2019年6月24日

CREST「人工知能」領域 成果公開シンポジウム第2回

深層学習のための Co-Designフレームワーク

篠田浩一
(東京工業大学)

国立研究開発法人
科学技術振興機構

CREST

イノベーション創発に資する人工知能基盤技術の創出と統合化
(通称 CREST AI領域、栄藤稔総括)

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

2016年12月～2019年 3月　スモールフェーズ（終了）
2019年　4月～2022年 3月　加速フェーズ　　(初年度)

共同研究者：

松岡 聡 (理研)

大西正輝 (産総研)

横田理央 (東工大)

村田剛志 (東工大)

中原啓貴 (東工大)

鈴木大慈 (東大)

目的

- 安全・安心なスマート社会
- 高度な映像技術→事故の防止、異常の早期発見
- 人の動きは複雑、予測が難しい→ 深層学習技術

課題 (Problem) (続き)

1. 大量の画像の実時間での解析
   Analyze a huge amount of images in real-time
2. 環境の変化に速やかに適応
   Rapidly Adapt to the changes in environmental conditions
3. 端末側での計算 → 通信量の削減
   Edge Computing　　Reduce traffics on Internet

これらの課題は密接に関連
These problems are deeply related with each other

→ 同時に最適化
   Simultaneous optimization

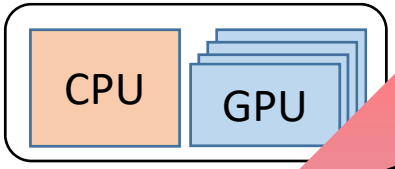
スモールフェーズ

2016年12月～2019年 3月

Approach－Co-Design －

横田

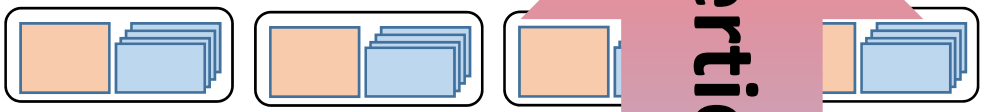
計算ノード
Compute Node



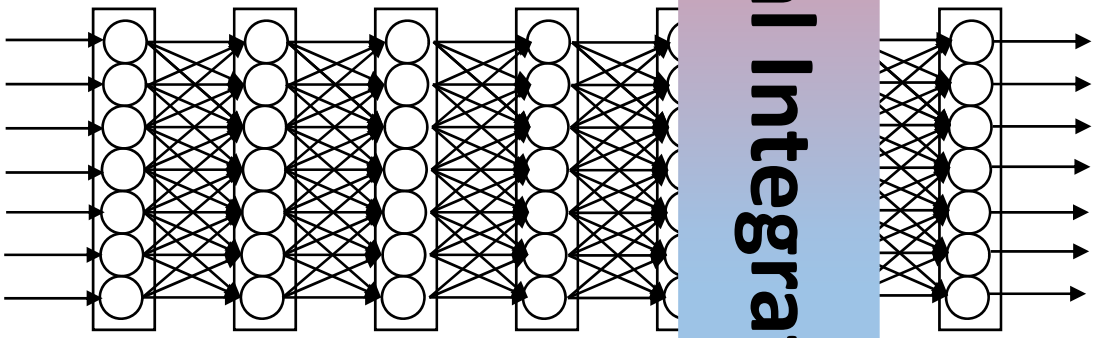
Architecture

松岡

並列化
Parallelization



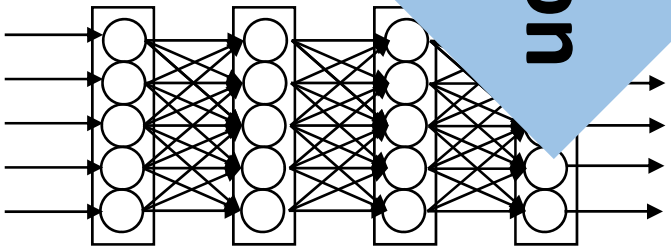
学習アルゴリズム
Learning Algorithm



Vertical Integration

篠田

小型化
Downsizing



Application

村田

Goal in Small Phase

Component	Speed	Memory
Compute node	50x	1/10
Parallelization	10x	
Learning Algorithm	10x	1/10
Downsizing		1/100
Total	> 1000x	< 1/1000

What we have **done** until Nov. 2018

Component	Speed	Memory
Compute node (Yokota G)	18x (50x)	1/5(1/10)
Parallelization (Matsuoka G)	1536x (10x)	?
Learning Algorithm (Shinoda G)	10x (10x)	?(1/10)
Downsizing (Murata G)		1/90(1/100)
Total	> 1500x	?

