implementation, in fact, both R-CNN and Fast R-CNN use selective search to find the regions proposals. The selective search algorithm not only is slow and time-consuming but it's also a fixed algorithm, that means that no learning is involved during that process, and that can lead to bad proposals. Therefore, researchers came up with an object detection algorithm that eliminates the selective search and lets the network learn the region proposals. The initial pipeline is similar to the old versions: an image is provided as an input to a convolutional network which provides a convolutional feature map, but now, instead of using selective search on the feature map to identify the region proposals, a separate network is used to predict the region proposals. These regions proposals are then reshaped using a RoI pooling layer and then is used to classify the image within the proposed region. All these algorithms are region based algorithms.

6.2 YOLO

YOLO or You Only Look Once, is an object detection algorithm much different from the Faster R-CNN and similar, in fact, instead of using regions to localize objects within the image, YOLO splits the image in an SxS grid, and from that grid it takes m bounding boxes, then, for each bounding box the network outputs a class probability and offset values for the bounding box. So, in YOLO, a single convolutional network predicts the bounding boxes and the class probabilites for these boxes. The bounding box with a class probability above a threshold value is selected and used to locate the object within the image.

7 Experiments and results

In the next section, we will cover all the phases involved in the process, including training, testing, validation and the metrics used to evaluate the performances. All of the following are made on our own dataset.

7.1 Faster R-CNN - model preparation

In both cases we did transfer learning from pretrained networks and do the training process only on the last layers in order to achieve our goal. The pipeline is the following:

- Initialize the Faster R-CNN with the selected backbone.
- Retrieve the number of input features for the classifier score in the region of interest head of the model.
- Instanciate a *FastRCNNPredictor* class using the number of input features just retrieved and replacing the number of classes with our (3).
- Replace the box predictor of the roi heads with the new instance of the *FastRCNNPredictor* class.

In summary, we initialize a pre-trained Faster R-CNN model with ResNet50 or MobileNet backbone and FPN architecture, we retrieve the number of input features from the ROI heads, and replace the box predictor to customize the model's output for a specific number of classes, then we select the trainable parameters and we train only those parameters.

7.2 Faster R-CNN - dataset preparation

We had to split our dataset in three different subsets: train, test and validation. To do that we used a script that allowed us to select the percentage of the split. After that we used a python script to generate the *annotations.json* file for each subset, we then had to write a custom *Dataset* class in order to be able to load our own dataset. For the Faster R-CNN an *annotation.json* file has to be in every folder used as dataset.

7.2.1 The Dataset class

The implementation of the Dataset class follows the guide lines provided by PyTorch, so what we had to do was to implement three main functions: __init__, __len__, and __getitem__. The code can be seen on the GitHub page (link) and will be explained in-depth during the presentation but one feature that's worth explaining about the dataset class is that it returns the loaded image and the corresponding target. The image returned is a transformation of the original image of the dataset, the transformations applied to each image are the following:

```
def get_transforms(train=False):
    if train:
        transform = A.Compose([
            A.HorizontalFlip(p=0.3),
            A.VerticalFlip(p=0.3),
            A.RandomBrightnessContrast(p=0.1),
            A.ColorJitter(p=0.1),
            ToTensorV2()
        ], bbox_params=A.BboxParams(format='coco'))
    else:
        transform = A.Compose([
```

```
            ToTensorV2()
        ], bbox_params=A.BboxParams(format='coco'))

    return transform
```

For these transformations we used the *albumentations* library, which is useful beacuse it applies the same transformation on both the image and the corresponding bounding boxes, maintaining the correlation between the two. Since we are doing object classification we need to specify also the format of the bounding boxes. The target, on the other hand, is returned as follows:

```
## The target is first loaded from the annotations
def _load_target(self, id):
  return self.coco.loadAnns(self.coco.getAnnIds(id))

/.../

## Snippet at the end of the __getitem__ function
target = self._load_target(id)
targ = {} # here is our transformed target
targ['boxes'] = boxes
targ['labels'] = torch.tensor([t['category_id'] for t in target],
     dtype=torch.int64)
targ['image_id'] = torch.tensor([t['image_id'] for t in target])
targ['area'] = (boxes[:, 3] - boxes[:, 1]) * (boxes[:, 2] - boxes[:, 0]) # we
     have a different area
targ['iscrowd'] = torch.tensor([t['iscrowd'] for t in target], dtype=torch.int64)
```

The last thing worth mentioning is that the boxes coordinates change from xywh to xyxy (xmin ymin xmax ymax). The dataset class is then used to load the three different subsets, and each subset is used in the *DataLoader* class of PyTorch to instantiate the different dataloaders.

