Project Proposal for AD4CV: Scan2Cap

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

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

Abstract

The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centred relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length. In this work we analyse a pipeline to generate description of 3D scenes using attention methods.

1. Introduction

The "Deep Learning" revolution has enabled great progress in the field of captioning or image description. However, this task has been studied in a sophisticated way for 2D images using different attention mechanisms. Transferring this task to generate a description of an object representation in point clouds or 3D data, there has been no work so far to the best of our knowledge.

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 Data Set [2]. The dataset consists of 1513 RGBd Scans of ScanNEt[3] 5 unique Objects descriptions for all existing objects in a scene. The work of [2] will also be used as guideline in this project.

The extraction of features on point clouds is presented by [6] applies 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 [5].

Methods for image captioning using visual attention are described by [8], [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 [7]is also of interest for this project.

3. Methods and Concept

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.

[2] Dave Zhenyu Chen, Angel X. Chang, and Matthias Niener. Scanrefer: 3d object localization in rgb-d scans using natural language.

[3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niener. Scannet: Richly-annotated 3d reconstructions of indoor scenes.

[4] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning.

[5] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds.

[6] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space.

[7] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction.

[8] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention.