Favoriser l'émergence des compétences recherche et innovation avant la thèse

L. Fesquet [a,b], X. Lesage [a,c], C. Merio [a,d], A. Naimi [a], S. Engels [a,d]

[a] Univ. Grenoble Alpes, CNRS, Grenoble INP*, TIMA, F-38000 Grenoble, France
[b] Univ. Grenoble Alpes, CNRS, Grenoble INP*, Phelma, F-38000 Grenoble, France
[c] Orioma, F-38430 Moirans, France
[d] STMicroelectronics, F-38920 Crolles, France

Contact email : laurent.fesquet@univ-grenoble-alpes.fr

L'enseignement supérieur et, en premier lieu, les formations d'ingénieurs utilisent de plus en plus souvent un référentiel de compétences pour évaluer leurs étudiants. Parmi les compétences requises, la compétence recherche est probablement l'une des plus difficiles à transmettre. En effet, même les étudiants de niveau Master ou les élèves-ingénieurs sont naturellement peu exposés à une activité de recherche car ils sont toujours en cours de formation et ne possèdent pas tout le bagage scientifique nécessaire pour s'y adonner. Pourtant, il est possible au travers de projets de leur faire découvrir les facettes nécessaires à l'innovation et à la découverte. La méthode est illustrée par un projet de conception d'un circuit microélectronique.

I. Introduction et contexte de la formation par la recherche

La formation par la recherche est usuellement comprise comme les travaux de thèse menés par les doctorants. Cette formation est souvent riche d'enseignements car elle permet de développer une technicité pointue, mais amène aussi progressivement le doctorant à définir sa propre trajectoire dans son activité de recherche. Il gagne ainsi en autonomie et s'offre même le luxe d'être imaginatif et créatif. C'est grâce à ce dernier ressort que nous pouvons devenir innovants, inventifs, voire révolutionnaires.

Si l'activité en laboratoire sous la houlette d'un directeur de thèse se prête bien à cet exercice, la situation est tout autre lorsque l'on souhaite développer cette compétence auprès de nos élèves-ingénieurs et, plus généralement, auprès des étudiants d'un niveau Master. Les obstacles sont en effet nombreux. Il est difficile de glisser cette compétence dans le syllabus d'une formation car elle ne répond finalement à aucune matière, ni programme scientifique. Nous allons essayer de voir, dans ce qui suit, comment il est malgré tout possible d'utiliser des enseignements et notamment des projets étudiants pour introduire un soupçon de formation par la recherche sans pour autant déroger au programme officiel de la formation.

II. La formation par la recherche avant la thèse

A proprement parler, cette formation est quasiment inexistante à l'exception des stages effectués en laboratoire par nos étudiants. En effet, l'enseignant qui aurait l'idée saugrenue

Article publié par EDP Sciences et disponible sur le site https://www.j3ea.org ou https://doi.org/10.1051/j3ea/20241006

d'initier à la recherche (hors stages) nos étudiants de Master et nos élèves-ingénieurs va immédiatement se heurter à plusieurs difficultés et pas des moindres.

La première difficulté, qui apparaît, sont les connaissances limitées de nos étudiants. En effet, ils sont encore en formation et ils doivent acquérir et s'approprier des connaissances nouvelles. Au regard du processus de cognition nécessaire à cette acquisition de connaissances et de savoir-faire, il est souvent compliqué, voire impossible, de demander à nos étudiants d'avoir un regard critique sur les contenus fraîchement emmagasinés mais rarement assimilés.

La seconde difficulté est la nécessité de respecter les programmes de formation car nous préparons nos étudiants à entrer dans la vie active avec les formations professionnelles que nous dispensons. Il serait donc mal venu de s'abstraire de l'environnement socio-économique et de développer un enseignement spécifique en vue de former les étudiants à résoudre une problématique de recherche tout aussi spécifique. Ce ne serait en aucune façon rendre service à nos étudiants et à nos partenaires industriels qui espèrent trouver des ingénieurs et des étudiants diplômés utilisables et adaptables à leurs besoins.

