Bouclage de Pertinence Cross-Modal Image et Texte

Stage de fin d'études d'Ingénieur ou de Master 2 – Année académique 2023-2024

Mots clés : Deep Learning, Active Learning, Relevance feedback, Cross-Modal Learning.
Encadrant : Dr. Olivier Buisson de l'INA (Institut National de l'Audiovisuel).
Chercheur associé au projet de Recherche : Dr. Alexis Joly de l'Inria (Institut national de recherche en informatique).

Contexte

La mission principale de l'Institut National de l'Audiovisuel (INA) consiste à sauvegarder et promouvoir le patrimoine audiovisuel français à travers la vente d'archives et la gestion du dépôt légal. Ce dernier permet aux chercheurs en sciences sociales, notamment dans le cadre de projets de recherche pluri-disciplinaires, de travailler sur les collections de l'INA. À ce titre, l'INA capte en continu 180 chaînes de télévision et radio et stocke plus de 25 millions d'heures de contenu audiovisuel. L'INA assure également des missions de formation, de production et de recherche scientifique. Le service de la Recherche de l'INA mène depuis plus de 20 ans des travaux de recherche dans le domaine de l'indexation et de la description automatique de ces fonds selon l'ensemble des modalités : textes, sons et images. Le service accueille des stages de Master et d'ingénieur ainsi que des doctorants en co-encadrement avec des laboratoires nationaux d'excellence. D'autres sujets de stage sont également proposés dans l'équipe : https://www.ina.fr/institut-national-audiovisuel/equipe-recherche/stages.

Sujet

L'accroissement du nombre de programmes audiovisuels à archiver au sein de l'INA impose de nouvelles contraintes d'accès. Pour satisfaire les besoins des différents types d'utilisateurs (documentalistes, clients, chercheurs, journalistes, le grand public, …), il devient essentiel de fournir des outils de recherche et d'exploration à très large échelle (sur des centaines de milliers d'heures de vidéo).
L'Apprentissage Profond (Deep Learning) permet de décrire efficacement les contenus visuels et sonores par l'extraction de représentations vectorielles. Ces représentations vectorielles peuvent être ensuite compressées et indexées afin d'être analysées et recherchées plus rapidement. La recherche par similarité en particulier permet de retrouver des objets visuels ou sonores à l'échelle de très grands corpus et d'offrir de nouvelles formes d'accès aux utilisateurs. Dans le domaine visuel, les derniers réseaux appris en Self-Supervised, comme DinoV2 [Oquab2023], offrent des performances impressionnantes en termes de qualité de résultats. Au sein des archives de très grandes tailles et spécialisées comme celles de l'INA, ces modèles restent cependant insuffisamment précis et les résultats d'une recherche par similarité sont encore trop bruités et/ou ambiguës.
Nos utilisateurs expriment ainsi le souhait de pouvoir définir beaucoup plus finement les objets ou concepts qu'ils recherchent, comme par exemple une sous-classe d'un type d'objet, ou encore le style d'images ou vidéos contenant l'objet recherché.
Pour permettre aux utilisateurs d'exprimer finement leurs requêtes et de retrouver efficacement des centaines de résultats, il est essentiel qu'ils puissent interagir avec le système de reconnaissance et que celui-ci soit suffisamment réactif à très large échelle.

Au cours de ces quatre dernières années, pour modéliser les requêtes de l'utilisateur, l'INA et l'Inria mènent donc des travaux couplant processus d'Apprentissage Actif [Settles2010] et Incrémentaux du type Bouclage de Pertinence [Li2013, Wu2022, Wu2022b] basés sur des représentations visuelles d'images provenant de réseaux de neurones, exemple : Vision Transformer [Dosovitskiy2021, Oquab2023] voire Clip [Radforda2021].
Le système développé fonctionne de la manière suivante :

- Après une première recherche dans un corpus d'images par une requête textuelle, l'utilisateur sélectionne positivement des résultats (les contenus pertinents selon la recherche courante pour l'utilisateur) et d'autres de manière négative.
- A partir de ces exemples étiquetés, un système d'identification peut créer par apprentissage un modèle plus riche que la requête initiale.
- Puis à partir de ce nouveau modèle, une nouvelle recherche est lancée et l'utilisateur peut à nouveau sélectionner des exemples d'apprentissage.

Le système de l'INA et de l'Inria de Bouclage de Pertinence permet donc très rapidement d'affiner et d'aider à trier ces très grands ensembles de résultats.

Pour le moment, ce système exploite des réseaux du type Clip [Radforda2021, Mehdi2022, Carlsson2022] afin que l'utilisateur puisse initier sa recherche par une requête textuelle. Puis, la modélisation durant le Bouclage de Pertinence est réalisée seulement avec les représentations visuelles des images indexées.
La modélisation n'exploite pas l'information textuelle des annotations fournies par les documentalistes qui sont associées à certaines des images que nous indexons avec les réseaux de neurones visuels.
A partir d'un modèle textuel du type Clip [Radforda2021, Mehdi2022, Carlsson2022], il serait possible d'indexer, dans le même espace vectoriel que les extractions visuelles, les annotations textuelles associées aux images fournies par les documentalistes. Il y aurait donc deux types de modalités associés à une image dans le même espace vectoriel : l'information visuelle et l'information textuelle provenant des annotations des images.

