3D Indoor Scene Long Tail Segmentation

Team Name: NetDB603

Team Member: R10921129 楊秉蒼、R10921A08 蕭可宣、R10921A13 劉羽忻、R10942083 鄭淑綾、R10921069 沈郁鈞

Abstract

Because 3D devices have become more affordable recently, there are more and more 3D datasets released. One of them is ScanNet200, which is a benchmark for 3D semantic segmentation and instance segmentation. In this work, we use ScanNet200 to train a 3D U-Net to perform 3D semantic segmentation and compute mIoU as an evaluation metric. However, the dataset is highly imbalanced, so we adopt several augmentation techniques and implement some adjustments to the loss function to mitigate the effect of data imbalance.

Experiment Results

Method	mIoU Score

Conclusion

In this work, we implement a deep learning scheme for large-scale 3D scene segmentation using 3D U-Net. First, we adopt Mix3D augmentation to generalize beyond the contextual priors in the training set. By doing so, it's harder for the model to rely on scene context alone, and instead infer semantics from local structure as well. Furthermore, we implement instance augmentation for rarely-seen class category instances to address imbalance data distribution, and also break overly specific context dependencies for recognition. In experiments, we have demonstrated the effectiveness of the proposed method.

Methodology

❖ Spatio-Temporal ConvNets

We adopt a sparse 3D U-Net implemented with Minkowski Engine as our baseline. Since the sparse convolution provided by Minkowski Engine is similar to standard convolution, conventional 2D CNN-based segmentation model can be transformed to 3D version without too much effort. We also apply basic data augmentation in our baseline setting, including random scaling, rotation, color jitter, etc.

❖ Mix3d augmentation

The given dataset is limited, so we want to increase the dataset diversity. We use Mix3D augmentation, which combines two augmented scenes to form a new training sample. From the initial N scenes, it can generate $O(N^2)$ new scenes with novel contexts. Therefore, the model can better learn from local structure instead of relying on the global context too much.

❖ Instance augmentation

Since rare classes are not only infrequently observed but are often small objects, they often overfit to recognizing both the surrounding context and the object. For each scene, the augmentation samples instances from infrequently seen class categories, and places them in potentially physically valid locations. The augmentation breaks overly specific context dependencies for recognition.

❖ Focal loss

Due to the severe data imbalance in the dataset, we use the focal loss to reweight the contributions of different classes to the loss according to their frequency. The focal loss also gives a lower weight to samples the model is confident with. Hence, it makes the model focus on hard samples and those that are seldom seen. However, it seems that it does not work well in this case.

❖ Logit adjustment

Logit adjustment is the method we use to conquer the difficulty in training with long-tailed distribution. This technique revisits the classic idea of logit adjustment based on the label frequencies. Although we try to use logit adjustment in this case, there is no significant improvement.

References

1. Christopher Choy, JunYoung Gwak, and Silvio Savarese. "4d spatio-temporal convnets: Minkowski convolutional neural networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.
2. Nekrasov, Alexey, et al. "Mix3D: Out-of-context data augmentation for 3D scenes." 2021 International Conference on 3D Vision (3DV). IEEE, 2021.
3. David Rozerberszki, Or Litany, and Angela Dai. "Language-Grounded Indoor 3D Semantic Segmentation in the Wild." Proceedings of European Conference on Computer Vision (ECCV), 2022.
4. Menon, Aditya Krishna, et al. "Long-tail learning via logit adjustment." International Conference on Learning Representations. (ICLR) 2021.