Minkowski SAR-UNet3D for Point Cloud Semantic Segmentation

Siddhi Brahmbhatt Vikas Mehta Suvobrat Ghosh

Abstract

Due to its numerous applications in domains like computer vision, robotics, autonomous driving, and due to the advancements in deep learning techniques, 3D point cloud is attracting significant attention from the computer vision community. In this work, we propose and analyse two encoder-decoder models trained to perform semantic segmentation on the S3DIS dataset. The model has been trained using sparse tensor representation of the quantized point cloud of S3DIS dataset. To implement the model architectures, we used Minkowski Engine - which is an an open-source auto-differentiation library for sparse tensors that provides extensive functions for high-dimensional convolutional neural networks. The first model is a ResUNet-a inspired architecture for sparse 3D data, and second model is the ResUNet-a inspired architecture with a non local self attention module to encode the sparse tensor features. We achieved a mIoU score of 0.55 and 0.54 respectively on the test data using these two models.

1. Introduction

Point cloud data has a variety of applications in the contemporary world with onset of self-driving cars, robotics and AR applications. Point clouds are much more difficult to comprehend since unlike images, they are not continuous and there is no defined order in their representation. Deep Learning applications on Point Cloud are heavily inspired by the contemporary networks on 2D images. For semantic segmentation of images, conventional approaches like U-Net [10] and it's variants [11] have proven to be extremely effective in capturing the information in images and representing them through a decoder back in the input space. Latest approaches include methods like ResUNet-a [5] which have been very effective in segmentation of large scale images with less number of parameters.

One of the challenging tasks in 3D space includes semantic segmentation of point clouds. Such an application directly influences decisions in the application areas mentioned above. For dealing with this problem, the use of encoder-decoder based models has been the golden standard. To approach such a problem, VoxelNet [18] uses a voxel based grouping and continuous convolution in an encoder-decoder system. Meanwhile PointNet++ [9] used multi layer perceptrons in 3D space with distance based convolutions using farthest point sampling in the process of encoding.

In this work, we propose a deep learning approach for the point cloud semantic segmentation problem inspired by the success of sparse tensor convolution based processing in Minkowski convolutional neural networks [3], and the transformer based approaches like Point Transformer [17]. Self attention layers in the transformer networks help in capturing long range dependencies in vast point cloud datasets. Building up on the intuitions behind these two kinds of networks, and the contemporary segmentation networks in 2D space, we propose a deep learning network that uses a self attention layer to encode long range dependencies in the sparse tensor representation of the point cloud features. As a backbone architecture, we used a sparse convolution based ResUNet-a inspired model. From this work we show that this approach is incredibly effective for semantic segmentation, and for encoding information effectively. Such a network has fewer parameters and can also be applied to LIDAR based robotic systems. It can also be augmented with Reinforcement Learning based systems for Automated decision making in the real world.

To summarize, we make the following contributions:

1. Train a ResUNet-a inspired model using sparse tensor representation of the voxelized point cloud of S3DIS dataset.
2. Add a self-attention layer to encode sparse tensor features in the above mentioned ResUNet-a backbone architecture.
3. Compare our results with the existing methods for point cloud semantic segmentation.

2. Related Work

For processing point cloud data using deep learning, there are three main approaches:

Multiview based method: Here, 3D data is represented by multi-view 2D images, which can be processed based on 2D CNNs. Popular approaches using this paradigm include MVCNN [12] which uses multiple views from different points in the 3D space passing them through contemporary 2D convolutional models which share parameters. Further, through a pooling mechanism it encodes the information from the different views to generate predictions on the 3D data. For segmentation, SnapNet [2] proposes a network which takes random views in point cloud space with their 2d projection and passes them through convolutional layers in a similar fashion. It then and backprojects them to the pointcloud for semantic segmentaion. However, these methods are exteremely slow and have a dependence on lower dimension projection and viewpoints. Hence, they are ineffective in real time applications. Also, it is often difficult to choose enough proper viewpoints for multiview projection.

