**Compas'2023** : Parallélisme/ Architecture/ Système
LISTIC/USMB - Annecy France, 4-7 juillet 2023

# Prototypage rapide pour l'inférence par réseaux de neurones convolutifs sur cibles matérielles hétérogènes

Sean Marotta[1], Alessandro Carlini[2], Vincent Brost[1], Cyrille Migniot[1], Manon Ansart[2], Michel Paindavoine[2], Julien Dubois[1]

Université de Bourgogne,
Laboratoire ImViA EA 7535[1] et LEAD CNRS UMR 5022[2]
21000 Dijon - France
Prénom.Nom@u-bourgogne.fr

---

**Résumé**
Ces dernières années, l'Intelligence Artificielle (IA) est devenue de plus en plus populaire dans les problématiques de vision par ordinateur. Néanmoins, les réseaux neuronaux convolutifs (CNN) nécessitent une quantité importante de calculs pendant la phase d'inférence. Bien que les GPU soient traditionnellement utilisés pour l'inférence d'IA, les FPGA représentent une alternative notable pour l'embarqué. Ce type de circuit présente une latence plus faible, une meilleure efficacité énergétique et une plus grande flexibilité de personnalisation. Toutefois, le temps de développement sur FPGA peut s'avérer très long. De plus, l'intégration de processeur au sein des FPGA permet certes de gagner en flexibilité, mais complexifie la conception. Cependant, les outils de prototypage rapide commerciaux permettent d'accélérer les temps de développement mais proposent également des solutions à l'intégration de phase d'inférence. Après comparaison de différents CNN, nous avons sélectionné EfficientNet-B0 car il possède des caractéristiques appropriées aux implantations embarquées. Le modèle, d'abord entraîné sous un schéma classique avec le dataset MNIST sous TensorFlow, est quantifié via l'outil Vitis AI de Xilinx, puis ré-entraîné sur ce même dataset pour pallier à la diminution de précision en détection. Une phase de partition matérielle entre le processeur et la partie logique est ensuite réalisée manuellement. Certaines couches du réseau sont implantées sur le processeur en raison de leur irrégularité ou pour faciliter les échanges de données avec le système global. Les phases convolutives plus régulières sont naturellement implantées sur la partie logique reconfigurable. Des accélérateurs matériels dédiés à l'inférence sont disponibles dans les outils de développement Xilinx, et servent de base pour l'implantation de cette phase régulière. Le paramétrage de l'accélérateur est réalisé en fonction de l'apprentissage, de la quantification et des contraintes utilisateur. Pour les images du dataset, le système atteint des fréquences supérieures à 250 images par seconde pour un seul accélérateur, soit 45 041 blocs de logique combinatoire (LUT) et 56 604 bascules, sans perte de précision. Les résultats démontrent de bonnes performances non seulement en temps d'exécution, mais aussi en termes de prédiction, sur un temps de conception raisonnable.

**Mots-clés :** Prototypage rapide, FPGA, Vitis AI, quantification, accélérateur matériel.

---