Couplage d'informations sémantiques et spatiales pour guider l'apprentissage de représentations d'images via des modèles neuronaux

Établissement : Université Paris Cité
Unité : Laboratoire d'Informatique Paris Descartes (LIPADE), équipe Systèmes Intelligents de Perception (SIP)
Directeur de thèse : Camille Kurtz, MCF-HDR

Résumé
Ce PRD est ancré dans le domaine de la vision par ordinateur et plus précisément la reconnaissance des formes. Analyser et interpréter une image numérique est une tâche qui consiste à extraire des informations visuelles à partir de son contenu au moyen d'algorithmes et de méthodes informatiques. Le fil conducteur de ce PRD repose sur la définition de représentations d'images de plus haut niveau, plus proches de la sémantique et du raisonnement humain. Les représentations d'images sont une des clés essentielles de la vision artificielle car elles permettent, à travers de nouveaux espaces de représentation des données visuelles, d'améliorer la capacité des algorithmes à raisonner pour différentes tâches de traitement et d'analyse (e.g. segmentation, reconnaissance, classification), avec comme objectif ultime de réduire le fossé sémantique entre les caractéristiques de bas niveau extraites des pixels et la perception humaine du contenu imagé. Nous abordons ici cette question principalement sous l'angle de la recherche d'images similaires par le contenu (CBIR), où l'on dispose d'une image « requête » et l'on souhaite interroger le contenu d'une base de données pour retrouver des images comportant des caractéristiques visuelles communes. Si les approches classiques, maintenant largement fondées sur l'optimisation de réseaux de neurones convolutionnels (CNNs) [Dub22], permettent d'obtenir des résultats à l'état de l'art dans différentes situations, elles souffrent néanmoins de certaines limites, en particulier lors de l'analyse de scènes complexes (e.g. composées d'objets multiples et portant une sémantique riche), pouvant conduire à des résultats qui ne sont pas toujours pertinents pour les besoins applicatifs de l'utilisateur (problème de l'*intention gap*). Ces limites sont principalement dues aux stratégies employées pour optimiser les modèles neuronaux, qui conduisent à des représentations ne prenant pas suffisamment en compte la richesse de la structure spatiale et de la sémantique des objets composant la scène. Souvent fortement supervisée (comme l'apprentissage d'un modèle neuronal pour une tâche de catégorisation), ces approches nécessitent par ailleurs pour l'entraînement une masse importante d'images annotées afin d'apprendre un modèle généralisable. Dans ce PRD (voir descriptif détaillé), nous proposons (1) d'explorer et de définir de nouvelles stratégies pour apprendre des représentations composites qui intègrent des informations de descriptions spatiales complexes entre couples de régions (et interne à chaque région) et (2) d'étudier la manière d'intégrer des informations sémantiques a priori (issues par exemple d'ontologies) pour contrôler plus finement l'optimisation des représentations issues des CNNs, conduisant à des descriptions plus fines des scènes considérées.

Description détaillée
Projet de recherche
Le but de cette thèse est de proposer un cadre unifiant permettant d'intégrer des informations sémantiques et spatiales pour guider / contrôler plus finement l'optimisation de modèles neuronaux convolutionnels, conduisant à l'apprentissage de représentations d'images de plus haut niveau. Il s'agit d'un problème de plus en plus abordé en vision par ordinateur, par exemple sous la tâche de la reconnaissance de triplets (sujet, prédicat, objet) via des CNNs [Dai2017, Pey2017]. Ces développements ont été rendus possibles par la constitution de grandes bases de données annotées contenant des relations visuelles comme Visual Genome [Kri17] ou SpatialSense [Yan19]. Bien que ces approches constituent des avancées notables, elles restent limitées pour la description de scènes structurées car elles reposent sur des objets (sémantiques) décrits par leurs labels et localisés par leur boîte englobante (ou leur barycentre), et par l'usage de stratégies d'entraînement qui prennent difficilement en compte les relations spatiales complexes entre les objets. Bien qu'appréhendées de manière implicite, les relations spatiales et sémantiques entre les différents objets d'intérêt composant une scène jouent un rôle primordial dans notre perception de

