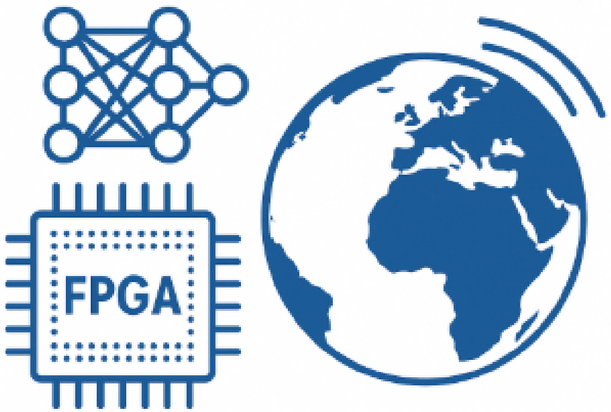
UQAR
Université du Québec
à Rimouski

FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Accélération de l'inférence en périphérie : Implémentation matérielle sur circuits FPGA du modèle de détection d’objets YOLOv5 avec l’outil Vitis AI

Étudiant : Achraf Chakroun

Directeur de recherche : Mohammed Bahoura



23 Avril 2025

UQAR
Université du Québec
à Rimouski

FIE FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Plan de la présentation

1. Introduction
2. Problématique
3. Objectifs
4. Méthodologie
5. Résultats
6. Conclusion et perspectives

UQAR
Université du Québec
à Rimouski

FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Introduction

23 Avril 2025

UQAR
Université du Québec
à Rimouski

Introduction

1. L'intelligence artificielle est de plus en plus présente dans des systèmes embarqués :
•Caméras intelligentes
•Véhicules autonomes
• Objets connectés

FIE FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE | ENTREPRENEURIAT | UQAR

2. Ces dispositifs ont besoin de modèles efficaces en temps réel, avec peu de ressources matérielles.

UQAR
Université du Québec
à Rimouski

FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE | ENTREPRENEURIAT | UQAR

Problématique

23 Avril 2025

UQAR
Université du Québec
à Rimouski

FIE FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Problématique

- Les CPU et GPU sont performants, mais peu adaptés au edge computing en raison de leur consommation élevée.

Comment peut-on atteindre un compromis optimal entre performance d'inférence, faible consommation énergétique et compatibilité embarquée ?

UQAR
Université du Québec
à Rimouski

FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Objectifs

23 Avril 2025

UQAR
Université du Québec
à Rimouski

FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

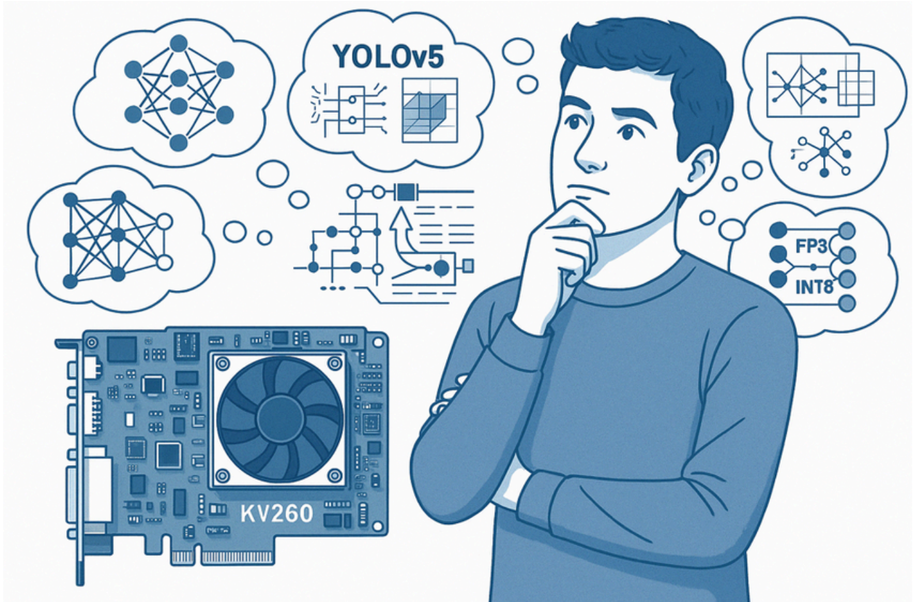
Objectifs

- Optimisation du modèle de détection d'objet Yolov5;
- Réduction de la latence tout en maintenant une précision acceptable;
- Déploiement du modèle quantifié sur une architecture matérielle optimisée (FPGA);
- Validation des performances par application en temps réel;