Enfin, l'évaluation des compétences plutôt que des connaissances et du savoir-faire va, si l'on n'y prend pas garde, dans le sens d'un appauvrissement de nos formations et d'une réduction du métier d'ingénieurs à un simple jeu de compétences comme la capacité à travailler en équipe ou d'ordonnancer des tâches. Heureusement, il reste la compétence dite « métier » qui demeure vague pour le profane et qui n'a de sens que pour le technicien qui connaît parfaitement sa profession. Ainsi, ajouter une dimension « recherche et innovation » semble une gageure supplémentaire.

III. Le projet étudiant vecteur de développement d'une compétence recherche

Le projet est un moment particulier dans la vie étudiante car il met l'apprenant en pause par rapport à l'acquisition de savoirs nouveaux et lui donne l'opportunité de développer des savoir-faire comme c'est le cas en TP, mais avec un temps d'analyse et de réflexion plus conséquent. Ce laps de temps constitue une opportunité pour l'enseignant de distiller subrepticement la compétence recherche. En effet, le projet se doit de rester dans les clous de la formation et constituer une base solide pour l'acquisition des savoir-faire usuels. Il n'est donc pas souhaitable de tout bouleverser à ce stade mais il est judicieux d'ajouter le zeste de connaissances supplémentaires qui permettra de remettre en cause les acquis. En procédant de la sorte, on invite l'étudiant à entamer une réflexion sur les connaissances qu'il a apprises, à analyser les limites du savoir dispensé et à s'interroger sur de possibles remédiations. Tout ceci peut paraître bien théorique mais il est temps d'illustrer par un exemple concret comment nous pouvons insidieusement faire réfléchir nos étudiants à une problématique recherche. Ainsi, ils feront un premier pas dans le monde de la recherche et de l'innovation et, pour peu qu'ils travaillent en équipe, cocheront les cases de nombreuses compétences demandées dans les écoles d'ingénieurs et à l'université.

IV. Un réseau de neurones convolutif piloté par les données

Les réseaux de neurones sont des sujets à la mode qui intéressent nos étudiants. Il est donc normal que les étudiants de la filière « Systèmes Electroniques Intégrés » de l'école Phelma de l'institut polytechnique de Grenoble aient envie, dans le cadre d'un projet de conception de circuits intégrés, de développer leur propre réseau de neurones. Le choix du réseau de neurones s'est porté sur un réseau de neurones convolutif. Ces derniers sont simples à concevoir, faciles à étudier et bien documentés. Ainsi, nous aurions pu nous contenter de mener à son terme le projet en choisissant un réseau pré-entrainé que nous

aurions pu implémenter au travers du flot numérique (du RTL au *Layout*) qui est enseigné à l'école. Ce travail est suffisamment conséquent et formateur pour que nos étudiants en tirent profit dans le cadre de leur formation professionnelle. En effet, les étudiants peuvent se contenter d'appliquer les recettes proposées par les enseignants puisqu'il s'agit de mettre en pratique les connaissances acquises.

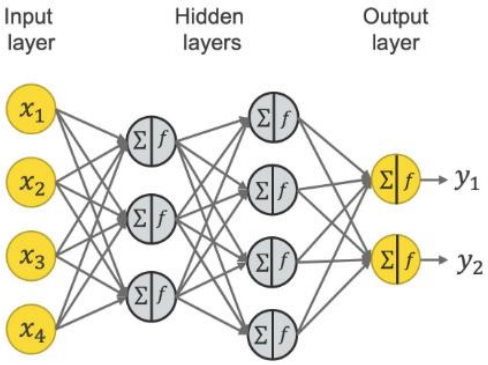

Fig.1. Un réseau de neurones usuel.

