Mask R-CNN Application: Instance Segmentation in Driving Scenes

Shaw Lu, Michelle Zhang, Chen Luo

WN18 CS230 Final Project
Mentor: Patrick Cho

{shawlu, zhangmx, chenl3}
@stanford.edu

OBJECTIVE

- Investigate state-of-the-art Mask R-CNN model in instance segmentation
- Apply Mask R-CNN (Facebook-2017 [1]) model to newly-released Mapillary dataset [2] in driving context
- Train and analyze performance of Neural Network application in an iterative process
- Aim at understanding traffic scenes that applies to self-driving perception

MAPILLARY DATASET

- Authenticity: Road-side and ground view on multiple locations around the world with a variety of weather, season, time of day
- Fine-granularity: 37 object classes with pixel-wise instance-level annotations
- Diversity: Diverse set of resolutions, aspect ratios and viewpoints from mobile-uploaded images by users

Target classes are defined as the intersection of Mapillary and MS COCO datasets:

Bird	Person	Bicyclist	Motorcyclist
Bench	Car	Fire Hydrant	Traffic Light
Bus	Motorcycle	Truck	Background

The 20,000 fine-annotated pictures in Mapillary is pre-processed, filtering out images without classes of interests.

Table 1: Split of datasets

Train	Dev	Test
16384	1024	1024

METHODOLOGY

Mask R-CNN Framework

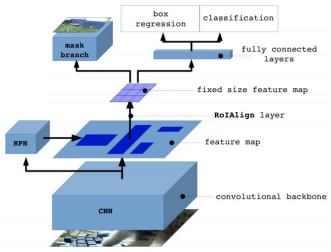

Fig. 1: The Mask R-CNN framework

- Extends Faster R-CNN with pixel-level segmentation
- Decouples classification (class prediction)/ bounding box regression (object detection), and binary mask generation (segmentation)

Implementation Details:

- Use ResNet-101 and Feature Pyramid Network (FPN) as CNN backbone
- Define loss as
  $$L = L_{class} + L_{box} + L_{mask}$$
  - Both RPN and Mask R-CNN classifier head contribute L_{class} and L_{box}
  - Mask loss is defined only positive ROIs which IoU > 0.5.
- Leverage transfer learning from pre-trained weights on MS COCO datasets
- Train from ResNet stage 5 and up (5+) to achieve higher efficiency

Hyperparameters Tuning

Started with Facebook configuration [3], we explored hyperparameter settings:

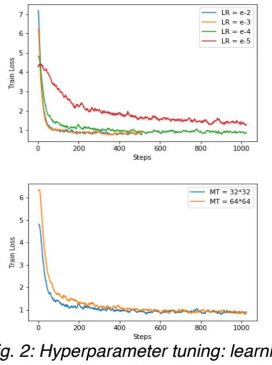

Fig. 2: Hyperparameter tuning: learning rate (top) and mask thresholds (bottom)

Table 2: Performance comparison

H-parameter	Value
Learning Rate	0.001
Mask threshold	32 x 32
Weight decay	0.0001
Momentum	0.9
Mini-batch size	8
Image size	1024 x 1024
Steps per epoch	128

RESULTS

Evaluation is defined by COCO metrics, where averaged precision (AP) is averaged over *IoU* thresholds [0.50, 0.95], with increment of 0.05.

Table 3: Performance comparison

Network	Classes	*MT*	*AP*	AP_{50}	AP_{75}
Facebook	80	-	35.7	58.0	37.8
Baseline	12	32^2	29.6	47.1	36.2
DreamNet	12	32^2	36.8	58.1	45.2
Baseline	12	64^2	40.6	47.1	36.2
DreamNet	12	64^2	49.6	72.4	61.2
Baseline	38	32^2	5.6	12.1	6.3
DreamNet	38	32^2	16.0	26.6	19.9

Fig. 3: DreamNet results on Mapillary test set

DISCUSSIONS

- Significantly improved precisions of 12 classes with large mask thresholds.
- Results of 38 classes need further exploration to investigate overfit/ underfit problem.
- Mask threshold needs to be revisited based on applications.

Reference: [1] K. He, G. Gkioxari, P. Dollas, R. Girshick, "Mask r-cnn," 2017. [2] G. Neuhold, T. Ollmann, S. R. Bul, P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," 2017. [3] "Mask r-cnn for object detection and segmentation,"