7.3 Faster R-CNN - training

Now that we load correctly our data and we have modified the pretrained models in order to correctly use those data we can perform the actual training. For the Faster R-CNN, the main function that does the actual training is the *train_one_epoch* that iterates over the dataloader, takes the images and targets and gives them in input at our model set in train mode; it then prints some useful informations. The Faster R-CNN model provided by PyTorch automatically compute the losses if we pass the image and the target, so all we had to do was to store the output of the model (the losses) and perform the backpropagation. In particular, the model returns the losses as follows:

- **Loss classifier**: this is the loss associated to the classificiation task of the Faster R-CNN, after generating region proposal using the RPN, the

Faster R-CNN takes these proposals and classifies them into different object categories. The loss classifier measures the difference between the predicted class probabilities and the ground-truth class labels of the proposed regions.

- **Loss box regression**: the Faster R-CNN models also performs bounding box regression to refine the initially generated region proposals. The loss box regression measures the discrepancy between the predicted bounding box and the ground-truth bounding box coordinates for each proposed region.
- **Loss RPN box regression**: the RPN generates regions proposals by regressing bounding box coordinates and predicting their offsets from predefined anchor boxes. The loss RPN box regression calculates the difference between the predicted bounding box offset and the ground-truth offsets for the bounding boxes. This loss is specific to the RPN and helps refine the proposed regions.
- **Loss objectness**: the RPN is responsible for generating region proposal by classifying each anchor box as either an object or background. The loss objectness measures the difference between the predicted objectness scores and the ground-truth labels for the anchor boxes.

The actual training is done in the training function, which calls the one epoch training for a fixed number of iterations (training epochs).

7.4 Faster R-CNN - testing

For the test part, we use a pre-existing function called *evaluate*, this function can be found in the PyTorch/vision GitHub (Pytorch/vision) and its function is to evaluate and return statistics about the model, we'll talk about the results in the **results** chapter. The function *evaluate* is used as follows:

```
test_dataset = ds.Dataset(root=data_path, type=opt.mode,
    transforms=ds.get_transforms(True))
test_loader = DataLoader(dataset=test_dataset, batch_size=1, shuffle=False,
    collate_fn=ds.collate_fn)

## model loading
model_trained = torch.load(opt.weights, map_location=torch.device('cpu'))
model_trained.eval()

## evaluation of the model
evaluate(model_trained, test_loader, 'cpu')
```

As shown, the evaluate function takes the model trained, the dataloader of the test set and the device on which perform the evaluation.

7.5 Faster R-CNN - validation

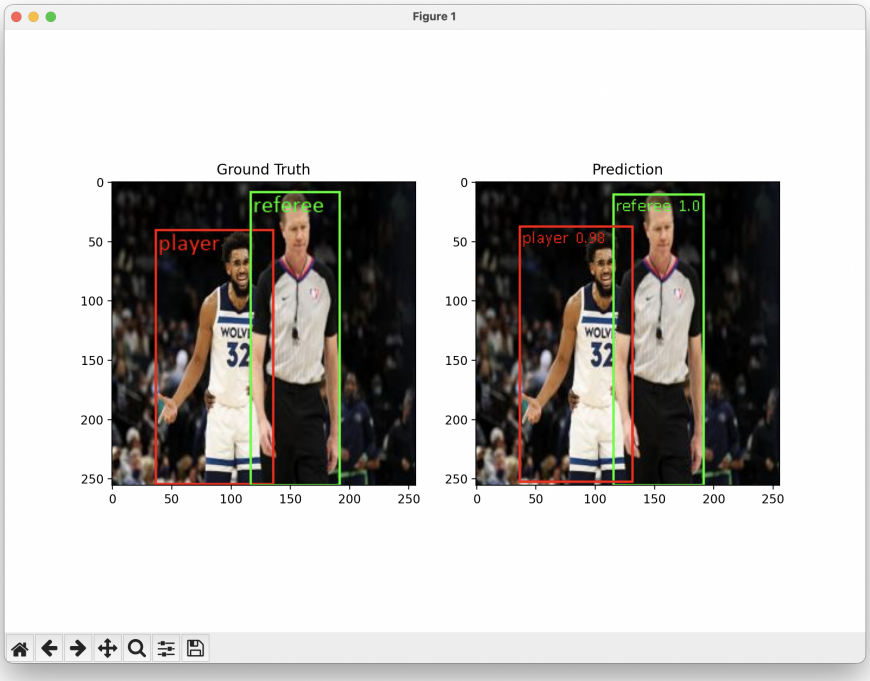
For the validation part we used an approach similar to the one used in the training phase, but this time we set the model in evaluation mode, we iterate over the validation set and we give every image of the set as an input to the model, we then plot both the ground truth and the output of the model. The output of the model in evaluation mode is much different from the output of the model in train mode. This time the output is:

```
# example of output
{'boxes': tensor[x, y, x, y], 'labels': tensor[2], 'scores': tensor[0.9915]}
```