UQAR
Université du Québec
à Rimouski

FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Méthodologie

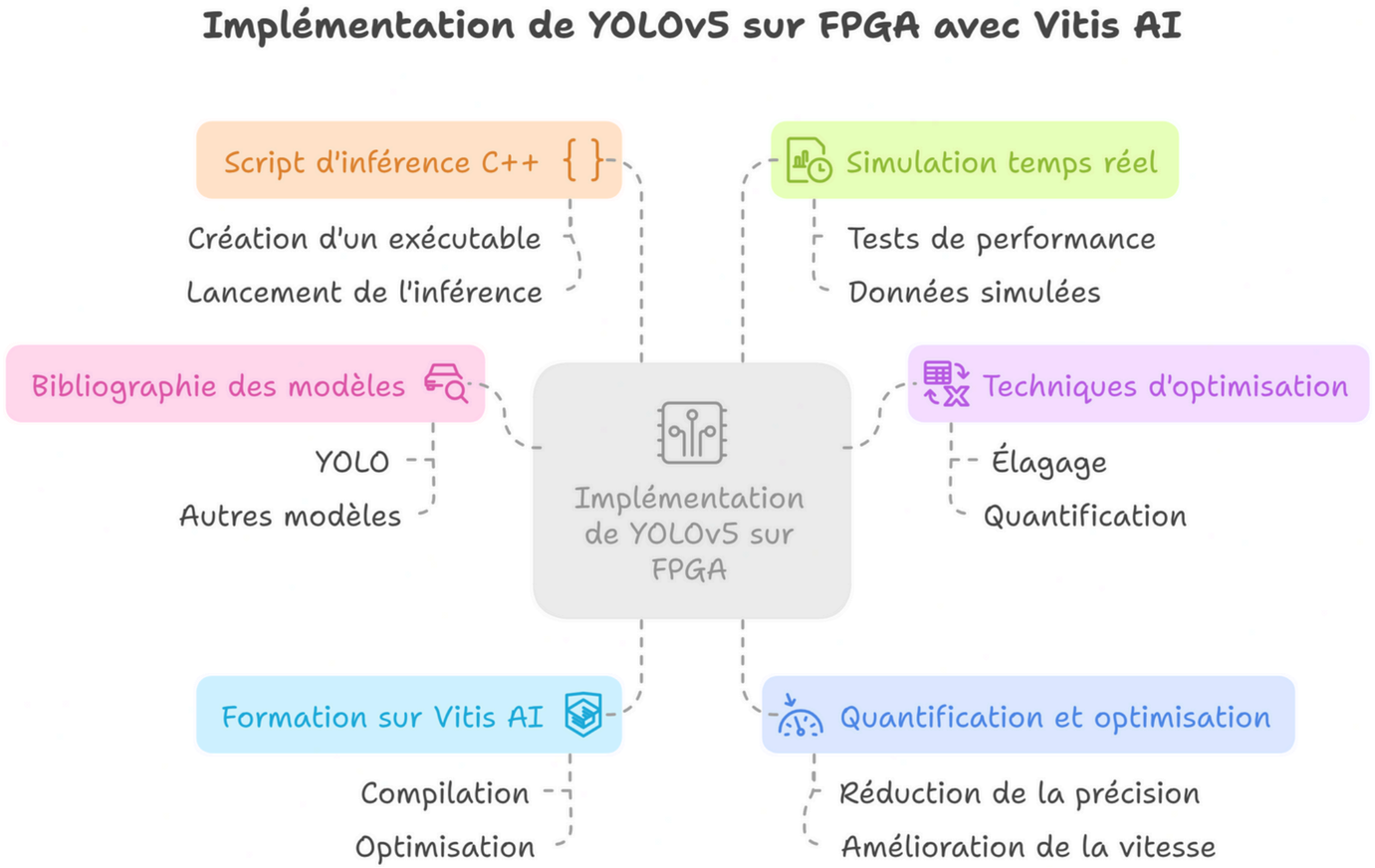


23 Avril 2025

UQAR
Université du Québec
à Rimouski

FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

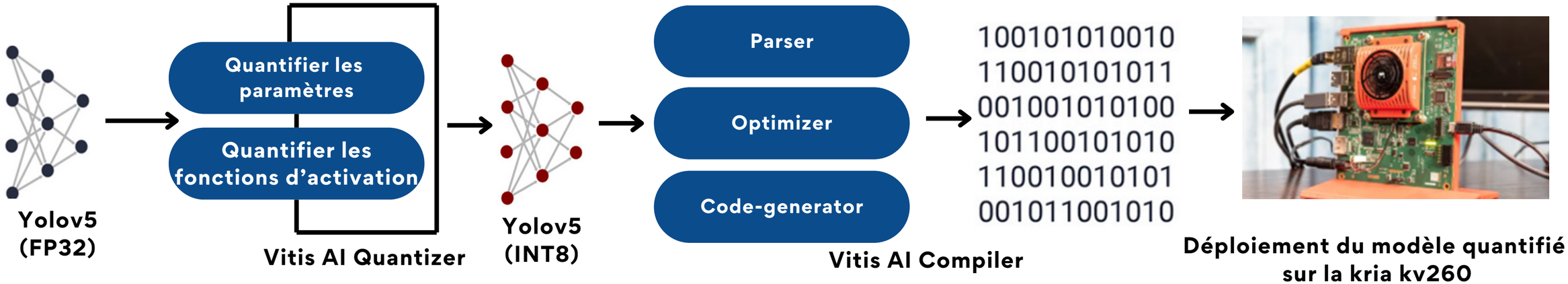
Méthodologie



UQAR
Université du Québec
à Rimouski

FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Méthodologie

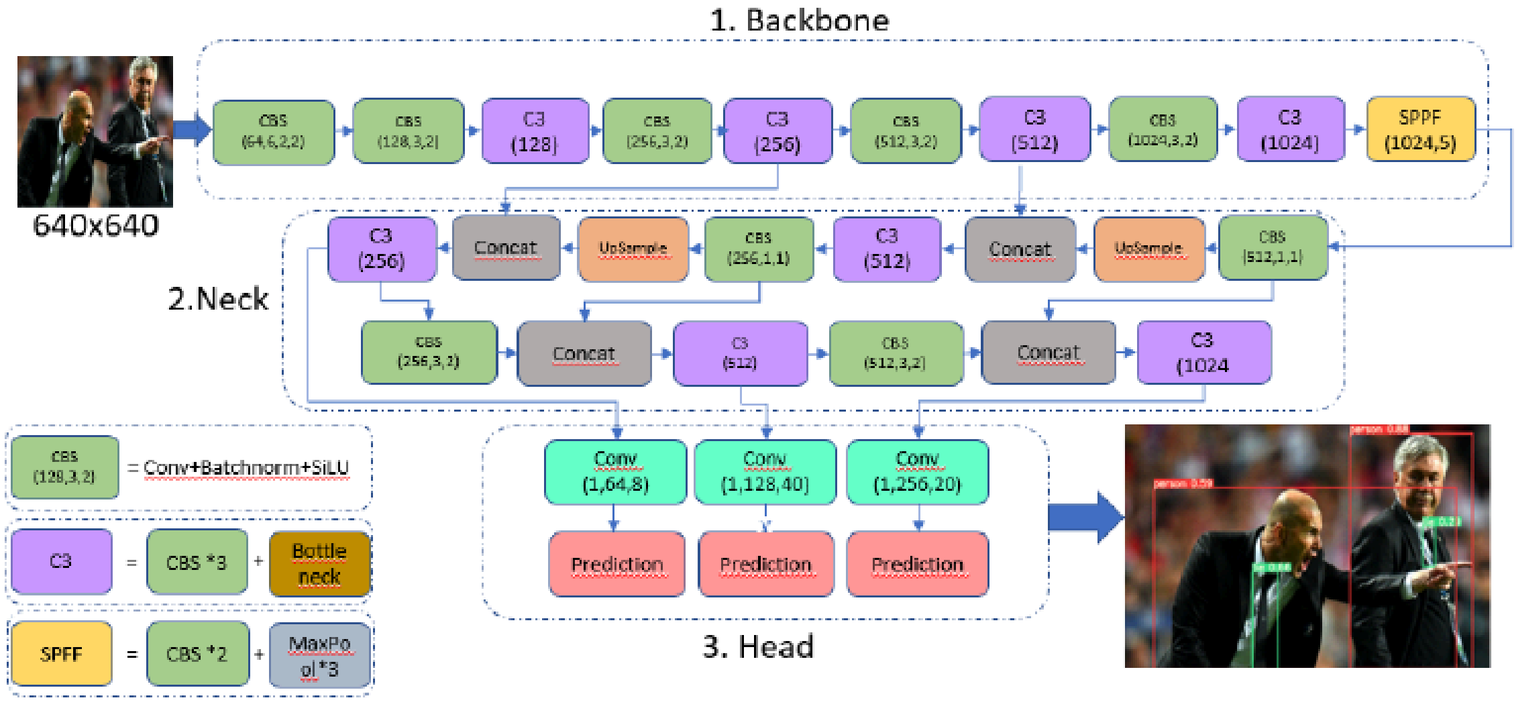


Workflow de l'implémentation du modèle de détection d'objet Yolov5 sur la Kria kv260

Bibliographie

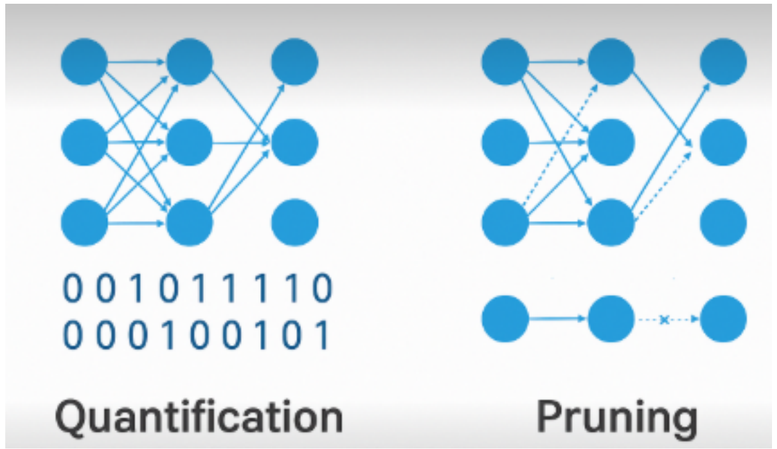
- Étude de différentes architecture de modèle de détection et de classification d'objets.

