Object Detection based on LLM

Université Sorbonne Paris Nord. LIPN - UMR CNRS 7030
Équipes A3[*]

This post-doctoral proposal is part of the IRISER LabCom, a Joint Laboratory in "Intelligence, Recognition, Surveillance, Reactive" (https://www-l2ti.univ-paris13.fr/iriser/) funded by the ANR. The IRISER LabCom aims to propose and fully control the behavior and performance of intelligent or embedded systems designed for artificial vision for the rapid and automated analysis of images/videos (of very large sizes, multispectral georeferenced high resolutions) captured aboard COSE aircraft, relying on strategies for processing visual information and machine learning. We are looking to recruit a post-doctoral in research and development in computer vision and machine learning for 12 months.

1 Context and objectives

The convergence of vision and language models stands as a critical and swiftly advancing domain in artificial intelligence, granting machines the capability to comprehend and produce content that amalgamates both visual and textual data, closely resembling human perception and communication. Vision-Language models primarily target the conversion of (image, text) pairs into textual outputs. This approach facilitates simultaneous processing and understanding of language (text) and vision (image) modalities, enabling the execution of sophisticated vision-language tasks like Visual Question Answering (VQA), image captioning, and Text-To-Image search.

Contemporary frameworks like simVLM [1], VirTex [2], CLIP [3], visualGPT [4], and visualBERT [5] employ diverse techniques such as contrastive learning, masked language-image modeling, and encoder-decoder modules. Despite their effectiveness, these methods can be highly resource-intensive due to the fusion of large models and training from scratch. To address this challenge, more resource-efficient approaches like Frozen PrefixLM [6], Flamingo [7], BLIP-2 [8], LLaVA [9], and VOLTRON [10] have been proposed, maintaining comparable performance levels.

While advancements in merging visual and language modalities have been remarkable, a significant drawback of current approaches lies in their limited perception ability, especially in tasks such as object detection, particularly in aerial images. This limitation underscores the need for a fresh research outlook within the visual-language domain. This perspective aims to explore novel methodologies for locating, identifying, and linking visual objects with language inputs to enrich human-AI interaction. Addressing this challenge, GLIP [11] integrates cross-attention fusion and contrastive learning at the object level, expanding upon the CLIP method for object detection. Conversely, ContextDET [12] proposes a solution for object detection across three representative scenarios: the language cloze test, visual captioning, and question answering. These models have

[*] Apprentissage Artificiel et Applications

revolutionized object detection, showcasing performance comparable to state-of-the-art methods and offering more targeted usage for context-conditioned detection.

Our primary objective is to present a solution within the domain of object detection, particulary in aerial images. In this pursuit, we aim to address the existing challenges and limitations by leveraging innovative methodologies. Our focus is on developing techniques that enhance the accuracy and efficiency of object detection algorithms, particularly in complex visual-language contexts. By combining cross-disciplinary insights from both computer vision and natural language processing, we strive to create robust and versatile solutions capable of accurately locating, identifying, and associating visual objects with corresponding language inputs. Through our research efforts, we seek to contribute to the advancement of object detection techniques, ultimately facilitating more effective human-AI interaction and enabling a wide range of applications in the field of aerial images. Additionally, we plan to implement RAG (Retrieval Augmented Generation) as a secondary approach. This module serves as a critical component for expanding the external knowledge base, enhancing the model's ability to query databases and execute fact-checking procedures, thereby increasing the depth and accuracy of the information it can provide.

2 Plan of Work

The research directions might be (but not restricted to):

- **Literature Review:**
  - Conduct an extensive review of existing vision-language models, focusing on their strengths and limitations in object detection tasks particulary in aerial images.
  - Explore recent advancements in retrieval-augmented generation (RAG) techniques and their potential applications in enhancing object detection.
- **Data Collection and Preparation:**
  - Gather diverse datasets comprising image-text pairs suitable for training and evaluating object detection algorithms in various real-world scenarios with focus on aerial images.
  - Preprocess the collected data to ensure compatibility with the selected vision-language model and RAG integration.
- **Model Development:**
  - Design and implement a novel vision-language model tailored for object detection tasks, incorporating elements from state-of-the-art frameworks.
  - Integrate RAG functionality to augment the model's knowledge base and enhance its ability to retrieve relevant information during object detection.
- **Training and Evaluation:**
  - Train the developed model on the collected datasets using appropriate optimization techniques and evaluation metrics for object detection.

    - Conduct comprehensive evaluations to assess the performance of the proposed approach compared to existing methods, considering metrics such as precision, recall, and mAP (mean Average Precision).

- **Integration and Deployment:**
    - Integrate the trained model into a user-friendly application or framework for practical deployment in real-world scenarios.
    - Conduct user testing and feedback gathering to iteratively improve the usability and effectiveness of the deployed system.

- **Documentation and Dissemination:**
    - Document the entire research process, including methodologies, findings, and insights gained, in the form of research papers and technical reports.
    - Disseminate the research outcomes through presentations at conferences, workshops, and publication in relevant journals to contribute to the wider research community.

Expected Outcomes:

- Development of a novel vision-language model capable of robust object detection in complex visual-language contexts.
- Integration of RAG functionality to enhance the model's knowledge base and improve retrieval and fact-checking capabilities.
- Empirical validation demonstrating the effectiveness and efficiency of the proposed approach compared to existing methods using aerial images, as well as non-aerial images, will be conducted.
- Contribution to advancing the state-of-the-art in object detection techniques particulary for aerial images, facilitating more effective human-AI interaction across various domains.

Prerequisites and competence

- End of PHD, in data science, statistics and/or artificial intelligence.
- Excellent experience in programming, especially with Python, Tensorflow/Keras or PyTorch.
- Big interest in and excellent understanding of machine and deep learning theory and applications.

To apply, simply attach:

- Your current Curriculum Vitae (CV),
- A portfolio of projects, if any,
- Your motivation for the position,
- Your latest university transcripts.

Send it all by email to azzag@univ-paris13.fr and mustapha.lebbah@uvsq.fr

3 Supervisors team

- Hanane Azzag, LIPN, CNRS UMR 7030, Université Sorbonne Paris Nord
- Mustapha Lebbah, DAVID Lab, UVSQ, Université Paris-Saclay

With colloaboration with Bilal Faye, PhD student at LIPN, CNRS UMR 7030, Université Sorbonne Paris Nord.

References

[1] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[2] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021.

[3] Alec Radford, Jong Wook Kim, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[4] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022.

[5] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[6] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[7] 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.

[8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[10] Zeba Mohsin Wase, Vijay K Madisetti, and Arshdeep Bahga. Object detection meets llms: Model fusion for safety and security. *Journal of Software Engineering and Applications*, 16(12):672–684, 2023.

[11] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[12] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models, 2023.