Thematic visual grounding

PhD topic proposal

General information

- Doctoral school: ED 130 - Informatique, Telecommunications et Electronique de Paris
- Duration: 3 years
- Supervision: Nicolas Loménie, Sylvain Lobry - (prenom.nom@u-paris.fr)
- Laboratory: Université Paris Cité, Laboratoire d'Informatique Paris Descartes (LIPADE), équipe Systèmes Intelligents de Perception
- Location: The candidate will have an office at 45 rue des Saints-Pères, 75006 Paris. Computing resources will be provided.
- Keywords: Deep learning, multimodal imagery, computer vision, natural language processing, medical imaging, remote sensing

Detailed proposal

Introduction

Visual grounding is a task that aims at locating objects in an image based on a natural language query. This task, along with image captioning, visual question answering or content based image retrieval links image data with the text modality. Numerous works have been produced in the last decade about visual grounding in the computer vision community [1]–[3]. These work most often consider both modality separatly, through a dedicated encoder (e.g. a convolutional neural network for images, a recurrent neural network for the text). Both encoded representations are then merged, potentially using attention mechanisms, to obtain a common latent representation.

Recently, text-image foundation models such as CLIP (Contrastive Image Language Pre-training, [4]) have changed the paradigm for visual grounding models [5]. Indeed, leveraging the shared semantics between language and images is a key element for the task.

While great amount of works have been produced in the computer vision community on the task of visual grounding on natural images, there is a lack of research works on this task for thematic domains such as medical imaging and remote sensing. In both of these domains, there is a need to precisely locate particular objects, following precise definitions, in images. In addition, the image of a particular scene (e.g. an organ in medical image, a geographical area in remote sensing) can be made through several acquisitions (e.g. an MRI stack or a time series). As such, we are interested in the question:

How can visual grounding be made domain specific?

In medical imaging, visual grounding is an important research task aiming at assisting medical doctors navigate the huge amount of visual data, for instance in radiology or histopathology. A recent trend in the field of computational pathology is to rely on a good feature extractor robust to stain variations or varous hostpital sites protocols [6] and to leverage on it to augment the assistance to clinical practice. In the field of medical imaging, the phrase visual grounding as

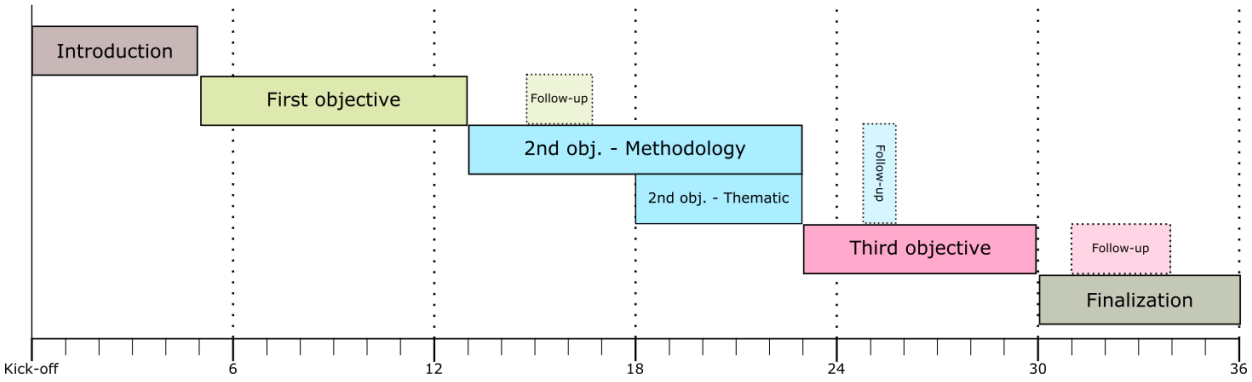

Figure 1: Gantt chart for the proposed planning. The follow-up boxes represent time planned for reviews and/or rebuttals linked to the publications of a given objective.

described in [7] allows to pave the way towards the clinical practice at horizon 2030 [8]. We already used Transformer modelling for medical images in the field of histopathology, hence our team would like to leverage on this visual architecture to integrate textual interactions[9].