- Étude des techniques d'optimisation des modèles d'IA (élagage et quantification) et Vitis-AI: Exploration des méthodes de simplification des modèles pour le déploiement sur FPGA.



Formation sur l’outil d’optimisation Vitis-AI

- Apprentissage des outils AMD/Xilinx pour compiler, optimiser et déployer les modèles



Rich Models in PyTorch, TensorFlow and ONNX

Open and Free on GitHub for All Developers

Advanced Optimization, Pruning Included

Retrainable with Custom Dataset

UQAR
Université du Québec
à Rimouski

FIE FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Développement d’un script d’inférence C++

- Développement d’un programme exécutable qui permet de lancer l’inférence sur la carte FPGA.

- Utilisation de librarie Gistreamer pour la capture vidéo temps réel en 720p

UQAR
Université du Québec
à Rimouski

FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Résultats

23 Avril 2025

UQAR
Université du Québec
à Rimouski

FIE FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Résultats

1. On obtient 65 FPS avec 1,2 W de consommation.
2. Très proche des performances GPU (118 FPS / 42 W).
3. Faible latence (16 ms/image).
4. Modèle plus léger et adapté au edge computing.

Tableau 1 – Évaluation des plateformes matérielles sur une base de données d'images

Métrique d'évaluation	Processeur (CPU)	Carte Graphique (GPU)	Kria kv260 (FPGA)
Puissance (W)	-	42.61	1.213
Image par seconde (FPS)	60	118	65
Temps d'execution (ms)	15.9	7.5	16
Quantification	32 FP	32 FP	8 Bits
Taille de l'image (pixels)	640x640	640x640	640x640

