Leveraging Self-Attention for Onboard Object Detection in Raw Satellite Images

Adrien Dorise[1,2], Marjorie Bellizzi[2]

[1] CNES, France
[2] IRT Saint-Exupéry, France

adrien.dorise@cnes.fr

Résumé

Avec l'augmentation de la puissance de calcul embarquée sur satellite, le traitement à bord en temps réel des images capturées par les satellites devient une réalité. Notamment, la détection d'objet à bord permet un gain en réactivité conséquent. Cependant, les images brutes sont bruitées et dégradées, impactant la performance des modèles de détection. Ces travaux explorent une nouvelle architecture de détection d'objet, en combinant les avantages des réseaux convolutionnels et des transformers. Ce nouveau modèle est testé sur FPGA, avec des performances prometteuses, ouvrant la possibilité d'utiliser le mécanisme d'attention à bord des systèmes spatiaux.

Mots-clés

Detection d'objet, IA bord, Self-attention, Données brutes

Abstract

With the increase in computing power on-board satellites, real-time processing of images captured by satellites becomes a reality. In particular, the detection of objects on board allows a significant gain in reactivity. However, raw images are noisy and degraded, impacting the performance of detection models. This work explores a new architecture for object detection, combining the advantages of convolutional networks and transformers. This new model is tested on FPGA, with promising performance, opening the possibility of using the attention mechanism on board space systems

Keywords

Object detection, Embedded AI, Self-attention, Raw data

1 Introduction

La croissance régulière des ressources de calcul embarquées entraine un changement de paradigme vers l'intégration de capacités de prise de décision directement au sein des plateformes satellites. Des tâches telles que la détection de nuage [3], la classification des images [15], la détection d'anomalies [9] et la détection d'objets [10, 16] sont de plus en plus déployées à bord afin de réduire les besoins en bande passante et la latence de la liaison descendante.
La détection d'objets a traditionnellement été dominée par les réseaux convolutifs [19]. Plus récemment, les architectures basées sur des transformers ont démontré une puissance de représentation améliorée en vision par ordinateur [17]. Cependant, les mécanismes d'attention posent des défis importants pour le déploiement à bord. Leur dépendance aux opérations softmax, aux multiplications matricielles de grande dimension et à la complexité quadratique par rapport au nombre de tokens rend l'intégration dans des environnements matériels à ressources limitées particulièrement exigeante [20].

Nous soutenons que les mécanismes d'attention peuvent considérablement améliorer la détection d'objets embarqués sur des systèmes spatiaux, à condition qu'ils soient reformulés pour satisfaire aux contraintes de déployabilité. Dans ces travaux, nous introduisons VisDPUFormer, une formulation d'attention adaptée aux accélérateurs FPGA. L'architecture TriCCOT (Tri-part Convolutional Conformal Transformer) proposée combine VisDPUFormer avec un Region Proposal Network (RPN) [18] et un prédicteur conforme [2] afin de fournir un pipeline de détection déployable et fiable.

2 Travaux antérieurs

La détection d'objets par télédétection a considérablement progressé. Les approches standard incluent des CNN à deux étapes [18] et des variantes YOLO à une étape [19]. Les détecteurs basés sur des transformers comme DETR [5], construits sur l'auto-attention [20], atteignent une grande précision mais avec un coût de calcul plus élevé.

Les déploiements sur FPGA dans la télédétection ciblent principalement les CNN [11, 4], tandis que l'accélération des transformers nécessite souvent des conceptions spécialisées [12]. De plus, des travaux ont montré que les images brutes disponibles à bord des satellites impactent les performances des modèles de détection [6, 8]. Notre hypothèse est que les méthodes d'attentions sont plus robustes aux dégradations. Récemment, des transformers convolutifs ont émergé [21], mais n'ont pas été testés sur des dispositifs FPGA. En reformulant l'attention à travers des projections convolutionnelles et des statistiques de canaux globaux, notre méthode aligne la modélisation de type transformers avec les pipelines d'accélération orientés CNN standard.

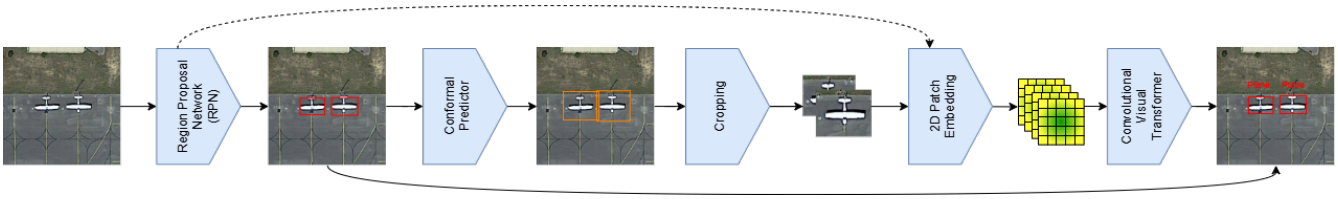

FIGURE 1 – L'achitecture de TriCCOT, composé d'un RPN, d'un prédicteur conforme, et de notre classifieur basé attention.

