# エッジデバイスへの組込Deep Learningの現状と課題

AlexNetが火を付けたDeep Learningブームは GoogleやMicrosoftのようなクラウドでの活用から徐々にエッジデバイスに浸透しつつある。ここでは、エッジデバイスでのDeep Learningを使用した推論の現状と、学習の課題について考察する。

株式会社Bee 最高技術責任者　中村 仁昭

## エッジでの推論

エッジデバイスでのDeep Learningを使用した推論はクラウド同様、多様なプレイヤーにより様々な方法を試行錯誤している。基本的には、並列化した計算ワークロードに適したハードウェアを使用し効率を上げて多くのユーザーを獲得しようとしている。代表的な例をあげてみよう。

## NVIDIAに代表されるGPUコンピューティング

Deep LearningブームでGoogleとならんで有名なNVIDIAは、車載向けのNVIDIA DRIVE PXやドローンなどより一般向けのJetsonを用意しエッジでGPUを使ったDeep Learningを進めている(fig.1)。一般的に使われるPythonベースのDeep Learningフレームワークは NVIDIAのCUDA環境を基盤として構築されているため、クラウドやオンプレで学習したフレームワークのモデルをそのままエッジで推論に使用できることが利点となる。

しかし、もう少しリソースの少ないIoTに向けたソリューションはまだ存在しない。MITのリサーチプロジェクトでDCNN専用プロセッサのEyerissに関するパネル展示をGTC2016で行なった程度であるが、論文にMIT教授でNVIDIAの研究員のJoel Emer氏が名を連ねているのを見ると専用チップソリューションに注意を払っていると思われる。

## FPGAによる効率化

昨年末ごろからXilinxが公開したBNN-PYNQプロジェクトを利用したZynq SoCをベースにしたPYNQ-Z1(fig.2)ボードが人気だ。PYNQはPythonでディープラーニングを実装しFPGAで並列化できる計算処理を高速に実行できる。また、Pythonから動的にFPGAで実行するPL(Programmable Logic)部分を変更することが可能なため、アプリケーションごとに最適な処理をFPGAにロードさせて実行することができる。PYNQではJupyter Notebookが起動しているためブラウザでアクセスすればすぐにDeep Learningを試すことができるため、大学の授業からDeep LearningをFPGAで試してみたいFPGA初心者まで色々活用されている例がWebで見られる。

Microsoftがクラウド上のDeep Learning基盤をFPGAベースのBrainwaveにしようとしている動きと同じく、エッジデバイスでもGPUでは速度が足りないためFPGAアクセラレータに切り替える動きが増えて来ている。汎用プロセッサーから効率に優れる専用チップに交替する動きとして興味深く見守る必要がある。

## Ideinにみられる既存デバイスの徹底活用

スタートアップ企業のIdeinがRaspberry PiのGPU、VideoCore IVを活用したデモを発表して注目を集めている(fig.3)。アセンブラやコンパイラ、数値計算ライブラリなどツールチェーンを自作することにより

fig.1

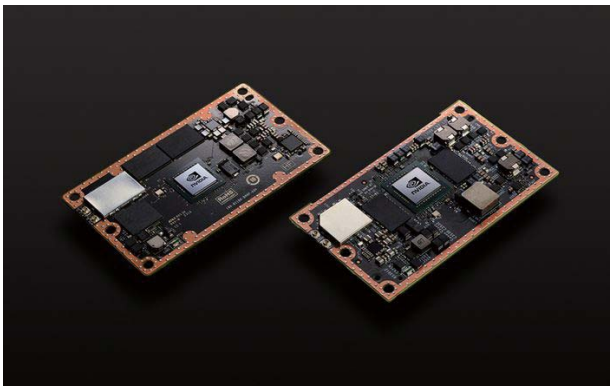
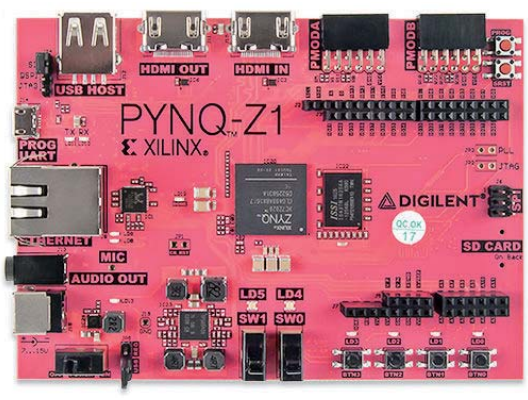