UQAR
Université du Québec
à Rimouski

FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Résultats

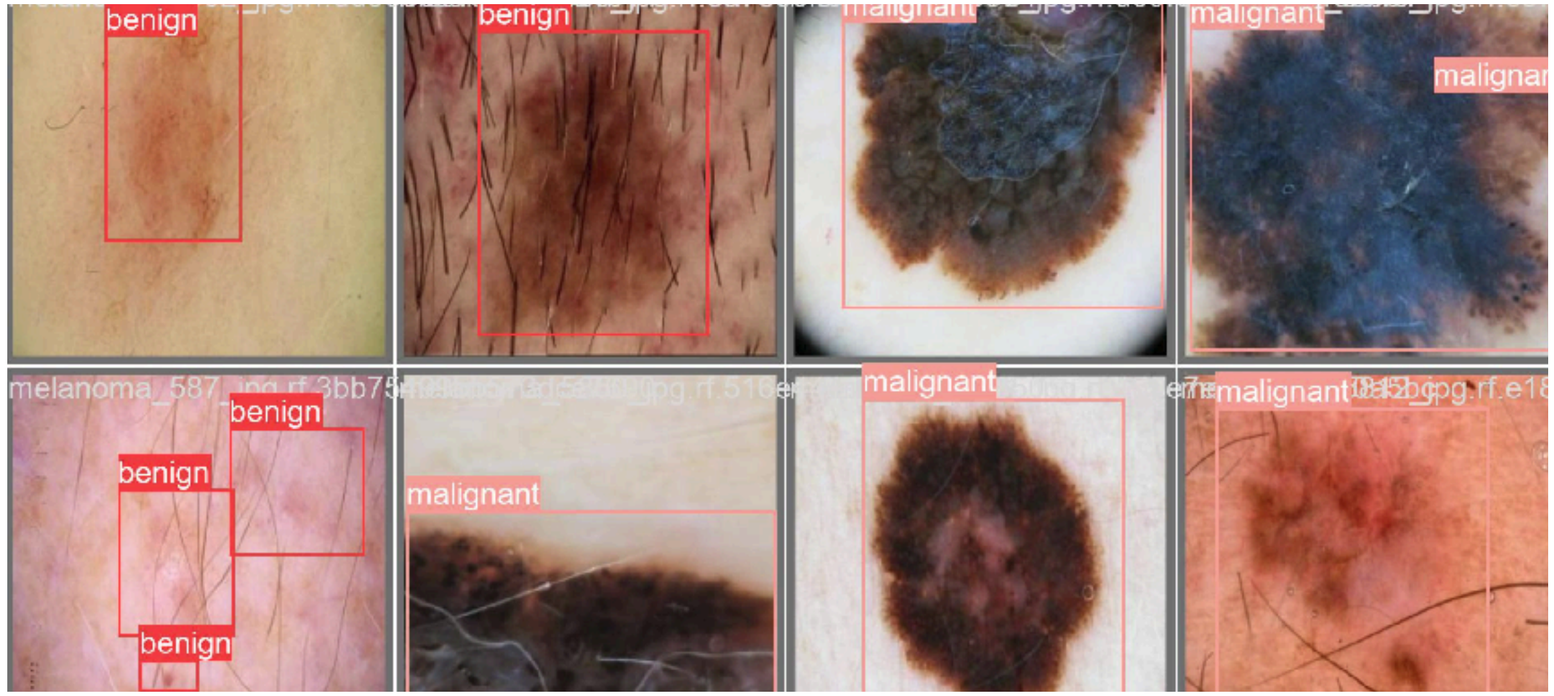


Figure 2 -Prédictions du modèle YOLOv5 sur des images de mélanome et de lésions bénignes

