Description vidéo par enrichissement textuel et sémantique

Mots-clés: description de vidéo, apprentissage profond, réseau neuronal convolutif, résolution de coréférences, graphe de connaissances, multi-tâches, multimodalité

Contexte et motivation

Les méthodes de description de vidéo visent à générer automatiquement des descriptions textuelles à partir de séquences / clips vidéo. C'est un problème bien connu dans le domaine de la vision par ordinateur ou de nombreuses méthodes ont été proposées dans la littérature [1]. Ce sujet est en lien avec l'évaluation de performance et en particulier les bases de données. De nombreuses bases de données ont été proposées à des fins d'évaluation de performance des méthodes de description de vidéo [2]. Les deux les plus utilisées sont les bases MSVD [3] et MSR-VTT [4]. Ces bases sont le plus généralement constituées à partir de vidéo web (e.g. Youtube) sur des catégories ciblées (e.g. cuisine, e-commerce, réseaux sociaux) ou en multi-catégorie. Les annotations textuelles y sont le plus souvent produites à la main via des services Web comme Amazon Mechanical Turk.

Néanmoins, les bases de données proposées dans la littérature présentent de sévères limitations [2]. L'encodage / qualité des vidéos y est hétérogène et les vidéos sont fournies sans multimodalité (sans ou avec peu de pistes audio, sous-titres et métadonnées). Elles ne sont pas à l'échelle et les jeux de test y sont fournis en mode boîte noire. Finalement, les descriptions y sont données a posteriori (sans temporalité) et se cantonnent à des informations visuelles sans contextualisation (e.g. quelle personne, quel lieu, quelle date, etc.). Comme évoqué en [1, 2], la constitution de bases de données à l'échelle, standardisées, avec information textuelle (pistes audio, sous-titres et métadonnées), structurées en jeux de tests et offrant une large vérité terrain est indispensable pour de futures recherches sur les méthodes de description de vidéo.

Ces limitations peuvent être en partie levées par recours à la captation télévisuelle (TV). En effet les flux vidéo TV, contrairement aux flux vidéo Web, permettent une capture standardisée, à l'échelle et avec information textuelle. Différents travaux ont été menés par le passé pour la constitution de bases de données TV pour la détection de segments vidéo [5], le journalisme des données [6] et le traitement automatique du langage naturel [7]. Néanmoins, le problème de génération automatique et fiable de description de vidéo, répondant aux exigences d'élaboration d'une vérité terrain, reste entier. Même si les méthodes de description de vidéo ont fortement gagné en maturité (et en particulier celles à base d'apprentissage profond [1]), elles ne peuvent ni garantir le niveau de robustesse nécessaire à l'élaboration d'une vérité terrain, ni franchir (sur la seule base d'une analyse visuelle) le fossé sémantique nécessaire à la contextualisation.

Objectif de la thèse

De façon à répondre à ce problème, nous proposons dans ce sujet de thèse d'explorer les approches à base de Traitement Automatique du langage naturel (TALN) pour la représentation sémantique et d'apprentissage profond afin de proposer un nouveau système pour la description contextuelle de vidéo. L'idée est de mettre en correspondance les données textuelles contenues dans les guides des programmes et les graphes de connaissances du Web des données, puis de lier ces données aux séquences vidéo qui leur sont associées grâce aux sous-titrages et leur horodatage. Au niveau textuel, le lien entre les différentes modalités soulève plusieurs défis scientifiques comme la résolution de coréférences [8, 9] entre différentes natures de textes (écrit, transcription orale) et le traitement multilingue [10]. Sur l'analyse vidéo, un

traitement multi-tâche [11, 12] et multimodale [13, 14] devra être engagé à des fins d'évaluation de performance [5, 6] pour la génération de la base de données et de la vérité terrain.

Ce travail se situe donc à la croisée entre l'apprentissage profond sur les vidéos et le TALN en s'appuyant sur les graphes de connaissances. Les principales contributions attendues sont :

- Proposition d'une méthode de génération d'un graphe de connaissances à partir du traitement des audiodescriptions
- Proposition d'un modèle d'apprentissage profond sur des descriptions vidéo intégrant un graphe de connaissances pour la génération des résumés
- Production d'une base de données de descriptions vidéo

Contexte et encadrement

L'étudiant(e) sera accueilli(e) au sein du Laboratoire d'Informatique Fondamentale et Appliquée de Tours (LIFAT) de l'Université de Tours (UT), cette direction de recherches prolongeant et renforçant des collaborations déjà en cours entre l'équipe BDTLN et l'équipe RFAI. Les travaux seront dirigés par Arnaud Soulet[1] (MCF-HDR), Donatello Conte (PR) et co-encadrés par Nathalie Friburger[2] (MCF), Mathieu Delalandre[3] (MCF).

Références

[1] S. Li and al. Visual to Text: Survey of Image and Video Captioning. Transactions on Emerging Topics in Computational Intelligence, vol. 3(4), pp. 297-311, 2019.
[2] M. Rafiq and al. Video Description: Datasets & Evaluation Metrics. IEEE Access, vol. 9, 2021.
[3] D.L. Chen and W.B. Dolan. Collecting highly parallel data for paraphrase evaluation. Annual Meeting of the Association for Computational Linguistics: Human Language Technologie, pp. 190-200, vol. 1, 2011.
[4] J. Xu and al. MSR-VTT: A large video description dataset for bridging video and language. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5288–5296, 2016.
[5] V.H. Le and al. A large-Scale TV Dataset for partial video copy detection. International Conference on Image Analysis and Processing (ICIAP), Lecture Notes in Computer Science (LNCS), vol 13233, pp. 388-399, 2022.
[6] F. Rayar and al. A large-scale TV video and metadata database for French political content analysis and fact-checking. Conference on Content-Based Multimedia Indexing (CBMI), 2022.
[7] P. Lison and J. Tiedemann. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. Language Resources and Evaluation Conference (LREC), 2016.
[8] R. Sukthanker, S. Poria, E. Cambria, R. Thirunavukarasu, Anaphora and coreference resolution: A review, Information Fusion, Volume 59, 2020, Pages 139-162, ISSN 1566-2535
[9] V. Ramanathan, A. Joulin, P. Liang and L. Fei-Fei, Linking People in Videos with ``Their'' Names Using Coreference Resolution, in Computer Vision -- ECCV 2014, 2014, Springer International Publishing, pp. 95-110
[10] Oliveira, I.L., Fileto, R., Speck, R., Garcia, L.P., Moussallem, D. and Lehmann, J., 2021. Towards holistic entity linking: Survey and directions. Information Systems, 95, p.101624.
[11] Z. Liu and al. Multi-Task Video Captioning with a Stepwise Multimodal Encoder. Electronics, vol. 11(17), pp. 2639, 2022.
[12] S. Chen and al. Video Captioning with Guidance of Multimodal Latent Topics. International Conference on Multimedia (MM), pp. 1838–1846, 2017.
[13] D. Ramachandram and G. W. Taylor. Deep Multimodal Learning: A Survey on Recent Advances and Trends. Signal Processing Magazine, vol. 34 (6), pp. 96-108, 2017.
[14] J. Summaira and al. Recent Advances and Trends in Multimodal Deep Learning: A Review. arXiv.2105.11087, 2021.

[1] https://www.info.univ-tours.fr/~soulet/
[2] https://www.info.univ-tours.fr/~friburger/
[3] http://mathieu.delalandre.free.fr/