In remote sensing, the task of visual grounding has been proposed by [10]. In this work, the authors take inspiration from [11] to build a visual grounding dataset from OpenStreetMap data on remote sensing images. In addtion, the authors propose a two-stream network leveraging attention to perform the visual grounding task. In [12] another dataset, RSVG, is built from a target detection dataset (DIOR [13]). Similarly, a two-stream method is proposed and evaluated on this proposed dataset. Among others, the task of visual grounding is used in [14] to build a grounded large vision-language for remote sensing data. These datasets and methods have in common to not be specifically tuned for one particular sensor. In addition, they cannot handle other remote sensing modalities such as SAR. Finally, they cannot work on time series.

To answer the question raised in this research proposal, we propose to consider two aspects: the specific semantics of the thematic domains and the particular domain distributions of the images. As such, we propose to decompose this research in three main research objectives:

1. **Visual grounding from multi-modal data:** this objective aims at simultaneously grounding an object on different data (e.g. optical and SAR) that may present different geometries.

2. **Visual grounding on stack of images:** we propose to take into account the fact that acquisition can be divided into several images (e.g. MRI stacks or multi-temporal data). Our objective is to perform the visual grounding task on this data, finding the best representation of a description in a stack of images.

3. **Pixel-level grounding:** For thematic images, there can be a need to have a precise location of objects of different shape. As such, we want to propose a new grounding task in which the output is not the bounding box of the object of interest, but a segmentation.

State of the art

In this work, we propose to tackle these three challenges. These objectives will be grounded in the current efforts of the SIP team of the LIPADE laboratory to build foundation models for thematic data, such as remote sensing.

Proposed planning

The PhD candidate is expected to progress through the 3 objectives according to the planning shown in Figure 1. If followed, this plan should allow the candidate to publish four journal articles and several top-tier conferences (in the remote sensing, medical, and computer vision communities).

1. KO $\Rightarrow$ KO + 4 months: **Training**
   Review of the literature, followed by a simple project to familiarize the candidate with thematic data. With the candidate, we will propose a simple model for visual grounding to be tested on existing datasets.
2. KO + 5 months $\Rightarrow$ KO + 12 months: **First objective**
   From existing datasets associating language and thematic data developed by the SIP team, the candidate will propose an adaptation in order to perform the visual grounding task. The contribution of this study will be to work on multi-modal data. The preliminary study will be reported at a thematic conference (e.g. IEEE IGARSS for remote sensing) and the full study in a thematic journal.
3. KO + 13 months $\Rightarrow$ KO + 22 months: **Methodological developments for the second objective**
   The candidate will propose (jointly with the supervision team) a methodology for the visual grounding task on stacks of images. These developments will be made generic, so they can be applied to medical imagery or remote sensing imagery.
4. KO + 18 months $\Rightarrow$ KO + 22 months: **Thematic developments for the second objective**.
   The candidate will propose thematic application (in remote sensing and medical imagery) for the methodological developments proposed in the second objective. For this part of the work, the objective will be to publish at a major computer vision conference (e.g. CVPR or ICCV).
5. KO + 23 months $\Rightarrow$ KO + 29 months: **Third objective**
   The candidate will work on a new formulation of the visual grounding task, in which the output is at the pixel level instead of the object level. For this, the candidate will leverage the datasets of the first objective. This work will be made first generic, and thematic adaptation will be proposed to the medical and remote sensing communities in at least one journal.
6. KO + 30 months $\Rightarrow$ KO + 36 months: **Finalization**

   - Submission of an article in a remote sensing journal and in a medical imagery journal summarizing the different aspects studied during the thesis.
   - Writing of the PhD manuscript and preparation of the Defense.