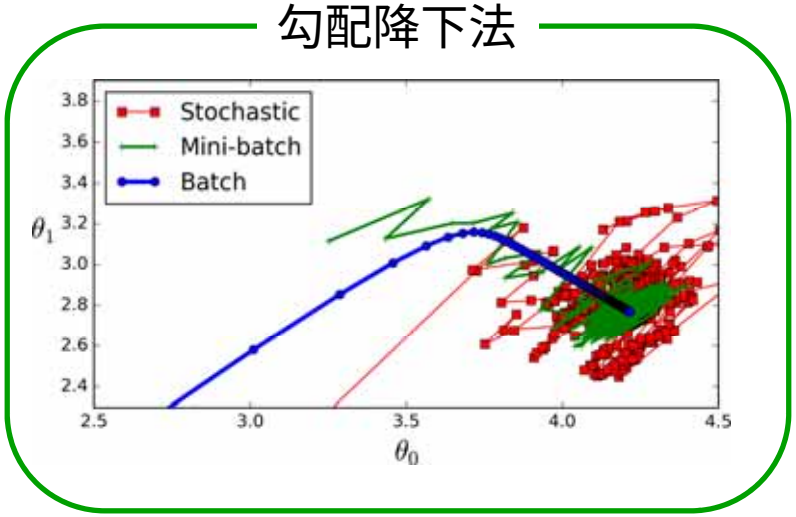
昨年度の成果

1. 計算ノード

個々の計算ノードにおける計算量を削減するための
行列構造化アルゴリズム

横田理央G

大規模並列深層学習の課題 (mini-batchの説明)



- 画像を一つずつ学習 : stochastic
- 画像を複数まとめて学習する : mini-batch

  ※画像を全て一気に学習: batch

- 画像数が増えるに従って安定した学習になる
- ただ，安定させすぎると汎化性能が悪くなる

自然勾配法 (Natural Gradient Descent)

Stochastic gradient descent $\theta_{t+1} = \theta_t - \eta \nabla J$

Momentum $\theta_{t+1} = \theta_t + v_t \qquad v_t = v_{t-1} - \eta \nabla J$

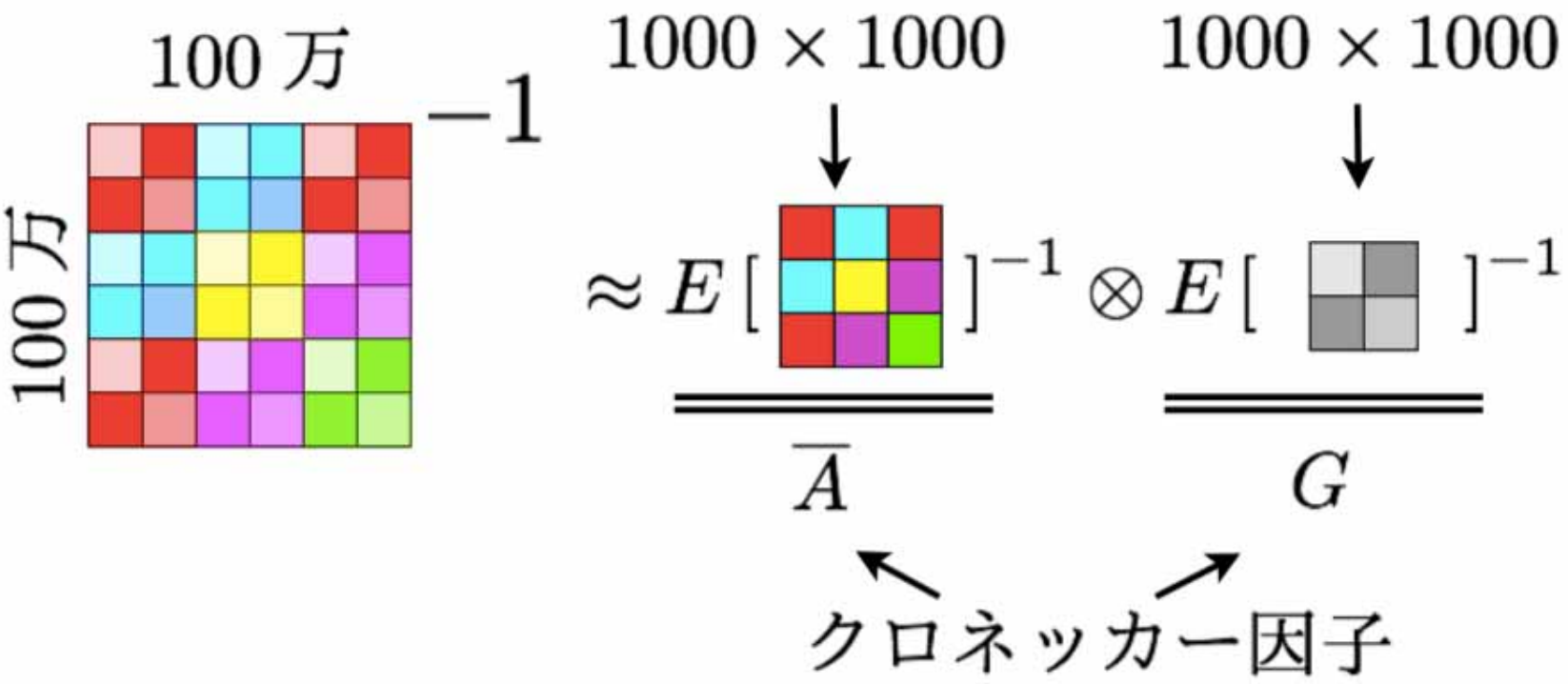
AdaGrad $\theta_{t+1} = \theta_t - \eta diag(F)^{-1/2} \nabla J \qquad F = \nabla J \nabla J$

