YOLOv3 Darknet for Person Detection from COCO Dataset

# What is the Use Case?

The use case involves detecting and identifying persons in a video titled "Driving Downtown - New York City 4K - USA" using a pre-trained object detection model. This is particularly useful for applications such as urban surveillance, pedestrian detection for autonomous vehicles, crowd monitoring, and enhancing safety measures in public spaces. The goal is to accurately identify instances of the "person" class from the COCO dataset, which contains a diverse set of everyday objects.

# Which Model is Used?

The model used for this task is YOLOv3 (You Only Look Once, version 3) implemented in the Darknet framework. YOLOv3 is known for its speed and accuracy in object detection tasks.

# Explain the Model in Short

YOLOv3 is a state-of-the-art, real-time object detection system that divides an image into a grid and predicts bounding boxes and probabilities for each grid cell. The network consists of 106 layers, making it deeper and more accurate than its predecessors, YOLO and YOLOv2. YOLOv3 uses a Darknet- 53 backbone, which is a convolutional neural network with 53 layers trained on ImageNet, providing robust feature extraction capabilities. The model predicts boxes at three different scales, enhancing its ability to detect objects of varying sizes.

# What is the COCO Dataset?

The COCO (Common Objects in Context) dataset is a large-scale object detection, segmentation, and captioning dataset. It contains over 200,000

labeled images with more than 80 object categories. The dataset is widely used in the research community for benchmarking and training deep learning models for various computer vision tasks. The images in COCO are diverse, capturing objects in their natural context, making it a challenging and realistic dataset for model training and evaluation.

1. **Which Class is Detecting?**

In this specific use case, we are focusing on detecting the "person" class from the COCO dataset. The COCO dataset includes 80 different classes, but our objective is to identify and localize instances of persons within the video "Driving Downtown - New York City 4K - USA". The "person" class is one of the most commonly used and significant categories, especially for applications related to human activity recognition and monitoring.

By employing YOLOv3 on this video, we can achieve efficient and accurate person detection, facilitating various real-world applications that require the identification of human presence in visual data.

1. **Frames Per Second (FPS)**

The Frames Per Second (FPS) is calculated by taking the number of rendered frames and dividing it by the seconds passed. In this use case, 100 frames were processed in a total of 83.6 seconds.

FPS = Number of frames/Second Passed FPS=100/83.6

FPS=1.20

Therefore, the FPS for detection of single frames is 1.20.

.