Espace multimodal pour la génération et la justification de liens sémantiques entre documents

Mots-clés : apprentissage profond ; apprentissage de représentation ; explicabilité en IA ; traitement automatique des langues ; multimédia ; recherche d'information multimodale

La thèse vise à proposer des techniques d'apprentissage d'espaces sémantiques multimodaux permettant d'établir des liens intra et inter-documents au sein de collections hétérogènes (textes, images et vidéos) à des fins de recherche et d'exploration d'archives. Elle s'attachera notamment à proposer des espaces permettant de fournir à un utilisateur une explication de la relation établie.

Dans un premier temps, on s'intéressera à l'apprentissage de représentations multimodales. On cherchera notamment à étendre l'approche proposée par Vukotić *et al*. (2018) en développant des techniques adaptées à la tâche d'exploration d'une collection hétérogène. Seront mises en œuvre des approches *end-to-end* permettant d'apprendre directement une mise en relation combinant plusieurs modalités (e.g., Nguyen *et al*., 2017). Ces approches combineront directement apprentissage des représentations des modalités, de la représentation multimodale et de la métrique dans ce dernier espace pour prédire l'existence d'un lien et sa pertinence, indépendamment des modalités mises en relation. Ces approches *end-to-end* permettront notamment l'adaptation du modèle, et donc des liens proposés pour explorer une collection, en fonction des intérêts de l'utilisateur mesurés à travers son parcours de navigation.

Dans un second temps, on s'attachera à favoriser l'explicabilité des liens créés. On cherchera donc à étendre les modèles *end-to-end* proposés pour comprendre la relation entre images, vidéos et textes, à l'instar des travaux récents en *image captioning* ou en *visual query answering (e.g.,* Xu et al. 2015*)*. On étudiera pour cela des mécanismes d'attention mettant en évidence les éléments qui justifient la mise en relation (Luong *et al*., 2015 ; Vaswani *et al*., 2017). Un verrou scientifique qu'il faudra lever pour cela est la combinaison de modèles de surface et de modèles syntaxiques dans une approche neuronale pour pouvoir focaliser l'attention sur le bon niveau du texte. On cherchera notamment à étendre les modèles d'attention classiques, applicables aux représentations de surface, de manière à intégrer des informations linguistiques à plusieurs niveaux (syntaxe, sémantique).

La thèse s'inscrit au sein de l'équipe LINKMEDIA de l'IRISA dans le contexte du projet collaboratif « Compréhension automatique multimodale du langage pour de nouvelles interfaces intelligentes de médiation et de transmission des savoirs » (ANR ARCHIVAL) impliquant Orange Labs, la Fédération des Maisons des Sciences de l'Homme (FMSH) et le Laboratoire Informatique & Systèmes (Aix-Marseille Univ.). L'équipe LINKMEDIA regroupe des chercheurs en analyse automatique de contenus multimédias, en traitement automatique des langues, en apprentissage de représentations multimodales et en recherche d'information multimédia.

Le candidat devra posséder de solides connaissances dans le domaine de l'apprentissage profond et avoir une appétence pour le traitement automatique des langues.

Contacts

Guillaume Gravier guillaume.gravier@irisa.fr
Pascale Sébillot pascale.sebillot@irisa.fr

Références

Phuong Anh Nguyen et al. *Vireo@ trecvid 2017: Video-to-text, ad-hoc video search and video hyperlinking*. In Proc. TRECVID Workshop, 2017.

Vukotić, V., Raymond, C., Gravier, G., A *Crossmodal Approach to Multimodal Fusion in Video Hyperlinking*. IEEE Multimedia 25(2):11-23, 2018.

K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. *Show, attend and tell: Neural image caption generation with visual attention*. In Proc. Intl. Conf. on machine learning, 2015.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. *Attention is all you need*. In Proc. Advances in Neural Information Processing Systems, 2017.

T. Luong, H. Pham, and C. D. Manning. *Effective approaches to attention-based neural machine translation*. In Proc. Intl. Conf. on Empirical Methods in Natural Language Processing, 2015.