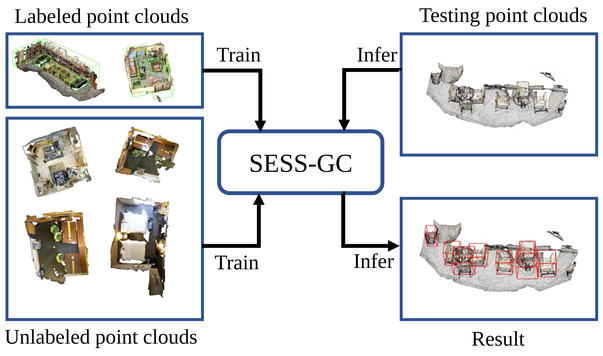
SESS-GC: Self-Ensembling Semi-Supervised 3D Object Detection Model with Graph-Level Consistency

Liu Yifan, Teng Yunlai, Zhu Diwen, Wang Jiyu

National University of Singapore

Introduction

In recent years, many deep learning-based point cloud-based 3D object detection methods have performed well on numerous datasets. Most of these methods are subject to strict supervision. However, strictly supervised learning methods require a large amount of well-annotated data, and usually their collection process is time-consuming and costly. Therefore, semi-supervised learning is a good alternative. We propose a novel semi-supervised 3D object detection model with graph-level consistency, containing a teacher and a student 3D object detection network.



The ScanNet Dataset

ScanNet dataset is a 3D laser point cloud of indoor scenes annotated in 20 categories, including *xyz* and label information, not including color information.

Preprocessing

- A random sub-sampling of each training point cloud is applied to the student and teacher networks to enhance the consistency of the feature space.
- Augmentation: flipping, rotation around the upright-axis and scaling for the student network and a stochastic flipping for the teacher network

$$F = \begin{cases} 1, & if\ \epsilon > 0.5 \\ 0, & otherwise \end{cases} \quad \epsilon \in [0,1]$$ flipping

$$R = \begin{bmatrix} \cos(\omega) & -\sin(\omega) & 0 \\ \sin(\omega) & \cos(\omega) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$ rotation

$$\hat{x} = T \cdot x$$ scaling

Loss Function

1. Supervised Loss

The supervised loss in VoteNet is used in training.

$$\boldsymbol{L_{detection} = L_{vote-reg} + \gamma_1 L_{obj-cls} + \gamma_2 L_{box} + \gamma_3 L_{sem-cls}}$$

2. Consistency Loss

The output of the student network and the output of the teacher network should be the same, so a consistency loss is required for regression.

$$\boldsymbol{L_{consistency} = L_{center} + L_{cls} + L_{size}}$$

3. Intra Loss

The feature vector of the object within the same class should have smaller distance than the object in the different classes, so the intra loss can be define as

$$\boldsymbol{L_{intra} = \frac{\mathcal{L}_{i,j}(1 - e^{-\frac{\|f_i - f_j\|_2}{2\sigma^2}})}{k^2}}$$

4. Inter Loss

The proposal features of teacher and student networks should have small distance, so the inter loss can be defined as

$$\boldsymbol{L_{inter} = \sum_{i=1}^{K} \|f'_s|_i - f'_t|_i\|_2}$$

5. Total Loss

Finally, the total loss is a weighted sum of all the losses described above:

$$\boldsymbol{L_{total} = \omega_1 L_{consistency} + \omega_2 L_{intra} + \omega_3 L_{inter} + \omega_4 L_{detection}}$$

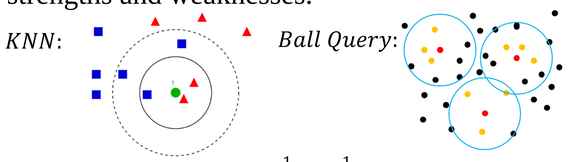
Graph Processing

There are still potential relationships between proposal features that have not been fully explored. Thus GCN and GAT layers are added to our model.

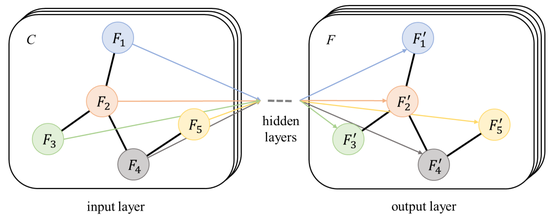
Graph Convolutional Neural Network:

Two-layer GCN is added to further model the relationships between proposal features.

Adjacent matrix is constructed by these two methods, they complement each other with their strengths and weaknesses.



$$X' = \hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}} X \Theta$$



In addition to the original graph convolutional operator, some variants of the original operator are also adopted.