En revanche, dans ce sujet, rien n'invite à creuser le sujet plus en profondeur. Il est donc temps d'ajouter quelques éléments à notre sujet « bateau » sur la conception d'un réseau de neurones. Premièrement, il est intéressant de réfléchir au comportement d'un tel réseau de neurones et de faire une première observation avec les étudiants. L'analyse du fonctionnement du réseau montre que les données, qui se propagent dans le réseau, mènent rapidement à un taux d'activation faible. Enfin, l'activation systématique des calculs par un signal d'horloge ne permet pas de profiter de ce faible niveau d'activation et conduit à une sur-activation du réseau. Ainsi, il apparaît que les réseaux neuronaux sont très exigeants en puissance de calcul et qu'ils consomment beaucoup. De plus, la taille de ces réseaux ne cesse d'augmenter sans que la problématique de leur consommation ne soit efficacement traitée. Dans ce contexte, toutes les techniques d'économie d'énergie sont les bienvenues dans le domaine de l'IA et de son utilisation dans les dispositifs connectés (*Edge AI*) en particulier.

Ainsi, la prévalence de nombreux zéros dans les cartes d'activation ou les matrices de poids des réseaux de neurones a été examinée dans divers systèmes, dont les réseaux neuronaux convolutifs (CNN) (1). Il en résulte des calculs inutiles, qui peuvent être évités en exploitant cette parcimonie. Ainsi, il est possible d'améliorer l'efficacité énergétique et la vitesse. En outre, plus les réseaux sont grands, plus ils sont parcimonieux (2-3) (*cf.* Fig.2). Enfin, le caractère épars des activations peut être cultivé en introduisant délibérément des valeurs nulles dans les cartes d'activation (4) ou dans les matrices de poids (5). Les implémentations de réseaux neuronaux convolutifs exploitant ces techniques sont désormais bien connues, comme Eyeriss (6), qui utilise le *clock gating* pour améliorer la consommation d'énergie, ou ZeNA (7) dans laquelle l'architecture parallèle a été conçue pour réduire l'énergie en tirant profit de calculs épars.

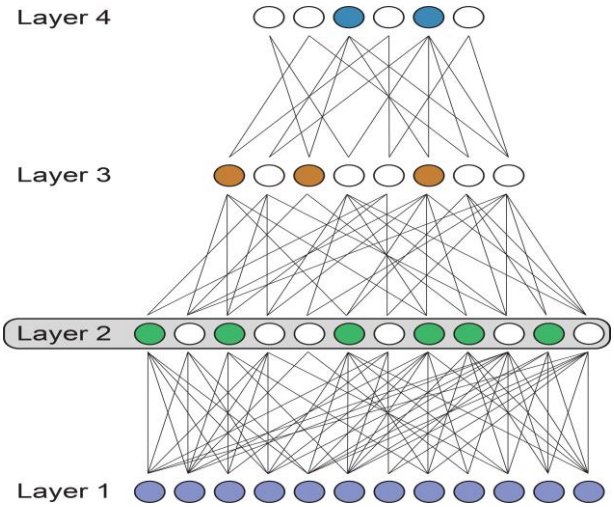

Fig.2. Les neurones activés (colorés) sont de moins en moins nombreux au fil des couches (3)

Un autre ingrédient a été ajouté à la sauce afin de réaliser une implémentation originale et potentiellement très efficiente énergétiquement. Nous avons eu recours à une conception asynchrone dite *Bundled-Data* qui est extrêmement proche d'un circuit synchrone (*cf.* Fig.3). En effet, cette classe de circuit présente un chemin de données similaire aux circuits synchrones mais l'arbre d'horloge a été partiellement remplacé par un contrôleur asynchrone qui fait avancer les données pas à pas dans le circuit, mais irrégulièrement d'un point de vue temporel. Ainsi, nous respectons la nécessité de former nos étudiants aux circuits synchrones en pointant du doigt leurs hypothèses temporelles sous-jacentes et leurs limites de performance (vitesse et énergie).

