Swift and Smart: A New Paradigm for Real-Time Garbage Segmentation

Robin Wang and Jun Wang
{qwang6, junw3}@cs.washington.edu
Paul G. Allen School of Computer Science and Engineering

Introduction

- Sorting waste into appropriate bins is a critical task, and improper trash disposal can have severe consequences (e.g., environmental, economic).
- Existing garbage classification algorithms struggle to accurately identify and sort waste that is mixed or contaminated with other materials.
- We introduce a real-time semantic segmentation solution featuring a two-stage, end-to-end pipeline.

Method

Balance 12 garbage classes (~700 images/class)

⬇️

Grounding DINO[1] (Zero-Shot object detection)

⬇️

Segment Anything Model[2] (instance segmentation)

⬇️

Integrate the result with Roboflow API

⬇️

Data augmentation (resizing, random cropping, horizontal flip, rotation, ...)

⬇️

ResNet50[3] + DeepLabV3[4],
MobileNetV3[5] + DeepLabv3,
ResNet50 + FCN (baseline)

⬇️

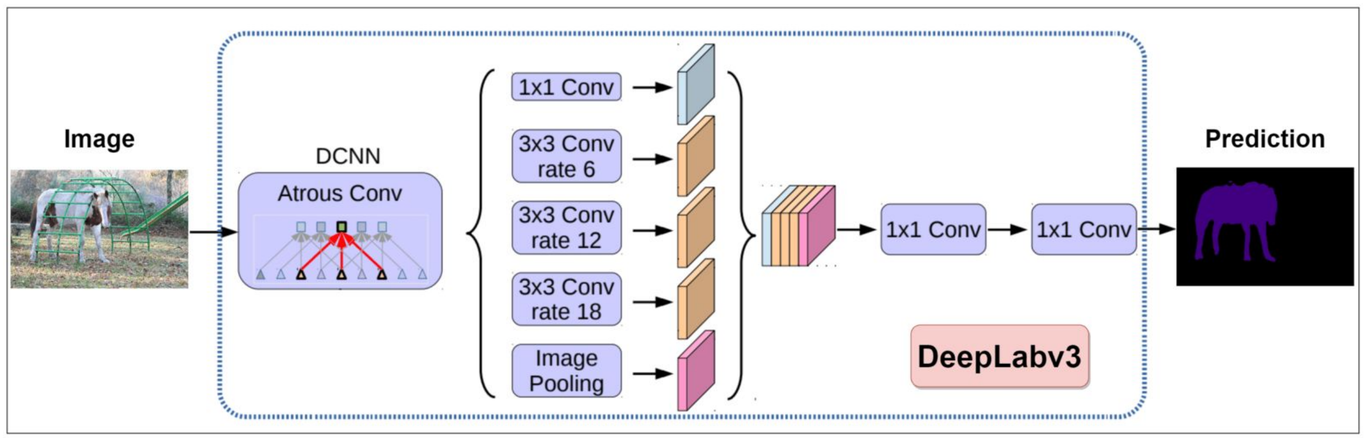
Grid search on batch size, learning rate, weight decay, step size, optimizer, learning rate decay

⬇️

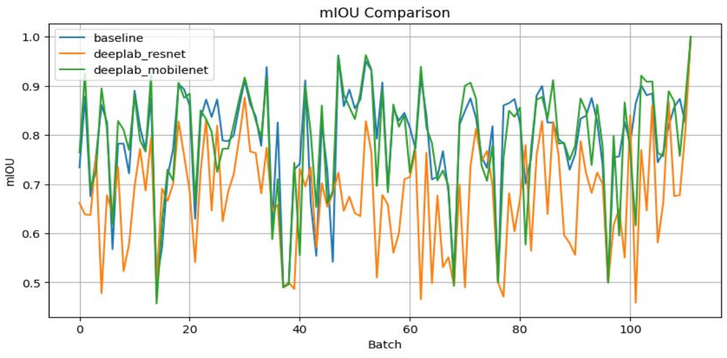
Measure MIoU, average recall and precision, FPS

