THALES

THALES ELANCOURT

PROPOSITION DE STAGE

TITRE DU STAGE :

Deep Learning optimization for embedded real time inference

Objectif du stage :

Depuis 2012, le « deep learning » s'est imposé comme une rupture technologique dans le domaine du « machine learning » et plus particulièrement dans la reconnaissance d'image. Cette rupture technologique basée sur la théorie des réseaux de neurones datant des années 80, combinée aux moyens de calcul actuels (utilisation de GPU) et la disponibilité de grandes bases de données (big data), permet de franchir un gap de performance.

Les modèles développés notamment par les universitaires sont souvent extrêmement gourmand en termes d'utilisation de mémoire et de nombre d'opérations. Ce stage vise à optimiser et simplifier les réseaux de neurones afin d'améliorer l'inférence temps réelle pour des applications embarquées. Plusieurs axes de travail se dégagent :

Optimisation dès la phase d'apprentissage :

- En optimisant le design du réseau de neurones [1, 2] ou en apprenant un design optimal [9]
- En apprenant à « mimiquer » un gros réseau de neurones avec un réseau plus petit [3,4]

Optimisation à postériori :

- Par quantification (calcul en FP16, int8, int4,… plutôt que FP32) [5,6]
- Par des méthodes de « pruning », c'est-à-dire en supprimant les parties du réseau de neurones qui portent peu d'informations [7,8]

Intégré au pôle apprentissage et intelligence artificielle de Thales LAS France, le stagiaire évoluera dans un cadre dynamique et motivant. Il devra faire preuve d'un très bon niveau scientifique et technique avec des qualités d'analyse, de logique, de rigueur, de synthèse, ainsi que de pragmatisme dans le choix des solutions envisagées.

[1] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *arXiv preprint arXiv:1801.04381*.
[2] Ma, N., Zhang, X., Zheng, H. T., & Sun, J. (2018). ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. arXiv preprint arXiv:1807.11164.
[3] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
[4] Yim, J., Joo, D., Bae, J., & Kim, J. (2017, July). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Vol. 2)
[5].Lin, D. D., Talathi, S. S. & Annapureddy, V. S. Fixed Point Quantization of Deep Convolutional Networks. CoRR abs/1511.06393, (2015).
[6].Gupta, S., Agrawal, A., Gopalakrishnan, K. & Narayanan, P. Deep Learning with Limited Numerical Precision. CoRR abs/1502.02551, (2015).
[7]Han, S., Mao, H. & Dally, W. J. Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding. CoRR abs/1510.00149, (2015).
[8] Molchanov, P., Tyree, S., Karras, T., Aila, T. & Kautz, J. Pruning Convolutional Neural Networks for Resource Efficient Transfer Learning. CoRR abs/1611.06440, (2016).
[9] Tan, M., Chen, B., Pang, R., Vasudevan, V., & Le, Q. V. (2018). MnasNet: Platform-Aware Neural Architecture Search for Mobile. arXiv preprint arXiv:1807.11626.

Durée du stage : 6 mois	**Dates :** Printemps - été 2019

Tuteur(s) du stage :

Gilles Henaff : gilles.henaff@fr.thalesgroup.com

Profil du stagiaire :

Stagiaire Master 2 / Ecole d'ingénieurs
Connaissances en traitement d'images et du signal, apprentissage machine et deep learning
Connaissances informatiques : Maîtrise de python.
Une première expérience dans l'un des frameworks suivant seraient un plus : TensorFlow et PyTorch
Anglais (lu)