Vehicle Detection with YOLO

Jeffrey Gu, Boning Zheng
{jeffgu, b7zheng}@stanford.edu
CS 230 (Deep Learning), Stanford University

Stanford Computer Science

Introduction

- Being able to accurately detect vehicles from videos has many practical applications, including autonomous vehicles
- Current state of art methods for doing this include YOLO, which is capable of real time detections
- Most object detectors are not optimized for video detections and do not take into account temporal information from the video
- Techniques such as sequential non-maximum suppresion aim to improve video detections by using neighboring frames to improve weak detections [1]

Dataset

- We trained our model with the UA-DETRAC dataset consisting of traffic videos and their annotations
- The dataset consists of 60 videos of urban traffic with a total 140K frames, 8250 vehicles and 1.21 million labeled bounding boxes [1]

Fig 1. Image samples from different environments

- The video data was preprocessed into 416x416 images before feeding into YOLO, along with their list of annotated ground-truth object labels

Methodology

For this task, we first trained different variations of the YOLO object detection architectures [2] to perform the object detections, including YOLOv2 and Tiny-YOLO. Below is a summary of the YOLOv2 architecture. The architecture for Tiny-YOLO is similar, but only with 8 convolutional layers in the bulk of the network.

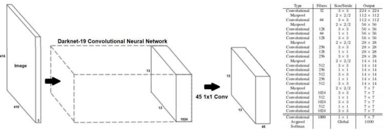

Fig 2. YOLOv2 network architecture

In addition, we use sequential NMS instead of NMS as a postprocessing technique

$$\begin{aligned} i' &= \operatorname*{argmax}_{i_{t_s},\dots,i_{t_e}} \sum_{t=t_s}^{t_e} s_t[i_t] \\ s.t. &\quad 0 \le t_s \le t_e < T \\ s.t. &\quad IoU(b_t[i_t], b_{t+1}[i_{t+1}]) > 0.5, \forall t \in [t_s, t_e) \end{aligned}$$

Fig. 3. The Seq-NMS analogue of object score[3].

Sequence NMS iterates three steps:

1. Find the max sequence subject to the constraint that adjacent frames must be similar (IoU > 0.5)
2. Weak detections in the sequence are then rescored
3. Frames close to the max sequence are then suppressed

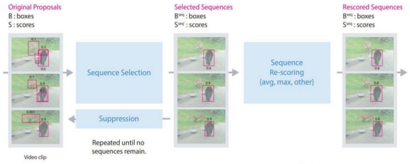

Fig. 4. Sequential NMS algorithm overview [3].

Results

For training and testing, we split our data into a 90/10 train/test ratio. Our test data contains 6 videos from a variety of environments (day/night, rainy/clear) to test the performance of the algorithm under different conditions.

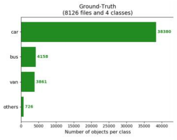

Fig. 4. Ground-truth test labels

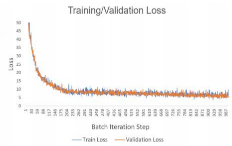

Fig. 5. Training loss curves for YOLOv2

To compare performance between models, we use the average precision (AP), which is the area under the precision/recall curve.

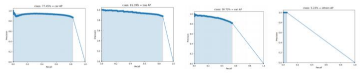

Fig. 5. Precision/recall curves for YOLOv2 (car, bus, van, others)

Model	Hyperparameters	AP: Car	AP: Bus	AP: Van	AP: Other	mAP
Tiny YOLO	S = 13, B = 7, lr = 1e-5, optim = Adam	0.66	0.47	0.44	0.01	0.62
YOLOv2	S = 13, B = 7, lr = 1e-5 to 1e-6, optim = Adam	0.81	0.77	0.51	0.05	0.77

Preliminary results indicate that Sequential NMS postprocessing does worse than normal NMS on a toy subset. Further testing , debugging, and tuning is needed to confirm these results.

Video	Time	Vehicle Orient.	AP: Car	AP: Bus	AP: Van	AP: Other	mAP
MV_20051	Day	Vertical	0.8063	0.9955	0.6088	N/A	0.81
MV_39861	Night	Diagonal	0.584	0.8439	N/A	N/A	0.61
MV_40181	Day	Horizontal	0.9031	0.8225	0.6621	N/A	0.88
MV_40732	Cloudy	Horizontal	0.9431	0.8633	0.3012	N/A	0.88
MV_41063	Day	Diagonal	0.8264	0.7063	0.6001	N/A	0.80
MV_63552	Day	Diagonal	0.7822	N/A	0.7606	0.0024	0.78

Conclusions

- Sequential NMS does not appear to improve the performance of YOLO
- More testing/debugging is needed to confirm this conclusion
- Detection accuracy is higher during the day time compared to night time
- Detection is more accurate for horizontal side-view vehicles than vertically front/back views

Future Works

- Implement real–time sequential NMS if sequential NMS proves to be fruitful
- Investigate and implement techniques that have been show to work better at extracting temporal information for video detection, such as tubelets
- Adapt the techniques mentioned above for real-time video detections

References

[1] L. Wen, D. Du, Z. Cai, Z. Lei, M. Chang, H. Qi, J. Lim, M. Yang, and S. Lyu, "UADETRAC: A new benchmark and protocol for multi-object detection and tracking," arXiv CoRR, vol. abs/1511.04136, 2015.
[2] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. In 2017 IEEE Conference on Computer Vision and Pattern Recognition,CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 6517–6525, 2017
[3] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang, "Seq-nms for video object detection," arXiv preprint arXiv:1602.08465, 2016.