Natural gradient descent $\theta_{t+1} = \theta_t - \eta F^{-1} \nabla J$

- 自然勾配法(NGD)は曲率を考慮して学習係数を調整
- NGDはSGDよりも10倍早く収束
- Fisher行列*F*の計算に時間がかかる

１ステップあたりの計算時間が10倍

Kronecker因子分解による行列構造化



$$\theta_{t+1} = \theta_t - \eta F^{-1} \nabla J \quad \longrightarrow \quad F^{-1} = E(A^{-1} \otimes G^{-1})$$
$$\approx E(A)^{-1} \otimes E(G)^{-1}$$

- 1 stepあたりの計算時間を大幅に削減
  ⇒通常の勾配降下法とstepあたり同じ計算時間に

ImageNet fast training Project

2018/12/2 Assert

Top 10 Arxiv Papers Today in Computer Science

2.06 Mikeys

#1. Second-order Optimization Method for Large Mini-batch: Training ResNet-50 on ImageNet in 35 Epochs

Kazuki Osawa, Yohei Tsuji, Yuichiro Ueno, Akira Naruse, Rio Yokota, Satoshi Matsuoka

Large-scale distributed training of deep neural networks suffer from the generalization gap caused by the increase in the effective mini-batch size. Previous approaches try to solve this problem by varying the learning rate and batch size over epochs and layers, or some ad hoc modification of the batch normalization. We propose an alternative approach using a second-order optimization method that shows similar generalization capability to first-order methods, but converges faster and can handle larger mini-batches. To test our method on a benchmark where highly optimized first-order methods are available as references, we train ResNet-50 on ImageNet. We converged to 75% Top-1 validation accuracy in 35 epochs for mini-batch sizes under 16,384, and achieved 75% even with a mini-batch size of 131,072, which took 100 epochs.

more | pdf | html

Figures

None.

昨年度の成果

2. 並列化

ノード間の通信処理を削減するための高並列アルゴリズムと資源スケジューリングによる全体最適化

松岡G

昨年度の成果

3. 学習アルゴリズム

知識の構造を活用した
高速な深層学習アルゴリズム

篠田G

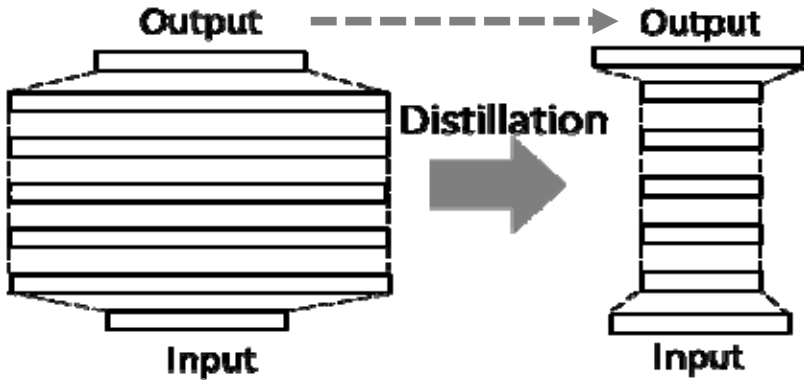
- 知識の階層構造の活用
  → 少量の教師ラベルつきデータで効率的に性能向上

- 知識の隣接関係を利用：ソフトターゲットを用いる
  → 性能を落とさずにコンパクト化
  - 大量のラベルなしデータも利用
  - Student-Teacher 学習による知識蒸留(Knowledge Distillation)

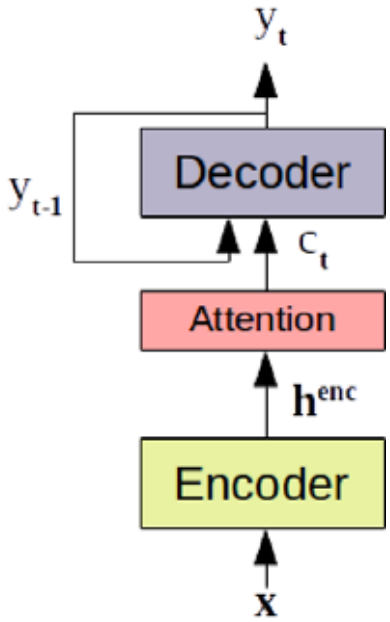
- 因果関係を抽出する能動学習
  → 学習・認識処理の高速化
  - Attention モデル

