AIxBoard で OpenVINO™ を使用して YOLOv8 物体検出モデルを高速化する

この記事は、Medium に公開されている「Accelerating YOLOv8 Object Detection Model on AIxBoard with OpenVINO™」の日本語参考訳です。原文は更新される可能性があります。原文と翻訳文の内容が異なる場合は原文を優先してください。

1. はじめに

AIxBoard で OpenVINO™ ツールキットを使用して YOLOv8 分類モデルのデプロイメントと評価を行ってみましょう。この記事では、YOLOv8 物体検出モデルの高速化に注目します。

最初に、この記事で紹介しているサンプルコードのリポジトリーをダウンロードし、OpenVINO™ の推論エンジンを使用した YOLOv8 向けの開発環境をセットアップします。

GitHub* リポジトリー (英語)

git clone コマンドを使用してコードのリポジトリーをクローンします。

2. YOLOv8 物体検出 OpenVINO™ IR モデルをエクスポートする

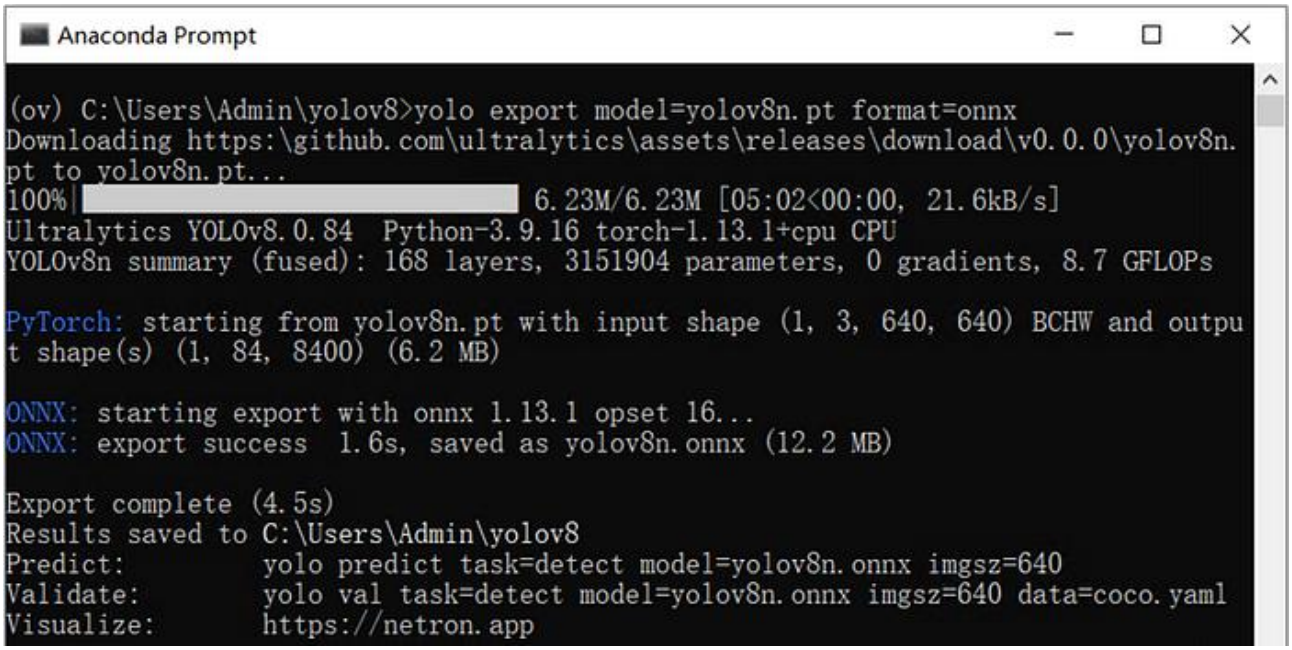
YOLOv8 には、以下の表に示すように、COCO データセットでトレーニングされた 5 つの異なる物体検出モデルがあります。

