2015 年 6 月 8 日 情報システム基盤学合同輪講発表資料

DianNao: A Small-Footprint High-throughput Accelerator for Ubiquitous Machine-Learning

著者： Tianshi Chen,Zidong Du,Ninghui Sun,Jia Wang,Chengyoug Wu,Yunji Chen,Olivier Temam
出典： *ASPLOS'14,March 1-5,2014,Salt Lake City,Utah,USA*
発表者： 1553004 大場百香

1 はじめに

特定の処理に特化したコンピュータであるアクセラレータは画像処理に特化した GPU や暗号化に特化した SSL アクセラレータなど様々な所で応用されている。アクセラレータを使うことによって高速な処理を低消費電力で行うことが出来る。また、人間の学習能力と同様の機能の実現を目指した機械学習は、最先端のアルゴリズムである CNN(Convolutional Neural Network[1])、DNN(Deep Neural Network[2]) などの大規模なニューラルネットワークを使うことで更に活躍の場を広げている。しかし大規模ニューラルネットワークを利用するには膨大なコンピュータ資源が必要なので、機械学習に利用できるアクセラレータが必要となってくる。従来のニューラルネットワークアクセラレータ研究は、ニューラルネットワークの計算部分をより速く効率的に実行することに焦点をおいてきた。大規模ニューラルネットワークの計算においてボトルネックとなるメモリ転送についてはあまり考慮されていない。

そこで本研究では、メモリ転送の回数を減らすこと、できるだけ効率よく実行しエネルギーを減らすことに焦点をおいた大規模ニューラルネットワークに適応できるアクセラレータを設計し、性能を評価する。

2 ニューラルネットワークの構造

脳内の多数のニューロンが行っている、様々な情報処理の仕組みをコンピュータ内に実現したのがニューラルネットワークである。ニューラルネットワークは、主に入力層、中間層、出力層で構成されている。

2.1 Convolutional Neural Network(CNN)

画像認識などに応用されている CNN は convolutional 層と pooling 層の 2 種類の層を交互に積み重ねた構造をもつ多層ニューラルネットワークである。

・convolutional 層

input feature map(入力ニューロンの集合) を output feature map(出力ニューロンの集合) に畳み込む。畳み込みとは適当な大きさの領域に含まれる各値を重み付けして足し合わせることである。また、複数のカーネルを使って画像の異なる特徴を別々の feature map として抽出する。(図 1)

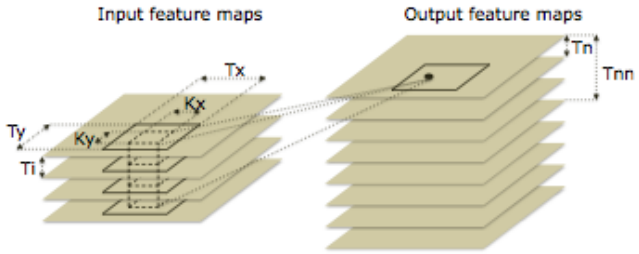

図 1: convolutional 層

・pooling 層

畳み込み処理で出力された feature map のサイズを max pooling 等により縮小する。これによってより新たな feature map を得ることができる。

・classfier 層

すべての feature map を集計する。

3 アクセラレータの設計

メモリからのデータ転送効率を良くすることで高いスループットを実現し、ニューラルネットワークの計算に必要最低限のハードウェアのみの設計にすることで小型化、省エネルギー化したアクセラレータを実装する。

3.1 メモリアクセスの高効率化

アクセラレータは主に、コントロールロジック（CP)、入力ニューロンを保持する入力バッファ(NBin)、パイプライン化された演算器 (NFU-1,2,3)、シナプス結合荷重を保持するバッファ(SB)、出力ニューロンを保持する出力バッファ(NBout) で構成されている。DMA エンジン (direct memory access)[3] をバッファに取り付けることでメモリとバッファのデータ転送とニューラルネットワークの計算処理をオーバーラップできるようになり、メモリアクセスの性能を向上させることができる。また、バッファ上のデータを再利用することでメモリアクセスのエネルギーを削減できる。

