Visual Spatial Question Answering

Stage M2/Ingénieur en Informatique

Informations générales

- Mots-clés : vision par ordinateur, IA, Visual Spatial Question Answering, relations spatiales
- Durée du stage : 6 mois (gratification standard)
- Structure d'accueil : Université de Paris, Laboratoire d'Informatique Paris Descartes (LIPADE), équipe Systèmes Intelligents de Perception
- Adresse : 45 rue des Saints-Pères, 75006 Paris
- Encadrement : Sylvain Lobry, Camille Kurtz, Laurent Wendling - (prenom.nom@u-paris.fr)

Sujet du stage

Contexte

De nos jours, l'augmentation constante des masses de données visuelles (Internet, entreprise,...) requiert l'utilisation d'outils de plus en plus performants pour rechercher des informations pertinentes et questionner des bases d'images structurées à partir de requêtes en langage naturel. C'est un problème difficile en vision par ordinateur car il intègre de nombreux aspects liés à la construction et à la réponse à des règles linguistiques qui peuvent requérir aussi la modélisation de critères statistiques (comme le calcul du nombre de bâtiments dans une image satellite) et de relations spatiales (à côté, le long de, à gauche, autour,...).

La tâche du *Visual Question Answering* (VQA) a été proposée récemment dans la communauté de vision par ordinateur [1]. L'objectif est de répondre à une requête (formulée sous forme de question ouverte et non contrainte) portant sur une image donnée. Un exemple d'une telle requête est montré dans la Figure 1. Les principaux travaux dediés à ce problème portent sur l'amélioration globale des modèles (le plus souvent basés sur l'apprentissage profond), sur des bases de données créées à partir d'images naturelles et acquises dans de bonnes conditions [2]. Cependant, des travaux récents proposent d'adapter cette tâche à l'aide aux personnes malvoyantes [3] ou à l'extraction d'information depuis des images satellitaires [4]. Dans ce dernier cadre, il apparaît important de prendre en compte l'information spatiale de manière explicite.

Le positionnement spatial se fonde souvent sur des approches qualitatives (comme le barycentre ou le rectangle englobant) qui proposent des représentations symboliques de l'information spatiale principalement pour l'étude de configurations qui font intervenir la topologie entre les objets. Pour la plupart, elles étendent facilement des relations élémentaires (à droite, au-dessus,...). Néanmoins, bien que ces approches qualitatives soient en mesure de détecter un large panel de configurations et proposer une représentation souvent pertinente des scènes, elles permettent difficilement de modéliser le caractère graduel de ces relations. De nombreux travaux pour la modélisation de relations spatiales quantitatives ainsi que leurs descriptions ont été développés par notre équipe [5, 6, 7, 8].

Travail a réaliser

Le sujet du recherche porte sur la tâche du *Visual Spatial Question Answering*. La première partie du stage sera axée sur l'intégration de relations spatiales dans le modèle de VQA pour générer la ré-

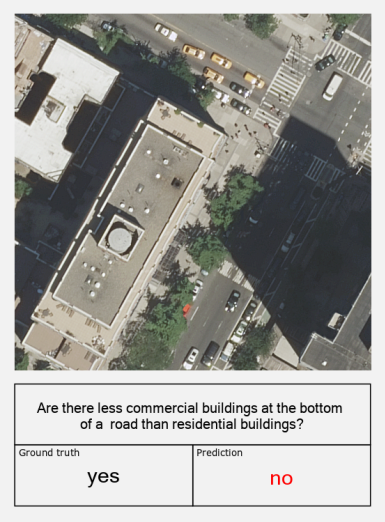

Figure 1 – Exemple de requête et résultat obtenu avec le modèle proposé par [4].

ponse à des questions complexes d'un point de vue spatial (e.g. combien y a t-il de bâtiments autour de cet hôpital ?). Dans cet axe, l'objectif sera donc d'intégrer la modélisation spatiale quantitative dans un modèle fondé sur l'apprentissage profond. Une deuxième partie pourra porter sur une approche intégrant des quantificateurs linguistiques pour affiner la requête (peu, beaucoup, fortement) ou la réponse obtenue grâce à l'aspect graduel de nos modèles de relations spatiales quantitatives.

Ces recherches seront appliquées au domaine de la télédétection et de l'imagerie satellitaire. Les données seront issues d'OpenStreetMap qui est la plus grande base de données géographique au monde. Suivant l'avancement, d'autres données pourront être employées pour évaluer le modèle de VQA, par exemple directement extraites d'images satellitaires via une pré-étape de segmentation sémantique pour étudier les objets géographiques élémentaires (bâtiments, routes, ...).

Profil recherché pour le/la candidat/e

Nous recherchons un(e) étudiant(e) en Master 2 (ou équivalent Ingénieur) informatique, ayant des compétences en analyse d'images, vision par ordinateur et en programmation Python. Une expérience avec un *framework* d'apprentissage profond est un plus.

Sources bibliographiques

[1] Stanislaw Antol et al. "VQA : Visual question answering". In : *IEEE international conference on computer vision*. 2015, p. 2425-2433.

[2] Qi Wu et al. "Visual question answering : A survey of methods and datasets". In : *Computer Vision and Image Understanding* 163 (2017), p. 21-40.

[3] Danna Gurari et al. "Vizwiz grand challenge : Answering visual questions from blind people". In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, p. 3608-3617.

[4] Sylvain Lobry et al. "RSVQA : Visual Question Answering for Remote Sensing Data". In : *IEEE Transactions on Geoscience and Remote Sensing* (2020).

[5] Pascal Matsakis et Laurent Wendling. "A New Way to Represent the Relative Position between Areal Objects". In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21.7 (1999), p. 634-643.

[6] Pascal Matsakis et al. "Linguistic description of relative positions in images". In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B : Cybernetics* 31.4 (2001), p. 573-88.

[7] Michaël Clément et al. "Directional Enlacement Histograms for the Description of Complex Spatial Configurations between Objects". In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.12 (2017), p. 2366-2380.

[8] Michaël Clément, Camille Kurtz et Laurent Wendling. "Learning spatial relations and shapes for structural object description and scene recognition". In : *Pattern Recognition* 84 (2018), p. 197-210.