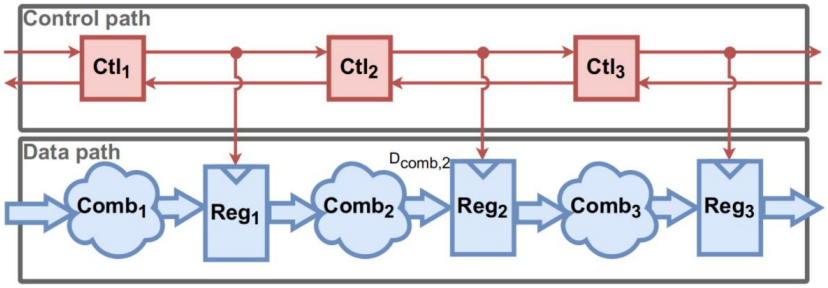

Fig.3. Principe d'un circuit *Bundled-Data.*

Pour bénéficier pleinement de cette stratégie asynchrone, Merio *et al*. (8) proposent une méthode asynchrone qui permet de sauter dynamiquement les calculs inutiles, d'économiser l'énergie de ces derniers et de rendre les circuits plus rapides puisque l'on ne fait pas les calculs inutiles ! Nous avons appliqué cette méthode aux neurones de notre réseau convolutif en mettant en œuvre cette stratégie d'élagage des calculs inutiles. Ainsi, les calculs avec des valeurs ou des poids proches de zéro se sont vus élagués dynamiquement ainsi que les entrées ayant les mêmes caractéristiques. Cette approche a permis de réduire la consommation d'énergie de façon drastique tout en augmentant le débit moyen. La Fig.4 montre que la réduction d'énergie est proportionnelle aux calculs inutiles, représentés ici par le vocable *Ghost Token Ratio*, qui correspond à un modèle des circuits asynchrones basé sur un réseau de Petri exploitant des jetons (*Token*) et des jetons fantômes (*Ghost Token*). La modélisation formelle d'un circuit asynchrone est hors contexte dans cet article mais elle permet aux étudiants de toucher du doigt de nouvelles

notions relatives aux modèles de circuits. Enfin, il est également mis en évidence que le débit moyen augmente dès lors que la quantité de calculs inutiles croît. En effet, le non-traitement des données négligeables permet de court-circuiter (*bypass*) de nombreux neurones et de gagner le temps de traitement leur incombant. Ainsi, le temps de calcul s'en trouve réduit et le débit augmenté.

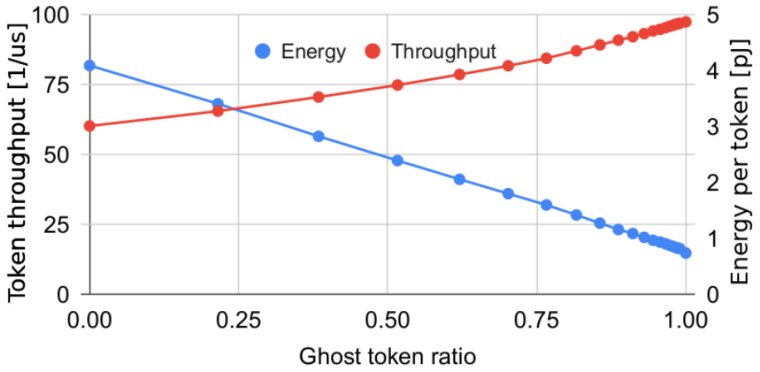

Fig.4. Résultats (énergie et débit) de l'étude menée sur un réseau de neurones convolutif

V. Bilan et conclusion

Les étudiants de la filière « Systèmes Electronique Intégrés » de Phelma sont amenés à concevoir des puces analogiques ou numériques au cours de leur scolarité. Cet enseignement pratique effectué en seconde année d'école est stratégique pour l'acquisition des savoir-faire élémentaires de cette filière de formation professionnelle. Le projet en lui-même représente déjà un défi pour les étudiants, mais aussi pour les enseignants. Ainsi, exposer nos étudiants à une activité de recherche et d'innovation parait être une mission très délicate. Pourtant, avec un peu d'astuces et de savoir-faire enseignant, il est possible de distiller une petite dose de cette compétence recherche !