celle-ci. Souvent définies de manière imprécise et ambiguë, leur exploitation dans des processus de reconnaissance automatisés demeure aujourd'hui délicate [Wan23] et assurément insuffisante pour combler ce qui est souvent désigné dans la littérature par le terme de fossé sémantique.
Un premier verrou scientifique réside dans la modélisation des relations spatiales et sémantiques entre des objets imagées. Freeman a défini un groupe de 13 relations spatiales (dont certaines ont été déclinées par la suite) qui font référence depuis 1970. Au sein de l'équipe Systèmes Intelligents de Perception (SIP) du LIPADE, nous avons défini dans des travaux fondateurs, de nouveaux modèles théoriques pour représenter de nouvelles relations spatiales complexes comme l'enlacement et l'entrelacement [Clé17] qui élargissent ces modèles initiaux, dans la lignée des descripteurs de positions relatives comme l'histogramme de forces [Mat99]. Un premier axe de recherche sera d'étendre ces relations pour proposer de nouveaux groupements de relations spatiales en lien avec l'imbrication de couples d'objets pouvant être composées, pour chacun d'eux, de plusieurs composantes connexes. Concernant l'aspect sémantique, notre axe d'étude reposera sur l'usage d'ontologies permettant de modéliser des connaissances et des relations sémantiques (hyperonymie, synonymie, etc.) entre les labels pouvant caractériser les objets de la scène.
Un second verrou scientifique repose sur l'intégration de telles informations spatiales et sémantiques pour guider l'optimisation de modèles neuronaux convolutionnels conduisant à l'apprentissage de représentations d'images. Ce verrou a été récemment étudié dans une thèse [Riv22] sous l'angle de la tâche de segmentation, en contraignant un modèle de type *Unet* via une prise en compte dans la fonction de perte d'informations spatiales entre les objets segmentés. Ce modèle est valable pour des paires d'objets mais nécessite un a priori fort sur les configurations spatiales reconnaissables (avec des données étiquetées dans ce sens). Notre ambition est ici d'explorer ce verrou sous l'angle du CBIR en couplant sémantique et informations spatiales pour être en mesure de comparer des scènes comportant des nombres variés d'objets et des configurations spatiales complexes. De manière plus générale, nos objectifs s'inscrivent dans une tendance actuelle en IA visant à baisser le niveau de supervision des algorithmes (e.g. approches faiblement / auto-supervisées) afin de minimiser l'emploi de masses de données annotées et l'a priori sur les classes reconnaissables.

Les contributions scientifiques et applicatives porteront principalement sur des problématiques liées à l'analyse de scènes structurées (à partir d'images naturelles), à l'imagerie médicale et à la télédétection.

Programme initial de travail

Le travail de thèse consistera à étudier comment des informations sémantiques et spatiales peuvent être employées pour contraindre l'entraînement d'un CNN via des stratégies de supervision différentes de celles classiquement employées. Le travail demandé se décomposera en trois étapes principales :

1. Une première étape sera axée sur la modélisation de la configuration spatiale des objets composant une scène, de manière à pouvoir quantifier la ressemblance entre deux images. Pour ce faire, les travaux pourront débuter par l'étude des descripteurs de formes et de positions relatives, récemment proposés par l'équipe SIP [Del22], et leur extension à des notions d'élasticité voire d'espacement entre paires d'objets imagés. Concernant la phase de segmentation, ces recherches pourront s'appuyer sur des résultats préliminaires déjà obtenus permettant de transférer des modèles de segmentation pré-entrainées sur des grandes bases comme MS-COCO via des informations a priori sur le contenu de l'image (*bouding box*, labels sémantiques et cartes d'attention post-hoc) [Del21]. La sémantique des objets de l'image pourra également être prise en compte a priori, à l'instar des approches de détection de triplets comme [Dai17] qui prennent en compte des caractéristiques linguistiques données par la nature des objets et leur configurations habituelles (e.g. une chaise est généralement sous un bureau).
2. Une deuxième étape sera liée à l'entraînement des réseaux de neurones à partir de ces informations via l'exploration de différentes stratégies d'apprentissage. Nous proposons d'exploiter des fonctions de coût comme les *triplet loss* (ou *ranking loss*) pour apprendre aux modèles à rassembler (ou ordonner) des configurations spatiales et sémantiques similaires et à éloigner les exemples dissimilaires, cette notion de similarité pouvant être quantifiée à partir de descripteurs de positions relatives (issues de la première étape de travail), sur un jeu d'entraînement. L'intégration

de sémantique pourra également être opérée via une prise en compte des relations hiérarchiques entre les labels, en s'inspirant par exemple du modèle HAPPIER (*Hierarchical average precision training for pertinent image retrieval*) [Ram22]. En considérant qu'un résultat n'est pas « binairement » pertinent par rapport à une requête, une telle stratégie permet d'intégrer des informations liées aux relations sémantiques (hiérarchiques) entres des labels associés aux images pour l'évaluation de résultats de CBIR, conduisant à une métrique plus proche de la perception. Une telle stratégie peut être implémentée dans une fonction de coût pour l'optimisation des CNNs.