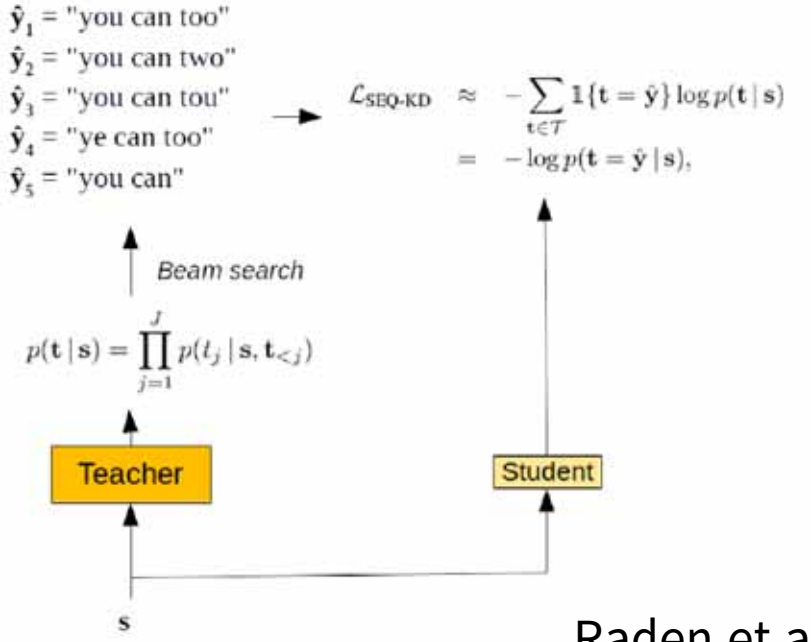
知識蒸留(Knowledge Distillation)

- Student-Teacher 学習を用いた知識蒸留
  1. 大きな教師モデルを教師付き学習
  2. 教師モデルの出力を教師として小さな生徒モデルを学習



- 注意機構付きEnd-to-End 時系列モデルに適用
- モデルサイズを1/10に（誤り率は7%増加)





Raden et al, ICASSP2019

昨年度の成果

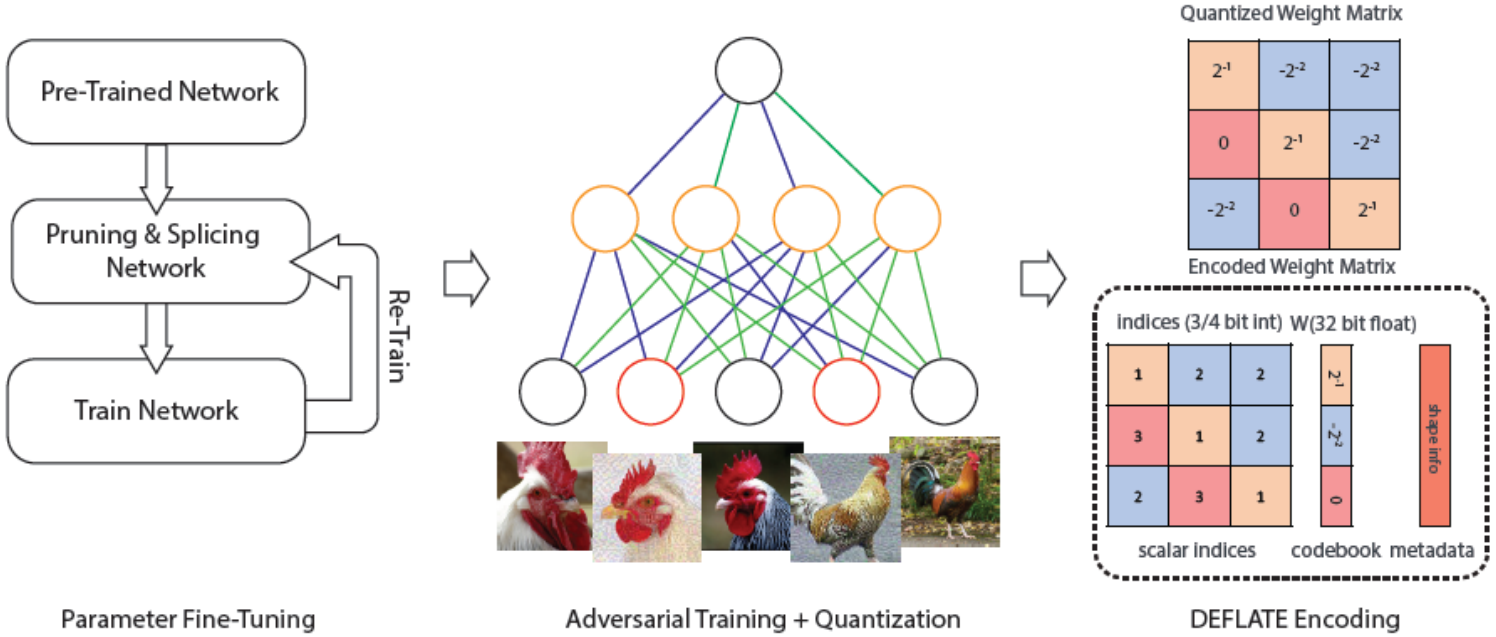
4. 小型化

リアルタイム認識・解析のための
Deep Net 構造のコンパクト化アルゴリズム

村田G

頑強なDeep Netサイズ圧縮

量子化において、敵対的な訓練と組み合わせることで精度を落とさずサイズ圧縮



Methods	Comp. Rate	Top-1 Acc.	Top-5 Acc.	Size (MB)
Original	-	0.58	0.80	240
Knoll, 2012 (P+H)	38x	0.58	0.80	6.3
Han, 2016 (P+Q+H)	35x	0.58	0.80	6.9
Zhou, 2017 (P+Q)	89x	0.58	0.80	2.69
Ours (P+Q+D)	**90x**	**0.59**	**0.81**	**2.64**

