M2 Internship topic

Transformer-based Multimodal Object Detection for Remote Sensing Imagery

Zuheng MING and Fangchen FENG

Laboratoire de Traitement et Transport de l'Information (L2TI)

February 9, 2023

1 General context

Object detection is an important task in computer vision with various applications in different scientific fields. The object detection algorithm attempts to jointly determine the location of each object in an image and the class to which that object belongs. Recent advances in deep neural networks significantly improve detection performance with large-scale datasets and accurate annotations [1]. Aerial images (remote sensing data) are special types of images captured from airplanes, drones, or satellites with different altitudes and resolutions. Object detection on aerial images is of particular interest for both civilian and military applications. However, recent research reveals that several challenges, such as lacking labeled samples for small objects, still remain with current techniques for aerial images compared with object detection in natural scenarios [2]. A possible track of interest to take up the challenge is to explore the multimodal characteristics (e.g. RGB, Infrared, LiDAR and Hyperspectral image) of remote sensing imagery [3]. Indeed, remote sensing imagery collected from multimodality is becoming available as imaging technology improves. The multimodal object detection learning complementary information between different modalities can effectively enhance the detection accuracy in RSI. However, most of the current object detection techniques are solely designed and applied for a single modality such as RGB [4]. Therefore, our motivation is to propose a multimodal object detection model based on the current framework that could achieve high detection accuracy. On the other hand, Vision Transformers (ViT) [5] using the self-attention mechanism for modeling the context of the local regions in an image or the relationship of frames in the sequential data such as video is becoming a most cutting-edge backbone of deep neural networks for computer vision tasks [6]. Integrating transformer in the current framework for object detection to gain a higher detection accuracy with good inference speed is another objective.

2 Topic description

In this internship, we are interested in multimodal object detection for remote sensing imagery using transformer-based methods. The work required during this internship will be organized as follows:

- First, review the state-of-the-art works on multimodal learning methods based on transformers for RSI object detection to familiarize the methods, evaluation metrics and the datasets used for this topic.
- In the second step, reproduce one of the selected works as the base of future work. The candidate will study and improve the current network based on CNNs and Transformers. In particular, to meet the challenge of detecting small objects in vast backgrounds.
- Third step, adapt the proposed framework to the multimodal remote sensing data and try to achieve high detection accuracy with high inference speed which should have a good accuracy-speed compromise.
- Final step, write the report and expected result in an academic publication at the end of the internship.

3 Practical aspects

- Supervision: Zuheng MING - Maître de conférences (L2TI, UR 3043) & Fangchen FENG - Maître de conférences (L2TI, UR 3043).
- Duration: 5 months starting from March 2023 with a standard internship grant.
- Location: Laboratoire de Traitement et Transport de l'Information, Université Sorbonne Paris Nord, Villetaneuse.
- Expected background: we look for a motivated M2 student with training in machine learning and/or deep learning, and an interest in computer vision
- Application: send by email your resume, motivation letter, transcription of master's degree, and at least one recommendation letter to `zuheng.ming@univ-paris13.fr` or `fangchen.feng@univ-paris13.fr`.
- Language: French or English
- Deadline of application: 7 February 2023

4 Required skills

- Master 2 in Computer Science or a related technical field.
- Experience working with Python, Pytorch or Tensorflow in Linux or Windows.
- Experience in computer vision, deep learning, and machine learning.

Preferred qualifications:

- Experience in training deep learning models based on GPU or Colab and the basic skills of Github.
- Experience in multi-modal learning (e.g., the images acquired in different types of camera sensors such RGB, LiDAR, Hyperspectral images,...) based on multiple modalities.
- Experience in learning from self-supervised learning.
- Publication at academic conferences, e.g., in conferences of computer vision, image procession, AI domain, or related technical fields.

References

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[2] Pierre Le Jeune and Anissa Mokraoui, "Improving few-shot object detection through a performance analysis on aerial and natural images," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 513–517.

[3] Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *arXiv preprint arXiv:2205.01380*, 2022.

[4] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu, "Learding2019learninging roi transformer for oriented object detection in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2849–2858.

[5] Alexey Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2020.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.