CEA List
Service d'Intelligence Artificielle pour le Langage et la Vision
Centre de Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact Jaonary Rabarisoa
Tél +33 (0)1 69 08 01 29
E-mail jaonary.rabarisoa@cea.fr

STAGE 2022 Réf : LVA-22-S3

Apport de l'attention spatio-temporelle pour la reconnaissance d'action deans une séquence vidéo

Présentation du laboratoire d'accueil

Basé à Paris-Saclay, le CEA List est l'un des quatre instituts de recherche technologique de CEA Tech, direction de la recherche technologique du CEA. Dédié aux systèmes numériques intelligents, il contribue au développement de la compétitivité des entreprises par le développement et le transfert de technologies.

L'expertise et les compétences développées par les 800 ingénieurs-chercheurs et techniciens du CEA List permettent à l'Institut d'accompagner chaque année plus de 200 entreprises françaises et étrangères sur des projets de recherche appliquée s'appuyant sur 4 programmes et 9 plateformes technologiques. 21 start-ups ont été créées depuis 2003.

Le Laboratoire de Vision et Apprentissage pour l'analyse de scène (LVA) mène ses recherches dans le domaine de la Vision par Ordinateur (Computer Vision) selon quatre axes principaux :

- La reconnaissance visuelle (détection et/ou segmentation d'objets, de personnes, de patterns ; détection d'anomalies ; caractérisation)
- L'analyse du comportement (reconnaissance de gestes, d'actions, d'activités, de comportements anormaux ou spécifiques pour des individus, un groupe, une foule)
- L'annotation intelligente (annotation à grande échelle de données visuelles 2D/3D de manière semi-automatique)
- Les modèles de perception pour l'aide à la décision.

Description du stage

La reconnaissance d'action dans une vidéo est une tâche de vision par ordinateur à la base de plusieurs applications (analyse sportif, vidéo projection, système de recommandation ...). La majorité des approches utilisent l'apprentissage profond à base d'architectures convolutives (*Convolutional Neural Networks*). Ces derniers temps les architectures à base d'attention (*transformers*) ont émergé comme une alternative performante aux CNN pour résoudre les tâches de vision. Pour l'analyse vidéo en particulier, les *transformers* offrent un moyen naturel de gérer le lien spatio-temporel entre les objets présents dans une séquence. Cependant, pour être efficace, ils nécessitent beaucoup plus de données annotées que les CNN pendant la phase d'apprentissage. L'étape de pré-apprentissage (initialisation des poids) devient primordiale pour réussir cette phase.

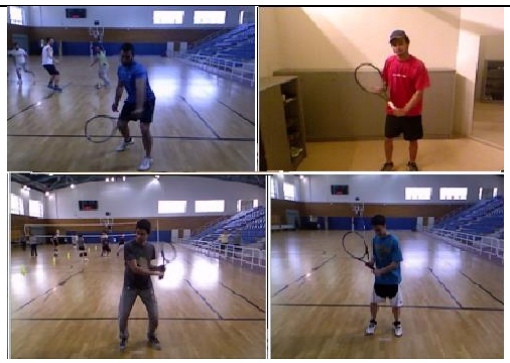
L'objectif de ce stage est d'étudier l'utilisation des réseaux de neurones à base de *transformer* pour la classification de séquence vidéo. Il s'agira principalement de répondre aux questions suivantes :

- Comment découper une séquence vidéo pour que les blocs d'attention soit les plus pertinents possible?
- Comment pré-apprendre le réseau afin de palier au manque de données annotées?

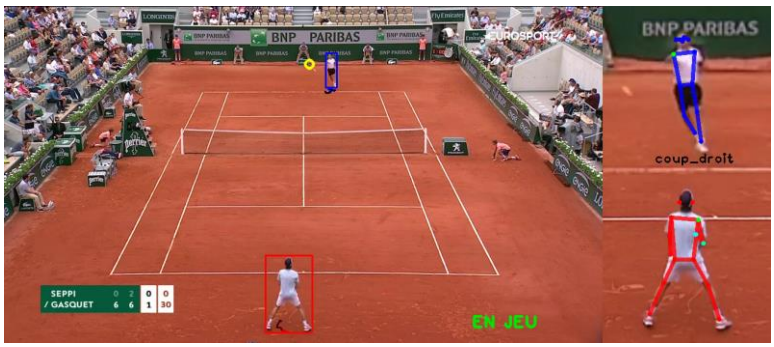
La méthode proposée sera évaluée sur des données d'analyse de geste sportif, comme le tennis par exemple. Des dataset public tel que Thetis [7] ou d'autres données interne au laboratoire pour mettre en place l'architecture basé transformer puis d'autres dataset seront utilisés pour mettre en évidence la stratégie de pré-apprentissage du réseau dans une situation d'adaptation de domaine.

CEA List
Service d'Intelligence Artificielle pour le Langage et la Vision
Centre de Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact Jaonary Rabarisoa
Tél +33 (0)1 69 08 01 29
E-mail jaonary.rabarisoa@cea.fr



Exemples du dataset video Thetis.



Exemples de reconnaissance de coup pendant un match de tennis.

Keywords

Action recognition, visual transformers, domain adaptation, sport analysis.

Références

[1] Temporal Contrastive Pretraining for Video Action Recognition. G. Lorre et al. WACV 2020.
[2] Spatiotemporal Contrastive Video Representation Learning. R. Qian et al. https://arxiv.org/abs/2008.03800
[3] An Image is Worth 16x16 Words. Transformers for Image Recognition at Scale. Alexey Dosovitskiy et al. https://arxiv.org/abs/2010.11929
[4] ViViT: A Video Vision Transformer. A. Arnab et al. https://arxiv.org/abs/2103.15691
[5] A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning. C. Feichtenhofer et al. CVPR 2021.
[6] Video Action Transformer Network. R. Girdhar et al. CVPR 2019
[7] http://thetis.image.ece.ntua.gr/

Niveau demandé :	Ingénieur, Master 2
Ce stage ouvre la possibilité de poursuite en thèse et ingénieur R&D dans notre laboratoire.	
Durée :	6 mois
Rémunération :	entre 700 € et 1300 € suivant la formation.
Compétences requises : - Vision par ordinateur - Apprentissage automatique (deep learning) - Reconnaissance de formes - Python, C/C++ - Maîtrise d'un framework d'apprentissage profond (en particulier Tensorflow ou PyTorch)	