P : Pruning
Q: Quantization
H: Huffman Encoding
D: DEFLATE

Knoll, 2012 and Han, 2016
required specific customized hardware

Wijayanto et al IEEE BigComp 2019

加速フェーズ

2019年4月～2022年3月

富岳: Game Changer

- **Fujitsu-Riken design A64fx ARM v8.2 (SVE), 48/52 core CPU**
  - ***HPC Optimized:*** Extremely high package high memory BW (1TByte/s), on-die Tofu-D network BW (~400Gbps), high SVE FLOPS (~3Teraflops), various AI support (FP16, INT8, etc.)
  - Gen purpose CPU – Linux, Windows (Word), other SCs/Clouds
  - Extremely power efficient – > **10x power/perf efficiency for CFD benchmark** over current mainstream x86 CPU
- **Largest and fastest supercomputer to be ever built circa 2020**
  - > 150,000 nodes, superseding LLNL Sequoia
  - > 150 PetaByte/s memory BW
  - Tofu-D 6D Torus NW, 60 Petabps injection BW (10x global IDC traffic)
  - 25~30PB NVMe L1 storage
  - ~10,000 endpoint 100Gbps I/O network into Lustre
  - The first 'exascale' machine (not exa64bitflops but in apps perf.)

目標達成のために

- 10万ノードでスケールする超並列処理の実現
  - 二次最適化をデファクトに
  - モデル並列
- より高精細動画(HD → 4K → 8K)を対象に
- 深層学習アルゴリズムもスケールアップ
  - 大規模GANを用いたデータ増強
  - 構造的な知識を表現したグラフを入力に
  - 映像の時系列End-to-Endモデルの実現
- FPGA実装まで踏み込む
- 理論的サポート、シミュレーション環境の充実
- 実応用での評価（ベンチマーキング）

Co-Design：加速フェーズ

: NEW

10万
ノード

大規模
GAN

グラフ
DNN

時系列
モデル

計算ノード

CPU

GPU

並列化

学習
アルゴリズム

FPGA 実装

シミュレーション・認識

中原G

鈴木G

大西G

社会実装

大西G：実社会応用における評価（新規）

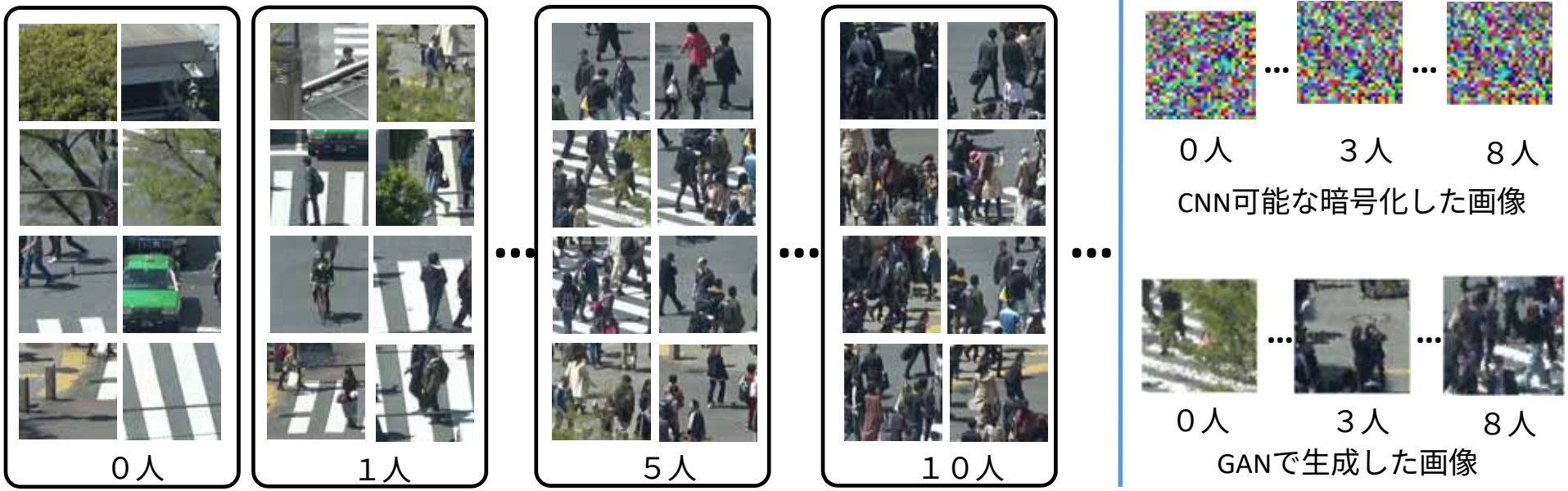
関門海峡花火大会での数万人規模の群集計測、新国立劇場での千数百人規模の避難訓練の計測など数多くの大規模な実証実験の経験を持つ。