where:

- **boxes**: this tensor represent the predict boundin boxes for the detected object(s) in the image.
- **labels**: is the tensor of predicted labels, for the detected object(s).
- **scores**: is the tensor of values of confidence scores for each bounding box.

During the validation process we decide to plot only the bounding boxes that have a score higher than a fixed threshold.
One output of the validation process:



7.6 Faster R-CNN - results

7.6.1 Training

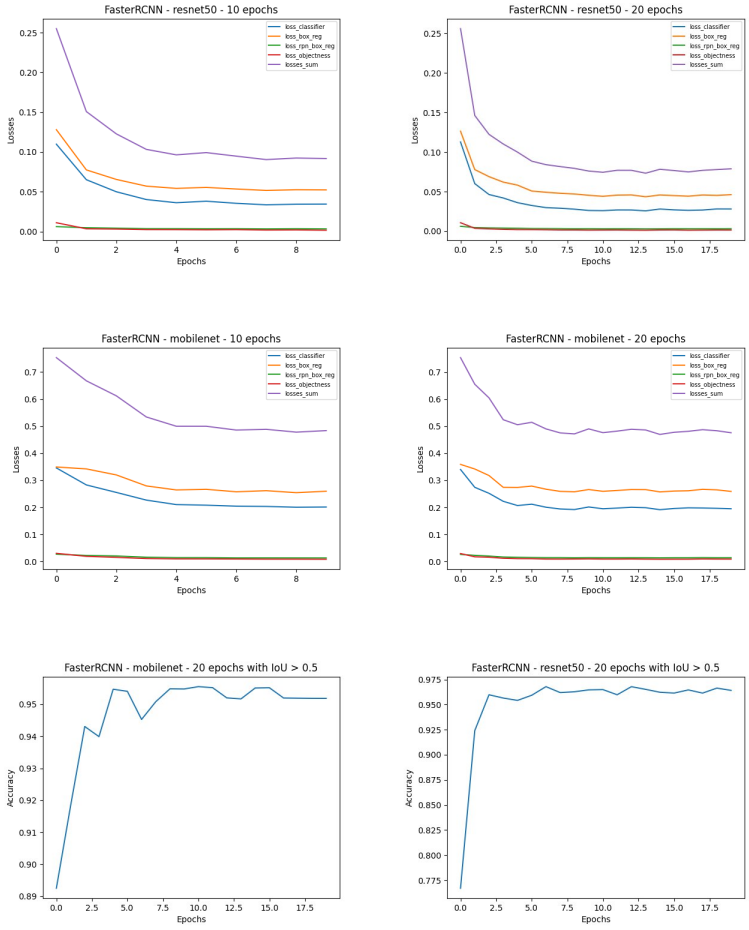


Figure 2: Losses and accuracy of IoU achieved during training

7.6.2 Testing

Statistics achieved during testing, using the *evaluate* function:

```
## ResNet50 10 epochs
IoU metric: bbox
Average Precision (AP) @[ IoU=0.50   | area=  all | maxDets=100 ] = 0.942
```

```
## ResNet50 20 epochs
IoU metric: bbox
Average Precision (AP) @[ IoU=0.50   | area=  all | maxDets=100 ] = 0.958
```

```
## MobileNet 10 epochs
IoU metric: bbox
Average Precision (AP) @[ IoU=0.50   | area=  all | maxDets=100 ] = 0.951
```

```
## MobileNet 20 epochs
IoU metric: bbox
Average Precision (AP) @[ IoU=0.50   | area=  all | maxDets=100 ] = 0.959
```

7.7 YOLO - model preparation

For YOLO we followed the guide provided by the official ultralytics GitHub page (ultralytics), so we first cloned the yolov5 GitHub repo, we installed the requirements and downloaded the Yolov5s weights.

