Robotics and Embedded Systems

BOSCH

Master Thesis

Leveraging Knowledge Graphs for Enhanced Multi-Modality Foundation Models

Background

Foundation Models (FMs) are large-scale machine learning model trained on a vast amount of diverse data and designed to serve as a general-purpose tool that can be adapted for various tasks. FMs usually integrate information from multiple sources or modalities, such as text and images, to understand and generate comprehensive and rich outputs. FMs has been increasingly applied in various tasks, such as visual entity recognition, visual question answering (VQA), autonomous driving, and robotics.

Recent advancements in multi-modal FMs include CLIP [6] (Contrastive Language-Image Pretraining) and ALIGN [4] (A Large-scale ImaGe and Noisy-text embedding) which train the FMs by aligning text and image embeddings. These models use contrastive learning to map text and images into a shared embedding space, enabling better understanding and retrieval across modalities. Additionally, some other approaches utilize transformer-based models and learn in an autoregressive manner. These approaches are more easy to scale. Some representative large-scale models are PaLI[2], Flamingo [1], etc. However, integrating structured data from Knowledge Graphs [3] (KGs) with visual and textual data remains an underexplored area despite the rich contextual information KGs can provide. Large-scale knowledge graphs, such as Wikidata [7] and PyTorch-BigGraph [5], contain rich structural information but also include noise. Integrating these knowledge graphs with foundation models presents a promising research direction.

Research Questions

One of the following research questions can be chosen.

- How can large-scale knowledge graphs be integrated into the foundation model learning pipeline?
- Can large-scale knowledge graphs assist with tasks: visual entity recognition and visual question answering?

Your Tasks

In this thesis, you will learn state-of-the-art methods to combine the information from KGs, images and text. You are supposed to

- train advanced knowledge graph embedding methods, like transE, transH, and some other GNN-based methods
- fine-tune the vision-language models (VLMs) like CLIP, Pali, and etc.
- perform ablation studies and run existing baselines.
- Develop your own methods to combine KG and foundation models.

Requirements

- High self-motivation and passion on research.
- Existing knowledge about KGs and VLMs will be a bonus

Advisors: Hongkuan Zhou (hongkuan.zhou@bosch.com), Xiangtong Yao (xiangtong.yao@tum.de)

Reference

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736, 2022.

[2] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James

Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023.

[3] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, Jos´e Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan

Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. ACM Computing Surveys, 54(4):1–37, July 2021.

[4] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. CoRR, abs/2102.05918, 2021.

[5] Adam Lerer, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. PyTorch-BigGraph: A Large-scale Graph Embedding System. In Proceedings of the 2nd SysML Conference, Palo Alto, CA, USA, 2019.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. CoRR, abs/2103.00020, 2021.

[7] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10):78–85, 2014