Model	size (pixels)	mAP^{val} 50-95	Speed CPU ONNX (ms)	Speed A100 TensorRT (ms)	params (M)	FLOPs (B)
YOLOv8n	640	37.3	80.4	0.99	3.2	8.7
YOLOv8s	640	44.9	128.4	1.20	11.2	28.6
YOLOv8m	640	50.2	234.7	1.83	25.9	78.9
YOLOv8l	640	52.9	375.2	2.39	43.7	165.2
YOLOv8x	640	53.9	479.1	3.53	68.2	257.8

次のコマンドを使用して YOLOv8n.onnx モデルをエクスポートします。

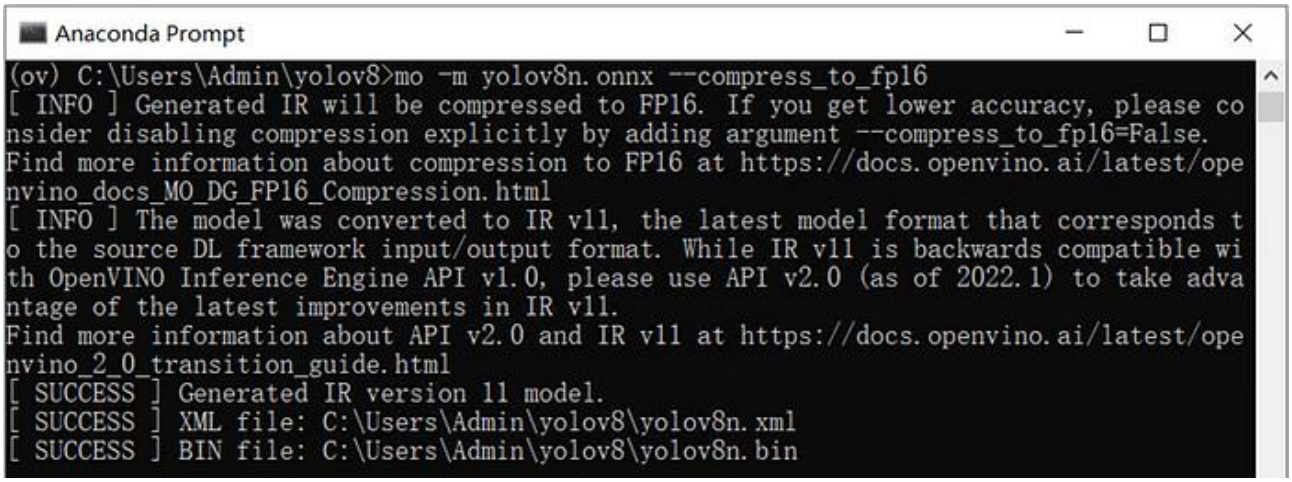
```
yolo export model=yolov8n.pt format=onnx
```

yolov8n.onnx モデルが生成されます。



次のコマンドを使用して、FP16 精度で OpenVINO™ IR 形式モデルを最適化してエクスポートします。

```
mo -m yolov8n.onnx --compress_to_fp16
```



