ST-Aware Mask R-CNN Segmentation for Automobile Dual-Camera Videos

Nicholas Tan, Hongshen Zhao

Motivation

Given the recent rise in prevalence of autonomous vehicles, reliable object detection and localization is becoming increasingly paramount for the safety of pedestrians and drivers alike as well as for making functional driving decisions on the road. This work uses, as input, the (unprecedented) dataset from the Kaggle challenge hosted by the 2018 CVPR workshop on autonomous driving (http://www.wad.ai/) to explore various deep neural network segmentation ideas. Video streams are given with the prompt of identifying movable objects, categorizing them, and locating where they are in the image at a pixel level.

Problem

This project aims to turn a stream of video input from two cameras mounted on a car into a stream of labeled masks for each frame. The masks highlight, at the instance level, pixels which belong to various categories of interest. There are six positive categories that our model supports: bicycle, pedestrian/person, truck, car, motorcycle, and bus. Our approach expands on previous work around Mask-RCNN [1] that is used for fast and simple single-image segmentation. Our contribution is to extend the work to input dual-video streams.

Dataset

The Kaggle dataset (https://www.kaggle.com/c/cvpr-2018-autonomous-driving) contains camera image frames taken from a non-stationary car driving on roads and highways. The dataset is unprecedented, with no open-source model previously trained on this data available to the public. The training data contains 100+GB of labeled RGB images that are 3384x2710 in size. There are 39222 image frames and labels. We portioned the data into 3000 images for training, 900 for validation, and 100 for testing, randomly sampled from the dataset. The images are organized in video streams each varying between 270-1530 frames. The frame rate for the streams is about 7.2Hz.

Method

Architecture

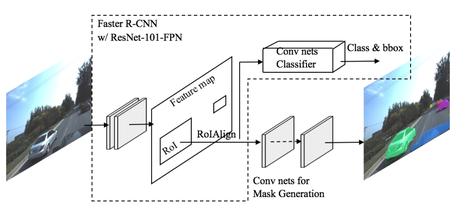

Figure 1: Mask R-CNN Network Architecture

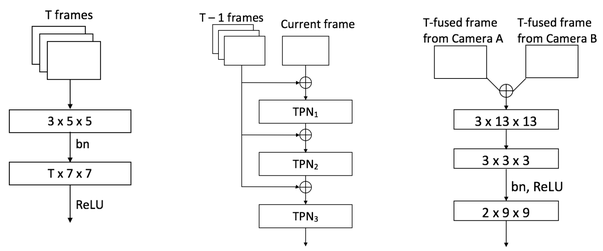

Figure 2: ST Pre-Networks: (a) TPN cell; (b) Temporal Pre-Network; (c) Spatial Pre-Network

The Mask R-CNN in our model uses the ResNet-101 [2] backbone with a Feature Pyramid Network (FPN) [3] head. The convolutional backbone architecture is used for feature extraction over an entire image. The head is used for bounding-box classification and regression, and mask prediction separately for each RoI.

Multi-task loss

$$L = L_{\text{cls}} + L_{\text{box}} + L_{\text{mask}}$$

$$L_{\text{cls}}(p, c) = -\log p_c \qquad L_{\text{mask}} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} (y_{ij} \log(p_c) + (1 - y_{ij}) \log(1 - p_c))$$

$$L_{\text{box}}(gtb, t^c) = \mathbb{1}_{c \text{ not background}} \lambda \sum_{i \in x,y,h,w} |gtb_i - t_i^c|_{\text{smooth}}$$

$$|d|_{\text{smooth}} = \begin{cases} 0.5d^2, & \text{if } |d| \le 1 \\ |d| - 0.5, & \text{otherwise} \end{cases}$$

Our loss function on each sampled RoI associated with ground truth class c is defined the same as the Multi-task loss for Mask R-CNN

Pass-through Initialization

$$\int_{\mathbf{R}^2} f(\mathbf{x} - \tilde{\mathbf{x}})(\delta\{d\tilde{\mathbf{x}}\} + W_l d\tilde{\mathbf{x}})$$

Our target filter weights should preserve the general structure of the original image, as the subsequent Mask R-CNN subsystem is pretrained on regular photograph-like images, and would thus perform best given photograph-like inputs. Our proposed filter weights would mimic the identity kernel for convolution, the Dirac-delta, to pass the image through each layer.

Metrics

IoU between a predicted instance A and a ground truth instance B:

$$IoU(A, B) = \frac{A \cap B}{A \cup B} \qquad \text{precision} = \frac{tp}{tp + fp}, \text{recall} = \frac{tp}{tp + fn}$$

We first calculate the AP and AR for all instances across all object classes; taking the mean across all video frames and a range of Intersection over Union thresholds IoU_t gives mAP and mAR. The range of Intersection over Union thresholds under our consideration is IoUt ∈ [0.5 : 0.05 : 0.95]

Experiments

Hyperparameter Tuning

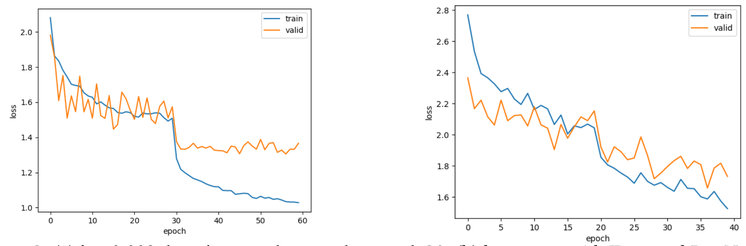

Figure 3: (a) λ = 0.003, learning rate decreased at epoch 30; (b) loss curve with Temporal Pre-Network

Noise Injection

	mAP	mAR
Original	0.2042	0.2108
With Guassian blur	0.1765	0.1867
With Guassian noise	0.1770	0.1796
With pixel suppression	0.1907	0.2108

Table 1: mAP and mAR with Image Processing Variations

Results

(a) (b) (c)

Figure 4: (a) MaskR-CNN result; (b) Mask R-CNN w/ TPN result; (c) Mask R-CNN w/ TPN and SPN result

Case	mAP_train	mAR_train	mAP_test	mAR_test
COCO pre-trained	0.0609	0.0816	0.0572	0.0851
Kaggle trained	0.194	0.225	0.206	0.233
w/ Temporal Pre-Network	0.149	0.168	0.104	0.123
w/ Spatial and Temporal Pre-Networks	0.270	0.237	0.145	0.179

Table 2: Comparing mAP and mAR @[0.5:0.95] IoU thresholds on Model Variations

References

[1] K. He, G. Gkioxari, P. Dolla´r, and R. B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.
[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
[3] T. Lin, P. Dolla´r, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016.