Apprentissage profond non supervisé de représentations spatio-temporelles pour la vidéo

Sujet

Identifier des actions ou une succession d'événements en fonction de l'expérience est une partie importante du processus de prise de décision humaine. Simuler ce processus par des machines en leur apprenant à localiser et identifier les événements en se basant sur des représentations internes de l'environnement pourrait être utile pour de nombreuses tâches de l'analyse vidéo telles que la reconnaissance, la détection et le suivi d'objets [1].

Les récents progrès de l'apprentissage profond et l'augmentation de la puissance de calcul des GPU spécialisés permettent d'envisager des architectures répondant à cette problématique. Cependant l'apprentissage profond supervisé nécessite un volume considérable de données étiquetées afin d'obtenir des résultats pertinents. Dans le domaine de la vidéo, rares sont les acteurs qui disposent d'un tel volume de données étiquetées.

Utilisant des vidéos non-étiquetées, nous souhaitons apprendre de manière non-supervisée un réseau profond encodant des représentations pour les vidéos [2]. L'objectif est de capturer la nature spatio-temporelle des vidéos dans un modèle unique, et non de traiter indépendamment les dimensions spatiales (images) et la dimension temporelle. Au même titre que les premières couches des réseaux convolutionnels 2D encodent des descripteurs locaux spécialisés pour les images, nous souhaitons apprendre des descripteurs spatio-temporels permettant de modéliser les évènements vidéo. Une fois appris, ces descripteurs pourront être utilisés pour une tâche supervisée, telle que la reconnaissance d'actions.

Pour répondre à cette problématique, vous serez amené à étudier plusieurs modèles d'apprentissage profond, comme les auto-encodeurs variationnels (VAE) [3], les réseaux génératifs adversariaux (GAN) [4] ou encore les réseaux récurrents, comme les *long short term memory networks* (LSTM) [5] ; pour enfin en proposer de nouveaux répondant au problème.

Modalités de la thèse

Cette thèse est proposée sous la forme d'une convention CIFRE d'une durée de 3 ans au sein de la société Foxstream et du laboratoire LIRIS (équipe Imagine). Elle se déroulera du 1er octobre 2019 au 30 septembre 2022. L'équipe Imagine du laboratoire LIRIS (Laboratoire en Image et Système d'Information) a notamment pour objectif d'analyser le contenu des images de manière automatique en vue de segmenter les régions d'intérêt, d'extraire des caractéristiques par le biais de descripteurs visuels afin, par exemple, d'identifier les objets contenus dans l'image, des actions dans une vidéo et de les caractériser.

Le groupe Foxstream est composé des sociétés Foxstream, Foxstream Inc. et Cossilys21. Foxstream est une société d'édition logicielle, fondée en 2004, spécialisée dans l'analyse et le traitement automatique en temps-réel du contenu d'images vidéo. Foxstream offre des solutions capables d'extraire et de transmettre une information pertinente à partir d'un flux vidéo. Foxstream est présent essentiellement sur le marché de la sécurité (vidéosurveillance), et sur le marché de la gestion de flux (comptage, files d'attente, etc.) pour des aéroports, commerces, *etc.* Sa filiale Foxstream Inc. est basée à Miami, USA.Cossilys21 est une société de haute technologie ayant pour vocation l'innovation et la production de systèmes intelligents de vidéo protection. Depuis plus de 20 ans, Cossilys21 s'impose comme référence sur le marché de la vidéo-protection notamment dans le secteur bancaire pour lequel Cossilys21 équipe de grandes banques nationales et régionales. Cossilys21 intervient également sur de nombreux secteurs d'activités comme le *retail* ou encore l'industrie.

Profil recherché :

Nous recherchons un.e candidat.e motivé.e ayant des bases solides en informatique et mathématiques, avec une expérience en traitement d'images, vision par ordinateur, apprentissage. Étant donné que le projet s'attaque à un problème qui n'a pas encore été exploré, le doctorant aura la possibilité de participer à de grandes conférences dans le domaine de la vision par ordinateur et intelligence artificielle, comme ICCV, ECCV, CVPR, ECAI et IJCAI. Ces participations lui donneront la possibilité d'interagir avec des acteurs majeurs de l'industrie et de tisser des liens professionnels qui pourront être utilisés comme point de départ pour de futures opportunités de travail.

Compétences attendues :

Nous recherchons un.e candidat.e connaissant :

- Langage C++ et Python
- Bibliothèque OpenCV / Qt
- Outil de versioning (GIT)

Il devra également maîtriser le SE Linux et le langage BASH

La connaissance des éléments suivants serait un plus :

- Framework PyTorch et TensorFlow.
- Framework Python Anaconda
- Maîtrise de langue anglaise.

