4K-03

# 物体検出アプリケーションによる新型エッジデバイスの性能評価

上山 大和† 近藤 鯛貴† 竹田 大将† 佐藤 裕幸†
岩手県立大学大学院ソフトウェア情報学研究科†

## 1 序論

通信技術の発展によりあらゆる分野で組み込みコンピューティングが活用されるようになった．エッジコンピューティングの需要が高まり，リアルタイム性が要求されると通信遅延が問題として注目されはじめる．そこで高性能化していくエッジデバイスに処理を任せて，通信の回数を減らそうという考えが生まれた．高性能化したエッジデバイスで実行させる処理はより高度かつ複雑に進化していき，今ではエッジデバイスは人工知能の推論モデルの主要キャリアとしても活躍している．

エッジデバイスによる推論モデルの実行の需要が増えると，低消費電力性かつ高性能なエッジデバイスが求められるようになった．推論モデルがリアルタイム実行可能なエッジデバイスの需要に対して，NVIDIA はGPU付きのエッジデバイス，NVIDIA Jetson シリーズを提供しており，NVIDIA Jetson シリーズは推論モデルの高速化に効果的な CUDA が使用できる GPU[1]を備えている．

本稿では推論モデルをベンチマークとして複数のエッジデバイスで実行することで性能評価を行う．

## 2 推論モデル

推論モデルには自然言語処理や画像認識，データ分析など用途に合わせて多様なモデルが存在する．本稿では，エッジコンピューティング需要が特に高いと思われる画像認識分野の物体検出を行う推論モデルを取り上げる．ベンチマークとして採用する物体検出は YOLOv7 である．

### 2.1 エッジデバイスにおける物体検出

エッジデバイスにおける物体検出は，カメラで撮影した画像をサーバに送信し，サーバ側で検出させることが主流だった．しかし，サーバとの通信が通信遅延や情報漏洩などのリスクにつながる可能性が考えられ，リアルタイム性が求められるようになり，リアルタイムで推論可能な高性能エッジデバイスの需要が高まった．

推論モデルに適したエッジデバイスには，エネルギー効率に優れた FPGA を基にしたものやディープニューラルネットワークに特化したアーキテクチャを持つ ASIC を基にしたものなどがあるが，本稿では，NVIDIA の GPU サポートが得られる Jetson デバイスをいくつか用意し，性能評価を行う．GPU サポートを得られるエッジデバイスを採用した理由は，アーキテクチャが市販のコンピュータに近く，汎用性に優れているためである．汎用性が高いため，新しい物体検出アプリケーションが提案されても Jetson デバイスは対応することができる．本稿では，性能評価により Jetson デバイス間の比較や特性を考察する．

### 2.2 YOLOv7

YOLO とは，画像全体から単一のニューラルネットワークによってバウンディングボックスとクラス確率を 1 回の推論で予測する物体検出の手法のことである[2]．他の物体検出の手法と異なる点は画像そのものから 1 回の推論で処理できる点にある．本稿でベンチマークとして採用する YOLOv7 は 2022 年に発表された YOLO の後継的存在である．YOLOv7 は従来の手法よりエッジデバイスでの推論を意識した存在である．YOLOv7 の提案者はパラメータと計算量の削減をすることで YOLOv5 と比較して最大 120%速いと主張している．また，YOLO の中でも高い精度を実現しているとも主張している[3]．今回使用する YOLOv7 は PyTorch によるフレームワークである．

## 3 物体検出によるエッジデバイスの性能評価

本稿の主な調査対象は NVIDIA Jetson デバイスの中でも新しい NVIDIA Jetson AGX Orin であり，ほかのデバイスは Orin の比較対象である．

性能評価には表 1 のデバイス群を利用する．ベンチマークには YOLOv7 を採用し，学習済みモデルは MS COCO データセットを使用し，性能評価に

Performance Evaluation of New Edge Devices with Object Detection Applications
†H. Kamiyama, †T. Kondo, †H. Takeda, †H. Sato
‡Graduate School of Software and Information Science, Iwate Prefectural University

