# ÉLÉMENT DE PORTFOLIO 04

# Publication

## 1 DÉFINITION DE CET ÉLÉMENT

**Titre de l'élément : Edge intelligence**

**URL de l'élément :** `https://hal.science/hal-04042615`

## 2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

L'IA et en particulier les réseaux de neurones profonds ont connu une forte résurgence dès l'année 2012. Il s'agit d'architectures paramétriques multi-couches dotées d'unités neuronales interconnectées. L'avantage majeur de ces architectures réside dans leur capacité à apprendre à représenter les données de manière hiérarchique. L'un des modèles les plus performants est le réseau de neurones convolutif (CNN) et le transformer. Ces modèles ont une particularité : il est observé que plus la précision attendue d'un réseau est élevée, plus sa taille et le nombre de données nécessaires pour l'entraîner deviennent importants. En effet, l'augmentation de la précision des réseaux de neurones a un coût ; à titre d'exemple, le réseau VGG16 nécessite plusieurs millions de paramètres pour réaliser une classification d'image alors que les réseaux transformers GPT les plus récents (en lien avec le succès actuel de ChatGPT) nécessitent des centaines de milliards de paramètres. Ce nombre a un impact sur les systèmes électroniques ou embarqués qui exécutent ces réseaux de neurones : une consommation énergétique importante, un temps d'exécution significatif ainsi qu'un espace mémoire et un nombre de données d'apprentissage nécessaires non-négligeables. Ce constat est acceptable dans le cadre d'utilisation de systèmes comme les gros serveurs de calculs, mais pour des systèmes embarqués qui sont de plus en plus ubiquitaires, cela est problématique. Il est nécessaire de satisfaire la contrainte de précision des réseaux de neurones mais aussi les tailles des ensembles d'apprentissage pour entraîner ces modèles, les contraintes de consommation énergétique, de temps d'exécution et d'espace mémoire.

Une tendance actuelle en apprentissage machine vise à rendre ces modèles frugaux et légers tout en préservant leur haute précision. En effet, la variabilité ou le manque des données annotées nécessite la mise en place de solutions d'apprentissage frugaux en données annotées (notamment actif et continu), comme en biomédicale ou la disponibilité des experts en annotations, à savoir les médecins, est très limitée. De plus les compressions algorithmiques applicables aux réseaux de neurones visent à réduire la taille et l'impact de l'inférence de ces modèles sur les systèmes électroniques. Ces efforts répondent aux besoins d'implémentation d'algorithmes d'analyses puissants au sein d'unités matérielles contraintes et largement déployées dans les applications embarquées, telles que les microcontrôleurs ou les FPGA. D'autre part, la préservation de la vie privée et la sécurité des données nécessitent la mise en place de solutions d'entraînement et d'inférence de ces réseaux de neurones In Situ (i.e. en les appliquant localement sur des dispositifs embarqués), ce qui permet de rassurer les utilisateurs à propos de leurs données personnelles (comme pour les applications domotiques). Les contraintes de transmission sont également importantes notamment pour des applications ou la bande passante est faible, voire inexistante, comme pour les missions spatiales lointaines.

Nous avons étudié et développé des solutions originales sur les problèmes susmentionnés (voir exemples de publications [1–11]) et ceci dans le cadre de collaborations bilatérales avec Thales, l'IGN, CEA, Essilor, etc. Ces travaux ont donné lieu à la soutenance de 3 thèses (entre 2018-2020) et d'autres thèses sont en cours.

## 3 PRÉSENTATION DE CET ÉLÉMENT

Le premier axe de nos travaux concerne l'apprentissage actif dont l'objectif est de construire itérativement des modèles d'apprentissage frugaux en données annotées [1]. Nous avons réalisé cela à l'aide de fonctions d'acquisitions permettant de sélectionner un "pool" de données d'apprentissage le plus petit possible dont l'usage — lors de l'entraînement des réseaux de neurones — permet d'obtenir des performances comparables à un apprentissage totalement supervisé. Le deuxième axe concerne l'apprentissage continu dont l'objectif est d'apprendre des réseaux de neurones en ingérant frugalement des incréments de données (comme en streaming). L'un des

problèmes de l'apprentissage continu est l'oubli catastrophique caractérisé par l'incapacité d'un réseau de neurones à mémoriser les anciennes taches lorsqu'il est entraîné sur de nouvelles. Dans cet axe de recherche, nous avons mis en place des solutions permettant d'atténuer l'oubli catastrophique [11] en apprenant les paramètres des réseaux de neurones dans des espaces orthogonaux par rapport aux données des anciennes taches, ce qui permet de réduire significativement l'interférence entre les paramètres des différentes taches en classification d'images.

