第4回 AIエッジコンテスト レポート

チーム名 Vertical_Beach

lp6m, medalotte

1　開発フロー

DNN の HW アクセラレーションには，Xilinx 社から提供されている DPU（Deeplearning Processing Unit）コア [4] および統合開発環境である Vitis-AI を使用した．Vitis-AI は caffe，tensorflow 等の DNN フレームワークを用いて設計された DNN モデルを量子化し，DPU 向けにデプロイすることができる．[1]

[1] Vitis-AI v1.3 にて pytorch への一部対応が追加された．

2　DNN モデルの学習

2.1　モデル

コンテストの課題であるセマンティックセグメンテーションを行う DNN モデルとして我々は resnet18-FPN を使用した．モデルは Xilinx 社から提供されるチュートリアル [3] に含まれるものを流用した．FPN（Feature Pyramid Network）は，低解像だが意味的に強い (semantically strong) 特徴と高解像だが意味的に弱い (semantically weak) 特徴の両方を使用することで物体検出及び領域分割のタスクにおいて高い精度を挙げられることが知られている．

2.2　損失関数

参考にしたチュートリアルでは損失関数として SoftmaxWithCrossEntropy が使用されていた．コンテストで提供される学習画像を使用して学習を行ったが，テスト画像に対する mIoU スコアは 0.50 程度に留まり，処理速度部門における基準値である 0.60 を上回ることができなかった．そこで我々は領域分割タスクにおいて精度を向上させる損失関数として提案されている Lovasz-Loss[1] を採用した．Lovasz-Loss 関数は，第 1 回 AI エッジコンテストのセグメンテーション部門において第 2 位のチームも使用していたことから [2]，精度向上に効果的であると考えた．Lovasz-Loss は予測領域と正解領域の IoU を指標とする Jaccard-Loss をさらに拡張したものであり，tensorflow と pytorch 向けに公式に実装が公開されている．流用したチュートリアルは Caffe を用いてモデルが定義されており，Caffe 上で Lovasz-Loss 関数を自前で実装するのは困難であった．モデルを pytorch に変換して pytorch 上で学習を行い，学習済みの重みを caffe 向けに変換することでこの問題を解消した．

2.3　学習結果

コンテストで提供される学習用画像 2243 枚の 8 割を学習用，2 割を検証用に分割し学習を行った．解像度 480*960 の画像に対して GPU 上で推論を行った結果，IoU スコアは 0.62 となり，Lovasz-Loss 関数を使用したことで精度が大幅に向上した．

Image Size	DPUTask [ms]		Score
	150/300MHz	200/400MHz	
256*512	35	30	0.539
320*640	60	55	0.579
352*704	72	65	0.593
384*768	81	70	0.608
480*960	128	110	0.616

3　ハードウェア最適化

Xilinx 社から提供される DPU IP コアは，画素や入出力チャネルに対する並列数が異なる，複数の種類の IP コアが提供されている．より並列数の高い IP コアを使用することで処理性能が向上するが，回路規模および消費電力が増加する．また，DPU コアの一部のレイヤーのサポートを無効にすることでリソース使用率を低減することができる．

コンテストの評価ボードである Ultra96V2 に搭載可能な DPU コアとして B2304 を採用した．デフォルトで有効になっている DepthWiseConvolution および Pool Average のレイヤーは今回設計したモデルでは必要ないため無効にした．

さらに，DPU の動作周波数を高めることにより，DPU における推論実行時間を短縮することができる．論理合成のストラテジを Flow_AreaOptimized_high，配置配線のストラテジを performance_ExtraTimingOpt に変更することで動作周波数を 150MHz/300MHz から 200MHz/400MHz[2]に高めてもタイミング制約を満たし，FPGA ビットストリームの生成を行うことができた．

表 3 に動作周波数と入力画像サイズごとの DPU における推論実行時間･およびスコアを示す．DPU の推論時間は入力画像サイズに概ね比例し，動作周波数を向上させることによって推論処理が約 1.2 倍高速化されることがわかる．

[2]DPU コアはベース周波数の 2 倍の周波数を DSP に接続するため，動作周波数はこのような表記になっている．

4　ソフトウェア実装

5 性能評価

6　おわりに

参考文献

[1] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4413–4421, 2018.

[2] 森 大輝）MTLLAB（谷合 廣紀. 第1回 ai エッジコンテストレポート. https://static.signate.jp/competitions/138/summaries/AIEdgeContest_Segmentation_2_MTLLAB.pdf.

[3] Xilinx. Ml-caffe-segmentation-tutorial. https://github.com/Xilinx/Vitis-AI-Tutorials/tree/ML-Caffe-Segmentation-Tutorial.

[4] Xilinx. Zynq dpu v3.1 product guide. https://www.xilinx.com/support/documentation/ip_documentation/dpu/v3_1/pg338-dpu.pdf.

付録