Institut de Recherche en Informatique de Toulouse

Stage 2019-2020 - Master 2 ou élève Ingénieur 3ème année

Prédiction de la difficulté de compréhension de contenu audiovisuel : approche basée sur des données textuelles faiblement annotées

Prediction of the difficulty to understand audiovisual content: approach based on weakly annotated textual data

Domaine : Analyse, indexation et compréhension de contenus audiovisuels (audio, vidéo, texte)
Thématique : Traitement automatique des langues – Analyse conversationnelle et interaction
Lieu du stage : IRIT, Université Toulouse III - Paul Sabatier, Toulouse (https://www.irit.fr)
Durée : 5 à 6 mois de stage (début février ou mars 2020)
Contacts :
Equipe SAMoVA (https://www.irit.fr/recherches/SAMOVA/)
Isabelle Ferrané (isabelle.ferrane@irit.fr) - 05.61.55.60.55
Equipe MELODI (https://www.irit.fr/-Equipe-MELODI-)
Tim Van de Cruys (tim.vandecruys@irit.fr) – 05.61.55.77.13

Contexte : L'exploitation avancée de grands volumes de **documents audiovisuels** passe par la compréhension de leur contenu. L'analyse automatique de ces contenus peut être réalisée sous plusieurs angles, en fonction des modalités considérées.

- L'analyse de la **composante audio** permet d'extraire des informations (descripteurs audio) concernant l'environnement sonore (zones de musique, de parole ou de bruits environnants, …), les locuteurs et les tours de parole (Vallet et al., 2012).
- L'analyse de la **composante vidéo** permet d'extraire des informations (descripteurs visuels) concernant le cadre (intérieur, extérieur, nuit, jour, …) ou les intervenants (foule, personne présente en premier plan ou groupe de plusieurs personnes, …) (Bost et al., 2015).
- L'analyse de la **composante textuelle**, à travers les sous-titres ou bien les transcriptions automatiques à disposition, permet d'extraire des informations sémantiques (descripteurs texte) qui permettent d'enrichir la caractérisation du contenu basée sur les modalités audio et vidéo (Lison and Tiedemann, 2016).

Objectif : Dans ce stage, on cherche de caractériser les contenus de films selon leur niveau de **difficulté de compréhension.** Vue que le niveau de difficulté de compréhension est principalement lié à la composante linguistique, on explorera les possibilités offertes par le domaine du **traitement automatique des langues** pour extraire les informations pertinentes, qui pourraient donner des indications sur la tâche envisagée.

Sujet de stage : L'objectif de ce stage est de prédire de manière automatique la difficulté de compréhension de séquences vidéo à travers leurs sous-titres ou transcriptions. Dans ce but, on appliquera des méthodes supervisées basées sur les plongements de mots (« word embeddings »). Ces représentations sont généralement obtenues par apprentissage non-supervisé réalisés à partir d'un volume très important de textes, et permettent de représenter les mots sous forme vectorielle (vecteur de N dimensions à coefficients réels associé à chaque mot) afin de mieux caractériser leur sens (Mikolov et al., 2013). En les intégrant dans un modèle de réseau de neurones supervisé, il est possible de construire des représentations vectorielles pour de plus grandes sections de texte, capable de prédire les descripteurs pertinents (Joulin et al., 2017). Pour l'entraînement des plongements de mots, nous utiliserons un corpus de textes ciblé par rapport à la tâche envisagée (sous-titres de documents de fictions ou de transcriptions automatiques de vidéos issues du web). L'application d'un modèle de classification nécessite également des données labellisées. Dans ce stage, nous avons pour objectif de construire un ensemble d'entraînement fournissant un premier niveau d'annotation approximatif, c'est-à-dire correspondant à des « données faiblement annotées », à la différence d'une vérité terrain exacte et précise. Pour cela, nous explorerons le paradigme de programmation de données («data programming» ; Ratner et al., 2017)[1] . Le but est de labelliser de manière automatique une grande quantité de données par l'application de fonctions de labellisation, éventuellement bruitées ; un modèle génératif

1 https://www.snorkel.org/

effectuera alors un débruitage des données en analysant les fonctions de labellisation comme variables latentes. Ceci permet de labelliser de manière assez rapide une grande quantité de données avec une exactitude satisfaisante. Les fonctions de labellisation s'appuieront sur des traits linguistiques (par rapport au lexique, syntaxe, etc.). Ils pourront également s'appuyer sur les travaux antérieurs réalisés lors de stages précédents (Petiot, 2018 ; Berdeaux, 2019). Ce travail pourrait potentiellement se faire en collaboration avec la société Archean Labs[2] et le laboratoire commun ALAIA[3], afin de comparer différentes approches possibles.

Compétences : Ce stage s'adresse à un(e) étudiant(e) de niveau M2 ou 3ème année d'Ecole d'ingénieurs, ayant de bonnes connaissances en programmation objet (python) sous Linux. Des compétences en reconnaissance de formes et apprentissage automatique sont également attendues. La connaissance des méthodes de traitement d'images, ou traitement de l'audio sont un plus pour bien comprendre les objectifs visés à terme de fusion des descripteurs audio, vidéos et texte. Un bon niveau d'anglais est également requis pour la lecture et compréhension d'articles scientifiques en lien avec les différentes thématiques de recherche.

Références

- X Bost, G Linares, S Gueye Audiovisual speaker diarization of TV series - Acoustics, Speech and Signal Processing (ICASSP), 2015.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, pp 427–431
- Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* pp. 923–929.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781
- Jim Petiot, Exploitation de données textuelles pour la recherche de Topics et la caractérisation de contenus de fiction : approche non-supervisée et semi-supervisée. Stage M2 IARF, 2018.
- Alexandre Berdeaux, Classification supervisée de thèmes de dialogues de film en contexte de données faiblement annotées, Stage M2 IARF, 2019.
- Ratner, Alexander, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. "Snorkel: Rapid training data creation with weak supervision." *Proceedings of the VLDB Endowment* 11, no. 3 (2017): 269-282.
- Félicien Vallet, Slim Essid, Jean Carrive, Gaël Richard, High-Level TV Talk Show Structuring Centered on Speakers' Interventions, TV Content Analysis: Techniques and Applications, Edited by Shiguo Lian Auerbach Publications 2012.

[2] http://www.archean.tech/archean-labs-en.html

[3] https://www.actuia.com/actualite/focus-sur-le-labcom-alaia-et-son-programme-de-recherche-dans-le-domaine-des-technologies-informatiques-pour-lapprentissage-des-langues-etrangeres/