S08 - Enhancing Visual Reasoning in Vision-Language Models (VLMs) through Dynamic Visual Feature Selection

6-month internship @ CEA List

Internship context

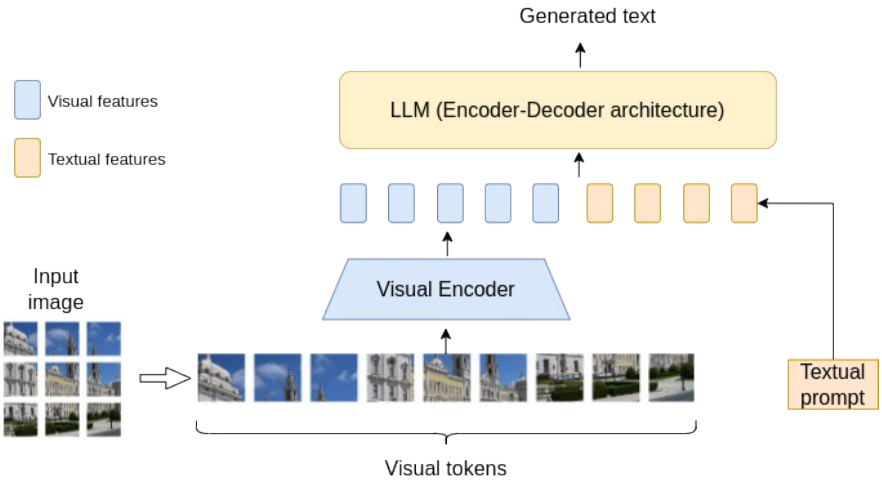
Based in Saclay (Essonne), the LIST is one of the two institutes of CEA Tech, the technological research division of the CEA. Dedicated to intelligent digital systems, its mission is to carry out technological developments of excellence on behalf of industrial partners in order to create value.

Within the LIST, the Laboratory of Vision and Learning for Scene Analysis (LVA) conducts research in the field of computer vision and artificial intelligence for the perception of intelligent and autonomous systems. The laboratory’s research themes include visual recognition, behavior and activity analysis, large-scale automatic annotation, and perception and decision models. These technologies are applied in major sectors such as security, mobility, advanced manufacturing, healthcare, and sports.

Missions

Generative Vision Language Models (VLMs) combine the understanding and generation of text in visual contexts. These models have demonstrated impressive performance on real-world visual question answering (VQA) benchmarks, suggesting their visual reasoning abilities. However, these benchmarks often mix pure visual reasoning tasks with tests of world knowledge, and typically involve questions requiring only a limited number of reasoning steps [2]. As a result, it is unclear whether a VLM's apparent success in visual reasoning tasks is truly due to its reasoning capabilities or simply leveraging its extensive world knowledge. Moreover, VLMs often struggle with fine-grained scene understanding and spatial reasoning, largely due to inefficient use of visual features [5].

This internship aims to tackle these limitations by developing a novel approach for VLMs, particularly those trained through instruction learning methods like LLAVA [1]. This architecture involves converting visual features, from a Visual Transformer model [3], into text embedding space before feeding them to a large language model (LLM) for text generation.



General Approach of Vision-Language Models (VLMs) via Instruction Tuning

Internship objectives

We propose leveraging the Chain-of-Thought (CoT) technique [4] to iteratively select the most relevant visual features during the text generation process. CoT involves generating step-by-step reasoning to break down complex tasks into simpler logical steps, which enhances model performance on tasks requiring complex reasoning. In our approach, we will begin by linking the reasoning steps in a textual chain to specific visual features within the image to provide a visual justification for each step. Afterward, the model will learn to directly select and process relevant visual features without relying on explicit textual reasoning steps, allowing for a more intuitive and efficient understanding of the visual context.

References

[1] Liu, H., Li, C., Wu, Q., & Lee, Y.J. (2023). Visual Instruction Tuning. ArXiv, abs/2304.08485

[2] Zhang, Y., Bai, H., Zhang, R., Gu, J., Zhai, S., Susskind, J., & Jaitly, N. (2024). How far are we from intelligent visual deductive reasoning? In COLM

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, abs/2010.11929

[4] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Xia, F., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. ArXiv, abs/2201.11903

[5] Zhang, J., Hu, J., Khayatkhoei, M., Ilievski, F., & Sun, M. (2024). Exploring Perceptual Limitation of Multimodal Large Language Models. ArXiv, abs/2402.07384

Qualifications

- Students in their 4th or 5th year of studies (M1, M2 or gap year)
- Computer vision skills
- Machine learning skills (deep learning, perception models, generative AI…)
- Python proficiency in a deep learning framework (especially TensorFlow or PyTorch)
- Scientific research experience will be appreciated

Job-related benefits

Joining the CEA List and the LVA as an intern means:

- Joining an organization that addresses societal challenges to build the world of tomorrow.
- Working in one of the most innovative research organizations in the world (ranked in the global top 100, top 3 in France).
- Discovering a rich ecosystem where the institute creates privileged links between the industrial and academic sectors.
- Conducting research in an environment where autonomy and creativity are recognized, and where valorizing results is encouraged (publication of scientific articles, patents, and sharing of open-source code whenever possible).
- Joining a young and dynamic team made up of research engineers, PhD students, post-doctoral researchers, and interns.
- Benefiting from an internal computing infrastructure equipped with around 300 state-of-the-art GPUs.
- Receiving a stipend between €1300 and €1400 per month.
- Having the opportunity to continue with a PhD or as a research engineer after the internship.
- Having the possibility of remote work, receiving a 75% (instead of 50%) reimbursement on public transportation costs, and benefiting from the "mobili-jeune" aid to reduce rent costs…

To apply, contact the laboratory with a CV and cover letter: lva-stages@cea.fr