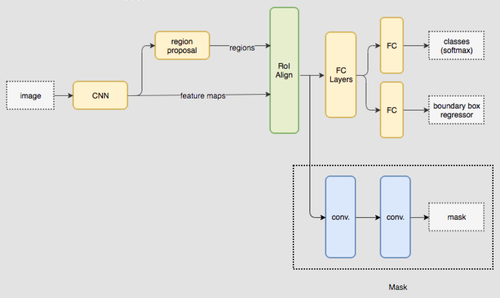
Vehicle Detection and Segmentation With Mask-R-CNN

Zhe Zhou, Jiaxuan Sun, Jen Wang
Georgetown University, Analytics – Data Science Program

Abstract

Recent advances in deep learning and computation infrastructure(cloud, GPUs etc.) have made computer vision applications leap forward very fast. Convolutional neural networks (CNN), the driver behind computer vision, are fast evolving with advanced and innovative architectures to facilitate computer vision applications especially those related to pattern recognition. Since 2018, Mask-R-CNN has been the latest state of art in terms of instance recognition and segmentation. In this project, we propose to use Mask-R-CNN network architecture to conduct vehicle recognition and segmentation. We use Matterport's (MIT) implementation of Mask-R-CNN with different sets of weights from transfer learning to train on our own synthetic dataset, then use the trained weights to run inference on new images.

Methodology



Mask-R-CNN is an extension of Faster-R-CNN that works in two stages with a total of four parts:

- **Stage 1** identifies Regions of Interest (ROI), in two parts. The image is fed into an FPN + ResNet50 "backbone", which outputs a feature map. A Region Proposal Network (RPN) then scans over this feature map, convolutionally evaluating multiple anchors simultaneously and identifying the ROI with the anchor and a simple foreground or background evaluation.
- **Stage 2** analyzes the regions considered foreground, generating masks for objects, and in parallel classifying objects to which the masks can be applied.

Configurations

This project was taken place on my local workstation with the following hardware and software configurations:

Hardware:
- CPU: Ryzen 2700x
- RAM: 32G
- GPU: RTX2070

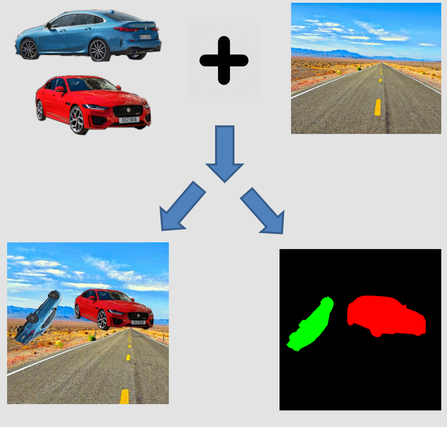
Software:
- CUDA 10.0
- CUDNN 7.5.0
- Tensorflow-GPU 1.13.1

Dataset

Self-created synthetic dataset was used for this project.

Dataset Generation:
Patterns and background images have to be prepared first manually. Then dataset-generation program would randomly take certain number of patterns and one background to generate sufficient number of COCO-like image sets. Here COCO-like image set means an image with its corresponding pattern mask.

The example below should explain better:

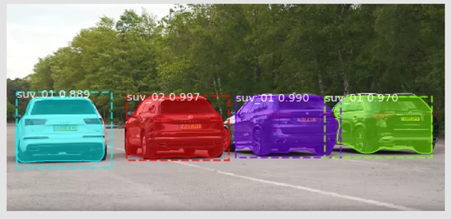


Results

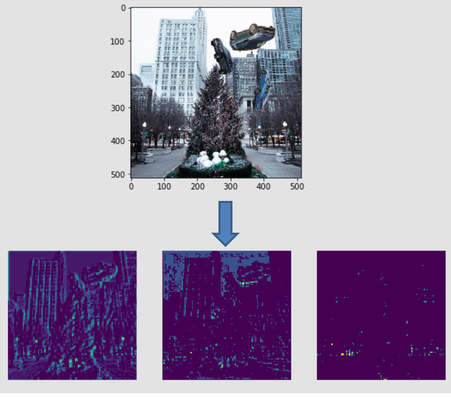
- Inference on synthetic Image (patterns trained):

- Inference on actual Image (pattern trained):

- Inference on actual Image (patterns not trained):



- Feature Visualization (Res2a Layer example):

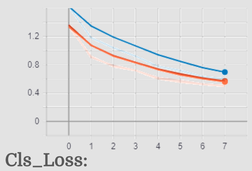


Results (Continued)

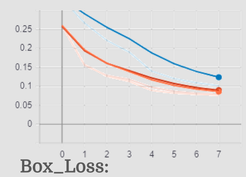
- Loss Function Visualization:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}}$$

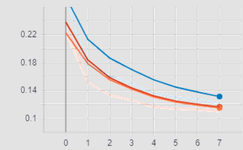
Overall_Loss



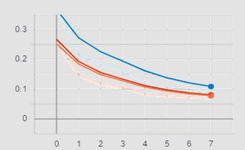
Mask_Loss:



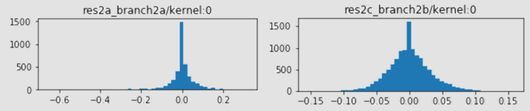
Cls_Loss:



Box_Loss:



- Weights Histograms (Res2a and Res2c showcase):



Conclusion & Further Improvement

For trained patterns, our model can reach 98% vehicle identification rate and 96% vehicle classification rate, which is pretty good. Given the fact we only have the time to prepare 4 specific models of cars, this high classification rate is not that representative. However, once we have more kinds and models of cars prepared, classification will play a more important role. What is more promising is that our model was able to identify and segment vehicles which have never been trained in the model, as shown on the left.

References

- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask R-CNN." arXiv preprint arXiv:1703.06870, 2017
- "A Brief History of CNNs in Image Segmentation: From R-CNN to Mask R-CNN" by Athelas