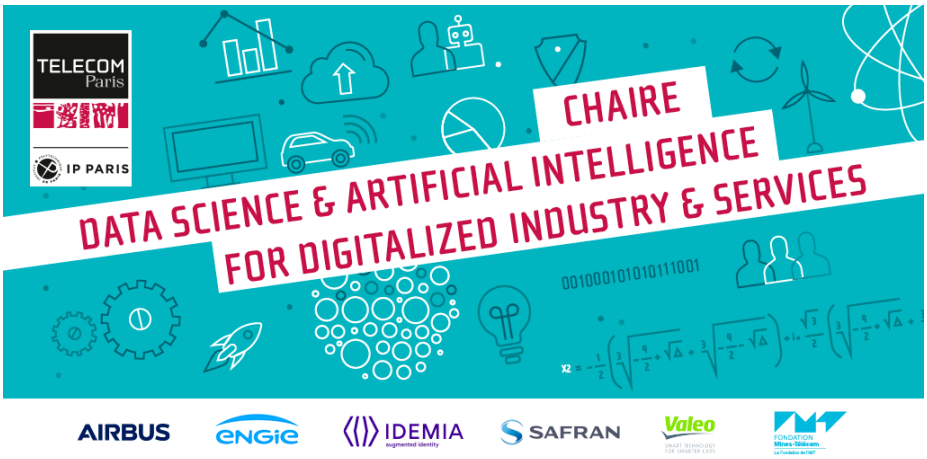


# *Offre de stage*

*Sujet : Efficiencing Deep Learning with quantization*

*Possibilité de poursuivre sur une thèse*

## La Chaire Data Science and Artificial Intelligence for Digitized Industry and Services

Portée par Florence d'Alché-Buc, enseignante-chercheur dans le département Image, Données, Signal de Télécom ParisTech, la chaire DSAI réunit cinq partenaires industriels : Airbus Defence & Space, Engie, Idemia, Safran et Valeo Finance. Son objectif général est de développer, en liaison étroite entre les Parties, une formation et une recherche de niveau international.

Ses quatre principaux axes de recherche sont :

1. Analyse et prévision de séries temporelles (Predictive Analytics on Time Series) ;
2. Exploitation de données hétérogènes, massives et partiellement étiquetées (Exploiting Large Scale and Heterogeneous, Partially Labelled Data) ;
3. Apprentissage pour une prise de décision robuste et fiable (Learning for Trusted and Robust Decision) ;
4. Apprentissage dans un environnement dynamique (Learning through Interactions with a Changing Environment).

## Description du stage

### Encadrement

Enzo Tartaglione, Maître de conférences à Télécom Paris, Institut Polytechnique de Paris

### Lieu et dates du stage

Telecom Paris, 19 Place Marguerite Perey, 91120 Palaiseau

Date de début du stage: début 2021

### Équipe(s) d'accueil de la thèse

Laboratoire LTCI, équipe MultiMedia (MM)

### Mots clés

Quantization, Deep learning

### Sujet détaillé

The ever-increasing number of parameters in deep neural networks poses challenges for memory-limited applications [1]. Besides, of great relevance covers the problem of numerical representation in deep neural networks-based computation, both at train and inference time. Strongly related to the problem of numerical representation is the problem of quantization: how should a set of continuous real-valued numbers be distributed over a fixed discrete set of numbers, to minimize the number of bits required, and to maximize the accuracy of the attendant computations? Moving from floating-point representations to low-precision fixed integer potentially reduces memory and energy consumption, especially in embedded devices, finding critical immediate application in real-time computation [2].

In the first stage of this proposal, the candidate will study and deploy state-of-the-art solutions to neural network quantization (like for example [3-8]), evaluating advantages and disadvantages in terms of execution time, final model size, memory requirement, computational complexity, testing them on GPU and CPU-based systems.

Besides such analysis, the candidate will design and measure the « redundancy » in the neurons/filters design, in the form of entropy for the quantized representation of the model. It has been observed that neural networks can be compressed minimizing an entropy proxy besides the standard loss function [9]. Towards this end, the candidate will design an efficient class/method to statically (ie. at inference time) measure

$$H(N) = -\sum_{n \in N} P(n) \log P(n)$$

where N is the quantized model, H() is the entropy, P() is a probability estimate and n is some subnetwork of N.

## Profil du candidat

Student holding a Master 2 researchM with:

- Knowledge of the deep learning working principles (back-propagation, gradient descent, convolutional layers, batch normalization),
- Theoretical/practical knowledge of standard deep learning approaches (architectures : ResNet, VGG ; optimization : SGD, Adam ; datasets: MNIST, CIFAR-x, ImageNet ; loss function : Cross-entropy, MSE ; regularization : weight-decay, dropout),
- Good level in programming (Python - PyTorch),
- Good level in English – the activity will be entirely in english.

## Candidatures

Send to enzo.tartaglione@telecom-paris.fr

- Curriculum vitae
- Personalized cover letter explaining the candidate's motivations (in the body of the email)
- Transcripts from previous years
- E-mail contact.

## Références

[1] Tartaglione, E., Lepsøy, S., Fiandrotti, A., & Francini, G. (2018, December). Learning sparse neural networks via sensitivity-driven regularization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 3882-3892).
[2] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., & Keutzer, K. (2021). A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*.
[3] Stock, P., Joulin, A., Gribonval, R., Graham, B., & Jégou, H. (2019). And the bit goes down: Revisiting the quantization of neural networks. *arXiv preprint arXiv:1907.05686*.
[4] Lou, Q., Guo, F., Liu, L., Kim, M., & Jiang, L. (2019). Autoq: Automated kernel-wise neural network quantization. *arXiv preprint arXiv:1902.05690*.
[5] Chen, T., Li, L., & Sun, Y. (2020, November). Differentiable product quantization for end-to-end embedding compression. In *International Conference on Machine Learning* (pp. 1617-1626). PMLR.
[6] Towards Accurate Post-training Network Quantization via Bit-Split and Stitching
[7] Wang, P., Chen, Q., He, X., & Cheng, J. (2020, November). Towards accurate post-training network quantization via bit-split and stitching. In *International Conference on Machine Learning* (pp. 9847-9856). PMLR.
[8] Nagel, M., Amjad, R. A., Van Baalen, M., Louizos, C., & Blankevoort, T. (2020, November). Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning* (pp. 7197-7206). PMLR.
[9] Tartaglione, E., Lathuilière, S., Fiandrotti, A., Cagnazzo, M., & Grangetto, M. (2021). HEMP: High-order Entropy Minimization for neural network comPression. *Neurocomputing*, *461*, 244-253.