Detect and track people aerially over Stanford

Priyanka Dwivedi (pdwivedi@stanford.edu), Youtube link: https://youtu.be/1BOUWzYEspY

Objective

With the cost of drones decreasing, there is a surge in amount of aerial data being generated. Aerial detection is a challenging problem due to small size of objects, variation in lighting and occlusion due to shadows. There is quite a bit of work on aerial detection of vehicles or buildings but very little on aerial detection of pedestrians. I have built a RetinaNet model to detect people aerially and then integrated RetinaNet into deep sort to track detections

Data Set Used

- Stanford Drone Data set consists of 60 videos shot through a drone over Stanford Campus. Each video includes annotations for pedestrians, bikers, cars and bus though 85%+ of annotations are for pedestrians and bikers.
- The data set is challenging since each class is only a few pixels wide and image resolution is around 1400 x 1100 pixels.
- Different videos were used for training vs test and validation.. Training set had ~35,000 annotations while test and validation had around 7,000 annotations each.

Modeling Approach

Retina Net Model: Is a single stage detector that uses a Feature Pyramid network to extract features at different image resolutions and focal loss for handling class imbalance. Retina Net model was trained on Stanford Drone Data Set

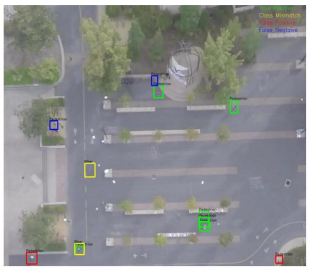
Deep Sort Tracker: Uses a combination of deep features and Kalman Filter for tracking. The detection from RetinaNet were fed into deep sort to assign an ID to a bounding box and track it over frames. `

Experiments and Learning

- Choice of Backbone Model - Experimented with ResNet50, ResNet 101, DenseNet121 and Mobilenet128. Resnet50 worked best
- Transfer Learning or not- Pretrained MSCOCO weights were available for ResNet50. Transfer learning from MS COCO or a previous checkpoint worked much better than training from scratch
- Adding a small anchor box - The anchor box of size 512 was replaced by an anchor box of size 16 as this better captured the annotations in the data set
- Other changes - Also tried random image augmentations, changes to NMS threshold and increasing image size. None of these mattered much

Results

- Best Model: ResNet50 backbone trained using transfer learning and small anchors. Mean Average Precision on Test Set: **0.632**
- Error Analysis was done to visualize true positive (green), class mismatch (yellow), false positive (red) and false negative (blue) bounding boxes as show in left image
- Deep Sort Tracker (right image) worked well in tracking when detections were spaced apart but suffered in crowded scenes



Error Analysis and Next Steps

- There were mispredictions between pedestrian and biker classes since they look similar aerially when there are no shadows. Tracking could help distinguish between them based on speed of motion. A possible next step would be to integrate feedback from tracking back into detection
- Objects can be occluded due to shadows or trees making them very difficult to detect aerially. On the flip side, shadows can sometimes also be confused as detection. To generalize this model better, it would be good to include data from multiple data sets

References

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollár, *Focal Loss for Dense Object Detection*

Nicolai Wojke, Alex Bewley, Dietrich Paulus. *Simple Online and Realtime Tracking with a Deep Association Metric*