GROUP 15: Report 2

Xiang Liu Han Wang

Abstract

The object of the task is to detect 3D objections in the driving scenes. The pipeline is composed of two stages, stage-1 of RPN(Region Proposal Network) architecture and stage-2 of proposal refinement. Our work is to try to improve the detection performance by implementing different approaches in the second stage based on the output data from stage-1. Specifically, we canonically transform the point clouds, incorporate local spatial features into feature learning for box proposal refinement, and replace cross-entropy loss with focal loss in classification tasks.

1. Introduction

Describe your problem and state your contributions. What is the shortcoming of the baseline method that you intend to solve?

The problem's input includes the coarse detections $(x, y, z, h, w, l, \theta)$ from stage-1, the RPN features with 128 channels, the coordinates of points, the intensity, and the ground-truth boxes. In the baseline, stage-2 adopts several layers of *Set Abstraction* made of a sampling layer, a grouping layer, and a PointNet layer. Then convolution layers are used for bounding box predictions and getting the confidence score, given the features extracted from *Set Abstraction* level. However, there are some shortcomings of the baseline, like the poor ability to deal with the imbalanced density of point clouds and the rare exploitation of spatial relationships between points. Therefore, the problem is to figure out other methods to refine the coarse detection results from stage-1, making it closer to the target. The approaches we try are as follows:

1) Combine local spatial features derived by canonical transformation and the global semantic features received from stage-1 to learn for proposal refinement.

2) Replace the cross entropy loss in classification with focal loss.

2. Related Work

Survey the related work. What has been done in this line of work? Where do your contributions stand in comparison?

Hint: Novelty in your contributions is not a requisite but highly valued. If you intend to re-implement existing modules, go into detail on how they work and what they intend to solve.

Several attempts have been made to overcome the shortcomings. PointRCNN [3] is introduced to improve the 3D object detection by fully exploiting the spatial features and adopting a novel bin-based loss function. In this work, we combine the global feature and the local spatial feature by canonical transformation to get a new feature, fed into the *Set Abstraction* layers and *CNN* (Convolutional Neural Network) to get the final predictions.

Besides, the class imbalance is a prevalent problem in object detection, as there are much more background samples than foreground samples. To mitigate this, we replace the basic cross-entropy loss with focal loss introduced by Lin *et al.* [1].

3. Method

Describe your idea and how it was implemented to solve the problem.

3.1. Integrate local spatial feature

We transform all the points' coordinates from stage-1 to the corresponding CCS (Canonical Coordinate Systems) and obtain new coordinates p, where the origins are located at the center of the proposals, the $x-axis$ is toward the head direction of the proposals, and parallel to the ground plane, the $y-axis$ is the same as the LiDAR system and the $z-axis$ is perpendicular to $x-axis$ and $y-axis$. And every point is assigned a segmentation mask m about whether it is a foreground or background point. Moreover, we take the point intensity r and point distances d from the center of LiDAR coordinate system into account to integrate depth information into the feature. The SharedMLP layers are applied to extract the local spatial feature from (p, m, r, d), combined with the global feature from stage-1 to get a new feature. Finally, *Set Abstraction* layers are adopted to further extract the feature, fed into convolutional layers to generate the regression box parameters and the confidence score.

3.2. Focal Loss

A Focal Loss function addresses class imbalance during training in tasks like object detection. Focal loss applies a modulating term to the cross-entropy loss to focus learning on hard, misclassified examples. It is a dynamically scaled cross-entropy loss, where the scaling factor decays to zero as confidence in the correct class increases. Intuitively, this scaling factor can automatically down-weight the contribution of easy examples during training and rapidly focus the model on hard examples.

Formally, the Focal Loss adds a factor $(1 - p_t)^\gamma$ to the standard cross entropy criterion. Setting $\gamma > 0$ reduces the relative loss for well-classified examples ($p_t > .5$), putting more focus on hard, mis-classified examples.

$$FL(pt) = -(1 - p_t)^\gamma log(p_t) \tag{1}$$

4. Results

Show evidence to support your claims made in the introduction. Compare your proposed method to the baseline from Problem 1.

