Enhancing Autonomous Navigation Few-Shot 3D Object Detection Using Depth Supervision

Ishita Mangla
Department of Computer Science
Stanford University
imangla@stanford.edu

Ishan Sabane
Department of Electrical Engineering
Stanford University
ishancs@stanford.edu

Tejas Y. Deo
Department of Mechanical Engineering
Stanford University
tejasdeo@stanford.edu

Motivation and Objective

Meta and transfer learning are promising approaches for improving the performance of few-shot object detection models on unseen tasks. We use the MAML (Model Agnostic Meta Learning) approach to train an N-shot K-way classifier on 2D bounding box data for predicting 3D bounding boxes on unseen tasks. While YOLO, [1] SSD [2] and Fast RCNN [3] are promising backbones trained on large amount of images for 2D/3D object detection, recent trends in transformers have enable single phase object detection easier. We will extend our few-shot object detection objective [4] to datasets such as KITTI [5], Nuscenes [6], and FSOD [4].

Experimental Setup

In our first set of experiments, we use the DETR model (ResNet backbone) available on HuggingFace with the KITTI [5] dataset. We test our model for few shot 2D object dectection where we implement 3-way 10-shot object detection.

Dataset: The KITTI dataset is very large in size, encompassing seven classes. As a dataset designed for self-driving cars, each image within it may feature instances of all seven classes, with multiple occurrences of each. To approach this challenge as a meta-learning problem and facilitate the generation of diverse tasks, we initially developed a preprocessing script. This script involved iterating through all the images and their corresponding text files, creating separate folders for each class. Additionally, a labels folder was generated to exclusively house the ground truth 2D and 3D bounding box information for the respective labels. We kept the same directory structure as that of the Omniglot Dataset with just the addition of another labels folder inside each class folder.

Figure 1: Kitti Images with 2D & 3D ground truth bounding boxes generated from the utils file using camera calibration matrix

Following the data pre-processing phase, we created two files: the KITTI custom dataloader and a dataloader utils file. The latter not only incorporates functions for the dataloader but also takes in the predicted 2D and 3D bounding box values and projecting these values onto the images using the camera calibration matrix. These GitHub repositories helped us advance on this end [7], [8]. We make use of the load_data.py script provided in Homework 1 for the dataloader and extend it to our use case.

Backbone: The DETR Model uses the Resnet 50 Model as the CNN backbone to produce image features. These are passed to a Transformer based Encoder-Decoder Model. The decoder is provided a fixed set of learnable queries and then the output hidden states are passed to the simple MLP bounding box and classification heads.

MAML: To implement the MAML framework, we make use of the code provided in Homework 2 and extend it to few shot object detection. We make use of inbuilt loss functions from the DETR model to train the inner and outer loop of MAML. To evaluate the performance, we calculate the IoU for each bounding box and match it with the ground truths. Finally, we make use of Average Precision for IoU threshold of 0.5 and average it across classes. Our GitHub repo: https://github.com/Tejas-Deo/CS330_final_project

Results

To implement a model which takes monocular images along with 2D bounding boxes in the support set and then predicts 3D bounding boxes for the respectively classes in query set, we begin by modeling the two tasks individually and then integrating them later. As a first experiment, we wanted to make sure we could fine-tune KITTI data end-to-end for our inner training loop.

2D Object Detection:

We run MAML for 3-way 10-shot task and randomise the object categories for each task from the kitti dataset. We keep 3 categories for the testing phase. We run MAML for 100 outer steps and 5 inner steps. We set inner learning rates to be learnable while keeping the outer learning rate as 0.001. We use the Bipartite Matching loss for bounding boxes and Cross Entropy for classification. We get a very high loss due to differences in class labels and bounding box structure (8.27). While compute intensive, we plan to increase the number of epochs in the future experiments.

Next Steps

We are currently fixing the 2D object detection by speeding up the dataloader to increase epochs. Moving ahead, we focus to augment our models to understand 3D bounding boxes from the predicted 2D bounding boxes by using concepts from prior works in monocular 3D object detection [9, 10]. This requires newer prediction heads and different loss functions for the inner and outer steps. We plan to do ablation studies to understand effects of different backbones, depth processing techniques and primarily effectiveness of various N-way K-shot methods.

References

[1] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[2] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[3] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[4] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *CVPR*, 2020.

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[7] Visualize kitti objects in videos. https://github.com/HengLan/Visualize-KITTI-Objects-in-Videos.

[8] Kitti object data transformation and visualization. https://github.com/kuixu/kitti_object_vis/tree/master.

[9] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017.

[10] Jason Ku, Alex D Pon, and Steven L Waslander. Monocular 3d object detection leveraging accurate proposals and shape reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11867–11876, 2019.