

卒研中間発表

~KV260におけるYOLO物体検出モデル最適化手法の再現実験と有効性評価の検討~

電気通信大学 I 類

コンピュータサイエンスプログラム 吉永研究室

東盛 祥大

はじめに

◆ 背景

- ・ドローンや監視カメラなどのエッジデバイス上で物体検出(入力に画像や動画を与え、予め学習しておいた物体を認識する)を行う需要が高まっている。
- ・エッジデバイスでは計算資源や消費電力が限られるため、並列処理や専用回路による高速・省電力な処理が可能なFPGAへのモデル実装が研究されている。
- ・検出精度を重視する場合
モデルサイズ・計算量が大きい
→ 推論速度が低い
- ・推論速度を重視する場合
モデルサイズが小さく、特徴抽出能力が低い
→ 検出精度が低い



図1：物体検出の様子

◆ これまでに取り組んだ手法

- ・モデルの並列処理化による推論速度改善(参考文献[1])
- ・モデルの構造変更による検出精度改善(参考文献[2])

関連研究

- ◆ YOLOv3-Tinyを対象に最適化手法を提案した先行研究(参考文献[3])
- ◆ 主な手法
 - ・ **反復的プルーニング**
重要度の低いフィルタを段階的に削除し、再学習
→ モデルを軽量化
 - ・ **QAT(Quantization-Aware Training)**
FP32からINT8への量子化を考慮して学習
→ 量子化による精度低下を抑制
 - ・ **メモリマッピングコントローラー**
BRAM・URAMなどのオンチップメモリにデータを効率的に配置
→ メモリアクセスを効率化し、推論速度を向上

関連研究との比較

表1 先行研究の結果とこれまでの実験結果

	物体検出モデル	データセット	検出精度 mAP(%)	検出精度 accuracy(%)	推論速度(FPS)
ベースモデル1	YOLOX-Nano	COCO 2017	17.4	—	110.4
ベースモデル1 +構造変更	YOLOX-Nano	COCO 2017	19.1	—	81.1
ベースモデル2	YOLOv3-Tiny	COCO 2017	7.9	—	4.7
ベースモデル2 +並列処理化	YOLOv3-Tiny	COCO 2017	7.9	—	19.1
先行研究の 手法	YOLOv3-Tiny	PASCAL VOC 2012	—	93.8	24.3

- ◆ 先行研究と本研究では使用するデータセットや評価指標が異なる
→ 報告された性能を単純に比較することはできない

実装

◆ 本研究の目的

先行研究の手法を本研究と同一の実験条件に適用し，検出精度および推論速度への影響を評価する．

◆ FPGAボード：AMD Kria™ KV260 Vision AI Starter Kit

AMD Zynq™ UltraScale+™ MPSoC

— Arm Cortex-A53クアッドコアAPU

→ 推論前処理・後処理

— PL(Programmable Logic)

└ DPU(Deep Learning Processing Unit)

