3D Visual Grounding with Transformers: Project Proposal

Philipp Foth Bastian Wittmann
Technical University of Munich
{philipp.foth, bastian.wittmann}@tum.de

1. Introduction

The aim of this project is to improve the performance of the 3D visual grounding task by utilizing the concept of the transformer block [6]. Transformer-based architectures have achieved state-of-the-art results in related areas like 2D object detection [1] and natural language processing. Therefore, the incorporation of transformer blocks might lead to promising results and outperform existing 3D visual grounding architectures.

2. Related Work

Point-Based 3D Object Detection. Qi *et al*. proposed VoteNet [4], which utilizes a PointNet++ [5] backbone to extract rich feature representations, which are clustered via a voting module to form object proposals. Recently, Liu *et al*. [3] introduced a novel transformer-based group-free 3D object detection method that outperforms existing 3D object detection architectures like VoteNet and H3DNet [7] on the ScanNet V2 dataset. Their approach is also based on a PointNet++ backbone and utilizes stacked transformer blocks in which the proposed object representations attend to the point features via a cross attention module. The outputs of each transformer block layer are fed into a decoder head that predicts the relevant outputs for 3D object detection. The outputs at each layer are used together as an ensemble in order to determine the final prediction.

3D Visual Grounding. The ScanRefer [2] architecture utilizes VoteNet for the task of 3D object detection. The target object description input is processed separately in multiple stages. The resulting language features are fused with the 3D object features and forwarded to the localization module that outputs the confidences for the proposed bounding boxes.

3. Proposed Project

We have divided our project in two main tasks. The first is to incorporate the state-of-the-art 3D object detection method from [3] into the ScanRefer [2] pipeline. The second task is to modify not only the object detection module, but include the description processing in the transformer architecture as well. This would allow the proposed object representations to attend to the target object description in addition to the point cloud features.

As a first step, we will make sure the available code for ScanRefer and the transformer-based object detection works in our environment and try to replicate their reported results. In the next step, we will combine the ScanRefer architecture with the transformer-based object detection method in a manner that strives for simplicity by, for example, using only the 3D coordinates as inputs for the object detection to avoid retraining (see Fig. 1). Additionally, we will use only the output of the last transformer block for the fusion with the target object description features. After this step has been completed, we will make use of additional point features such as normals and colors and retrain the model, which can lead to improvements as shown in [2].

A major part of the second task will be to propose different architectures and ideas for the incorporation of the description tokens into the transformer block. Subsequently, the most promising approaches will be selected and specified more in detail on paper. The proposed models will finally be implemented, trained, and evaluated.

4. Timeline and Milestones

- **Presentation 1**: We would like to have a working infrastructure and implementation of our first task, a ScanRefer pipeline with its object detection module replaced by the transformer-based object detection approach. We do not expect to show any impressive results at the first presentation.
- **Presentation 2**: We expect to have a trained model for task 1, including results from a ScanRefer Benchmark submission. Additionally, we will present our selected architecture for the second task.
- **Final report and poster**: By the very end, we would like to have a trainable prototype implementation of an architecture solely based on transformers. Ideally, we would be able to present some initial results.

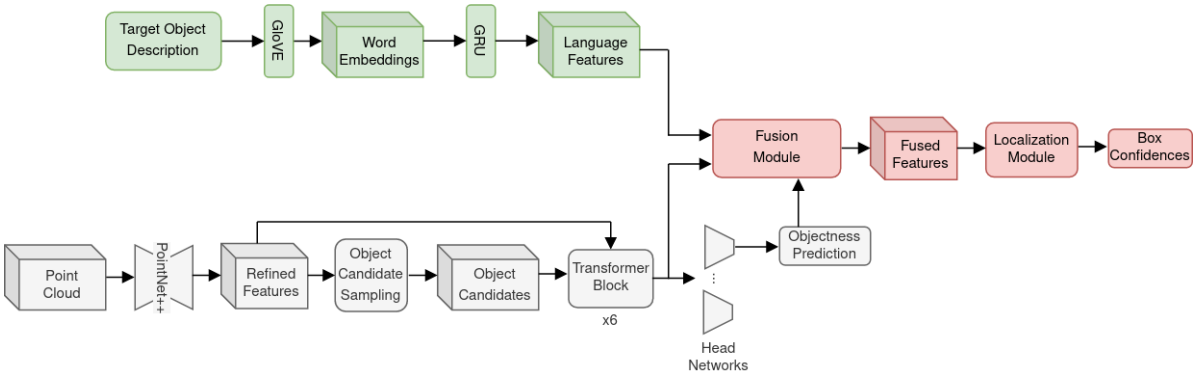

Figure 1. Implementation of transformer-based 3D object detection in ScanRefer pipeline.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing. 1

[2] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *16th European Conference on Computer Vision (ECCV)*, 2020. 1

[3] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers, 2021. 1

[4] Charles R. Qi, Or Litany, Kaiming He, and Leonidas Guibas. Deep hough voting for 3d object detection in point clouds. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9276–9285, 2019. 1

[5] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 5105–5114, Red Hook, NY, USA, 2017. Curran Associates Inc. 1

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 1

[7] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 311–329, Cham, 2020. Springer International Publishing. 1