Saathi

7/9 June 23

Topic: Swin 3d: A pretrained Transformer backbone for 3D Indoor scene Understanding

Abstract:
Pretrained backbones with fine-tuning have been widely adopted in 2D vision and natural language processing tasks and demonstrated significant to task specific networks. In this paper, we represent a pretrained 3D backbone, named Swin3d, which first outperforms all state of the art methods in downstream 3D indoor scene understanding tasks. Our backbone network is based on the 3D swin transformer and carefully designed to efficiently conduct self attention on sparse voxels with linear memory complexity and capture the irregularity of point signals via generalized contexual relative positional embedding. Based on this backbone design, we pretrained a large Swin3d model on a synthetic structured 3D dataset that is 10 ~~ta~~ times larger than the ScanNet dataset and fine tuned the pretrained model in various downstreams real world indoor scene understanding tasks. The results demonstrate that our model

pretrained on the synthetic dataset not only exhibits good generality in both downstream segmentation and detection on real 3D point datasets, but also surpasses the state of the art methods on downstream tasks after fine tuning with +2.3 mIoU and +2.2 mIoU on S3DIS Area 5 and 6 fold semantic segmentation, +2.1 mIoU on ScanNet Segmentation (val), +1.9 mAP @0.5 on ScanNet detection +8.1 mAP @0.5 on S3DIS detection. Our method demonstrates the great potential of pretrained 3D backbones with fine tuning for 3D understanding tasks.

Conclusion

We present a pretrained 3D model backbone- swin 3D for indoor scene understanding, whose scalability, transferability and superior performance have been validated through extensive experiments. We believe that the capacity of Swin 3D can be extended extended further in the following directions. First it would be interesting to revisit self- supervised pretraining schemes using our backbone and maximize its capability with more real and synthetic data, including outdoor 3D data. Second, as point

clouds are usually accompanied by high resolution, multiview images supplied by 3D capture devices, it is promising to leverage image data and incorporate views both pretrained image backbones and 3D backbones to enhance the efficacy of 3D learning.