Hint: If you have implemented multiple modules, isolate their roles in the outcome by providing ablation studies. Show us how they affect the results with example figures. You can use your visualization code from Project 1 for this task. *Tip*: While you should report your final score on the test set, any further ablation studies required should be conducted on the validation set as you are limited to only 10 submissions.

4.1. Integrate local spatial feature

Firstly, we only consider (p, r, d) as the local spatial feature, and the experiment shows that its performance is about the same as the baseline. At the end of the train, epoch 35th, the training loss of baseline (Fig. 1) is 0.3191, while the training loss of the new method (Fig. 2) is 0.3155. Figure. 2 starts from epoch 6 because of the recovery from a checkpoint. Here (Fig. 3 and Fig. 4) visually show the predictions of the baseline and method 1, respectively, also demonstrating a similar performance that gets a high IOU about the close objects and a relative low IOU about the distant objects.

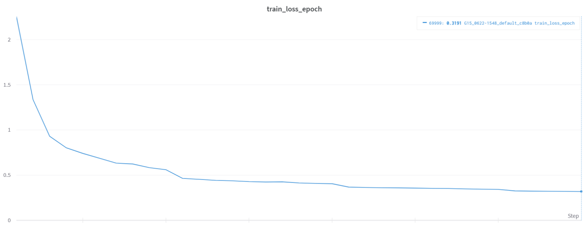

Figure 1. Train Loss Epoch of Baseline

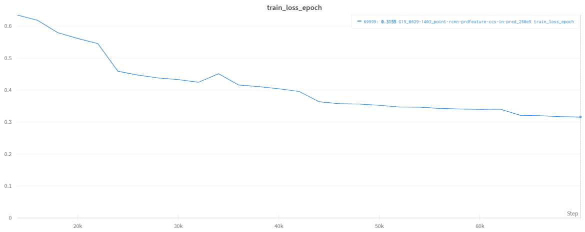

Figure 2. Train Loss Epoch of Method 1.1

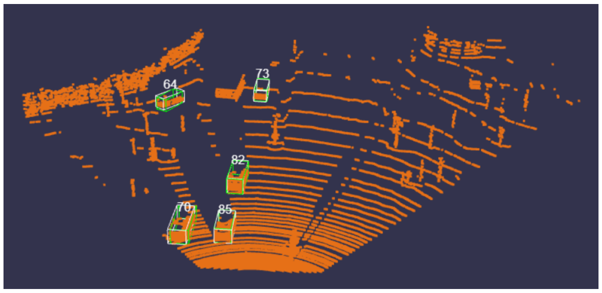

Figure 3. Prediction of Baseline, Epoch 35

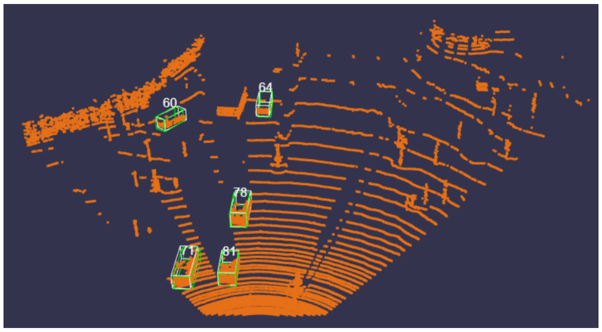

Figure 4. Prediction of Method 1.1, Epoch 35

Moreover, we consider the prediction mask and generate the local spatial feature given (p, m, r, d) without changing the cost function. In epoch 5, the training loss of this method is 0.2829, less than that of baseline, 0.6869. As expected, the prediction results perform better than the baseline (Fig. 5 and Fig. 6). When detecting close objects, our method gets higher IOUs (75, 73, 68) than baseline (61, 65, 74). Moreover, our method successfully detects all distant objects despite low IOUs. We believe we will get a better final result if we complete the training process.

Based on the previous study, we also try to redefine another loss function, bin-based box refinement. This method fails to get the prediction boxes with high IOU even about the close objects (Fig. 7). There, unfortunately, might be something wrong in the code.

