Christian-Albrechts-Universität zu Kiel
AI and Visual Computing Lab

Bachelorthesis

Predicting Gaussians: 3D Scene Reconstruction using Transformers

Anton Wagner

First Assessor:
Prof. Dr. Sören Pirk
Second Assessor:
Ma.Sc. Helge Wrede

Christian-Albrechts-Universität zu Kiel
Institute of Information Technology
Information Technology
Neufeldtstraße. 6, 24118 Kiel

März 2024

Abstract

This work attempts to use a novel combination of existing point cloud architectures and new embedding networks to encode a 3D gaussian scene with missing parts into a tokenized representation that can be understood and continued with a transformer architecture. It also aims to decode the generated tokens back into new gaussians to fill in the missing parts of the scene.

Contents

1. Related Works

This Chapter will discuss the origins and inspirations for different components of the final network, as well as other works that use 3D Gaussian Splatting. Relevant sources will be: [KKLD23] [ZYG+23] [YTR+22] [YRW+21]

I will also mention the similarity to vision transformers such as: [DCLT19] [YZW+21] [MWL+24]

2. Introduction

This chapter will explain 3D Gaussians and their ability for novel HD view synthesis at high framerates [KKLD23] It will also give a preliminary explanation as to why i chose a transformer architecture. [VSP+23]

2.1. Motivation

[An Image showing a hole in an otherwise good Scene]
This section will highlight the large interest being shown towards gaussians and will explain how gaussian scenes end up with holes that need to be fixed.

2.2. Goals

This section describes my goal of trying to have an existing gaussian scene be understood by a transformer and then continued. Elaborating on the idea of getting a "good" latent space representation of gaussians.

2.3. Challenges

This section will describe the non structural nature of gaussians and the difficulties of regressing them directly [ZYG+23]. It will also show a diagram of the attention layer and it's quadratic memory/computational complexity, describing the need to reduce the sequence length.

3. Network Architecture

This chapter will start of with a diagram and description of the rough structure of the entire network.

3.1. Visual Embedding

This section will describe the Visual Embedding Net with a more closed up diagram. Both the novel encoder and the decoder (same as used in [ZYG+23]) will be elaborated on. There will also be some pictures showing the capabilities of the Visual to encode and decode gaussians while retaining good image quality.

3.2. DVAE

This chapter will introduce the DVAE [Rol17] as a way comprehend a local pointcloud and convert it to a token and vice-versa.

3.2.1. Sampling/Grouping

This section will elaborate on the underwhelming performance of the commonly used Furthest-Point + KNN Sampling and will compare and contrast it with Random-Point + KNN Sampling.

3.2.2. DGCNN

This section will explain the DGCNN [WSL+19] and how it enables the DVAE to understand local geometries

3.2.3. Discretization

This section will introduce the vocabulary/codebook and how Gumbel-Softmax [JGP17] is used to discretize the resulting logits before sampling from the vocabulary.

3.3. Transformer

This section will talk about the transformer as a whole.

3.3.1. Encoder

This subsection will describe the self-attention mechanism and the positional encoding used in the encoder-blocks of the transformer.

3.3.2. Query Generator

This subsection will describe the Query generator as a way to turn the memory tokens generated by the encoder into positional information of where the missing content is located.

3.3.3. Decoder

This subsection will describe the cross-attention mechanism used to combine the memory and missing content tokens into useful output tokens.

4. Training

This chapter will explain the training regime used for the transformer. Describing both methods: 1. Training everything together vs 2. Training every component seperately

5. Results

This chapter will look at some results, evaluating them both qualitatively and quantitavely.

5.1. Overall Evaluation

This subsection will focus on the final results obtained.

5.2. Vocabulary Analysis

This subsection will analyze the different tokens learned by the DVAE.

6. Limitations

This chapter will highlight the limitations currently present in both the architecture and also results obtained

7. Further Work

7.1. Improvements

This chapter will somewhat speculate on ways that the architecture could be improved and goals i have in mind for future work on this particular problem

7.2. Prospects

This chapter will talk about the things that would become possible if I can achieve a good latent space representation of Gaussian Scenes (such as style-transfer, etc).

A. Source Code Excerpts

Bibliography

[DCLT19] DEVLIN, Jacob ; CHANG, Ming-Wei ; LEE, Kenton ; TOUTANOVA, Kristina: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019

[JGP17] JANG, Eric ; GU, Shixiang ; POOLE, Ben: *Categorical Reparameterization with Gumbel-Softmax*. 2017

[KKLD23] KERBL, Bernhard ; KOPANAS, Georgios ; LEIMKÜHLER, Thomas ; DRETTAKIS, George: 3D Gaussian Splatting for Real-Time Radiance Field Rendering. In: *ACM Transactions on Graphics* 42 (2023), July, Nr. 4. `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`

[MWL+24] MIAO, Wei ; WANG, Lijun ; LU, Huchuan ; HUANG, Kaining ; SHI, Xinchu ; LIU, Bocong: ITrans: generative image inpainting with transformers. In: *Multimedia Systems* 30 (2024), Januar, Nr. 1. `http://dx.doi.org/10.1007/s00530-023-01211-w`. – DOI 10.1007/s00530–023–01211–w. – ISSN 1432–1882

[Rol17] ROLFE, Jason T.: *Discrete Variational Autoencoders*. 2017

[VSP+23] VASWANI, Ashish ; SHAZEER, Noam ; PARMAR, Niki ; USZKOREIT, Jakob ; JONES, Llion ; GOMEZ, Aidan N. ; KAISER, Lukasz ; POLOSUKHIN, Illia: *Attention Is All You Need*. 2023

[WSL+19] WANG, Yue ; SUN, Yongbin ; LIU, Ziwei ; SARMA, Sanjay E. ; BRONSTEIN, Michael M. ; SOLOMON, Justin M.: *Dynamic Graph CNN for Learning on Point Clouds*. 2019

[YRW+21] YU, Xumin ; RAO, Yongming ; WANG, Ziyi ; LIU, Zuyan ; LU, Jiwen ; ZHOU, Jie: PoinTr: Diverse Point Cloud Completion with Geometry-Aware Transformers. In: *ICCV*, 2021

[YTR+22] YU, Xumin ; TANG, Lulu ; RAO, Yongming ; HUANG, Tiejun ; ZHOU, Jie ; LU, Jiwen: Point-BERT: Pre-Training 3D Point Cloud Transformers with Masked Point Modeling. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022

[YZW+21] YU, Yingchen ; ZHAN, Fangneng ; WU, Rongliang ; PAN, Jianxiong ; CUI, Kaiwen ; LU, Shijian ; MA, Feiying ; XIE, Xuansong ; MIAO, Chunyan: *Diverse Image Inpainting with Bidirectional and Autoregressive Transformers*. 2021

[ZYG+23] ZOU, Zi-Xin ; YU, Zhipeng ; GUO, Yuan-Chen ; LI, Yangguang ; LIANG, Ding ; CAO, Yan-Pei ; ZHANG, Song-Hai: *Triplane Meets Gaussian Splatting: Fast and Generalizable Single-View 3D Reconstruction with Transformers*. 2023

List of Figures

List of Tables

Quellcodeverzeichnis

Nomenclature

z.B. . . . . zum Beispiel

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Die eingereichte schriftliche Fassung der Arbeit entspricht der auf dem elektronischen Speichermedium.

Weiterhin versichere ich, dass diese Arbeit noch nicht als Abschlussarbeit an anderer Stelle vorgelegen hat.

Kiel, den March 21, 2024

Anton Wagner