COSE474-2024F: Final Project Proposal
From Geometry to Semantics: A Deep Learning Approach to 3D Object Descriptions

Sehyeon Park

1. Introduction

In recent years, numerous generative models have been developed to create 3D models, thanks to the advancements in generative AI. These techniques are being used across various domains such as gaming, film, and virtual reality. However, there has been less focus on models that can convert arbitrary 3D objects into natural language descriptions based on their geometric information. This project proposes a custom neural network model that analyzes 3D mesh data and generates natural language descriptions of the 3D objects.

2. Problem definition & chanllenges

The problem is to generate natural language sentences that explain a 3D object based on its .obj mesh file data. In other words, we need to build a model that can analyze the geometric traits of the object and then convert this information into text.

This problem is far more complex than 2D image classification because 3D mesh data has a much more complicated structure than 2D images. Specifically, understanding the relationships between vertices, edges, and faces in 3D mesh data and translating that into text will be a novel and challenging task. In addition, since each 3D mesh has a different number of vertices, edges, and faces, we need to handle these irregularities carefully. This requires designing the model in detail to process varying input sizes while still extracting meaningful features from each mesh.

3. Related Works

MeshCNN(Hanocka et al., 2019) is one of the CNN-based models designed to process 3D mesh data at the edge level, effectively learning the geometric structure of the mesh.

PointNet++(Qi et al., 2017) is another popular method for processing 3D data. Unlike MeshCNN, PointNet++ processes 3D point clouds.

LLaMA(Touvron et al., 2023) demonstrates outstanding performance in text generation and natural language processing tasks.

CLIP(Radford et al., 2021) excels in learning the relationship between images and text, achieving strong performance in image-text mapping tasks.

4. Datasets

The ShapeNet(Chang et al., 2015) dataset will be used. ShapeNet is a large-scale dataset containing over 50,000 3D models from various categories such as furniture or vehicles. It provides 3D mesh representations of objects, making it ideal for tasks like classification in the project.

5. State-of-the-art methods and baselines

PointNet++ processes 3D data using point clouds and performs well in several tasks, but it does not directly handle geometric structures, leading to missing potential information when converting meshes to point clouds. CLIP handles 2D images by learning image-text mappings, but it is not suitable for directly processing 3D data, as it requires converting 3D meshes into 2D images, using projection methods, which can lose important structural details.

In contrast, MeshCNN directly processes 3D mesh data by learning geometric features at the edge level, making it more effective in preserving mesh structure and being a better baseline method for this project.

6. Schedule & Roles

Week 1 : Prepare the ShapeNet dataset and set up the environment for MeshCNN and LLaMA models.
Week 2 : Preprocess 3D mesh data for MeshCNN.
Week 3 : Train MeshCNN and optimize the model performance.
Week 4 : Design and experiment with prompts in LLaMA.
Week 5 : Write the final project paper.

References

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., and Yu, F. ShapeNet: An

Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

Hanocka, R., Hertz, A., Fish, N., Giryes, R., Fleishman, S., and Cohen-Or, D. Meshcnn: a network with an edge. *ACM Transactions on Graphics (ToG)*, 38(4):1–12, 2019.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.