Unsupervised approach to improving player detection in broadcast videos

Chris Andrew

Problem: Object detection approaches tend to have low performance on soccer videos.

Possible causes:

Players are in different scales, orientations, lighting conditions in different images.

Players are generally moving in videos which makes it hard to get to get proper images(blur).

Players occlude each other and tend to be in very close proximity to be able to properly distinguish.

Major Causes:

Object detection datasets are biased towards person images that have minimal movement and small variation in pose, orientation, viewpoint, etc.

Player detection is more close to instance detection.

Multiple instances of the ‘person’ class appear in sport videos.

We need to be able to differentiate between players of different teams, audience, coaches/managers, referees.

General object detectors are trained to focus on the common features among different samples of the ‘person’ class, we need to focus on the differentiating features instead.

Lack of data is a major problem.

- Synthesizing data to get better results
  - An approach published in ICCV 2017 uses a simple approach to synthesize data for instance detection.
  - They argue that detection models focus more on patch level realism than on the image as whole.
  - We can use different backgrounds to generate instances with a higher diversity in patch level realism.
  - Pixel artifacts from using different backgrounds affect the performance of detectors when using synthesized images.
  - Models can be made robust against these pixel artifacts.

Dwibedi, Debidatta, Ishan Misra, and Martial Hebert. "Cut, paste and learn: Surprisingly easy synthesis for instance detection." *Proceedings of the IEEE International Conference on Computer Vision*. 2017. [paper]

There is a large domain shift between object detection and player detection.

It can be argued that general object detection datasets do not provide large enough variance to be able to detect players accurately.

The conditions under which general detection data is collected and player data is collected is very different.

Most detectors are trained using MS Coco and Pascal VOC which contain a large number of pedestrians for person detection and can introduce an inherent bias.

We need to adapt detection models to perform better on player detection.

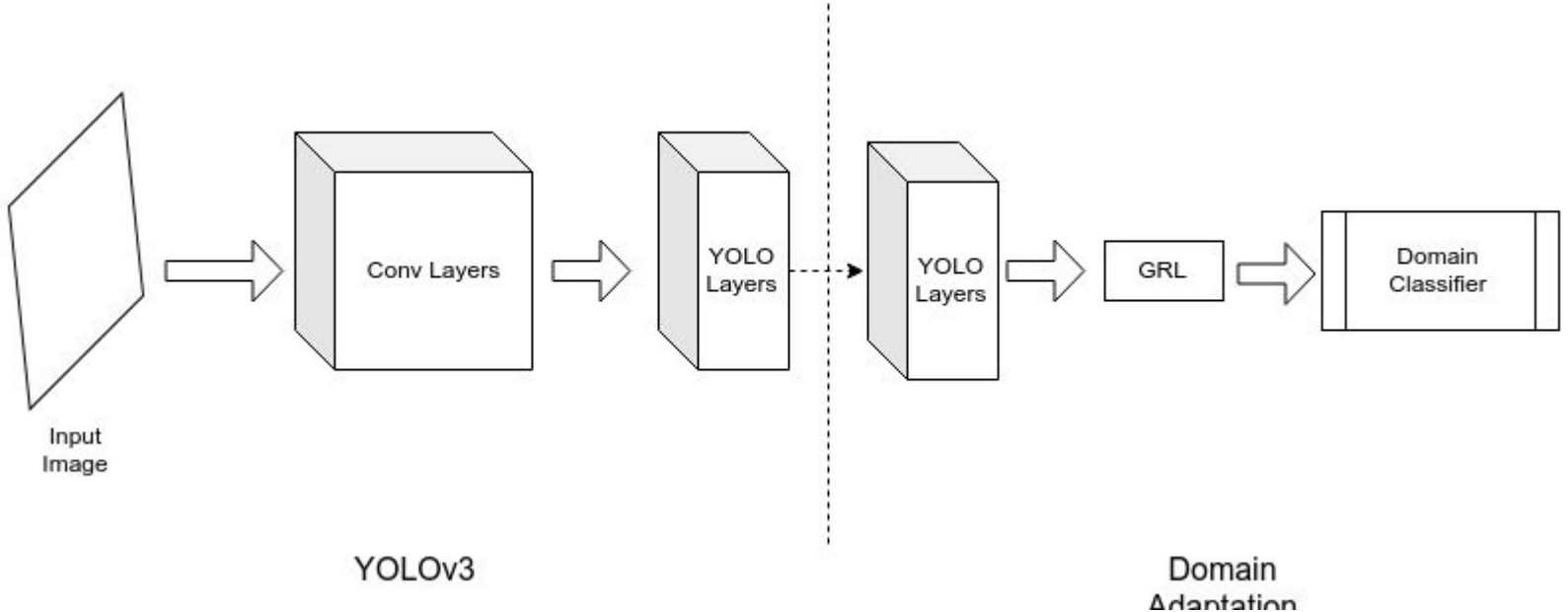
Domain adaptation has a number of methods to help us bridge the gap between the two domains.

A popular approach is using domain classifiers and Gradient Reversal to adapt model weights to a new domain. (ICML 2015)

This approach could be used to detect instance level variations and adapt detection to perform better on sports videos.

Ganin, Yaroslav, and Victor Lempitsky. "Unsupervised Domain Adaptation by Backpropagation." *International Conference on Machine Learning*. 2015. [paper]

Proposed Model



Domain classifier

It will be a convolutional model that takes the X * Y * 255 representations produced by YOLOv3.

Each of the X, Y channels are a prediction made by the model.

We pass this through the domain classification model and calculate loss as cross entropy loss.

Gradient reversal layer will negate the gradients so that instead of learning differentiating representations for the two domains, the model adapts the representations for the two.

Training will be done using the synthesized data without need for manual annotations. We will also combine the classification loss with the loss of the domain adaptation model.

Each batch during training will contain the image pairs from different domains.

Expected results

There have been reports of 5-10% increase in results using the Gradient Reversal approach. However they have not been trained for instances as is the case here.

We also ensure that the synthesised data has patch level realism and the model is trained to be robust against pixel artifacts.

Synthesis allows to have a larger training dataset, which should perform well as compared to other approaches.

Expect to see above 10% jump in accuracy.