

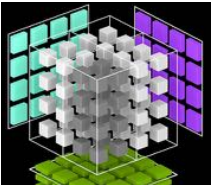
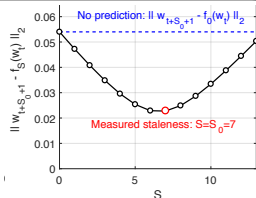

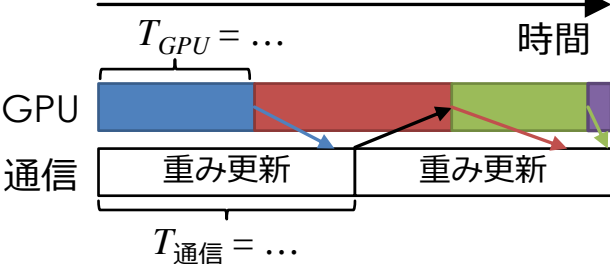
平成 30 年度 特別研究員 DC1
研究課題「**深層学習の精度を考慮した
自動性能最適化フレームワークの構築**」

大山洋介

東京工業大学 情報理工学院 松岡研究室

研究背景

→ 高速な深層学習(DL)には**学習速度**と**学習の質(精度)**が必須

	学習速度 (計算・通信速度)	学習の質 (精度)
最適化	<p>DLに特化したGPU アーキテクチャ (例: Tensor Core*)</p> 	<p>学習アルゴリズム やモデルの改善 [研究実績a]</p> 
<p>低精度な計算・通信 = 速度と精度のトレードオフ</p> <p>低精度浮動小数点型を用いた 符号: 1bit 仮数: 2bit 通信高速化 [研究実績b] 指数: 5bit</p> 		
性能予測	<p>非同期DLの 性能モデリング [研究実績1]</p> 	

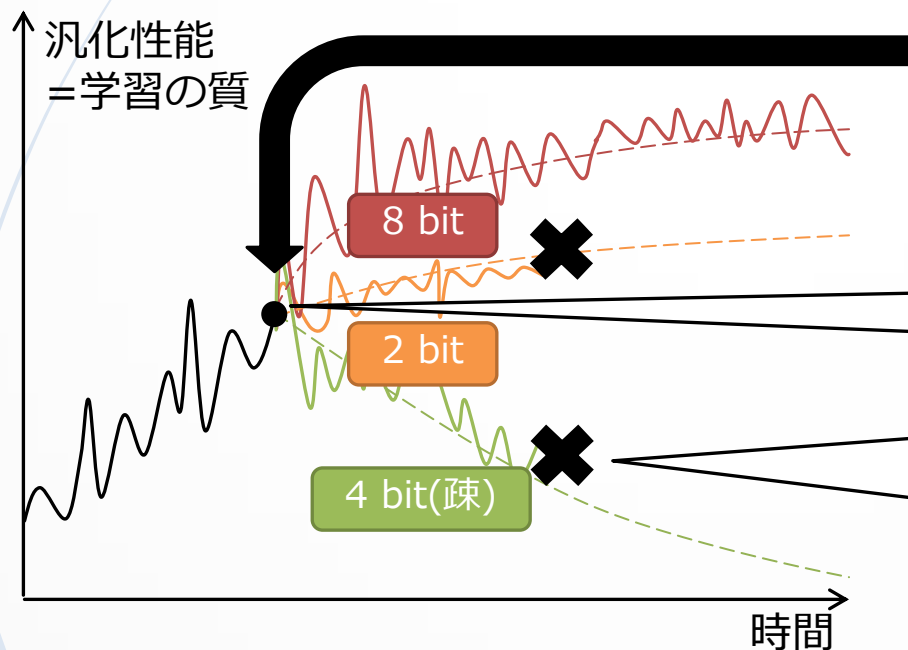
「学習の質」の性能モデリング・性能予測が不十分であるため、
学習速度とのトレードオフの最適点を予測・保証することができない

