Offre de Stage

Compression de réseaux multi-tâches

Mots clés: Quantization, Computer Vision, Multi-Task, Action Units, Facial Expression Recognition, Emotion detection

contact
114 Boulevard Malesherbes, 75017, Paris

ad@datakalab.com

ey@datakalab.com

kb@datakalab.com

lf@datakalab.com

Environnement

Datakalab est une startup basée à Paris (17ème arrondissement) spécialisée dans des algorithmes d'apprentissage profond à faible consommation, efficaces en termes d'exécution, respectueux de la vie privée et fonctionnant entièrement en embarqué. Ses travaux de recherche ont donné lieu à des publications dans les meilleures conférences et journaux du domaine (T-PAMI, NeurIPS, ICCV, CVPR, AAAI)

Le stage sera encadrée par Kévin Bailly directeur de la recherche de Datakalab et Maître de conférences, HDR, à l'ISIR, Edouard Yvinec et Arnaud Dapogny, chercheurs en IA à Datakalab.

Contexte

L'apprentissage profond multi-tâches consiste à utiliser un même réseau pour formuler plusieurs prédictions. Un exemple de problème intrinsèqement multi-tâches est la détection d'Action Units (AUs [1,6,7]) du visage, lequel consiste à prédire à partir d'une image un ensemble d'activation de muscles du visage, formant une base de décomposition des expressions faciales. Par exemple, une expression de joie est généralement codée comme l'activation des AUs 12 (remontée des coins des lèvres) et 6 (remontée des pommettes).

Par ailleurs, les réseaux profonds permettent d'atteindre des performances supérieures, bien souvent au prix d'une complexité de calcul accrue, qui rend leur déploiement difficiles et a fortiori sur micro-controleur. Pour compenser ce problème, il existe de très nombreuses approches pour compresser un réseau telle que, par exemple, la quantification (les opérations en arithmétique à virgule flottante sur 32 bits sont remplacées par des opérations à virgule fixe sur un plus faible nombre de bits), par élagage (des opérations sont retirées du graphe de calcul) ou par distillation (la connaissance d'un réseau est transférée dans un réseau de plus petite taille). L'équipe de recherche de Datakalab s'est principalement intéressée aux approches dites sans données (data-free) qui utilisent uniquement l'information contenue dans le réseau [2,3,4,5].

Les approches de compression envisagées ci-dessus rencontrent généralement des problèmes lors de la quantification ou de l'élagage de réseaux multi-tâches: il est en effet difficile de préserver l'équilibre et la précision de chaque tâche dans le réseau, lesquelles peuvent avoir des dynamiques très différentes, et leurs prédictions être impactées par des problèmes de déséquilibre dans les données ou de bruit sur les annotations.

Objectifs du stage

Des travaux préliminaires visant à quantifier l'impact de la compression par rapport au bruit sur les données (cadre mono-tâche) et quantification de réseaux multi-tâches pour la détection d'AUs [6,7] ont été réalisés par Datakalab. Un premier travail consistera à prendre en main ces études et visera à établir un diagnostic qualitatif et quantitatif visant à pointer du doigt pourquoi il est difficile de quantifier ces réseaux. L'étudiant proposera des solutions innovantes pour pallier au problème et validera ces méthodes, d'abord dans un contexte de prédiction d'AU sur les jeux de données de l'état de l'art [6,7], puis dans un cadre grande échelle, par exemple en utilisant des architectures inspirées de [8].

Profil et compétences recherchées

Edudiant de Master ou Grande École. Compétences requises :

- Machine Learning / Deep Learning
- Vision par ordinateur

- Programmation Python et librairie deep learning (tensorflow ou pytorch)
- Excellentes capacités relationnelles et rédactionnelles (français et anglais)

Modalités de candidature

Pour postuler à ce stage, le candidat est invité à communiquer par mail (cf. liste des contacts associés à cette fiche de stage) :

- Son CV
- Ses résultats académiques des deux dernières années universitaires
- (optionnel) Un lien vers un des ces projets en machine learning (lien GitHub / GitLab ou Colab)

Références

[1] G. Tallec, A. Dapogny, and K. Bailly. Multi-order networks for action unit detection. TAC, 2022

[2] RED: Looking for Redundancies for Data-FreeStructured Compression of Deep Neural Networks, 2021, NeurIPS, Yvinec Edouard, Dapogny Arnaud, Cord Matthieu and Bailly, Kévin

[3] RED++: Data-Free Pruning of Deep Neural Networks via Input Splitting and Output Merging, 2022, TPAMI, Yvinec Edouard, Dapogny Arnaud, Cord Matthieu and Bailly, Kévin

[4] To Fold or Not to Fold: a Necessary and Sufficient Condition on Batch-Normalization Layers Folding, 2022, IJCAI, Yvinec Edouard, Dapogny Arnaud, Cord Matthieu and Bailly, Kévin

[5] REx: Data-Free Residual Quantization Error Expansion, arXiv preprint arXiv:2203.14645, E Yvinec Edouard, Dapogny Arnaud, Cord Matthieu and Bailly, Kévin

[6] Z. Xing, Y. Lijun et al. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. IVC, 2014

[7] M. S. Mohammad, M. Mohammad H et al. Disfa: A spontaneous facial action intensity database. TAC, 2013

[8] I. Kokkinos. UberNet: Training a 'Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory. CVPR