fig.2




Raspberry PiのCPUでの推論と比較して10倍から30倍の高速化を実現している。GoogLeNet(Googleが配布している画像認識モデル)による認識時間が0.7秒など従来のRaspberry Piでは考えられない性能を見せている。

今のところデモ動画が公開されているだけであるが、将来的にはライブラリの一部をオープンソース化することを考えているとのことで、公開された場合プロトタイプ開発で頻繁に使われるのではないだろうか。

## USB接続のNeural Compute Stick

1年前にIntelに買収されたMovidiusが最近発売したDeep Learning専用チップMyriad 2を搭載したUSB 3.0のスティックNeural Compute Stick(fig.4)でRaspberry Piのようなエッジデバイスでの画像処理を高速化させる試みが話題だ。また、Myriad 2を高速化したMyriad Xも発表されエッジでのDeep Learning活用を加速することが見込まれる。

現在のところ、Neural Compute Stickを使用するSDKがLinux版しかないため、もっと非力な環境(例えばArduinoなど)では使えないが、自由にアクセスできるようになれば、USBホスト機能さえあれば簡単にDeep Learningが使えるようになり採用が進むと思われる。

## エッジでの学習

以上のようにエッジでの推論利用について見て来たが、将来的にはエッジでの学習が必要になる。全てのエッジデバイスが収集するセンサデータをクラウドなりに集めて処理することが不可能になるからである。理由としては下記の3点があげられる。

* ネットワークの帯域やコストの問題
* クラウドの処理能力の限界
* セキュリティの懸念

これらのことから、必然的にエッジデバイスが推論を駆使して制御を行ないセンサーデータを学習し、学習結果をクラウドにフィードバックし集約、更新の形で再度配布するような形態が考えられる。しかし、エッジでの学習は例がほぼなく、次で述べるような課題が存在する。

## エッジデバイスでの学習の課題

まず、学習に必要なリソース(主に計算量とメモリ量)が膨大であることがあげられる。例えば、NVIDIA Jetson TX1で比較的小さなネットワークをChainerで学習させると4GBもメモリがあるにもかかわらず学習途中でメモリが足らずOOM Killerでアプリケーションが殺される現象になる。Jetsonのような比較的富豪な環境でも学習は困難であることから、さらに小さな環境では絶望的である。リソース問題についてのソリューションはまだ見えていない。

また、学習結果を集約する手法が確立されていないことが懸念点としてあげられる。学習結果のモデル(パラメータ)はかなりの大きさになるため、そのままネットワークに流すことは出来ない。精度を落さずモデルを1/2〜1/3に圧縮する手法が提案されているが、エッジからクラウドに集約するのに適した差分情報のような小さなデータに関する研究はまだないようである。

これらの課題を認識し解決しようとする動きはDeep Learningで有名なPreferred Networksの提案するエッジヘビーコンピューティングの一連の資料で確認できる。ただ、まだ課題と解決案の方向性を提示するのみで具体的な解決策には至っていない。これらの課題を解決する手法を論文などを観察しつつ自らも考えて行きたい。

fig.3

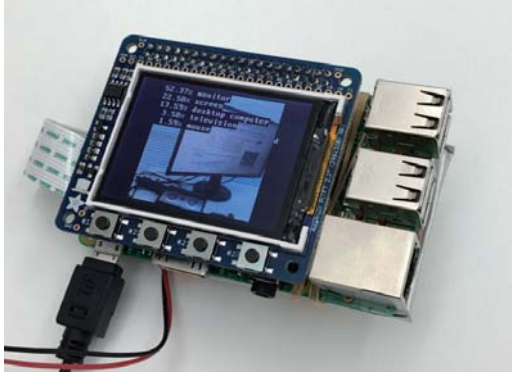


fig.4