Le déploiement de ces solutions sur des dispositifs à bas coût (smartphones, etc) nécessite la compression des réseaux de neurones. Dans cette perspective, nous avons mis en place des méthodes de compression et d'accélération des réseaux de neurones profonds (e.g. [2]). La démarche employée est basée sur l'élagage de connexions neuronales à l'aide de plusieurs critères (par magnitude, en modélisant explicitement la latence, en exploitant des contraintes d'orthogonalité et de consistance topologique, etc.). Les usages en classification d'images et d'actions en vidéos ont permis d'obtenir des taux d'élagage très élevés tout en maintenant une précision proche des réseaux de neurones initiaux non-élagués.

La compression des réseaux de neurones à travers l'application des techniques d'élagage, ou encore de quantification ou de distillation de connaissances engendre une réduction de son coût d'inférence. La réduction en consommation énergétique, temps d'exécution et espace mémoire des réseaux de neurones doit être prise en compte pour respecter au mieux les contraintes des différents cas d'usages utilisant des dispositifs à bas coût. Dans cet objectif, nous avons conçu et évalué un nouveau modèle [4] caractérisant ces contraintes dans le cas de l'implémentation d'un réseau de neurones convolutifs compressé ou non. La démarche a pour objectif de guider les concepteurs de système embarqué au plus tôt dans le processus de conception.

Enfin, nos cas d'usages sont multiples et concernent essentiellement des applications de vision par ordinateur et reconnaissance des formes comme la conduite automatisée, télédétection, vidéosurveillance, etc. L'un des cas d'usages que nous avons étudié concerne l'estimation du mouvement (flux optique) dans les séquences vidéos à l'aide des réseaux de neurones profonds entraînés en mode auto-supervisé [6] (c.a.d. en étant frugaux en annotations). Une des contributions proposées dans ces travaux concerne une méthode originale permettant d'estimer le flux optique en prenant en compte les changements d'illumination dans les scènes traitées. D'autres cas d'usages concernent (i) la vidéosurveillance ou nous avons développé des méthodes de reconnaissance d'actions dans les séquences vidéos à l'aide des réseaux convolutifs légers sur des graphes [7], et aussi (ii) l'analyse des séries d'images multi-temporelles en télédétection, et ceci pour la segmentation et l'estimation de l'occupation des sols [9], la classification d'images [5] et des navires [8] ainsi que la détection de zones ayant subi des changements (destructions) suite à des catastrophes naturelles (comme les tornades et les tremblements de terre) [1, 10].

# 4 RÉFÉRENCES BIBLIOGRAPHIQUES


[1] Sebastien Deschamps and Hichem Sahbi. Reinforcement-based display selection for frugal learning. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1186–1193. IEEE, 2022.

[2] Robin Dupont, Mohammed Amine Alaoui, Hichem Sahbi, and Alice Lebois. Extracting effective subnetworks with gumbel-softmax. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 931–935. IEEE, 2022.

[3] Thomas Garbay, Orlando Chuquimia, Andrea Pinna, Hichem Sahbi, Xavier Dray, and Bertrand Granado. Distilling the knowledge in cnn for wce screening tool. In *2019 Conference on Design and Architectures for Signal and Image Processing (DASIP)*, pages 19–22. IEEE, 2019.

[4] Thomas Garbay, Khalil Hachicha, Petr Dobias, Wilfried Dron, Pedro Lusich, Imane Khalis, Andrea Pinna, and Bertrand Granado. Accurate Estimation of the CNN Inference Cost for TinyML Devices. In *2022 IEEE 35th International System-on-Chip Conference (SOCC)*, 2022.

[5] Mingyuan Jiu and Hichem Sahbi. Nonlinear deep kernel learning for image annotation. *IEEE Transactions on Image Processing*, 26(4) :1820–1832, 2017.

[6] Rémi Marsal, Florian Chabot, Angélique Loesch, and Hichem Sahbi. Brightflow : Brightness-change-aware unsupervised learning of optical flow. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2061–2070, 2023.

[7] Ahmed Mazari and Hichem Sahbi. Mlgcn : Multi-laplacian graph convolutional networks for human action recognition. In *The British Machine Vision Conference (BMVC)*, 2019.

[8] Quentin Oliveau and Hichem Sahbi. Learning attribute representations for remote sensing ship category classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(6) :2830–2840, 2017.

[9] Tristan Postadjian, Arnaud Le Bris, Hichem Sahbi, and Clément Mallet. Investigating the potential of deep neural networks for large-scale classification of very high resolution satellite images. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4, 2017.

[10] Hichem Sahbi. Interactive satellite image change detection with context-aware canonical correlation analysis. *IEEE Geoscience and Remote Sensing Letters*, 14(5) :607–611, 2017.

[11] Hichem Sahbi and Haoming Zhan. Ffnb : Forgetting-free neural blocks for deep continual learning. In *The British Machine Vision Conference (BMVC)*, 2021.