3 TriCCOT : Tri-part Convolutional Conformal Transformer

TriCCOT est un modèle de détection d'objet en trois étapes : Une détection par boîte englobante utilisant un Region Proposal Network (RPN) convolutionnel, un affinage des boîtes englobantes grâce à la prédiction conforme, et une classification des objets grâce à notre mécanisme de self-attention. La Figure 1 montre les différents modules de TriCCOT.

Region Proposal Network RPN : Le RPN est un module qui effectue une détection d'objets sans classification avec la création des boîtes englobantes. Dans ces travaux, nous nous inspirons de modèles de type YOLO [19] qui propose une architecture pyramidale [14] couplée à un système d'ancrage sous forme de quadrillage.

Prediction conforme : Les boîtes englobantes peuvent couper des objets ou manquer de contexte. Nous appliquons donc la prédiction conforme [1] pour obtenir des boîtes calibrées :

$$\mathbb{P}\left\{Y_{\text{new}} \in \widehat{C}_{\alpha}(X_{\text{new}})\right\} \geq 1-\alpha. \tag{1}$$

Cela garantit qu'en moyenne, une fraction $1-\alpha$ d'objets est entièrement couvert, tout en restant rapide à inférer.

VisDPUFormer : Un transformer sur FPGA : L'architecture des transformers n'étant pas compatible avec les FPGA, nous avons redéfini le mécanisme d'attention. Dans un premier temps, les vecteurs d'embeddings sont remplacés par des matrices 2D de taille prédéfinie. Ainsi nous gardons la capacité des transformers de traiter des patches de tailles variées, tout en contraignant les configurations possibles pour faciliter le calcul matriciel sur FPGA. Ensuite, le mécanisme d'attention est défini comme un système de filtrage des attributs :

$$\mathbf{A} = \sigma_{\text{hs}}(\mathbf{Q} \odot \mathbf{K}) \tag{2}$$

Avec σ_{hs} la fonction hardigmoid, $\mathbf{Q}$ la matrice de *Query*, $\mathbf{K}$ la matrice de *Key*, et $\odot$ la multiplication matricielle élément par élément. Le tenseur résultant $\mathbf{A} \in \mathbb{R}^{B\times C\times H\times W}$ agit comme une porte d'attention bornée. La sortie de ce mécanisme d'attention est obtenue par multiplication avec la matrice *Value* $\mathbf{V}$:

$$\mathbf{U} = \mathbf{A} \odot \mathbf{V} \tag{3}$$

4 Résultats

Performances de détection : TriCCOT est testé sur le dataset public DIOR [13], dont les résultats sont disponibles

TABLE 1 – Résultats de détection de TriCCOT.

Dataset	mAP@50	mAP@50-95	Precision	Recall
DIOR original	75.3%	61.8%	72.8%	54.15%
DIOR dégradé	70.2%	56.1%	83.6%	86.7%

TABLE 2 – Performance des modules de TriCCOT sur une Versal VCK190 pour un patch de taile 640x640 pixels.

Module	Param.	Latence	FPS	Conso. (W) ↓	Pxls/s/W↑
RPN	2.1M	85.5ms	-	-	-
Pred. conforme	N.A	7.2ms	-	-	-
VisDPUFormer	400k	11.7ms	-	-	-
TriCCOT	2.5M	106.7ms	28.2	23	502 924

Table 1. Afin d'évaluer les performances en condition dégradée, un bruit gaussien ainsi que du flou sont ajoutés sur les images [7]. On constate que les performances de TriCCOT sur DIOR sont très bonnes, avec un mAP@50 de 75.3%, confortant la validité de l'approche. Ensuite, malgré une légère baisse de performance sur les images bruitées, passant à 70.2% de mAP@50, TriCCOT propose toujours des performances satisfaisantes, avec une forte précision dans ses prédictions (83.6%).

Performances hardware : Les performances hardware de TriCCOT sont évaluées sur la Xilinx Versal Core dont les résultats sont visibles Table 2. Grâce aux modifications apportées au mécanisme d'attention, le module RPN ainsi que VisDPUFormer exploitent pleinement le DPU de la Versal, avec une inférence temps réel, montant jusqu'à 28 FPS pour un patch 640x640 pxls. De plus le faible nombre de paramètres permet de limiter la consommation, avec la Versal qui n'a jamais dépassé les 23 watts lors des inférences.

Ces résultats montrent que l'approche de TriCCOT est valide pour traiter des images brutes directement à bord des satellites et ouvrent de nouvelles possibilités pour le traitement bord.

