Compte-Rendu de l'atelier ISIS-MaDICS sur l'indexation video

23 juin 2017
Marseille

Présents :
Maxence Ahlouche ; Alexandre Benoit ; Vincent Claveau ; Hervé Le Borgne ; Yi Ren ; Hedi Ben Younes ; Jenny Benoit-Pineau ; Patrick Lambert ; Bernard Merialdo ; Georges Quénot ; Vedran Vukotic ; Renaud Peteri ; Omar Jaafor

Inscrits absents :
Daniel Carvalho Aguiar ; Réda Khouani ; Boukaye Boubacar Traore ; Mathieu Cord ; Yossef Tamaazousti

TL;DR
Cet atelier organisé conjointement par les GdR ISIS et MaDICS avait pour objet de réunir les chercheurs s'intéressant à l'indexation des images, vidéo ou plus largement de données multimédias. Les sept exposés ont illustré la grande convergence de méthodes que connaît ce domaine : les réseaux de neurones sont désormais l'état-de-l'art.
Mais chacun s'est attaché montrer comment améliorer ou tirer parti au mieux des ces modèles :

- en leur adjoignant des descripteurs plus traditionnels,
- en améliorant leur ré-utilisation et diminuant le coût d'annotation de données,
- en réduisant le nombre de paramètres à apprendre
- en les utilisant pour des tâches non-supervisées

Plusieurs tâches et contextes d'utilisation ont également été présentés :

- indexation sémantique de video, recherche de vidéo
- question-réponse sur des images
- analyse de propriétés bactériennes
- construction de liens multi-modaux dans des collections multimédias

L'atelier s'est terminé par une discussion sur les possibilité de reconduire les collaborations autour de ces thématiques, au sein du GdR ISIS, et des partenariats possibles avec le GdR MaDICS.

Introduction - G. Quénot, V. Claveau

Sont rappelés l'historique du GT IRIM au sein du GdR ISIS et notamment les participations conjointes, le partage de données, descripteurs et méthodes entre les participants pour les tâches de la conférence-compétition TRECVid.
Il est mis en avant la pertinence de la co-labellisation de cet atelier par les deux GdR : données communes, méthodes communes (notamment autour du *deep learning*).

Présentation de TRECvid - G. Quénot

Rappel historique de TRECvid.
Le coordinateur du consoritum IRIM qui a réunit les laboratoires franaçais travaillant dans le domaine d'indexation Multimédia a présenté l'historique du challenge international TRECVID et s'est arrété notamment sur les tâches Semantic indexing et Adhoc video search (AVS).
Les tâches restent difficiles (les précisions sont faibles, voire très faibles pour la tâche AVS), avec de grandes différences selon les concepts à repérer, mais les systèmes progressent. L'arrivée du deep-learning en 2013 a produit un saut qualitatif notable, et désormais la plupart des systèmes reposent dessus. Les systèmes reposent aussi beaucoup sur des ressources externes (banques de données annotées, réseaux pré-entraînés).
Le groupe IRIM a travaillé sur différents aspects des systèmes (descripteurs, fusion, classification, re-ranking…) menant chaque fois à des publications.

MuCaLe-Net - Hervé Le Borgne

Problème attaqué : diminuer le coût d'annotation tout en assurant une universalité des représentations. Une des façons de faire est le transfer learning, c'est-à-dire réexploiter des données annotées. Relabellisation des images avec des classes plus générales (cardinal -> oiseau) puis apprentissage d'un réseau par niveau. Chaque réseau apprend des filtres différents, ce qui se vérifie en regardant la cross-corrélation entre couche de même niveau.

Multi-modal Tucker for visual question answering - Hedi Ben Younes

Visual question answering . Pour le problème principale qu'est celui de la représentation multimodale du texte (question) et de l'image. Pour cela, un modèle bilinéaire est utilisé, mais pour éviter d'apprendre trop de paramètres, une décomposition de Tucker est utilisée. En ajoutant des contraintes supplémentaires (rangs sur des tranches de torseur), le problème devient facilement apprenable. Les expériences sur le corpus VQA montrent que l'approche permet des résultats légèrement supérieurs à l'existant, tout en ayant moins de paramètres à apprendre.

Participation d'EURECOM à TRECvid - Bernard Merialdo

EURECOM a participé à la tâche d'Ad=hoc Video Search (AVS). C'est une tâche non supervisée (pas de données avec vérité terrain), en vocabulaire ouvert. Là encore, le problème se pose en terme multimodal : les requêtes sont en langage naturel, les documents sont des segments vidéos. Deux stratégies sont étudiées : soit tout est en mode texte (les vidéos sont transformées en texte avec des réseaux existants, comme NeuralTalk), soit tout en mode image (les textes sont transformés en image par exemple en utilisant Google image).

