Project Proposal

MVCNN for 3D Shape Recognition and Reconstruction

1 Abstract

Based on the paper "Multi-view Convolutional Neural Networks for 3D Shape Recognition" (MVCNN) by Su et al.[2] we want to implement a model which is able to cope with two tasks at the same time. On the one hand the model should be able to classify the object and on the other hand it should also be able to simultaneously reconstruct its 3D shape solely based on the input images.

The original MVCNN architecture classifies a 3D shape model based on a set of rendered 2D images which are processed by 2D convolutional networks to extract rich feature maps. A big advantage of this approach is that the model can exploit the expressiveness of powerfull pre-trained models like VGG-M. Furthermore, due to the view pooling operator, the model can take an arbitrary number of images from different camera angles as input.

We propose an extension that adds another reconstruction head that can be used to reconstruct a 3D model alongside the recognition task. We will take inspiration from different models that use multi-view images for recovering 3D representations and will create a model that takes parts from MVCNN and other 3D reconstruction models to create a hybrid that can perform both tasks. We plan on experimenting with shared layers between the two models we combine and will use the outcome of the classification task as additional features which might improve the performance of the 3D reconstruction task. We also intend to investigate if the performance of the classification task increases due to the shared layers in the training process. Additionally, we aim to test the performance of our model on several different input geometric representations (point clouds, voxels, etc.). The reconstructed 3D model can be represented in many different ways. Similar to existing methods we might use a 3D volume representation with a resolution of 32^3.

One possible inspiration for the task of 3D reconstruction from multi-view images is the Pix2Vox model from Xie et al.[3]. It utilizes an enoder-decoder architecture to generate coarse 3D volumes for each input image. Afterwards, those coarse volumes are fused with the help of a multi-scale context-aware fusion module to produce a fused volume which is then refined to generate the final output. To cope with the issue of limited GPU accessibility, we might make use of the Pix2Vox++/F architecture which has fewer parameters and hence a lower computational complexity. As an additional improvement, we will look into extending the voxel branch using the mesh refinement method proposed in [1] in order to attain mesh reconstruction results. In particular, we will attempt to utilize the cubify and mesh refinement options to extend Pix2Vox.

There exist PyTorch implementations of both models, MVCNN and Pix2Vox. We plan to implement our model in PyTorch as well and the library PyTorch3D might be used to render the images or to convert between representations.

2 Requirements

- ShapeNet dataset
- Library to render images with different 3D representations, likely PyTorch3D
- Extending (pre-trained) VGG16 with MVCNN and Pix2Vox heads
- Test our model on different geometric input representations
- Evaluate classification performance before and after adding the reconstruction head

3 Team

- Andrew Desousa (03733072)
- Florian Donhauser (03695568)
- Bastian Wittmann (03732309)

References

[1] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[2] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proc. ICCV*, 2015.

[3] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, Shengping Zhang, and Xiaojun Tong. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. *CoRR*, abs/1901.11153, 2019.