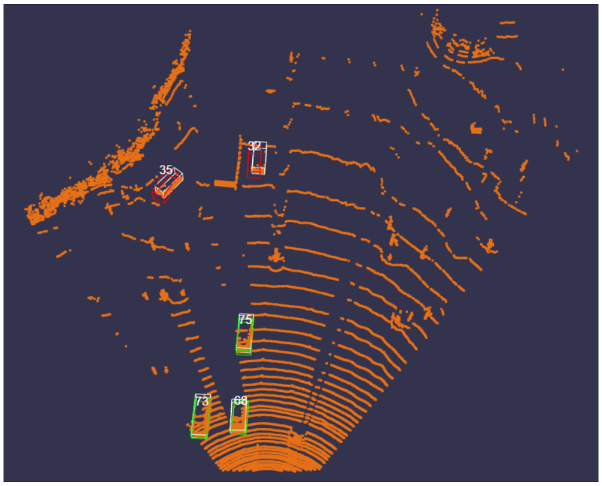

Figure 5. Prediction of Method 1.2, Epoch 5

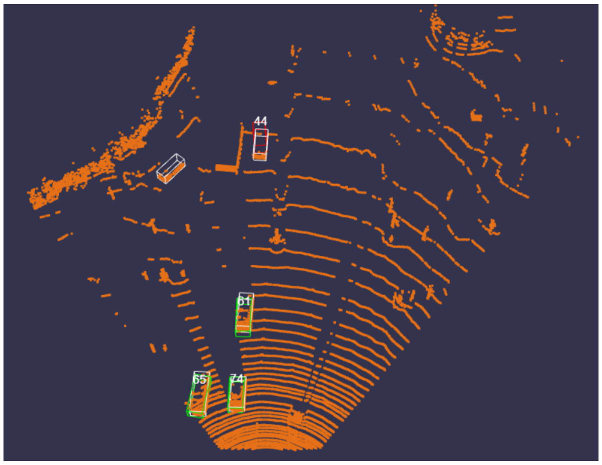

Figure 6. Prediction of Baseline, Epoch 5

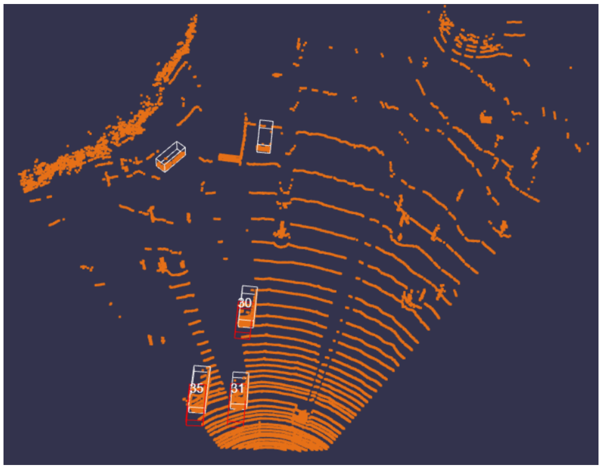

Figure 7. Prediction of Method 1.3, Epoch 15

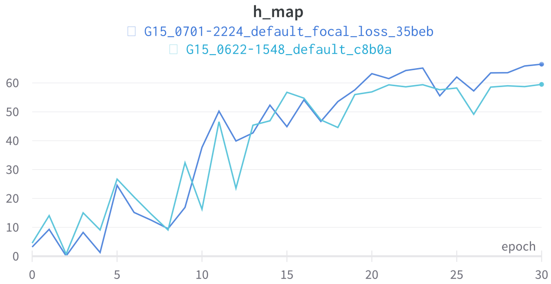

Figure 8. Hard examples comparison of focal loss and cross entropy loss

4.2. Focal Loss

As shown in Fig. 8, the model with focal loss performs better in detecting hard examples, as focal loss puts more weight on optimizing *w.r.t.* hard examples.

5. Conclusion

Due to limited computational resources, we have not strictly conducted exhaustive experiments to show our methods' effectiveness. However, as these are mature ideas that are verified by early works, thus they are expected to work empirically.

Besides, we also found an interesting idea called Pyramid R-CNN [2] that could be incorporated into our model, which is built upon PointRCNN [3].

References

[1] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[2] Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. Pyramid r-cnn: Towards better performance and adaptability for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2723–2732, 2021. 3

[3] Qiang Zhou and Chaohui Yu. Point rcnn: An angle-free framework for rotated object detection. *Remote Sensing*, 14(11):2605, 2022. 1, 3