実社会応用における評価

- データベースの整備
  - ニアミスと群集移動のデータベースを作成
- 監視カメラ映像処理への応用
  - 逆走や混雑過多の認識で手法を評価
- 車載カメラ映像処理への応用
  - ニアミス対象と危険度の認識で手法を評価
- 時系列モデリング
  - 時空間の三次元畳み込みによる行動認識
- 公開方法の検討
  - 個人情報を含む映像を公開する方法の検討

監視カメラ映像処理の応用例

監視カメラ映像で混雑状況をアノテーションしたデータベースを作成し、
例えばResNet-50でどの程度認識できるかを評価する（下図は頭頂部の数）



プライバシーの観点で公開するのが難しいデータセット → 作成 公開用データセット

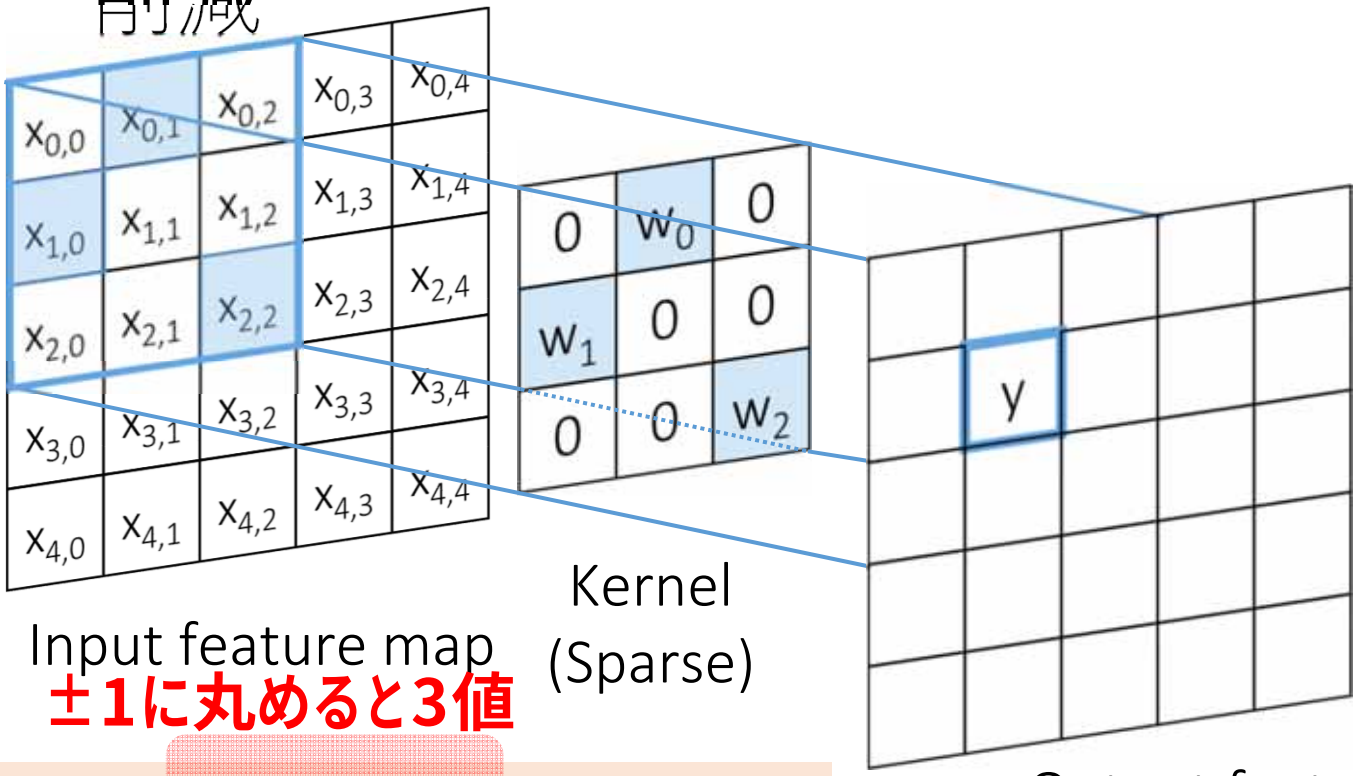
中原G (新規)

多値論理, 計算機アーキテクチャが専門。
深層学習専用ハードウェアの研究に取り組み、FPGAによる2値化・３値化・混合精度による高性能推論回路を実現。

FPGA実装向け深層学習回路

- FPGAエッジコンピューティング
  FPGAの柔軟性を活かし, 最新の深層学習アルゴリズムをコンパクトかつ高性能な推論回路に実装
- 頑健化・小型化
  他グループとの連携で得られた知見を活用
- モデルシミュレーション
  FPGA上にシミュレーションアクセラレータを実現し、高速化

