GraspDETR: DETR based Grasp Planning

Description

The original proposal of the transformer model [5] was for natural language processing, leveraging the attention mechanism to capture contextual information by assigning higher weights to the most relevant positions. Today, the transformer network has found wide applicability in visual tasks, including vision transformer (VIT) [3] and Swin transformer [4]. DETR [2] utilizes the Transformer structure with a bipartite graph-based object detection network, and is widely used to locate objects in an image. In the field of robotics, grasp planning [1] is a critical task that involves identifying a set of grasps based on given images. This study seeks to explore how the DETR network can be utilized to generate grasping instead of merely object location in images, with the objective of integrating powerful vision technologies and knowledge transfer into the domain of robotics.

Tasks

- Literature review of Transformer based vision network.
- Design the GraspDETR model structure for generating grasp candidates with a new cost function.
- Evaluate the proposed model using the prepared dataset in the simulation and real work robotics
- Compare the proposed network with other state-of-the-art approaches.
- Optional: submit the result to a top conference

References

[1] Hu Cao, Guang Chen, Zhijun Li, Qian Feng, Jianjie Lin, and Alois Knoll. Efficient grasp detection network with gaussian-based grasp representation for robotic manipulation. *IEEE/ASME Transactions on Mechatronics*, 2022.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Technische Universität München

TUM School of Computation,
Information and Technology

Lehrstuhl für Robotik, Künstliche
Intelligenz und Echtzeitsysteme

Supervisor:
Prof. Dr.-Ing. Alois Knoll

Advisor:
Jianjie Lin, M.Sc.

Research project:
GraspDETR

Type:
BA/MA

Research area:
DETR, Grasp Planning

Programming language:
Python

Required skills:
Very good mathematical background, programming in Python, Deep learning

Language:
English

For more information please contact us:

E-Mail: jianjie.lin@tum.de

Internet: www6.in.tum.de