* <https://www.nvidia.com/en-us/data-center/volta-gpu-architecture/>より引用

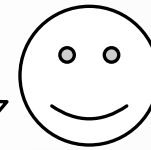
研究計画

研究内容1: 並行学習によるアルゴリズムの自動選択

- **アイデア:** 計算資源量を一時的に増加させることで学習の質を直接比較する
- **提案手法:** 性能モデル[研究実績1]や学習曲線の予測モデル[参考文献4]を用いて汎化性能の向上具合を予測し、最良のアルゴリズムを自動的に選択する
- **インパクト:** 速度と精度のトレードオフの最適点が自動的に選択される
 - 深層学習の研究サイクルが加速する



新たなモデルに対して
通信型bit数を**自動で**
最適化したい



研究者

スパコン・クラウド環境を用いて
計算資源量(ノード数, GPU数, …)を
追加

性能モデル[研究実績1]により

- 並行学習の開始間隔
- 不要ジョブの停止タイミング
- 個々の最適な資源量 …

を最適化

研究計画

研究内容2: 計算カーネルの自動性能モデリング

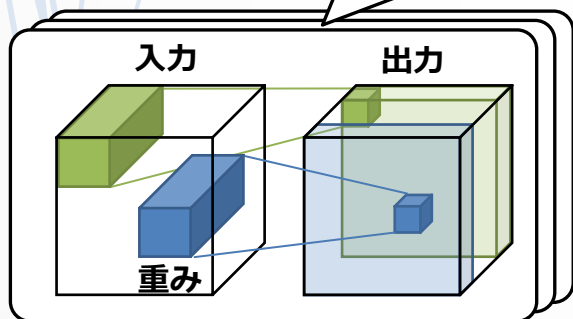
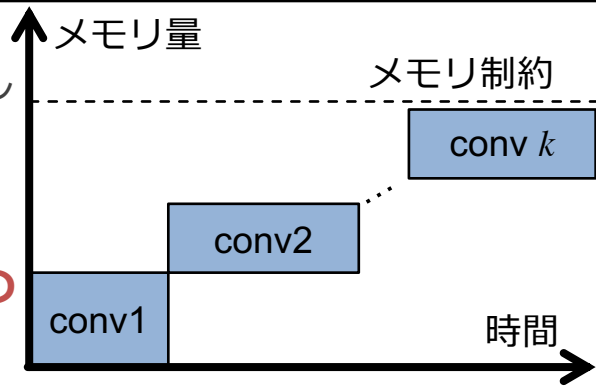
- **アイデア:** 計算アルゴリズムとDNNモデル設計のco-design
- **提案手法:** モデル設計パラメータに対する自動性能最適化基盤
- **インパクト:** 研究者に対して速度・精度の両方を最適化したモデルを提示

畳み込み層の設計パラメータ

- 入力テンソル幅、マップ数
- フィルタ幅、フィルタ数、stride、padding、dilation
- 入出力データ型、演算型
- ミニバッチサイズ
- 計算アルゴリズム …

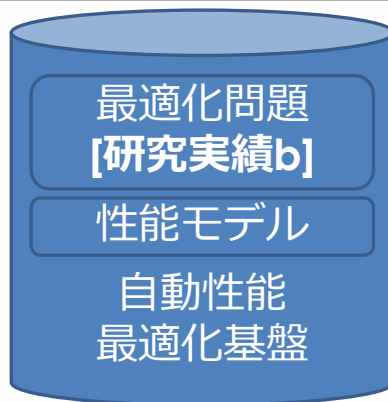
ミニバッチサイズ・メモリ使用量を最適化問題に帰着し畳み込みを最大2.3倍高速化
[研究実績b]

→ 現存フレームワークでは計算アルゴリズムとモデルのco-designが不十分



例: 畳み込み演算

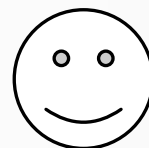
ベンチ
マーク



速度・精度の両方を最適化したモデルの提示



各パラメータの精度についてのsensitivity



研究者

研究の立ち位置

学習アルゴリズム

計算手法

学習・計算基盤

JST CREST「社会インフラ映像処理のための高速・省資源深層学習アルゴリズム基盤」
(研究代表者: 篠田浩一(東工大情報理工学院))