エッジ推論デバイス向け軽量化:
重み3状態CNNによるパラメータ・計算量削減



Input feature map

Kernel (Sparse)

Output feature map

skip
$X_{0,1} \times W_0$
skip
$X_{1,0} \times W_1$
skip
+) $X_{2,2} \times W_2$
y

±1に丸めると3値

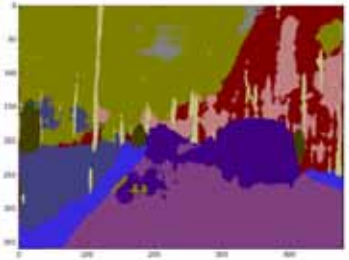
$$w = \begin{cases} Sign(w_{hid},) & |w_{hid}| > \rho \\ 0, & Otherwise \end{cases}$$

ρ: しきい値 (定数)

→　モデル・アプリケーション毎に適したスパース率・ビット精度を探索

重み3状態CNNに適した FPGA回路実装と応用事例への適用

- 物体認識(YOLOv2)をFPGA(Intel Arria10)に実装
  GPU(RTX2018Ti)よりも電力75%削減3倍高速化
  →電力性能効率12倍達成



→ 領域認識・姿勢推定などへの適用
小型FPGA向け実装技術の開発

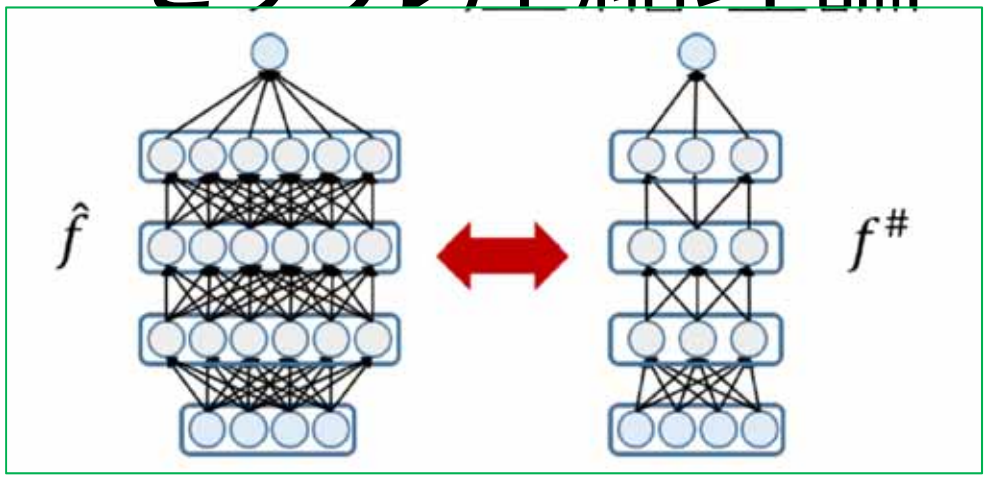
鈴木G (新規)

機械学習の汎化誤差理論・数理最適化理論を推進
（理研AIP深層学習理論チームリーダー）。
加速フェーズから理論面のバックアップを担当。

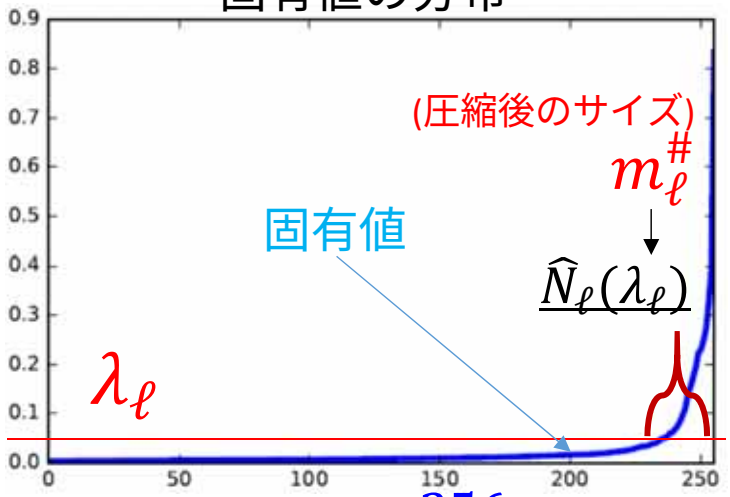
深層学習理論 (汎化誤差，圧縮率，収束率)

- 確率的最適化
  理論的保証のある新しい確率的最適化・分散計算アルゴリズム
  (ICML2013,ICML2014,AISTATS2017,NIPS2017,NeurIPS2018,AISTATS2019)
- 深層NNの圧縮および汎化誤差解析
  深層NNを最適なサイズに圧縮する方法論，“なぜ深層か”の理論
  (AISTATS2018,ICML2019,ICLR2019)
- 適応的深層NNモデリング
  ResNet型ネットワークの自動構築・学習方法を提案
  (AISTATS2018,ICML2018)

モデル圧縮理論



中間層における分散共分散行列の
固有値の分布



$m_\ell = 256$
(もとのサイズ)

