Technische Universität München – Faculty of Informatics
Chair for Computer Aided Medical Procedures (Prof. Nassir Navab)
Practical Course: Machine Learning in Medical Imaging

# Self-supervised Multimodal Representation Learning

## 1 General Info

**Project Title**: Self-supervised Multimodal Representation Learning
**Supervisors**: Azade Farshad, Yousef Yeganeh
**Contact Email**: azade.farshad@tum.de, y.yeganeh@tum.de

## 2 Project Abstract

Utilizing various modalities, such as images, text, and more, has become crucial for addressing real-world challenges. For instance, CLIP [7] is a large-scale model that identifies shared representations between text and images. There are emerging fields dedicated to tackling the difficulties of multimodal machine learning, where one or more modalities may be unbalanced, absent, noisy, lacking annotated data, or have unreliable labels. Our goal is to examine how the model performs in the face of these obstacles and identify methods to enhance the learned representations of the data.

One such approach involves leveraging knowledge from a resource-rich modality to benefit a resource-poor modality, through the transfer of knowledge between modalities, including their representations and predictive models. We aim to comprehend these models and employ counterfactual modeling to investigate the impact of each modality on straightforward downstream tasks.

This project will consist of multiple phases, and based on the progress made, we may conduct either a general analysis or address specific limitations.

## 3 Background and Motivation

Recent advancements in self-supervised representation learning have led to significant performance improvements in various domains. For instance, the development of contrastive learning techniques such as SimCLR [2] and MoCo [5] have demonstrated remarkable results in learning visual representations. Similarly, BERT [3] and GPT [8] have revolutionized the natural language processing landscape by showcasing the power of self-supervised learning in the textual domain. However, most of these approaches are designed to learn representations within a single modality, and their extensions to multimodal scenarios are non-trivial. Some recent studies have attempted to bridge this gap by proposing multimodal self-supervised learning frameworks such as CLIP [7], ViLBERT [6], and LXMERT [10], which learn joint representations for images and text. Despite these successes, there is still much room for improvement and exploration in the realm of self-supervised multimodal representation learning[1],[4], [9].

## 4 Technical Prerequisites

- Good background in machine learning and deep learning
- Experienced in PyTorch
- Experienced in Python
- Familiar with MONAI Framework

## 5 Benefits

- Weekly supervision and discussions
- Possible novelty of the research
- The results of this work are intended to be published in a conference or journal

## 6 Work packages and Time-plan

* The dates are adopted from the previous year and are not finalized yet.

| | Description | # Students | From | To |
|---|---|---|---|---|
| WP1 | Familiarizing with the literature. Scoping of datasets. | 4 | 10.05 | 17.05 |
| WP2 | Implementing the baselines on toy datasets. Download of real dataset(s). | 4 | 17.05 | 31.05 |
| WP3 | Improving the baselines and validation on relevant dataset(s). | 4 | 31.05 | 14.06 |
| | Midterm Presentation | 4 | 14.06 | 23.06 |
| WP4 | Implementing the model | 4 | 14.06 | 07.07 |
| WP5 | Finalizing the results and evaluation | 4 | 07.07 | 21.07 |
| | Final Presentation | 4 | 21.07 | 28.07 |

Table 1: Project Timeline

## References

[1] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, pages 1–32, 2021.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Learning Representations*, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2020.

[5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[6] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[8] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

[9] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Multimodal co-learning: challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239, 2022.

[10] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5103–5114, 2019.