第7章 モダンAI計測制御のキモ

1万円GPUボードJetson Nanoに組み込んで動かす

鎌田 智也

エッジAI用「推論」ライブラリの世界

● 学習の次は推論

モダンAI計測制御実験を行うために，DIGITSで作ったロボット・アームをセグメンテーションするDNNモデルをJetson Nanoに組み込んで動かしていきます．

DNNモデルを動かすには，「学習（learning）」と「推論（inferenceあるいはインファレンスという）」の2つのフェーズがあります（**図1**）．

PC上でモデルを開発したとき，学習データセットを使ってモデルを「学習」させました．

一方，テスト時には画像を入力してモデルを使って実際にセグメンテーション処理を行いました．このようなモデルにデータを入力して実際に動作させてみることを「推論」といいます．

● 計測制御向けのエッジAI推論の仕組みが絶賛進化中

Jetsonを含めた組み込みコンピュータを使ったエッジで，教師ありモデルのAI処理を行わせる場合，エッジ上で学習を行わせたいケースというのは極めてまれで，大抵はあらかじめ別のコンピュータ上で学習させた学習済みモデルを組み込んで推論処理のみを行わせることができれば十分です．そのようなニーズに対応するため，エッジで推論処理を高速に行わせるための仕組みが公開され盛んにアップデートされています．

この章では，NVIDIAのGPUボード（Jetson Nanoを含めたJetsonシリーズ）で高速に推論を行わせるための仕組みTensorRTを利用できるライブラリを使って，作成してきたAIモデルをJetsonに実際に組み込んで動作させます．

モダンAI計測制御のキモ「推論」の実装

● 推論は安価で小型なエッジ側で行うのが吉

AIモデルを動かすには「学習」と「推論」という2つのフェーズがあります．

同様に，ディープ・ラーニングAIの開発工程も，「学習」と「推論」のステップに分けて行うことが一般的です．

例えば**図1**のように，膨大な学習データセットを用意し，DNNモデルを設計してハイパフォーマンスなGPUコンピュータを使ってモデルの学習を行わせる開発ステップがあって，その後に，できたモデルをターゲット・マシンに組み込んで推論させる開発ステップ（デプロイ=deployともいう）があります．

DNNモデルの学習の開発段階では，膨大な計算を何度も反復計算しなければならないため，とても時間がかかります．特に教師あり学習を利用したAI開発において，学習は開発段階にのみ行えばよいので，消費電力など気にすることなく高性能で高機能なコンピュータを使って開発します．

その一方，学習済みモデルを使った推論は，できるだけ低消費電力かつ小型のコンピュータで動いてくれたら好都合です．安くて低消費電力で小型のエッジ（末端）コンピュータで処理を完結できれば，幅広い応用が期待できます．エッジで完結できればクラウドとの通信の必要もありません．高度なAIディープ・

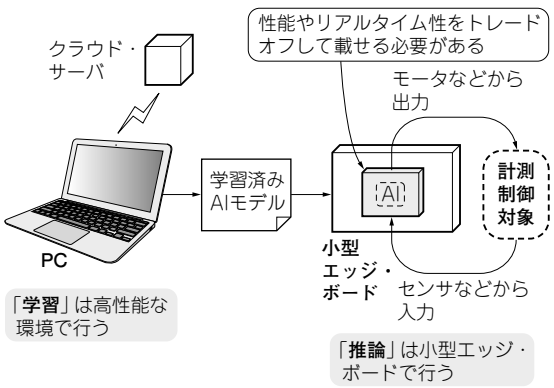

図1　計測制御などのAIでは高性能GPUコンピュータなどで「学習」させた学習済みモデルを小型エッジ・ボードに載せ替えて「推論」させるのが一般的