MASTER'S THESIS

Multimodal Transformer Models for Human Action Classification

Problem description:

The abundance of large-scale video datasets featuring human activities has enabled researchers to gain insight into human behaviors though visual analysis. However, the cues used by humans during task execution go beyond the visual modality. For example, it is challenging to determine the force required to cut a tomato while preparing a salad through visual means only. Humans use multiple modalities, such as pose, gaze, muscle activity, and tactile information to interact with their environment effectively. To translate this embodied intelligence to robots, it is necessary to develop models that can capture the relationships between these diverse sensory modalities and the human task. To address this challenge, researchers have gathered a multi-modal dataset, named ActionSense [1], of human participants performing tasks in a kitchen scenario.
The student will study the application of deep learning techniques to classify human actions through multiple sensory modalities. The aim is to design a multi-modal transformer [2, 3] that can receive inputs from various sensory sources and classify the actions performed by a human. The student will then conduct an ablation study to analyze the model's performance when fed with one or several modalities. Optionally, the student will be encouraged to create a plug-in decoder model that outputs a modality that is different from the input modality.

Tasks:

- Task1. Literature review on multi-modal transformers models.
- Task2. Download and preprocessing of the ActionSense dataset.
- Task3. Design and implement a transformer-based neural network for human action classification given different input modalities.
- Task4. Conduct an ablation study to analyze the model's classification performance.
- Task5. (Optional) Develop a decoder to generate tactile and muscle activity data from vision.

Bibliography:

[1] Joseph DelPreto and Chao Liu et al. ActionSense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. In *Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022.

[2] A Vaswani and et al. Attention is all you need. In *Advances in neural information processing systems*, 2017.

[3] Rohit Girdhar, Mannat Singh, and et al. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Supervisor: M. Sc. Esteve Valls Mascaro (esteve.valls.mascaro@tuwien.ac.at)
M. Sc. Daniel Sliwowski (daniel.sliwowski@tuwien.ac.at)

(D. Lee)
Univ.-Professor