5 Conclusion

Nous avons présenté VisDPUFormer, un mécanisme d'auto-attention conçu pour améliorer la déployabilité des modèles de transformers sur les plateformes FPGA. Nous avons intégré VisDPUFormer avec un réseau de proposition de région (RPN) basé sur YOLO et un prédicteur conforme. Ce nouveau modèle appelé TriCCOT permet d'effectuer la détection d'objets sur l'imagerie spatiale brute. Les résultats expérimentaux montrent que TriCCOT performe efficacement, que ce soit en détection ou en vitesse d'inférence.

Dans de futurs travaux, TriCCOT sera testé sur d'autres cas d'études, comme la détection de navires en temps réel, ou la détection d'objets sur drone.

Références

[1] Leo Andeol, Thomas Fel, Florence de Grancey, and Luca Mossina. Confident object detection via conformal prediction and conformal risk control : an application to railway signaling. In Harris Papadopoulos, Khuong An Nguyen, Henrik Boström, and Lars Carlsson, editors, *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pages 36–55. PMLR, 13–15 Sep 2023.

[2] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction : A gentle introduction. *Found. Trends Mach. Learn.*, 16(4) :494–591, March 2023.

[3] Cesar Aybar, Gonzalo Mateo-García, Giacomo Acciarini, Vít Růžička, Gabriele Meoni, Nicolas Longépé, and Luis Gómez-Chova. Onboard cloud detection and atmospheric correction with efficient deep learning models. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17 :19518–19529, 2024.

[4] Jacob Brown, Colton Yates, Jeffrey Goeders, and Michael Wirthlin. Rareplanes detection using yolov5 on the versal adaptive soc. In *2025 IEEE Aerospace Conference*, pages 1–9, 2025.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[6] Roberto Del Prete, Gabriele Meoni, Manuel Salvoldi, Domenico Barretta, Maria Daniela Graziano, Nicolas Longépé, and Alfredo Renga. Enhanced maritime monitoring via onboard processing of raw multi-spectral imagery by deep learning. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 1713–1717, 2024.

[7] Adrien Dorise, Marjorie Bellizzi, Adrien Girard, Benjamin Francesconi, and Stéphane May. Explaining raw data complexity to improve satellite onboard processing. In *2025 European Data Handling and Data Processing Conference (EDHPC)*, pages 1–8, 2025.

[8] Adrien Dorise, Marjorie Bellizzi, and Omar Hlimi. Rethinking satellite image restoration for onboard ai : A lightweight learning-based approach, 2026.

[9] Thomas Goudemant, Benjamin Francesconi, Michelle Aubrun, Erwann Kervennic, Ingrid Grenet, Yves Bobichon, and Marjorie Bellizzi. Onboard anomaly detection for marine environmental protection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17 :7918–7931, 2024.

[10] Thomas Goudemant, Benjamin Francesconi, Houssem Farhat, Lionel Daniel, Olivier Thiery, Erwann Kervennic, Adrien Girard, and Seif Mzoughi. Détection de navires embarquable à bord de satellites. In *Actes de la 4ème Conference on Artificial Intelligence for Defense (CAID 2022)*, Actes de la 4ème Conference on Artificial Intelligence for Defense (CAID 2022), Rennes, France, November 2022. DGA Maîtrise de l'Information.

[11] Younis Ibrahim, Li Chen, and Tian Haonan. Deep learning-based ship detection on fpgas. In *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 454–459, 2022.

[12] Ehsan Kabir, Jason D. Bakos, David Andrews, and Miaoqing Huang. A runtime-adaptive transformer neural network accelerator on fpgas. *Microprocessors and Microsystems*, 120 :105223, February 2026.

[13] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images : A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159 :296–307, January 2020.

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017.

[15] Gabriele Meoni, Marcus Märtens, Dawa Derksen, Kenneth See, Toby Lightheart, Anthony Sécher, Arnaud Martin, David Rijlaarsdam, Vincenzo Fanizza, and Dario Izzo. The ops-sat case : A data-centric competition for onboard satellite image classification. *Astrodynamics*, 8(4) :507–528, 2024.

[16] Janne Mäyrä, Elina A. Virtanen, Ari-Pekka Jokinen, Joni Koskikala, Sakari Väkevä, and Jenni Attila. Mapping recreational marine traffic from sentinel-2 imagery using yolo object detection models. *Remote Sensing of Environment*, 326 :114791, 2025.

[17] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks ? In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn : Towards real-time object detection with region proposal networks, 2016.

[19] Ultralytics. YOLOv5 : A state-of-the-art real-time object detection system. https://docs.ultralytics.com, 2021. Accessed : 12/12/2025.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[21] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt : Introducing convolutions to vision transformers, 2021.