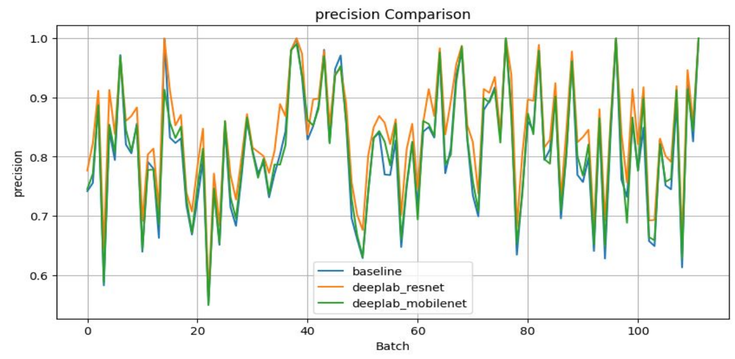
⬇️

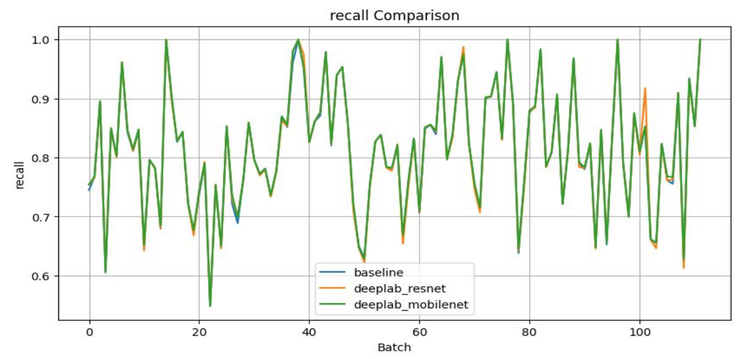
Evaluate results



Results







	ResNet50 + DeepLabv3	MobileNetV3 + DeepLabv3	ResNet50 + FCN (head)
Best MIoU	0.7911	0.7865	0.8011
Best Average Precision	0.8350	0.8147	0.8129
Best Average Recall	0.8490	0.8112	0.8039
Best FPS	118.48	218.93	170.85

Hyperparameters:

- 7 epochs (6k iterations)
- Batch size = 8
- AdamW optimizer
- Initial learning rate of 5e-5 and weight decay of 0.1 for ResNet50 backed models
- Initial learning rate of 1e-4 an weight decay of 0.01 for MobileNetV3 backed models.
- Training scheduler with a step size of 200 iterations and a decay rate of 0.1.

Findings:

- Adopting a light-weighted backbone or a simpler head improves inference speed.
- While the combination of ResNet50 and DeepLabv3 achieves the highest accuracy, it falls short in FPS and MIoU.
- MobileNetV3+DeepLabv3 reveals the strongest deployment applicability.

Conclusion

- MobileNetV3 + DeepLabv3 is the best performing model with the strongest deployment applicability.
- The quality and quantity of the dataset can be improved.
- Future works could explore various vision transformers (e.g., MobileViT, Transformer iN Transformer)
- Future works could determine a suitable number of classes based on user studies and real-world analysis.

References:
[1] Shilong Liu et.al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *arXiv preprint arXiv:2303.05499, 2023.*
[2] Alexander Kirillov et.al. Segment anything. In *arXiv preprint arXiv: 2304.02643, 2023*
[3] Kaiming He et.al. Deep residual learning for image recognition. In *CVPR, 2016*
[4] Liang-Chien Chen et.al. Rethinking atrous convolution for semantic image segmentation. In *arXiv preprint arXiv:1706.05587*
[5] Andrew Howard et.al. Searching for mobilenetv3. In *arXiv preprint arXiv:1905.02244, 2019.*