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 dans 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 génératif efficace, et non de traiter indépendamment les dimension 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 des modéliser les évènements vidéos. Une fois appris, les descripteurs de ce réseau génératif pourront être utilisés comme entrée d'un réseau discriminatif appris pour une tâche supervisée, telle que la reconnaissance d'action.

Pour répondre à cette problématique, plusieurs architectures peuvent être envisagées comme les auto-encodeurs variationnels (VAE) [3], les réseaux génératifs adversariaux (GAN) [4] ou encore les réseaux récurrents et tout particulièrement les *Long short term memory network* (LSTM) [5]. L'accent sera mis également sur l'aspect multi-échelle spatiale et/ou temporelle.

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 d'un laboratoire de recherche partenaire. Elle se déroulera du 1er octobre 2019 au 1er octobre 2022.

Foxstream :

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'activité comme le retail ou encore l'industrie.

Contact : Lionel Robinault l.robinault@foxstream.fr

Biblio

[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.