Salaire : 26k€ / an

Lieu de travail :

LIRIS - Université Lumière Lyon 2 (Bron) / Foxstream (Vaulx en Velin)

Contacts :

Carlos Crispim, carlos.crispimjunior@univ-lyon2.fr, +33(0) 4 78 77 31 15

Lionel Robinault : l.robinault@foxstream.fr

Bibliographie

[1] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In ICLR, 2016.

[2] Kiran, B., Thomas, D., &Parakkal, R. (2018). An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. Journal of Imaging, 4(2), 36.

[3] Diederik, P.K.; Max, W. Stochastic Gradient VB and the Variational Auto-Encoder. In Proceedings of the 2nd International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.

[4] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 8–13 December 2014; pp. 2672–2680.

[5] Hochreiter, S.; Schmidhuber, J. Long short-term memory. Neural Comput. 1997, 9, 1735–1780.

Unsupervised deep learning of spatio-temporal representations for video

Subject

Identifying actions or series of events depending on experience is a crucial part of human decision-making. Simulating this process with computers and teaching them to localize and identify events based on intern representations of the environment could be useful for many tasks in video analysis, such as recognition, detection and object tracking [1].

Recent progress in deep learning, as well as the increase on GPU computing power, allow us to consider architectures answering to this issue. However, supervised deep learning requires a huge volume of labelled data to obtain relevant results. In the video field, rare are the actors of this field that have such a volume of labelled data available.

Using non-labelled videos, we want to learn a deep network encoding video representation in an unsupervised way [2]. The goal is to capture the spatio-temporal nature of videos in a single model, instead of addressing independently spatial (images) and temporal dimensions as prior work. The same way the first layers of 2D convolutional networks encode local descriptors adapted to images, we want to learn spatio-temporal descriptors that allow us to model video events,. Once learned, these descriptors may be used for a supervised task, such as action recognition.

In response to these challenges, you would be brought to study several deep learning architectures, like: variational auto-encoders (VAE) [3], generative adversarial networks (GAN) [4] or recurrent networks, such as *long short-term memory networks* (LSTM) [5]; to propose new ones more adapted to the target problem.

Thesis modality

This thesis is part of a CIFRE convention for 3 years and it is expected to start on October 1st, 2019 until September 30th, 2022. It will take place at the Foxstream Company in Vaulx-en-Velin, as well as in LIRIS laboratory (Imagine team). The Imagine team of LIRIS (Laboratoire en Image et Système d'Information) has as objective to analyze the content of images in an automatic manner to segment regions of interest, to extract image features by visual descriptors with the goal of, for instance, identifying objects in an image and recognizing and describing actions.

The Foxstream Group is composed of 3 companies: Foxstream, Foxstream Inc. and Cossilys21. Foxstream is a software company, founded in 2004, specialized in real time Video Content Analysis (VCA).Foxstream solutions extract and transfer relevant information from video streams. Foxstream is mainly present in the security industry (CCTV), as well as in the flow management industry (counting, Queue Management System, etc.)for airports, shops, etc. Innovation is in Foxstream's DNA, therefore strong links have been built with the researchcommunity. Its subsidiary Foxstream Inc. is located in Miami, USA. Created more than 20 years ago, the French firm COSSILYS21 offers intelligent video-protection solutions. It equips major national banks, numerous regional banks, as well as shops. The COSSILYS21 firm is nowadays a reference in the banking sector.

Candidate Profile:

We are looking for a motivated candidate with a solid foundation in computer science and mathematics, with experience in image processing, computer vision, and deep learning. As the project tackles a problem that has not yet been explored, the Ph.D. student will have the opportunity to participate in major conferences in the field of computer vision and artificial intelligence, such as ICCV, ECCV, CVPR, ECAI, and IJCAI. These participations will give him the opportunity to interact with major players in the

industry and to build professional relationships that can be used as a starting point for future work opportunities.

Skill Requirements:

We are looking for a candidate that knows :

- C++ and Python languages
- OpenCV / Qt libraries
- Code versioning tools (GIT)
- LINUX OS
- BASH language

The following skills will be also appreciated:

- PyTorch et TensorFlow frameworks
- Python Anaconda framework
- Mastering of English language.

Salary: 26k€ / year

Workplace:

LIRIS - Université Lumière Lyon 2 (Bron) / Foxstream (Vaulx en Velin)

Contact:

Carlos Crispim, carlos.crispimjunior@univ-lyon2.fr, +33(0) 4 78 77 31 15

Lionel Robinault : l.robinault@foxstream.fr

Bibliographic references

[1] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In ICLR, 2016.

[2] Kiran, B., Thomas, D., &Parakkal, R. (2018). An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. Journal of Imaging, 4(2), 36.

[3] Diederik, P.K.; Max, W. Stochastic Gradient VB and the Variational Auto-Encoder. In Proceedings of the 2nd International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.

[4] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 8–13 December 2014; pp. 2672–2680.

[5] Hochreiter, S.; Schmidhuber, J. Long short-term memory. Neural Comput. 1997, 9, 1735–1780.