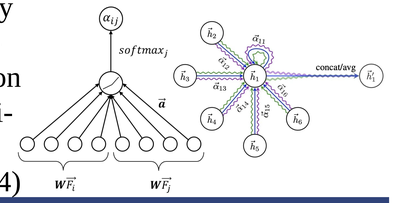
$$X' = \hat{D}^{-1}(I + A)X\Theta$$

$$X' = I + (\hat{D}^{-1}A)X\Theta \qquad X' = I + \hat{D}^{-1}(I + A)X\Theta$$

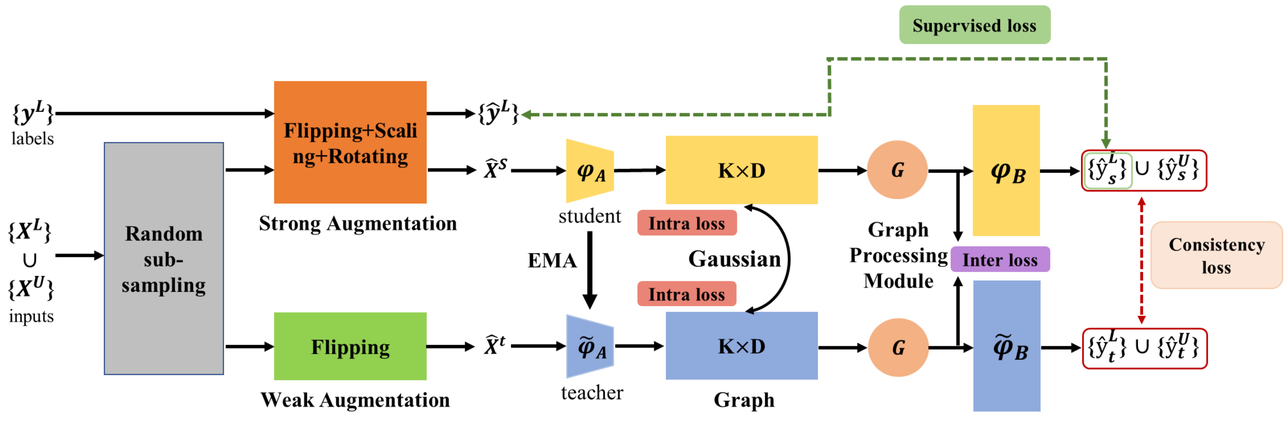
$$X' = I + \left(\hat{D}^{-\frac{1}{2}} A \hat{D}^{-\frac{1}{2}}\right)X\Theta \qquad X' = I + \hat{D}^{-\frac{1}{2}}(I + A)\hat{D}^{-\frac{1}{2}}X\Theta$$

Graph Attention Network:

For the GAT, we try two-layer and one-layer graph attention network with multi-head attention (the number of head is 4)



Model Architecture



Result

We try many combinations of the hyperparameters so as to find the optimal configurations for training. Besides, we also do extensive ablation studies.

In order to test as many configurations as possible, we just conducted experiments extensively on ScanNet, using 10% labeled data. Some results are shown below.

BQ: Ball Query, NoP: Number of points

Intra loss				Inter loss		Consistency loss (weight / ramp-up)	Graph processing module type	Method to get A	Operator type	Best Inductive performance (epoch NO.)	Abs. improve
Method	Sigma	Obj_score threshold	Weight	Option	Weight					mAP@0.25	Compared with SESS
N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	True (10/30)	N.A.	N.A.	N.A.	40.98 (90)	
K2	1	0.5	1	MSE	1	True (10/30)	N.A.	N.A.	N.A.	41.43 (192)	**0.45↑**
N.A.	N.A.	N.A.	N.A.	MSE	100	True (10/30)	N.A.	N.A.	N.A.	41.01 (190)	**0.03↑**
num1	10	0.5	10	N.A.	N.A.	True (10/30)	N.A.	N.A.	N.A.	42.02 (145)	**1.04↑**
num1	10	0.5	10	MSE	100	True (10/30)	N.A.	N.A.	N.A.	42.82 (165)	**1.84↑**
num1	10	0.5	10	MSE	100	False	N.A.	N.A.	N.A.	40.76 (171)	0.22↓
num1	10	0.5	10	MSE	100	True (10/30)	GCN	BQ (r1.2, NoP16)	Third	43.99 (147)	**3.01↑**
num1	10	0.5	10	MSE	100	True (10/30)	GCN	KNN (NoP16)	Third	44.10 (134)	**3.12↑**
num1	10	0.5	10	MSE	100	True (10/30)	GAT (1 layer)	BQ (r0.5, NoP8)	N.A.	43.94 (141)	**2.96↑**
num1	10	0.5	10	MSE	100	True (10/30)	GAT (2 layer)	KNN (NoP8)	N.A.	44.45 (183)	**3.47↑**

Conclusion

Most existing point cloud 3D object detection methods are heavily supervised, largely dependent on large datasets of well-annotated 3D scenes, which remain a major bottleneck due to the high expenses. We aim to address this problem based on the state-of-the-art method SESS, and improves the performance on the ScanNet benchmark. Using only 10% of labeled data on ScanNet, our model outperforms SESS. We have newly designed augmentation scheme, intra loss, graph processing module, and graph-level consistency loss, capable of forcing the network to model local information and generating more accurate detections.