Layer	N_x	N_y	K_x	K_y	N_i	N_o	Description
CONV1	500	375	9	9	32	48	Street scene parsing
POOL1	492	367	2	2	12	-	(CNN) [13], (e.g.,
CLASS1	-	-	-	-	960	20	identifying “building”, “vehicle”, etc)
CONV2*	200	200	18	18	8	8	Detection of faces in YouTube videos (DNN) [26], largest NN to date (Google)
CONV3	32	32	4	4	108	200	Traffic sign
POOL3	32	32	4	4	100	-	identification for car
CLASS3	-	-	-	-	200	100	navigation (CNN) [36]
CONV4	32	32	7	7	16	512	Google Street View house numbers (CNN) [35]
CONV5*	256	256	11	11	256	384	Multi-Object
POOL5	256	256	2	2	256	-	recognition in natural images (DNN) [16], winner 2012 ImageNet competition

図 2: benchmark

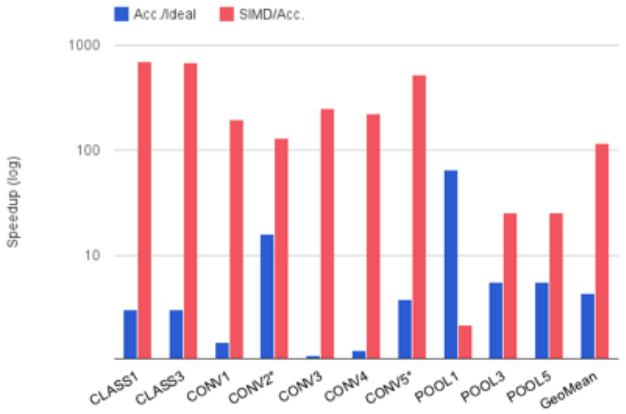

図 3: Speedup of accelerator over SIMD,and of ideal accelerator

3.2 アクセラレータの小型化/省エネルギー化

ニューラルネットワークの計算に必要なハードウェアのみを搭載することで余計なハードウェアがないため小型で省エネルギーを保つことができる。そして、エネルギーと実装面積を考慮して各バッファのサイズ、データ幅、演算器数を調整する。

3.3 レイアウト後のアクセラレータ

Synopsys が開発したツールを使ってアクセラレータのレイアウトを行った。また、このアクセラレータはスマホなどに組み込まれている Coretex-A-15 x4 より小面積で設計することができた。

アクセラレータシミュレータ、CAD ツールを用いて設計したアクセラレータのサイクル数時間、面積、エネルギーの測定を行う。また、SIMD 型 CPU と比較することで性能を評価する。

3.4 ベンチマーク

実際利用されている convolutional 層、pooling 層、classifier 層。(図 2)

4 実験結果

4.1 時間とスループット

図 3 より、赤のグラフは、SIMD 型 CPU と設計したアクセラレータの実行時間（SIMD 型 CPU の実行時間で正規化）を比較した結果を表している。全ての層が SIMD 型 CPU より演算スピードが速いことから時間短縮に成功した。また、青のグラフは、理想のアクセラレータと設計したアクセラレータの実行時間（設計したアクセラレータの実行時間で正規化）を比較した結果を表している。これより、conv3,conv4 は理想のアクセラレータの実行時間にかなり近い結果が得られたが、pool1,conv2 は理想からはなれている。pool1 は Ni=１２、conv2 は Ni=8 と設計したアクセラレータの入力 feature map(Ni=16) より少なく、フル活用されていないのでスループットが低いと考えられる。

4.2 エネルギー減少量

SIMD 型 CPU に比べて設計アクセラレータは約 21 倍エネルギー効率が良いという結果が得られた。

5 まとめ

本研究では、機械学習のアルゴリズムである大規模ニューラルネットワークに適応できるアクセラレータの設計を行い、小面積、高スループット、省エネルギーを実現することに成功した。今後の課題として、メモリからのデータ転送のオーバーヘッドが挙げられる。

参考文献

[1] Y.Lecun,L.Bottou,Y.Bengio,and P.Haffner. Gradientbased learning applied to document recognition. proceedings of the IEEE,86,1998.

[2] G.Hinton and N.Srivastava. Improving neural networks by peeventing co-adaptation of feature detectors. arXiv preprintarXiv:...,pages 1-18,2012.

[3] S. Chakradhar, M. Sankaradas, V. Jakkula, and S. Cadambi. A dynamically configurable coprocessor for convolutional neu- ral networks. In International symposium on Computer Archi- tecture, page 247, Saint Malo, France, June 2010. ACM Press.