4月号特集で使ったOrin向けプログラムを2倍に高速化

Jetson大実験…trtexec経由でのTensorRT最適化と推論の実行

土井 伸洋

写真1　自動運転向けアプリケーションが動作している
動画像に対して自分の車が走っているレーン領域を抽出する

本記事では，Jetsonシリーズ（エヌビディア）への深層学習モデルのマッピングと高速化について掘り下げます．2023年4月号 特集1「自動運転の学習データ作り＆Jetson研究」の第9章，第10章「Jetson大研究」を前提に話を進めますので，必要に応じて当該記事を参照してください．なお，当該記事を読まなくても楽しめると思います．

＜本記事の前提条件＞

- 対象：ドライブ・レコーダの動画像に対して，白線やレーン領域を抽出するアプリケーション・プログラム（**写真1**）
- 利用している深層学習モデル（TensorFlowで作成）
  - 入力：512×256×3
  - 出力：512×256×4
  - セグメンテーション・モデル
  - エンコード部：Resnet18
  - デコード部：U-net
- アプリケーション・プログラム
  - 深層学習モデルの処理結果視覚化，重畳
- 対象機器：Jetson AGX Orin（メモリ32Gバイト，**写真1**）
  - ベースパッケージ：Jetpack 5.0.2
  - 利用コンテナ：l4t-tensorflow:r35.1.0-tf2.9-py3

最適化ツールtrtexecを使ってみる

Jetson上で深層学習モデルを動作させる際に，ほとんどの場合は動作させる機器向けの最適化を行います．利用するのは，エヌビディアGPUおよびSoC（Jetsonシリーズのこと）向けの最適化ツールであるTensorRTです．このTensorRTには，利用方法が2つあります．

- 深層学習フレームワーク組み込み済みのTensorRT（TF-TRTやTorch-TensorRT）
- エヌビディアの提供するtrtexecコマンド

本稿では，4月号特集では用いなかった後者（trtexecコマンド）を用いて最適化を行います．

● コマンドライン・ツール trtexecのメリット

trtexecは，TensorRTによる最適化を実施するためのコマンドライン・ツールです．API経由でTensorRTによる最適化処理を行うTR-TRTやTorch-TensorRTとは異なり，Linuxのターミナル上で必要なコマンドを実行することで最適化処理を行います．API経由では実施できないような，高度な最適化を行うこともできます．また最適化結果は，フレームワークに沿ったフォーマットではなく，TensorRTに特化されたファイル形式（TensorRT engine）で出力されます．

● trtexecコマンドを用いた最適化フロー

trtexecコマンドを用いた最適化フローを次に示します（**図1**）．

1. 学習済みモデル（SavedModel）をONNX形式（.onnx）に変換する
2. ONNX形式のモデルをtrtexecに入力として与え，最適化された推論エンジン（.trt）を得る

入力には特集記事にてTensorFlowを用いて作成したSavedModel形式のモデルを使います．作業は