Project Report: Scan2Cap

Felix Wimbauer
Technical University of Munich
felix.wimbauer@tum.de

Nicolas Seppich
Technical University of Munich
nicolas.seppich@tum.De

Supervisor: Dave Zhenyu Chen
Technical University of Munich
zhenyu.chen@tum.de

Abstract

In this work, we aim to investigate the task of generating a description for a target object in context to its environment in the 3D domain. To this end, we propose a pipeline which combines concepts from 3D object detection and visal attention-based captioning. The proposed pipeline first uses VoteNet to extract feature vectors of the scene. It then combines this information with the features of the object of interest, which are extracted by PointNet++, and feeds this data into an LSTM captioning mechanism, that generates a caption of the object in context of the scene.

1. Introduction

Extracting a detailed and semantic correct understanding of the layout of a 3D scene is crucial for many tasks, e.g. in robotics for navigation and interaction with objects. This includes relating the 3D positions of the objects and their spatial extent, so that a semantically correct description of the objects' environment is generated. However, to the best of our knowledge, there has been no work so far to generate a description of an object representation in point clouds or 3D data.

Therefore, we are interested in implementing a pipeline to obtain a description for a given object in a 3D scan, using state-of-the-art point cloud feature extractors, object detectors, and a captioning mechanism to generate a semantic description for a given object in the 3D scene. This allows the object to be placed in a global semantic context within its environment.

2. Related Work

Our work will be based on the ScanRefer dataset [2]. This dataset consists of 1513 RGB-D scans of ScanNet [3] and contains approximately 5 unique object descriptions for each object in each scene. The work of [2] will also be used as guideline in this project.

The extraction of features on point clouds is presented by [7], who apply the feature extraction directly on the point cloud on a hierarchical level, allowing the extraction of local features in a global context. The task of object detection on point clouds is studied by [6].

Methods for image captioning using visual attention are described by [9], [4] and [1]. These methods have in common, that they generate a caption for the entire image. Since our goal includes using a bounding box for the object to be set in context to the scene, the work of [8] is also of interest for this project.

3. Architecture

Given a point cloud $p \in R^{N\times(d+C)}$ and an object in that scene, which is described by a target bounding box $b_{target} \in R^6$, our goal is to generate a meaningful caption for the object embedded in the context of the scene. To this end, we used three different pipelines, which are described in the following.

3.1. Baseline

To extract information from the point cloud, we use a PointNet++ [7] model. To give the network information about which object we are interested in, we add a new feature channel to each point that masks all points that lay within the bounding box of the object. To ensure that we receive meaningful features, we use weights pretrained for classification of the masked object. The idea behind this is that classification will not only use information from the masked objects, but also global features, for example from close-by objects.

To generate the caption, we use a classical LSTM with an appended fully-connected layer. The fully-connected layer acts as a word classifier and maps the hidden state of LSTM to our vocabulary. As input, the LSTM receives the feature vector extracted from the pointcloud and the word embedding vector of the previously generated word. The word embedding is taken from a pre-computed GloVe [5] word embedding matrix. This structure is similar to [9].

To infer information about the scene in general, we employ a VoteNet [6] network. We don't use the final object labels and bounding boxes, but the feature vectors $f_{object} \in R^{M\times128}$ generated by the *ProposalModule*.

In our initial pipeline, we combine those features through average pooling into $f_{scene} \in R^{128}$. At a later stage, it is possible to replace this step with attention-based pooling, similar to [1].

For the generation of the caption, we use a classical LSTM. As input, we concatenate the f_{target}, f_{scence} and word embedding vector of the previously generated word together. The word embedding is taken from a precomputed GloVe [5] word embedding matrix. The output of the LSTM is passed through a fully-connected layer and softmax function to obtain probabilities for the various possible next words (similar to [9]).

?? summarizes the project pipeline.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[2] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *arXiv preprint arXiv:1912.08830*, 2019.

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.

[4] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017.

[5] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[6] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9277–9286, 2019.

[7] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017.

[8] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.

[9] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

Date	**Milestones**
First Presentation	• Get familiar with dataset • Test different sub components of pipeline (e.g. PointNet++, VoteNet and captioning mechanism) • Start implementing the pipeline • Optional: first training
Second Presentation	• Finish implementation of pipeline • Training and hyperparameter tuning • Set up concept for attention mechanism • Optional: start implementing attention
Final Presentation	• Final results • Optional: attention mechanism

Table 1. Project milestones