圧縮性能および圧縮後のネットワークの汎化性能
はこの固有値の分布で特徴づけられることを証明

$$L(f^{\#}) \leq \hat{L}(\hat{f}) + \underbrace{\sum_{\ell=2}^{L} \sqrt{\lambda_\ell}}_{\text{bias}} + \underbrace{\sqrt{\frac{\sum_{\ell=1}^{L} m_\ell^{\sharp} m_{\ell+1}^{\sharp}}{n}}}_{\text{variance}}.$$

新しい圧縮手法の提案：
実データでの優れた性能

Model	Top-1	Top-5	# Param.	FLOPs
ResNet-50-1	72.89%	91.07%	25.56M	7.75G
ThiNet-70	72.04 %	90.67%	16.94M	4.88G
ThiNet-50	71.01 %	90.02%	12.38M	3.41G
NISP-50-A	72.68%	—	18.63M	5.63G
NISP-50-B	71.99%	—	14.57M	4.32G
Spec-ResA	72.99%	91.56%	12.38M	3.45G

(ResNet 50の圧縮)

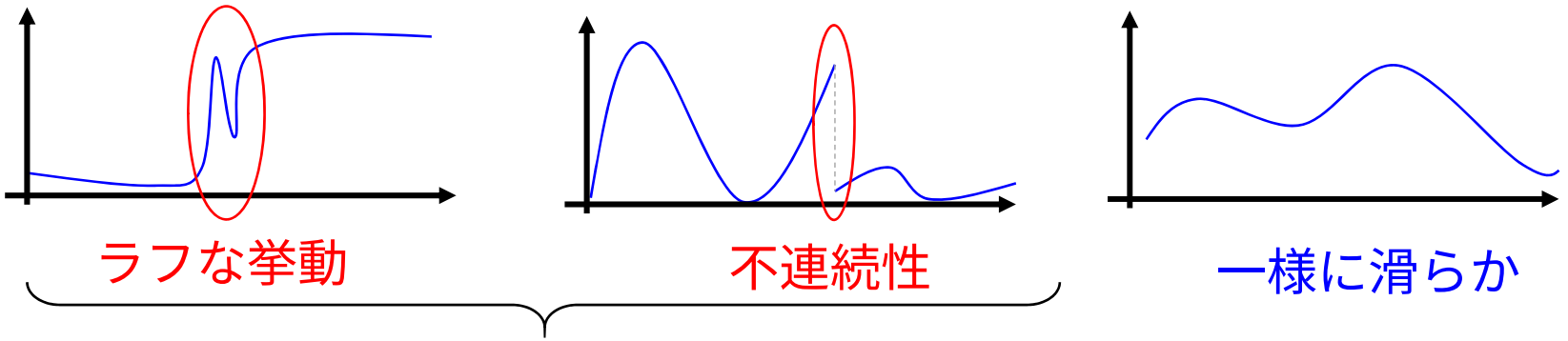
Approximation: [Suzuki: Fast generalization error bound of deep learning from a kernel perspective. AISTATS2018.]

Compression: [Suzuki, Abe, Murata, Horiuchi, Ito, Wachi, Hirai, Yukishima, Nishimura: Spectral-Pruning: Compressing deep neural network via spectral analysis, 2018]

[Suzuki: Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. ICLR2019]

なぜ深層学習が良いのか？

機械学習に現れる様々な形状の関数：



難しい　　**簡単**

ラフな部分に合わせると全体で過学習，スムーズな部分に合わせると過小学習

“適応力”が重要

定理

深層学習はBesov空間($B^s_{p,q}$)の元を推定するのにミニマックス最適レートを達成する．一方，カーネル法などの“浅い手法”は最適でない．

(浅層) $n^{-\frac{2s-2(1/p-1/2)_+}{2s+1-2(1/p-1/2)_+}} \gg n^{-\frac{2s}{2s+1}}$ (深層)

非最適　　**最適**

Our collaborators

TSUBAME3.0

NVIDIA

Agency for Science, Technology and Research
SINGAPORE

REAL WORLD BIGDATA
RWBC-OIL

AIST-Tokyo Tech
Real World Big-Data Computation
Open Innovation Laboratory
(RWBC-OIL)

DENSO IT LAB

CREST Deep

FUJITSU

R-CCS

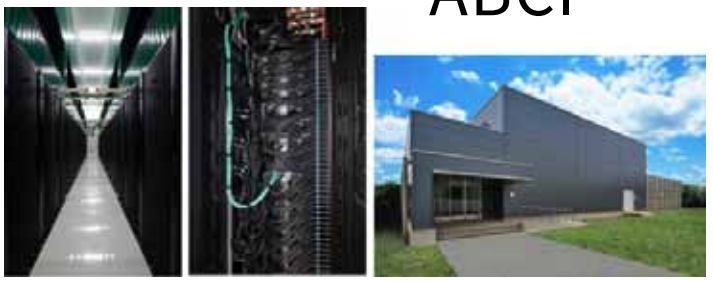
AIST

RIKEN

AIP

富岳

ABCI



今年3月にシンガポールでI2Rと
合同ワークショップを開催しました。