Background of the candidate

The candidate must have a strong background in Computer Science or Mathematics. A background in either computer vision, image processing or natural language processing (NLP) is welcome. Knowledge in Python, C or C++ and in a deep learning framework is a plus.

About LIPADE (host laboratory)

The Laboratoire d'Informatique de Université de Paris (LIPADE) is located at the UFR Mathématiques et Informatique of the University Paris Cité. The Intelligent Perception Systems

(SIP) team of LIPADE was created in 1990 by Professor Georges Stamon. The candidate will integrate the SIP team of the LIPADE.

The SIP team develops a priority axis on image analysis and interpretation with a particular emphasis on computer vision. The structural models (graph or pixel grid) as well as the statistical models (distributions) are addressed by integrating external knowledge to the image analysis workflow (complementary textual information and high-level expert knowledge for example). The main contributions of our team consist in developing new algorithms and processing methods for the analysis of complex images (such as multi-source image segmentation, 3D image reconstruction, object recognition, extraction of semantic information from images for indexing).

The research is done through three thematic axes: document imaging, remote sensing imaging and bio-medical imaging.

Bibliography

[1] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell **and** B. Schiele, "Grounding of textual phrases in images by reconstruction," **in***Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* Springer, 2016, **pages** 817–834.

[2] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell **and** M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[3] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu **and** M. Tan, "Visual grounding via accumulated attention," **in***Proceedings of the IEEE conference on computer vision and pattern recognition* 2018, **pages** 7746–7755.

[4] A. Radford, J. W. Kim, C. Hallacy **andothers**, "Learning transferable visual models from natural language supervision," **in***International conference on machine learning* PMLR, 2021, **pages** 8748–8763.

[5] W. Jin, S. Mukherjee, Y. Cheng **andothers**, "Grill: Grounded vision-language pre-training via aligning text and image regions," *arXiv preprint arXiv:2305.14676*, 2023.

[6] G. Wolflein, D. Ferber, A. R. Meneghetti **andothers**, "A good feature extractor is all you need for weakly supervised learning in histopathology," 2023. eprint: 2311.11772. **url**: https://github.com/georg-wolflein/histaug.

[7] Z. Chen, Y. Zhou, A. Tran **andothers**, "Medical phrase grounding with region-phrase context contrastive alignment," 2023. arXiv: 2303.07618.

[8] A. B. et al., "Computational pathology in 2030: A delphi study forecasting the role of ai in pathology within the next decade," 2023. **url**: https://doi.org/10.1016/j.ebiom.2022.104427.

[9] Z. Guo, Q. Wang, H. Muller, T. Palpanas, N. Lomenie **and** C. Kurtz, "A hierarchical transformer encoder to improve entire neoplasm segmentation on whole slide image of hepatocellular carcinoma," 2023. arXiv: 2307.05800. **url**: https://arxiv.org/abs/2307.05800.

[10] Y. Sun, S. Feng, X. Li, Y. Ye, J. Kang **and** X. Huang, "Visual grounding in remote sensing images," **in***Proceedings of the 30th ACM International Conference on Multimedia* 2022, **pages** 404–412.

[11] S. Lobry, D. Marcos, J. Murray **and** D. Tuia, "Rsvqa: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, **journvol** 58, **number** 12, **pages** 8555–8566, 2020.

[12] Y. Zhan, Z. Xiong **and** Y. Yuan, "Rsvg: Exploring data and models for visual grounding on remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, **journvol** 61, **pages** 1–13, 2023.

[13] K. Li, G. Wan, G. Cheng, L. Meng **and** J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS journal of photogrammetry and remote sensing*, **journvol** 159, **pages** 296–307, 2020.

[14] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan **and** F. S. Khan, "Geochat: Grounded large vision-language model for remote sensing," *arXiv preprint arXiv:2311.15826*, 2023.