Titre:

Fusion multi-niveaux pour la réponse automatique à des questions visuelles sur des images de télédétection.

Domaine scientifique : *Sciences et technologies de l'information et de la communication*

Contexte :
De grandes quantités d'images de télédétections sont aujourd'hui facilement accessibles grâce aux efforts venant des secteurs public et privé. Un exemple fort sont les satellites Sentinel lancés depuis 2014 dans le cadre du programme Copernicus de l'Union Européenne. Cette mission offre un accès libre à des images de natures différentes (multi-spectral et radar notamment) avec une grande couverture spatiale et un temps de revisite court.

Cependant, il peut être difficile d'extraire de l'information des images de télédétection. Cette interpretation est généralement faite par des experts, et implique souvent un travail manuel, qui devient un facteur limitant avec l'augmentation de la quantité de données produites. Ainsi, des méthodes automatiques ont été développées pour des applications d'intérêt général (par exemple: le suivi des feux de forêts) ou présentant un intérêt financier. Cependant, les informations contenues dans ces images peuvent être d'intérêt pour un publique bien plus large. Par exemple, les journalistes pourraient suivre d'une manière indépendante les guerres ou les effets du déréglement climatique. Les administrations locales pourraient utiliser cette information dans la prise de décision. Enfin, les citoyens sont aussi intéressés par leur environnement, comme le montre le succès d'initiatives telles que OpenStreetMap ou Google StreetView. Alors que les données sont là, le grand public n'a pas la compétence pour en extraire une information utile. Notre objectif est donc de permettre l'extraction d'information via des modèles permettant de répondre automatiquement à des questions posées (en langage naturel) à propos d'un ensemble d'images de télédétection (de différentes modalités).

Cette tâche de *visual question answering* (VQA) a été récemment proposée dans la communauté de la vision par ordinateur [1] et pour la télédétection [2]. Dans cette thèse nous nous intéresserons aux opérations permettant la fusion des caractéristiques extraites de la question, et celles des images de différentes modalités. L'objectif sera de proposer de nouvelles méthodes permettant de prendre en compte les différents niveaux d'information contenus dans les différentes modalités en lien avec la requête en langage naturel.

L'encadrement se fera conjointement avec Sylvain Lobry (LIPADE – Équipe SIP) et pourra faire l'objet de collaborations et de visites avec d'autres équipes de recherche à l'internationale travaillant sur ce sujet.

Description détaillée :

Projet de recherche :

Les méthodes de réponse automatique à des questions visuelles, à base de réseaux de neurones profonds, peuvent généralement se décomposer en quatre blocs :

1) Un bloc permettant l'extraction de caractéristiques visuelles ;
2) Un bloc d'extraction de caractéristiques textuelles ;
3) La fusion des caractéristiques textuelles et visuelles ;
4) La prédiction de la réponse.

Dans cette thèse, nous souhaitons nous intéresser principalement à l'étape de fusion (3^{e} bloc) et à ses conséquences sur la prédiction. Cette étape a pour objectif de permettre l'interaction entre un vecteur de caractéristiques visuelles et un vecteur de caractéristiques textuelles. Une interaction complète pourrait être obtenue en réalisant un simple produit matriciel entre l'un des vecteurs et la transposée de l'autre [3]. Cependant, cette opération implique des coûts computationnels bien trop élevés en pratique. Ainsi, il est fréquent d'utiliser des mécanismes d'attention afin de choisir à travers la question quelles sont les zones de l'image à prendre en compte pour répondre à celle-ci. Par exemple, des méthodes de décomposition [4] ou des architectures type bottom-up / top-down [5] sont souvent utilisées. Ces mécanismes d'attention ont aussi été explorés pour la réponse automatique à des questions portant sur des images de télédétection [6].

L'originalité de ce projet de recherche est que nous considérons la tâche de la réponse automatique à des questions visuelles à partir d'un ensemble d'images de télédétection de différentes modalités et de différentes résolutions. Ainsi, il n'est plus seulement nécessaire de choisir les parties spatiales d'une image qui seraient pertinentes pour répondre à la question, mais de choisir les parties spatiales pertinentes (éventuellement différentes pour chacune des modalités) des différentes images. Ces opérations permettraient par exemple de combiner les avantages de la précision spatiale d'images très haute résolution, tout en gardant les avantages de la couverture spatiale et de leur fréquence d'acquisition d'images plus faiblement résolues.

Programme initial de travail :

Nous proposons de décomposer le programme de recherche en trois étapes principales :

- Constitution d'une *baseline* : la tâche de réponse automatique à des questions portant sur des images de différentes modalités d'observation de la terre étant nouvelle, la première étape sera de prendre en main un nouveau jeu de données et de proposer une méthode basée sur une fusion des différentes modalités avant l'interaction avec le texte. Cette méthode servira d'étalon pour évaluer les propositions méthodologiques dans la suite de cette thèse.
- Proposition de nouvelles méthodologies permettant l'attention sur les différentes modalités de la base considérées. Une première méthode considérant des poids pour chacune des modalités sera proposée. Dans un second temps, nous nous intéresserons à la robustesse aux données manquantes en introduisant des interactions au niveau des poids d'attention.

- Proposition de méthodes permettant d'utiliser les poids d'attention pour fournir des informations supplémentaires sur les résultats. En effet, les informations sur les modalités et leurs zones spatiales considérées pour la réponse aux questions peuvent permettre d'interpréter ou d'expliquer les résultats du modèle.

Enfin, le sujet étant complexe et novateur, nous proposerons de généraliser les propositions méthodologiques issues de ces travaux à d'autres domaines thématiques de la vision par ordinateur telles que l'imagerie médicale ou encore l'imagerie de document à travers l'expertise de l'équipe SIP.

Profil du candidat :

Le candidat doit avoir de très bonnes connaissances dans les domaines de la reconnaissance des formes et de l'analyse d'images. Il doit aussi avoir un excellent niveau en programmation, notamment en Python. Des connaissances dans les domaines de la télédétection ou du VQA sont appréciées.

Bibliographie :

[1] Antol, Stanislaw, et al. "VQA: Visual question answering." Proceedings of the IEEE international conference on computer vision. 2015.
[2] Lobry, Sylvain, et al. "RSVQA: Visual question answering for remote sensing data." IEEE Transactions on Geoscience and Remote Sensing 58.12 (2020): 8555-8566.
[3] Chappuis, Christel, et al. "How to find a good image-text embedding for remote sensing visual question answering?" MACLEAN Workshop at ECML/PKDD 2021
[4] Ben-Younes, Hedi, et al. "Mutan: Multimodal tucker fusion for visual question answering." Proceedings of the IEEE international conference on computer vision. 2017.
[5] Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[6] Zheng, Xiangtao, et al. "Mutual attention inception network for remote sensing visual question answering." IEEE Transactions on Geoscience and Remote Sensing 60 (2021): 1-14.