Voxel based approach: It includes the use of 3D convolutions on voxelized data. Voxelization solves both unordered and unstructured problems associated with point clouds. This low resolution representation then can be passed through 3D segmentation layers with contemporary convolutional models. Segmentation tasks have been tackled using voxel based models like Segcloud [13], but have recently taken a backstep due to high computational and memory problem from fixed-sized voxels. Similar computational bottlenecks were encountered in classification tasks using models like VoxNet [7]. However, solving these limitations, MinkowskiNets [3] use sparse convolutions on sparse representaion of voxelised point cloud data and have achieved a good performance on a series of semantic segmentation and classification benchmark datasets.

Processing point cloud: Directly processing the point cloud is another approach to the problem. Breakthrough in this space was provided by the PointNet [8] architecture which uses shared MLPs for feature encoding of the 3D points and then segmentation is done through the network without a reduction in dimensional space. While PointNet is still considered as a baseline, several other architectures like PointNet++ [9], Dynamic Graph CNN (DGCNN) [15], 3P-RNN [16] have shown considerable improvements in performance for both classification and segmentation tasks. Contemporary approach in the field includes the Point Transformer [17] which uses self-attention as inspired from language models for encoding spatial context with points along with other multi layers for effective segmentation. This approach solves the limitations of the above mentioned approaches due to its simplicity in implementation and strong representation ability. Hence, attention mechanisms in networks processing point cloud is an active area of research.

3. Dataset

Stanford Large-scale 3D Indoor Spaces Dataset (S3DIS) [1] is a large-scale indoor RGB-D dataset consisting of over 215 million points, covering an area of over 6,000 m^2 in six indoor regions originating from three buildings (example raw point cloud visualization is shown in Figure 1). Each point in the scene point cloud is annotated with one of the 13 semantic categories (structural elements: ceiling, floor, wall, beam, column, window, door, and movable elements: table, chair, sofa, bookcase, board and clutter for all other elements) and 11 scene categories (office, conference room, hallway, auditorium, open space, lobby, lounge, pantry, copy room, storage and WC). The dataset is collected from 6 large-scale indoor areas that originate from 3 different buildings.

Figure 1. Visualization of raw point cloud of one of the copy rooms of area 5

Data Preparation: Out of the six areas of the S3DIS dataset, as per the convention we used area 5 as test set and the remaining areas as train set for our experiments. S3DIS dataset does not have any explicit validation dataset, so we used 30% of the test set as validation set. Before feeding the data to network, we firstly voxelize the point cloud (quantization size = 0.1 in our case) and convert it into a sparse tensor using Minkowski Engine. Sparse tensor is a high-dimensional extrapolation of a sparse matrix where non-zero elements are represented as a set of indices and associated values. The Minkowski Engine is an an open-source auto-differentiation library specially developed for working with sparse tensors. It supports all standard neural network layers such as convolution, downsampling, upsampling, as well as broadcasting operations for sparse tensors.

4. Model Implementation

We performed training experiments on two model architectures:

1. MinkowskiResUNet3D: ResUNet-a inspired architecture for the given sparse 3D data.
2. Minkowski Sparse Attentive ResUNet3D (MinkowskiSAR-UNet3D): MinkowskiResUNet3D architecture with a self-attention layer to encode sparse tensor features

We used Minkowski Engine (which is internally wrapped with PyTorch) to work with sparse tensor data as described in the data preparation section. We setup our environment on Google Colab for GPU compute (setup details in the submitted code). The setup includes installation of PyTorch and Minkowski Engine which takes approximately 30 minutes everytime one connects to the Colab runtime.

In this section, we explain the implementation details and results of the two aforementioned models. We also explain in the detail the transformer based self attention layer (often used in natural language processing tasks for encoding long distance information) implemented in our model.

4.1. MinkowskiResUNet3D

As mentioned in the introduction, encoder-decoder based models are known to work well for the task of semantic segmentation, specifically U-Net [10] [11] derived architectures. Point-Unet [6] has shown appreciable performance for semantic segmentation by directly processing point cloud data. MinkowskiUNet based networks have specifically shown to work well for semantic segmentation (taking sparse tensor representation of voxelized point cloud as input) as compared to the many standard point cloud processing networks like PointNet++ [9] variants and ScanNet [4], to name a few.