→ 推論処理



図2：KV260

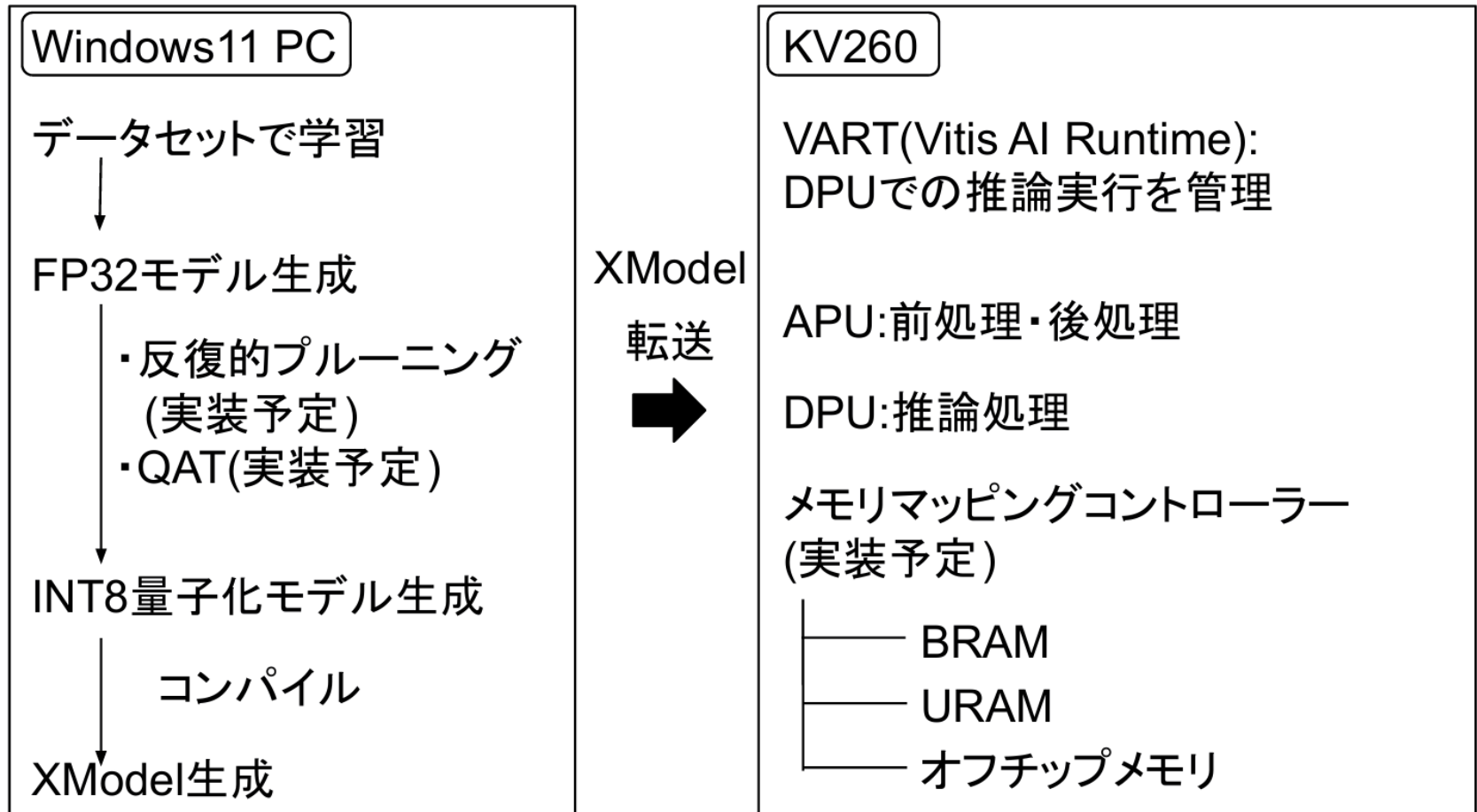


図3：学習からKV260上で推論実行までの流れ

今後の方針

- ◆ 先行研究の手法を実装
 - ・ 反復的プルーニング
 - ・ QAT
 - ・ メモリマッピングコントローラー
- ◆ データセットや評価指標などの実験条件を統一し、先行研究の手法による検出精度・推論速度への影響を評価
- ◆ 評価結果をもとに、先行研究の手法を改良し、評価する

参考文献

- [1] 前野龍輝, 「KV260におけるYOLO物体検出AI推論の実装と性能向上」, 電気通信大学 情報理工学域 情報・ネットワーク工学専攻, 学士論文, 2026
- [2] W. Lei, Y. Gao, M. Sun, X. Li, and X. Liang, "Edge-Constrained UAV Small-Object Detection with P2 Enhancement and Quantum-Inspired Lightweight Structure Search," arXiv preprint arXiv:2606.09081, 2026
- [3] J. S. Fata and W. M. Elmannai, "Low-Cost FPGA-Enhanced CNN Accelerator for Real-Time YOLO Object Detection and Classification," IEEE Access, vol. 14, 2026