3. Suivant l'avancement de la thèse, un dernière étape pourra être liée à la modélisation de la scènes à partir de graphes de régions, dans la continuité de [Clé18]. À partir de graphes de régions caractérisés issus des premières étapes, il s'agira ensuite d'apprendre automatiquement de nouvelles représentations d'images (*embeddings*) en étudiant ici l'usage des graphes CNN. L'objectif principal sera de découvrir automatiquement, à partir des données, quelles sont les relations spatiales les plus pertinentes pour caractériser une scène (ou une base de scènes) via une transcription sémantique des différentes relations calculée en fonction de l'agencement des régions.

Au delà de la tâche du CBIR, de tels modèles neuronaux pourront ensuite être exploités dans différents contextes liés à l'analyse ou la reconnaissance / classification de scènes structurées (potentiellement avec des approches du type *few-shot*), dans la lignée de travaux préliminaires menés au LIPADE. Des jeux de données classiques comme SpatialSense (2D [Yan19] voir 3D [Goy20]) ou VisualGenome [Kri17] serviront de base à cette étude.

Informations complémentaires

D'un point de vue méthodologique, des collaborations avec des collègues du LaBRI (M. Clément) et de l'University of South Dakota, USA (K.C. Santosh) intéressés par ces thématiques de recherche sont envisagées. D'un point de vue thématique, ces travaux pourront conduire à un rapprochement avec l'institut Pasteur (V. Meas-Yedid Hardy) pour des applications en imagerie médicale et histopathologique.

Profil du candidat

Le candidat doit avoir de très bonnes connaissances dans les domaines de la reconnaissance des formes et de la vision artificielle. Il doit aussi avoir un excellent niveau en programmation (par exemple en C, C++ et Python). Une très bonne aptitude à la communication (orale, écrite) en anglais est également attendue.

Publications de référence

[Clé17] Michaël Clément, Adrien Poulenard, Camille Kurtz, Laurent Wendling: Directional Enlacement Histograms for the Description of Complex Spatial Configurations between Objects. IEEE Trans. Pattern Anal. Mach. Intell. 39(12): 2366-2380 (2017)

[Clé18] Michaël Clément, Camille Kurtz, Laurent Wendling: Learning spatial relations and shapes for structural object description and scene recognition. Pattern Recognition 84: 197-210 (2018)

[Dai17] B. Dai, Y. Zhang, and D. Lin, "Detecting Visual Relationships with Deep Relational Networks," in CVPR, pp. 3298–3308, 2017.

[Del21] Robin Deléarde, Camille Kurtz, Philippe Dejean, Laurent Wendling: Segment My Object: A Pipeline to Extract Segmented Objects in Images based on Labels or Bounding Boxes. VISIGRAPP (5: VISAPP) 2021: 618-625

[Del22] Robin Deléarde, Camille Kurtz, Laurent Wendling: Description and recognition of complex spatial configurations of object pairs with Force Banner 2D features. Pattern Recognit. 123: 108410 (2022)

[Dub22] Shiv Ram Dubey: A Decade Survey of Content Based Image Retrieval Using Deep Learning. IEEE Trans. Circuits Syst. Video Technol. 32(5): 2687-2704 (2022)

[Goy20] Ankit Goyal, Kaiyu Yang, Dawei Yang, Jia Deng: Rel3D: A Minimally Contrastive Benchmark for Grounding Spatial Relations in 3D. NeurIPS, pp. 3258–3267, 2020

[Kri17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, Li Fei-Fei: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. Int. J. Comput. Vis. 123(1): 32-73 (2017)

[Mat99] Pascal Matsakis, Laurent Wendling: A New Way to Represent the Relative Position between Areal Objects. IEEE Trans. Pattern Anal. Mach. Intell. 21(7): 634-643 (1999)

[Pey19] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, “Weakly-Supervised Learning of Visual Relations,” in ICCV, pp. 5189–5198, 2019

[Ram22] E. Ramzi, N. Audebert, N. Thome, C. Rambour, and X. Bitot, “Hierarchical average precision training for pertinent image retrieval,” in ECCV, pp. 250–266, 2022.

[Riv22] Mateus Riva, Raisonnement relationnel spatial en apprentissage : apprentissage profond et groupement de graphes, Thèse de doctorat, Telecom Paris, 2022

[Wan23] Yang Wang, Huilin Peng, Yiwei Xiong, Haitao Song, Spatial relationship recognition via heterogeneous representation: A review, Neurocomputing, Volume 533, Pages 116-140 (2023)

[Yan19] Kaiyu Yang, Olga Russakovsky, Jia Deng: SpatialSense: An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition. ICCV 2019: 2051-2060