Building up on these ideas, and also knowing about the success of ResUNet-a [5] architecture on large scale image data, we propose a ResUNet-a inspired architecture with sparse convolutions for our voxelized and sparse point cloud data. We hypotheize that adding residual connections to U-Net like encoder-decoder architecture will help in better flow of information between different layers, which helps in better overall flow of gradients while training. Our model (exact details in the submitted code) has 3 sparse convolutional layers and corresponding 3 upsampling layers (convolution transpose layer in our case). The model has several skip connections, and also has residual connections as mentioned above. Intuitively, deeper layers of network capture high level features and it is a common paradigm in most of the computer vision tasks to gradually increase the number of filters as we go deeper into the network. We have followed the same paradigm and have increased the number of filters with depth. Additionally we have used batch normalization as it is known to assist in effective training of deep neural networks. Also, we used ReLU as the nonlinear activation.

4.2. MinkowskiSAR-UNet3D

Self attention [14] is helpful in the case of point clouds as they contain irregular and often hard to encode positional information and long dependence. Minkowski Engine quantizes point clouds with a distance based metric hence short term positional encoding information would be redundant with such a self-attention. Hence we focus on irregular dependence between points in a point cloud. For this we apply self-attention on quantized features of the point cloud.

Our implementation of self attention (as shown in Figure 2) involves encoding feature representations into keys, values and queries (exact details submitted in the code folder). With a dot product attention, the formulation involves keys and queries of the same vector dimension d_k. Hence the dot product of each key and query vector results in a single raw score. Each query vector is attended with all the key values including itself resulting in [*number_of_queries, number_of_keys*] shaped matrix representing the influence of each of the keys for each query vector. The query key dot product for each query is then passed through a softmax layer resulting in the normalized scores of each of the points that influence a particular point's query. These

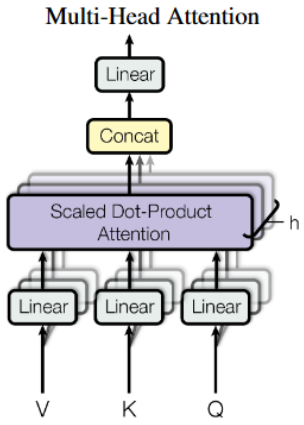

Figure 2. Self-attention layer with Query(Q), Value(V) and Key(K) and number of heads(h) [14]

scores are then multiplied with the value vectors for the respective keys and summed up to give a a single vector representation (for the query representing the attended values across the point cloud that influence it). For non-linearity the output is then encoded with a fully connected layer and a non-linearity activation function ReLu. The major advantage with such a proposition in contrast with a direct self attention layer is that the quantization drastically reduces the point cloud dimension and makes the process of encoding this information much faster and effective. The detailed model architecture by incorporating this self-attention layer in the backbone MinkowskiResUNet3D model is as shown in Figure 3.

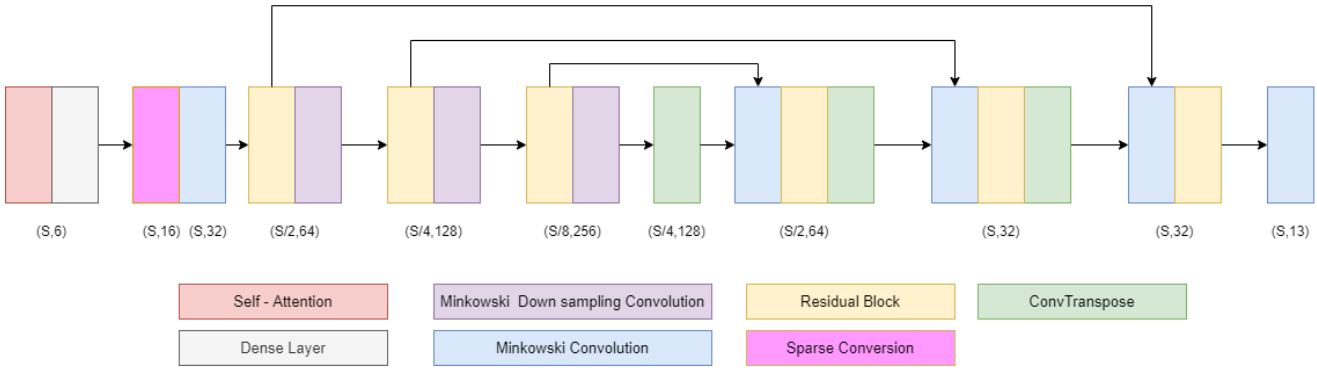

Figure 3. MinkowskiSAR-UNet3D architecture

5. Experimental evaluation

5.1. Evaluation criteria

For evaluation, we used the mean classwise intersection over union (mIoU) metric, which is a standard metric for semantic segmentation.