Participation à TRECvid Instant search (INS) - Jenny Benois-Pineau, Alexandre Benoit

Cette tâche de reconnaissance de la personne, objet personne, ou personne spécifique dans des lieux spécifiques, se décline en différentes sous-tâches selon que l'on utilise ou non les exemples videos. Le consortium IRIM a participé à cette tâche depuis sont lancement, c'est à dire 4 ans. Pour la reconnaissance de lieux, l'approche adoptée par le Labri repose sur une représentation traditionnelle avec les *Bags of Visual Words* . Une seconde approche du LISTIC repose sur GoogleNet entraîné sur MITplaces et des métriques de similarités définies spécifiquement a aussi été testée.
Le CEA-LIST et le LIMSI ont travaillé sur la reconnaissance de visages. Les differentes contributions ont été testées individuellement et fusionnées par le Labri. La particularité de participation cette année consistait en tests 'grandeur réelle" sur les données de l'année 2016 afin de sélectionner des combinaisons les plus prometteuses pour la soumission

Sur cette tâche, les approches sac de mots sont encore employées par la plupart des équipes avec quelques succès.

Analyse de micro- et nano-structures d'iridescence de bactérie marine - Renaud Péteri

L'iridescence étant causée par la structure microscopique d'un matériau, il doit être possible de reconnaître ce phénomène à partir d'image de ce matériau. Dans le cas de certaines bactérie, c'est la colonie qui s'organise de manière à être iridescente. Avec des outils de traitement d'image classique (morpho-mathématique, triangulation), il est possible d'extraire la structure de la colonie d'image MET. Des premières expérimentations ont montré l'intérêt de ces premiers résultats.
Le passage à des images issues de microscopie à fluorescence pose cependant de nouveaux problèmes rendant les méthodes précédentes non performantes. L'utilisation de CNN pré-appris (Google Inception V3) permet néanmoins de classer les images (ou des portions d'images) avec une grande précision.

Cross-modal approaches for video Hyperlinking - Vedran Vukotic

Cette tâche d'hyperlinking consiste à détecter des liens entre segments vidéos dès lors qu'ils partagent un éléments de similarité (pas simplement visuelle mais conceptuelle). La vidéo et la parole de la bande-son, transcrite, sont utilisées conjointement. Le texte est représenté par des embeddings agrégés (c'est la moyenne qui marche le mieux). L'image par des couches tirées de réseaux pré-entraînés. La multimodalité est gérée en apprenant des réseaux ; plusieurs possibilité ont été étudiés, et la plus performante est de générer une modalité à partir de l'autre et vice versa. C'est cette approche qui a permis de remporter la tâche cette année. L'utilisation de GANs permet également d'obtenir des représentation multimodales et de visualiser une modalité à partir de l'autre.

Discussion

Quel avenir pour l'action IRIM de l'axe 4 " Masses de données Images et Vidéo" du thème B ?

J. Benois-Pineau a mis en avant la proposition d'un nouvelle action MAIM (méthodes et apprentissage en multimedia), au sein de l'Axe 4 du thème B du GdR ISIS.

Le coeur de l'action « Méthodes et Apprentissage en Indexation Multimédia » porte sur la fusion et la combinaison des traitements de diverses modalités avec la modalité visuelle. Elle s'interessera aux liens en termes de traitement et d'indexation par apprentissage statistique et plus particulièrement par apprentissage profond. Les défis propres à cette problématique concernent notamment l'hétérogénéité des approches en fonction des modalités, les différences de performances obtenues selon les médias. Dans le contexte d'apprentissage profond les modes de fusion optimale restent à étudier afin d'obtenir des scores suffisamment élevés pour l'utilisation des outils dans des problèmes réels avec une quantité de données d'apprentissage souvent faible et des annotations bruitées.

Les sujets de recherche à traiter portent sur trois aspects :

1. Combinaison image, son, texte
2. Combinaison visuelle + capteurs, IoT…
3. Modalité visuelle multi-sources

G. Quénot fait remarquer que les tâches de TRECVid nécessite maintenant des connaissances multimodale, notamment sur le texte.

V. Claveau rebondit en indiquant que la participation à des tâches communes n'est peut-être pas aussi structurant que cela l'avait été à l'époque pour IRIM.

J. Benoit-Pineau répond que le partage de ressources, modèles, etc., reste encore important pour la communauté visée par MAIM. En effet l'action MAIM est proposée comme une évolution directe de l'IRIM. Forte de l'expérience de participation aux défits communs, ayant appréhendé les nouveuax outils d'apprentissage et classification supervisée, la communauté française en indexation multimédia a besoin d'entretineir des liens forts entre les équipes.

En effet, la participation à l'IRIM a permis d'écrire ensemble les livres chez l'éditeur prestigieux Springer? De monter des projets ANR, de se faire connaître et avoir une visibilité su rle plan de participation aux comités de programme des congrès majeurs tel ACM Multimédia, ACM ICMR, d'organiser des ateliers spécifiques aux congrès de la communauté de vision par ordinateur ECCV'2012, d'impliquer des nouvelles forces de recherche, en termes de jeunes chercheurs.