Master internship subject

Action Recognition by Knowledge Augmentation in Vision Language Model

Hosting institute

ICube Laboratory (The Engineering science, computer science and imaging laboratory) at the University of Strasbourg is a leading research center in Computer Science, with more than 300 permanent researchers, with the recently opened AI graduate school supported by the French government.

Work place and salary

The thesis work will take place in the MLMS (Machine Learning, Modeling & Simulation) research team of the ICube laboratory (The Engineering science, computer science and imaging laboratory) of the University of Strasbourg, a leading research center with more than 300 permanent researchers. The workplace is located on the hospital site of the laboratory, a 10-minute walk from the heart of downtown Strasbourg, listed as a UNESCO World Heritage Site.

650 euros net monthly

Supervisors

– director: Hyewon Seo (ICube, Univ. Strasbourg)

– co-supervisor: Diwei Wang (ICube, Strasbourg)

Staring date

February – April 2025.

Work description

Action recognition from video is highly important for assistive care robots, as it enables them to understand and respond appropriately to the needs and activities of the people they assist. Recent DL models for action recognition are moving toward more data-efficient, interpretable, and computationally optimized frameworks: The combination of transformer architectures, spatio-temporal attention, multimodal fusion, and self-supervised learning, just to mention a few. Meanwhile, the recent emergence of large-scale pre-trained vision-language models (VLMs) has demonstrated remarkable performance and transferability to different types of visual recognition tasks, thanks to their generalizable visual and textual representations. It has been confirmed by our recent study[1] [2], where our developed model learns and improves visual, textual,

[1]Wang D., Yuan K., Muller C., Blanc F., Padoy N., Seo H., "Enhancing Gait Video Analysis in Neurogenerative Diseases by Knowledge Augmentation in Vision Language Model", Lecture Notes in Computer Science (Proc. Medical Image Computing and Computer-Assisted Intervention), vol. 15005, pp 251–261, Springer, 2024.

[2] Wang D., Yuan K., Seo H., "GaVA-CLIP: Refining Multimodal Representations with Clinical Knowledge and Numerical Parameters for Gait Video Analysis in Neurodegenerative Diseases", under revision, 2024.

and numerical representations of patient gait videos based on a large-scale pre-trained Vision Language Model (VLM), for several classification tasks.

Motivated by these recent successes, we will extend our previous developed model and the multimodal representation for a new classification task – action recognition from video. Similarly to our previous method, we will adopt the prompt learning strategy, keeping the pre-trained VLM frozen to preserve its general representation and leverage the pre-aligned multi-modal latent space the prompt's context with learnable vectors, which is initialized with domain-specific knowledge.

We will proceed with the following steps:

1. **Data organization:** The datasets that are at our disposal will be rearranged and selected to ensure seamless use as training data.
2. **Knowledge distillation:** Per-class description will be collected and refined in a semi-automatic manner, which we will use to initialize learnable prompts.
3. **Adaptation of the model**: Based on the above knowledge, we will adapt our previously developed model to perform the new classification task. Whenever applicable, we will design specialized loss functions tailored to the specific nature of the new task. A number of ablation studies will help improve the performance and assess the impact of various components. This may involve testing a number of VLMs as backbone.
4. **Experiments**: The developed model will be parameter-tuned, tested and compared with the state-of-the-art models.

Candidate profile

− Solid programming skills in Python/C++

− Experience in Deep Learning (Transformer, CLIP, etc.)

− Good communication skills

Application

Send your CV and academic records (Bachelor and Master) to seo@unistra.fr, for (a) possible interview(s).