La question principale abordée dans ce stage est la suivante : L'intégration de l'information textuelle couplée à l'information visuelle, lors de l'apprentissage des modèles de Bouclage de Pertinence, peut-elle améliorer la qualité des résultats voire accélérer la convergence des recherches interactives des utilisateurs ?
Il en découle les trois questions suivantes :

- Comment réaliser l'apprentissage Cross-Modal des modèles de Bouclage de Pertinence ?
- Peut-on développer des méthodes d'Apprentissage Actif spécifiques au Cross-Modal ?
- Comment évaluer et avec quels corpus ces méthodes ?

Le but de ce stage est d'initier un nouveau thème de R&D qui devrait durer 3 à 4 ans. Durant cette période, nous envisageons d'étendre cette problématique de Bouclage de Pertinence Cross-Modal à la proposition interactive pour la reformulation de requêtes textuelles et d'annotations associées aux classes d'images créées par les utilisateurs. Pour cette problématique, nous envisageons d'exploiter les méthodes de Image Captioning [Xul2023] et les LLM.

Encadrement et contexte

L'encadrement de ce stage sera assuré par Olivier Buisson (https://scholar.google.fr/citations?user=rWunhTEAAAAJ&hl=fr). Le stagiaire sera amené à participer à la collaboration entre Olivier Buisson et Alexis Joly (HDR, Inria, https://scholar.google.fr/citations?user=kbpkTGgAAAAJ&hl=fr&oi=ao). Ces travaux de R&D s'inscrit dans la continuité de plus de 10 ans de collaboration entre l'Inria et l'INA. Quatre thèses CIFRE ont notamment été encadrées ou sont en cours depuis 2013 sous leur co-supervision. Par ailleurs, une plateforme de R&D nommée Snoop a été co-développée entre l'INA et l'Inria. Celle-ci est en cours d'expérimentation et de déploiement au sein de l'INA, et est aussi utilisée pour l'application de reconnaissance des plantes Pl@ntNet (http://identify.plantnet-project.org). Les acteurs institutionnels de ce thème de R&D, l'équipe Zénith de l'Inria et l'INA ont une expérience solide dans l'analyse de données multimédia et le passage à l'échelle et apporteront des compétences complémentaires sur le sujet. Les travaux de Zenith s'articulent autour de la gestion, l'analyse et de la recherche d'informations dans des données hétérogènes de très grandes tailles. Au sein de l'INA, le stagiaire rejoindra le service de la Recherche qui s'intéresse aux sujets de recherche en lien avec l'archivage audiovisuel. Le stage se déroulera sur une période de 4 à 6 mois, au sein du service de la Recherche de l'INA. Il aura lieu sur le site Bry 2, situé au 28 Avenue des frères Lumière, 94360 Bry-sur-Marne. Un ordinateur sera fourni ainsi qu'un accès au cluster de calcul GPU de l'Institut. Télétravail possible une journée par semaine.

Candidature

Envoyez par email et en PDF à l'adresse thcand@ina.fr, les documents suivants : CV et relevés de notes + liste des enseignements suivis en M2 et en M1 ou en école d'ingénieurs.

Profil recherché

- Étudiant en dernière année d'un bac+5 dans le domaine de l'informatique et de l'IA.
- Forte appétence pour la Recherche et la création technologique.
- Connaissance profonde du Deep Learning image et text.
- Maîtrise du développement en Python, et aussi le C++ serait un plus.

Bibliographie

[Carlsson2022] Carlsson F. et al., Cross-lingual and Multilingual CLIP, LREC, https://aclanthology.org/2022.lrec-1.739, 2022.
[Dosovitskiy2021] Dosovitskiy A. et al, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021, https://arxiv.org/abs/2010.11929.
[Li2013] Li et al., Relevance feedback in content-based image retrieval: a survey, Handbook on neural information processing, 2013.
[Mehdi2022] Mehdi C. et al., Reproducible scaling laws for contrastive language-image learning, https://arxiv.org/abs/2212.07143, 2022.
[Oquab2023] Oquab M. et al., DINOv2: Learning Robust Visual Features without Supervision, https://arxiv.org/abs/2304.07193, 2023.
[Radford2021], Radford Alec et al., Learning Transferable Visual Models From Natural Language Supervision, https://arxiv.org/pdf/2103.00020.pdf, 2021.
[Settles2010] Settles B., Active learning literature survey, University of Wisconsin, Madison, 2010.
[Xu2023] Xu, L. et al., Deep Image Captioning: A Review of Methods, Trends and Future Challenges, Neurocomputing, 2023.
[Wu2022] Wu X. et al, A SURVEY OF HUMAN-IN-THE-LOOP FOR MACHINE LEARNING, Future Generation Computer Systems, 2022, https://arxiv.org/abs/2108.00941.
[Wu2022b] Wu M. et al, Deep Active Learning for Computer Vision Tasks: Methodologies, Applications, and Challenges, Applied Sciences, 2022.