5.2. Results and comparisions

Using the aformentioned two models, we performed the training on sparse tensor representation of voxelized point cloud data of S3DIS dataset (train-test-val split explained in the data preparation section above). We used Adam optimizer with learning rate of 1e-5 and weight decay of 0.01 and trained the model to minimize cross entropy loss. Training the MinkowskiSAR-UNet3D model for 30 epochs resulted in the training plots as shown in Figure 3 and Figure 4.

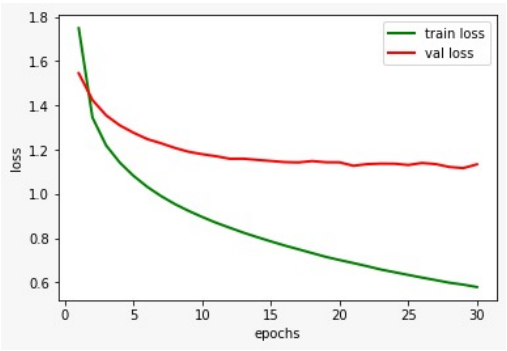

Figure 4. Training plot of MinkowskiSAR-UNet3D

Table 1. shows the mIoU score on test set (area 5) of our models compared to few other models we explored during the course of this project.

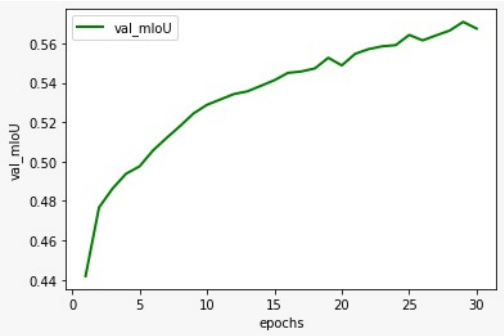

Figure 5. validation mIoU vs epochs for MinkowskiSAR-UNet3D

Method	mIoU on test set
PointNet [8]	0.41
SegCloud [13]	0.49
MinkowskiNet20 [3]	0.63
MinkowskiNet32 [3]	0.65
MinkowskiResUNet3D	**0.55**
MinkowskiSAR-UNet3D	**0.58**

Table 1. Results

6. Conclusion

In this paper, we proposed two novel architectures implemented using Minkowski Engine for semantic segmentation of point cloud data. While their performance is better than methods like PointNet and SegCloud, their performance is not yet par with MinkowskiNet20 and MinkowskiNet32. We feel that the primary intuitions behind our attempts (namely use of residual connections in U-Net derived models, and further introduction of self-attention layer) will yield good results by performing more experiments. In the future, we are considering exploitation of model capacity by increasing the depth of layers, introducing more parameters and experimenting with the model hyperparameters.

We would also like to generate our results using model ensembles with more computational resources.

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 2

[2] Alexandre Boulch, Joris Guerry, Bertrand Le Saux, and Nicolas Audebert. Snapnet: 3d point cloud semantic labeling with 2d deep segmentation networks. *Computers & Graphics*, 71:189–198, 2018. 2

[3] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 1, 2, 4

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 3

[5] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 1, 3

[6] Ngoc-Vuong Ho, Tan Nguyen, Gia-Han Diep, Ngan Le, and Binh-Son Hua. Point-unet: A context-aware point-based neural network for volumetric segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 644–655. Springer, 2021. 3

[7] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928. IEEE, 2015. 2

[8] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2, 4

[9] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1, 2, 3

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 3

[11] Nahian Siddique, Sidike Paheding, Colin P Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 2021. 1, 3

[12] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 2

[13] Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*, pages 537–547. IEEE, 2017. 2, 4

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[15] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 2

[16] Xiaoqing Ye, Jiamao Li, Hexiao Huang, Liang Du, and Xiaolin Zhang. 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 403–417, 2018. 2

[17] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021. 1, 2

[18] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 1