7.8 YOLO - dataset preparation

Since we already had the dataset with the labels in the YOLO format all we had to do was to split the dataset into the different subsets, then separate the images from the labels in each subset folder. Last but not least we had to do was to add a *dataset.yaml* file that is required for training the model.

```
train: /yolov5/data/images/train
val: /yolov5/data/images/test
# number of classes
nc: 2
# class names
names: ['player', 'referee']
```

7.9 YOLO - training

Running the train, once we are all set with the dataset we can actually run the train, using the command line:

```
python train.py --img 640 --batch 16 --epochs 20 --data dataset.yaml --weights
    yolov5s.pt
```

The training comprehends also the validation part and gives as output all the statistics of the both parts and also saves the last and the best weights of the model.

7.10 YOLO - inference

Since the train part comprehends also the validation, all we can do now is inference. This can be done by command line like before:

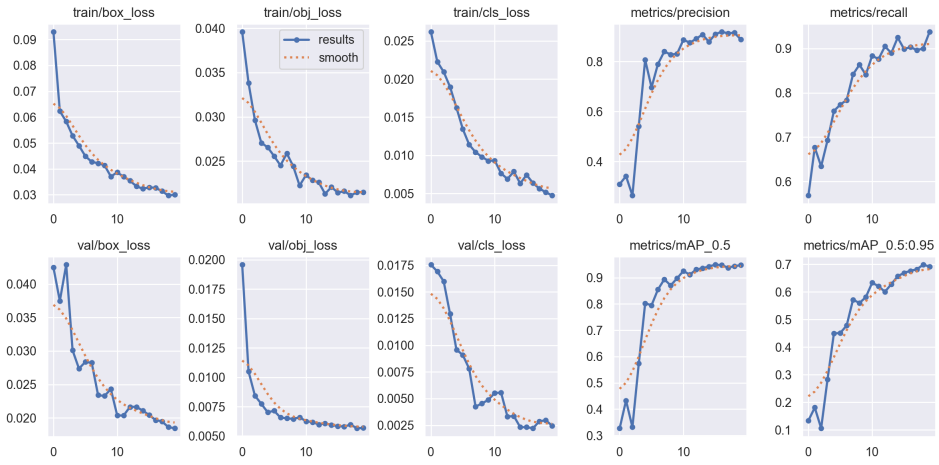
```
python detect.py --weights runs/train/exp/weights/best.pt --img 640 --conf 0.4
    --source image.jpg
```

or loading the weights of the model with pytorch and doing inference on it.

```
model = torch.hub.load('yolov5', 'custom', path='best.pt', source='local')
output = model(image)
```

7.11 YOLO - results

Results acquired during training and validation:



7.11.1 Tests

8 YOLO - ultralytics hub

Ultralytics offers a comprehensive service that enables you to streamline the process of trainin an user-chosen YOLO model using your dataset. It allows you to upload your dataset and to chose a model that you want to train on that dataset, it then opens a colab notebook and trains for 100 epochs. Once the training is complete, ultralytics provides you a range of statistics and evaluation metrics to access the performances of your model. With the ultralytics app, you also gain the ability to test your trained YOLO model in real time.

9 Advanced Programming - Few shot learning

We also deployed a method for training the model on a even smaller dataset (ca. 70 images), achieving good results. This part will only be cited during the presentation for the course Computer Vision & Deep Learning, since is part of the Advanced Programming course.

10 Conclusions

The project aimed to deploy and object detection model specifically designed for detecting NBA players and referees. Two popular object detection algorithms, Faster R-CNN and YOLO, were employed as the methods for this task. The Fast R-CNN algorithm is knowk for the precise bounding box prediction, while YOLO offers fast and efficient real-time object detection. The project involved several steps:

- Dataset collection and annotation.
- Model training.
- Model evaluation.

Despite having a relatively small dataset, the project managed to achieve decent results in NBA player and referee detection using the two models. This outcome highlights the inherent power and effectiveness of these models, even when trained on limited data. We are delighted with the results of our tests because despite the significant variations between the classes, the model performed well, even considering the considerable differences in the uniforms of players and the strikingly similar uniforms of referees. Therefore, the project's outcome not only demonstrates the effectiveness of these two models for object detection tasks but also emphasize the immense potential of these models when trained on a substantial amount of data.