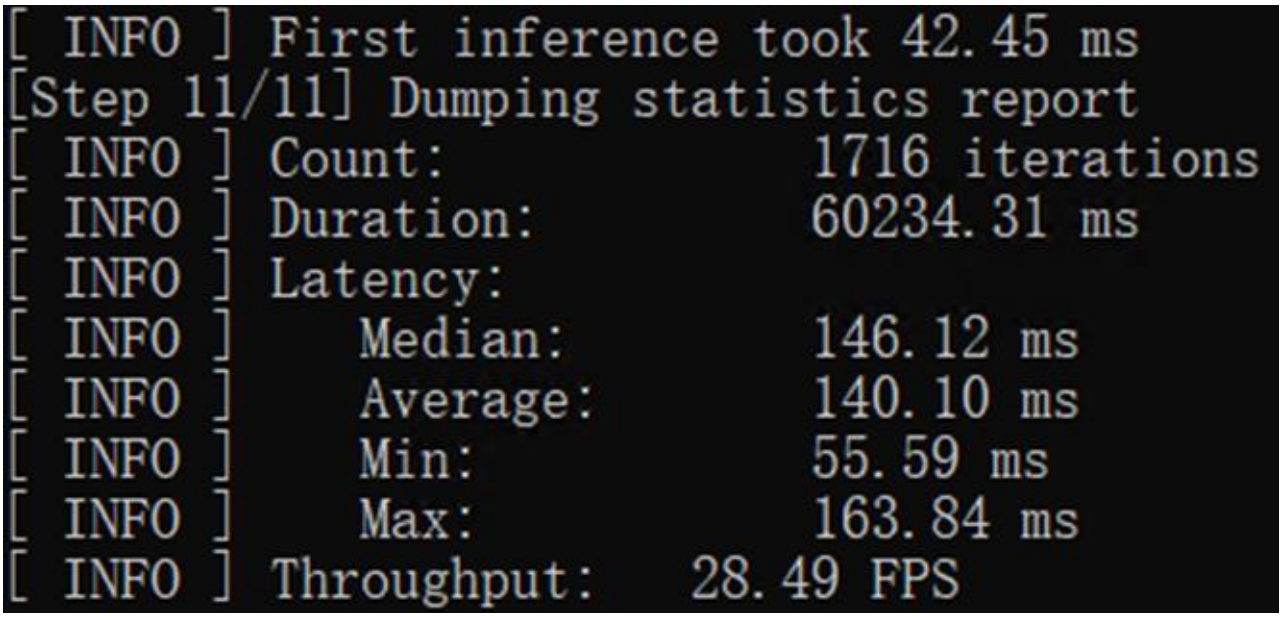
3. benchmark_app を使用して YOLOv8 物体検出モデルの推論パフォーマンスをテストする

benchmark_app は、OpenVINO™ ツールキットで提供されている、AI モデルの推論パフォーマンスを評価するためのパフォーマンス・テスト・ツールです。同期モードまたは非同期モードで、異なる計算デバイスで前処理や後処理を行うことなく、純粋な AI モデルの推論パフォーマンスをテストできます。

次のコマンドを使用します。

```
benchmark_app -m yolov8n.xml -d GPU
```

AIxBoard (英語) の統合 GPU での yolov8n.xml モデルの非同期推論パフォーマンスが表示されます。



4. OpenVINO™ Python* API を使用して YOLOv8 物体検出モデルの推論プログラムを作成する

次の図に示すように、Netron を使用して yolov8n.onnx を開きます。モデルの入力形状は [1,3,640,640]、出力形状は [1,84,8400] です。「84」は、cx、cy、h、w、および 80 クラスのスコアを表します。「8400」は、画像サイズが 640 の場合の YOLOv8 の 3 つの検出ヘッドの出力セル数を示します (80x80+40x40+20x20=8400)。

OpenVINO™ Python* API を使用した YOLOv8 物体検出モデルのプログラム例を次に示します。

コアのソースコードは次のとおりです。

`yolov8_od_ov_sync_infer_demo.py`

yolov8_od_ov_sync_infer_demo.py の結果を次に示します。

5. まとめ

AIxBoard (英語) の統合 GPU (24 実行ユニット) と OpenVINO™ を活用することで、YOLOv8 オブジェクト検出モデルのパフォーマンスを高速化できます。非同期処理 (英語) と AsyncInferQueue (英語) を使用すると、計算デバイスの使用率がさらに向上し、AI 推論プログラムのスループットを向上させることができます。

OpenVINO™ ツールキットとは

AI を加速する無償のツールである OpenVINO™ ツールキットは、インテルが無償で提供しているインテル製の CPU や GPU、VPU、FPGA などのパフォーマンスを最大限に活用して、コンピューター・ビジョン、画像関係をはじめ、自然言語処理や音声処理など、幅広いディープラーニング・モデルで推論を最適化し高速化する推論エンジン/ツールスイートです。

OpenVINO™ ツールキット・ページでは、ツールの概要、利用方法、導入事例、トレーニング、ツール・ダウンロードまでさまざまな情報を提供しています。ぜひ特設サイトにアクセスしてみてください。

https://www.intel.co.jp/content/www/jp/ja/internet-of-things/openvino-toolkit.html