表 1　性能評価を行う NVIDIA Jetson デバイス

| デバイス | AGX Orin | AGX Xavier | TX2 |
|---|---|---|---|
| CPU | 12core ARM v8.2 64bit CPU | 8core ARM v8.2 64bit CPU | Dual core Denver and Quad core A57 |
| GPU | 2048core Ampere GPU with 64 Tensor cores | 512core Volta GPU with 64 Tensor cores | 256core Pascal GPU |
| RAM | 32GB LPDDR5 | 32GB LPDDR5 | 8GB LPDDR4 |
| Storage | 64GB eMMC | 32GB eMMC | 32GB eMMC |
| OS | Ubuntu 20.04.5 LTS | Ubuntu 18.04.5 LTS | Ubuntu 18.04.6 LTS |
| CUDAver | CUDA11.4 | CUDA10.2 | CUDA10.2 |
| PyTorch | 1.12.0a0+2c916ef.nv22.3 | 1.7.0 | 1.8.0 |
| Power | 15W~60W | 10W~30W | 7.5W~15W |
| リリース | 2022 年 5 月 | 2018 年 10 月 | 2017 年 4 月 |

使う画像は PASCAL VOC2007 のトレーニングデータからランダムに 5 枚選び，サイズは長辺に 640pixel を持つ矩形に統一した．画像 5 枚の推論を 1 ラウンドとして 4 ラウンド連続で行う．推論時間を安定させるために最初のラウンドはウォームアップとして，残り 3 ラウンドの平均推論時間を結果とする．ベンチマーク中は Tegrastats でメモリ使用率を，ワットチェッカーで最大消費電力を計測する．

## 4　結論

表 2 に計測結果を示す．同じメモリ量を持つ Orin と Xavier でもメモリ使用率に大きく差が出ていることが読み取れる．搭載メモリ量の違いから換算しても，使用メモリ量に差が出ている．一般的に推論モデルは GPU の VRAM 上または RAM 上に展開されるため，メモリ量は大きければ大きいほど良いとされるが，CPU や GPU の性能が高い Orin は今回メモリを持て余した．また，Orin は Xavier の最大消費電力より 170%高いが，平均推論時間は 208%向上しており最大消費電力の上昇以上に推論時間が短縮されている．これは GPU アーキテクチャの違いが結果に表れたと考えている．対照的に TX2 は消費電力においては今回計測したデバイスの中では最も優れているのに非常に推論時間が遅く，また，メモリ使用率も最も高い．これは，設計が他より古い CPU と GPU が原因と考えられる．新しい Jetson デバイスの GPU には Tensor コアが搭載されており，Tensor コアは行列計算に適している．そのため比較すると TX2 は低消費電力を抑えている以上に計算性能が低く見えてしまう．

表 2 測定結果

| デバイス | AGX Orin | AGX Xavier | TX2 |
|---|---|---|---|
| 平均推論時間 | 20.9ms | 43.4ms | 233.8ms |
| 平均メモリ使用率 | 18.00% | 42.40% | 68.50% |
| 最大消費電力 | 47.0W | 27.7W | 10.2W |

本稿は，NVIDIA Jetson デバイスの中でも最高性能を誇る新型機種である AGX Orin の性能評価を行うために YOLOv7 をベンチマークとして測定を行った．その結果，CPU コア数の増大や GPU アーキテクチャの発展および Tensor コアの搭載が推論モデルに効果を発揮している結果が得られた．しかし，電力効率の観点から見ると AGX Orin はエッジコンピュータより一般的なコンピュータに近い電力消費を行っている結果を示したため，今後は Jetson のパワーモードを使い，使用可能な電力に制限をかけた状態でどれぐらいの計算性能を保てるのかを調査する．

## 参考文献

[1]A.A.Suzen, B.Duman, and B. Sen, Benchmark analysis of jetson tx2, jetson nano and raspberry pi using deep-cnn, in 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2020, pp.1–5.

[2]J.Redmon, S.Divvala, R.Girshick, A.Farhadi, You Only Look Once: Unified, Real-Time Object Detection,arXiv:1506.02640 [cs.CV].

[3]C.Y.Wang, A.Bochkovskiy, H.Y.M.Liao, YOLOv7:Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, arXiv:2207.02696 [cs.CV].