EGU24-10914, updated on 22 Apr 2024
https://doi.org/10.5194/egusphere-egu24-10914
EGU General Assembly 2024

Efficient adaptation of Foundation Models for Visual Grounding Remote Sensing task

Ali J. Ghandour[1], **Hasan Moughnieh**[1], Mohammad Hasan Zahweh[1], Hasan Nasrallah[1], Mustafa Shukor[2], Cristiano Nattero[3], and Paolo Campanella[3]

[1]National Center for Remote Sensing, CNRS, Lebanon
[2]Sorbone University, France
[3]WASDI, Dudelange, Luxembourg

Foundation models have demonstrated impressive proficiency across multiple domains, including language, vision, and multi-modal applications, establishing new standards for efficiency and adaptability. In the context of localization-based foundational models, the core strength of such models is their ability to precisely recognize and locate objects across a diverse set of objects in wide-area scenes. This precision is particularly vital in the Remote Sensing (RS) field. The multimodality aspect of these models becomes pivotal in RS, as they can process and interpret complex data, allowing for more comprehensive aerial and satellite image analysis.

Multimodality has emerged as a crucial and dynamic area in recent AI developments, finding diverse applications such as image captioning and visual question answering. More related to traditional visual tasks, Visual Grounding (VG) stands out, involving the localization of objects based on textual descriptions. Unlike conventional approaches that train models on predefined and fixed lists of objects, VG allows a model to locate any entity in an image based on diverse textual descriptions, enabling open-vocabulary predictions. Despite notable efforts in developing powerful VG models to solve general benchmarks, there is a need for more exploration into transferring these models to the remote sensing context.

This paper addresses this gap by delving into the task of visual grounding for remote sensing. Our initial exploration reveals that utilizing general pretrained foundational models for RS yields suboptimal performance. After recognizing these limitations, our work systematically investigates various parameter-efficient tuning techniques to fine-tune these models for RS visual grounding applications. The insights and methodologies presented in this paper provide valuable guidance for researchers seeking to adapt pretrained models to the RS domain efficiently. This adaptation marks a substantial advancement in the field, offering a significant stride toward enhancing the applicability of visual grounding in remote sensing scenarios.