UQAR
Université du Québec
à Rimouski

FIE FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Résultat

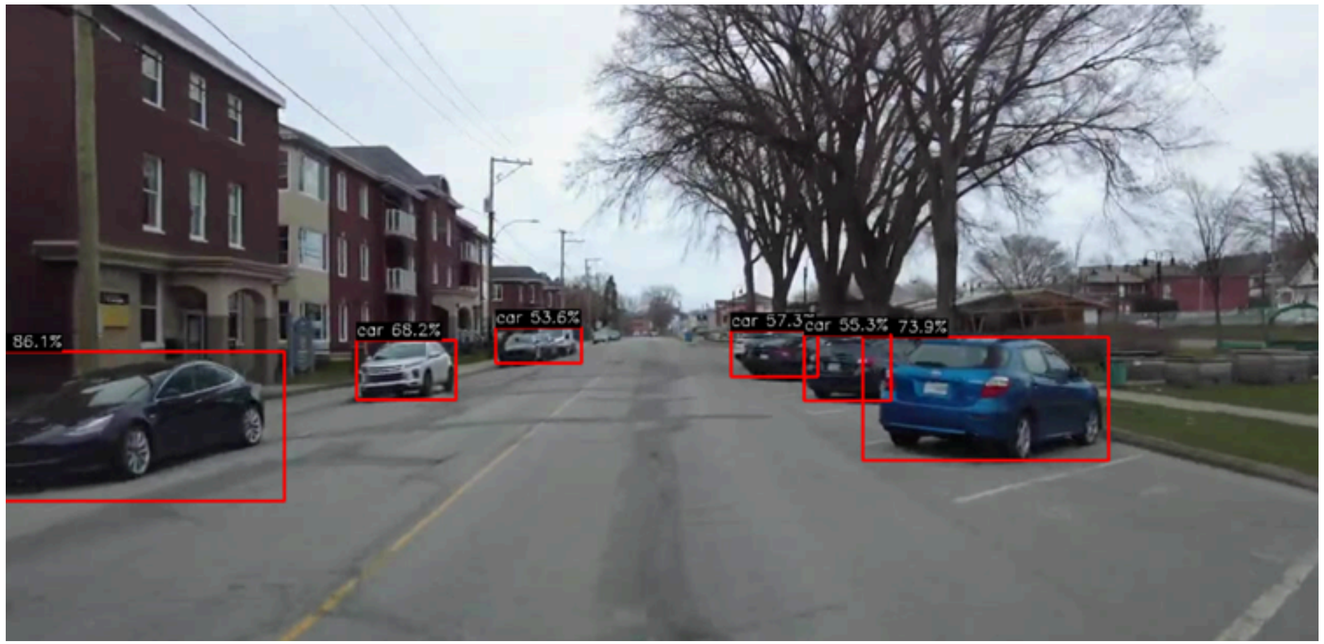


Figure 1 – Inférence temps réel de YOLOv5 optimisé sur FPGA Kria KV260 avec une séquence vidéo de Rimouski

- Latence moyenne basse : L'inférence en temps réel affiche une latence de 7,025 ms par image, équivalente à une cadence théorique de 142 FPS.
- Accès mémoire optimisé : Les transferts atteignent régulièrement 1 GB/s en lecture et 800 MB/s en écriture, sans goulots d'étranglement détectés.
- Limites CPU : Les étapes de pré- et post-traitement, exécutées sur le CPU (jusqu'à 18,1 ms), freinent la performance globale du pipeline.

UQAR
Université du Québec
à Rimouski

FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Conclusion

- Le FPGA est une plateforme efficace pour l'IA embarquée
- Réduction importante de la consommation énergétique.

Perspectives

- Fusion de couches
- Support de modèles plus complexes
- Applications réelles à venir

UQAR
Université du Québec
à Rimouski

FIE FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

Références

1. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” 2016. [Online]. Available: https://arxiv.org/abs/1506.02640.
2. J. Zhang, M. Tian, Z. Yang, J. Li, and L. Zhao, “An improved target detection method based on yolov5 in natural orchard environments,” Computers and Electronics in Agriculture, vol. 219, 4 2024.
3. AMD/Xilinx, “Vitis AI 3.0 user guide (UG1414),” https://docs.amd.com/r/3.0-English/ug1414-vitis-ai, AMD, 2023, accessed: 2025-03-31.
4. R. Gong, X. Liu, S. Jiang, T. Li, P. Hu, J. Lin, F. Yu, and J. Yan, “Differentiable soft quantization: Bridging full□precision and low-bit neural networks,” 8 2019. [Online]. Available: http://arxiv.org/abs/1908.05033.
5. K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, “Panet: Few-shot image semantic segmentation with prototype alignment,” 8 2019. [Online]. Available: http://arxiv.org/abs/1908.06391
6. T. Saidani, R. Ghodhbani, A. Alhomoud, A. Alshammari, H. Zayani, and M. B. Ammar, “Hardware acceleration for object detection using yolov5 deep learning algorithm on xilinx zynq fpga platform,” Engineering, Technology and Applied Science Research, vol. 14, pp. 13 066–13 071, 2 2024.
7. T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, “Pruning and quantization for deep neural network acceleration: A survey,” Neurocomputing,vol. 461, pp. 370–403, 10 2021.

UQAR
Université du Québec
à Rimouski

FORUM INNOVATION
INGÉNIERIE | INFORMATIQUE |
ENTREPRENEURIAT | UQAR

MERCI DE VOTRE ATTENTION

- Je suis disponible pour vos questions.

FUQAR, 23 Avril 2025