- 特定の学習アルゴリズムに関する研究
[篠田研, 松岡研]
- ネットワークの圧縮表現
[村田研]

- 構造化行列による計算量削減 [横田研, 松岡研]
- GPU・FPGAを用いた低エネルギー・低コストな学習 [松岡研]

- 深層学習フレームワークのout-of-core実行 [遠藤研]
- ジョブスケジューリングに関する研究 [松岡研]

申請者の
研究内容


ITLAB
デンソーアイティ
ラボラトリーとの共同研究


SPCL
チューリッヒ工科大
SPCLとの共同研究

低精度表現を用いた通
信の高速化 [既発表]

非同期深層学習の性能
モデリング [既発表]

研究内容 1: 投機的な並行学
習による高性能なアルゴリ
ズムの自動選択

研究内容 2: 自動性能モデリ
ングと動的な計算資源選択

特定のネットワーク・学習アルゴリズムに限定しない学習基盤の研究を行う

研究実績

▶ 国際会議における発表

1. ○**Yosuke Oyama**, Akihiro Nomura, Ikuro Sato, Hiroki Nishimura, Yukimasa Tamatsu, Satoshi Matsuoka, “Predicting Statistics of Asynchronous SGD Parameters for a Large-Scale Distributed Deep Learning System on GPU Supercomputers”, In proceedings of 2016 IEEE International Conference on Big Data (IEEE BigData 2016), Washington D.C., Dec. 5-8, 2016. [査読あり, 採択率: 18.68%]
 - a. Ikuro Sato, Ryo Fujisaki, ○**Yosuke Oyama**, Akihiro Nomura, Satoshi Matsuoka, “Asynchronous, Data-Parallel Deep Convolutional Neural Network Training with Linear Prediction Model for Parameter Transition”, In The 24th International Conference On Neural Information Processing (ICONIP 2017), Nov. 2017. [査読あり]

▶ 他 ポスター 2件、査読なし口頭発表 1件

▶ 国内学会・シンポジウム等における発表

5. ○**大山洋介**, 野村哲弘, 佐藤育郎, 西村裕紀, 玉津幸政, 松岡聡, “学習条件を考慮した大規模非同期ディープラーニングシステムの性能モデリング”, 情報処理学会研究報告, Vol. 2016-HPC-155, 2016.
6. ○**大山洋介**, 野村哲弘, 佐藤育郎, 松岡聡, “ディープラーニングのデータ並列学習における少精度浮動小数点数を用いた通信量の削減”, 情報処理学会研究報告, Vol. 2017-HPC-158, 2017.
- b. ○**Yosuke Oyama**, Tal Ben-Nun, Torsten Hoefler, Satoshi Matsuoka, “Less is More: Accelerating Deep Neural Networks with Micro-Batching”, Vol. 2017-HPC-162, 2017.

研究実績

▶ 受賞歴

c. 情報処理学会 コンピュータサイエンス領域奨励賞 (研究実績6, 2017年度)

▶ 特許等

7. ○大山洋介(30%), 佐藤育郎(25%), 西村裕紀(25%), 野村哲弘(10%), 松岡聡(10%),
“予測装置、予測方法および予測プログラム” (出願済み)

d. ほかに1件

▶ その他

8. 国立研究開発法人科学技術振興機構 AIPチャレンジ 採択 (2016/10~2017/3)

9. 東京大学情報基盤センター 若手・女性利用者推薦 採択 (2017/4~9)

10. 東京大学学際大規模情報基盤共同利用・共同研究拠点 萌芽型共同研究課題 採択 (2017/4)

11. チューリッヒ工科大学 Student Summer Research Fellowship 採用 (2017/7~8)

▶ 研究成果の一部は研究実績bとして発表済

e. “AI白書2017” (独立行政法人情報処理推進機構, 2017年) 執筆協力 (2017)