La méthode consiste à rester dans les clous du syllabus de la formation mais d'additionner une petite connaissance supplémentaire qui vient déstabiliser un *corpus* qui semble être acquis ou immuable. Dans l'exemple proposé, la découverte d'une conception sans horloge constitue cet élément de déstabilisation. Il oblige les étudiants à repenser leurs savoirs, à remettre en question leurs acquis. Cette approche est pratiquée à Phelma depuis de nombreuses années. Elle peut être assez facilement mise en œuvre avec tous les élèves pour peu qu'ils soient un peu curieux. Elle peut être appliquée à tous les champs disciplinaires et il est facile d'imaginer des sujets relatifs à la sécurité matérielle, à la sûreté de fonctionnement, aux méthodes de test ou aux réseaux de neurones.

Le travail relaté dans cet article a été mené par des étudiants ayant un bon niveau mais la méthode est déployée plus largement au travers de nombreux projets de conception de puces. Enfin, il est intéressant de noter que le travail d'optimisation effectué par nos étudiants a fait l'objet d'un article qui a été soumis pour publication dans une conférence internationale renommée.

Remerciements

Les auteurs remercient les collègues du CIME Nanotech qui soutiennent quotidiennement les activités de conception en microélectronique. Merci à Robin Rolland-Girod, Mohamed Ben Jrad et Abdelhamid Aitoumeri. Merci aussi aux collègues de Grenoble INP/Phelma qui échangent avec nous sur les bonnes pratiques pédagogiques.

Enfin, le GIP-CNFM est remercié pour son soutien au long cours et à l'aide qu'il apporte dans le contexte de « France 2030 » et du projet « ANR-23-CMAS-0024 INFORISM ».

Références

1. Z. Li, C. You, S. Bhojanapalli, D. Li, A. S. Rawat, S. J. Reddi, K. Ye, F. Chern, F. Yu, R. Guo, and S. Kumar, "The Lazy Neuron Phenomenon: On Emergence of Activation Sparsity in Transformers,", *arXiv:2210.06313 [cs, stat]* (2023).
2. T. Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste, "Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks", *arXiv:2102.00554 [cs]* (2021).
3. K. L. Hunter, L. Spracklen, and S. Ahmad, "Two sparsities are betterthan one: Unlocking the performance benefits of sparse-sparse networks,"*CoRR, **abs/2112**, 13896* (2021).
4. M. Kurtz, J. Kopinsky, R. Gelashvili, A. Matveev, J. Carr, M. Goin, W. Leiserson, S. Moore, B. Nell, N. Shavit, and D. Alistarh, "Inducing and exploiting activation sparsity for fast neural network inference," in *Proceedings of the 37th International Conference on Machine Learning*, **119 of ICML'20**, 5533–5543, JMLR.org (2020).
5. Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming", *IEEE International Conference on Computer Vision (ICCV)*, Oct. 22 2017 to Oct. 29 2017, Venice, Italy (2017).
6. Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *IEEE Journal of Solid-State Circuits*, **52**, 127–138 (2017).
7. D. Kim, J. Ahn, and S. Yoo, "ZeNA: Zero-Aware Neural Network Accelerator," *IEEE Design & Test*, **35**, 39–46, Feb. 2018. Conference Name: IEEE Design & Test (2018)
8. C. Merio, X. Lesage, A. Naimi, S. Engels, K. Morin-Allory, L. Fesquet, "Method for Data-Driven Pruning in Micropipeline Circuits", *31st IFIP/IEEE Conference on Very Large Scale Integration (VLSI-SoC 2023)*, October 16 - 18, UAE (2023).