Reversible Networksによるデータフロー型NN学習の検討

宮城竜大，高瀬英希，東京大学情報理工学系研究科コンピューティングシステム学研究室

研究背景

- 近年のNNの成功は，誤差逆伝播法（Backpropagation，BP）の高い計算効率と汎用性に立脚している
- しかし，膨大なメモリ容量/帯域幅が必要
- BPに代わる省メモリな学習アルゴリズムが求められる

先行研究，Reversible Networks

- 可逆性を利用して，逆伝播時に中間activationsを再構築
- Additive Couplingによる任意の非可逆変換$\boldsymbol{F}_t$
- 大規模モデル，データでも省メモリで学習可

提案手法，データフロー型BP

- Reversible Networksで中間activationsを保持せず再構築することで，データフロー型のBPが実現できる可能性
- モデルを深さ方向に分割しパイプライン処理
- 省メモリ，高いスループット，などが期待できる

（期待される）貢献

- 省メモリなNNの学習
- IoT，FPGAでの学習

展望

- 本当にできるのか実証
- BPとのトレードオフを正当化できるか
- モデリング，アーキテクチャの精査

🤗絶賛検討段階，WIP

研究背景

Neural Networks

- 複雑な関数を線形/非線形変換を交互に重ねて表現（＝近似）

$$\boldsymbol{x}^{(l+1)} = f\left(\boldsymbol{W}^{(l)}\,\boldsymbol{x}^{(l)}\right)$$

  - $\boldsymbol{W}^{(l)}$は線形変換，fは要素毎の非線形活性化関数

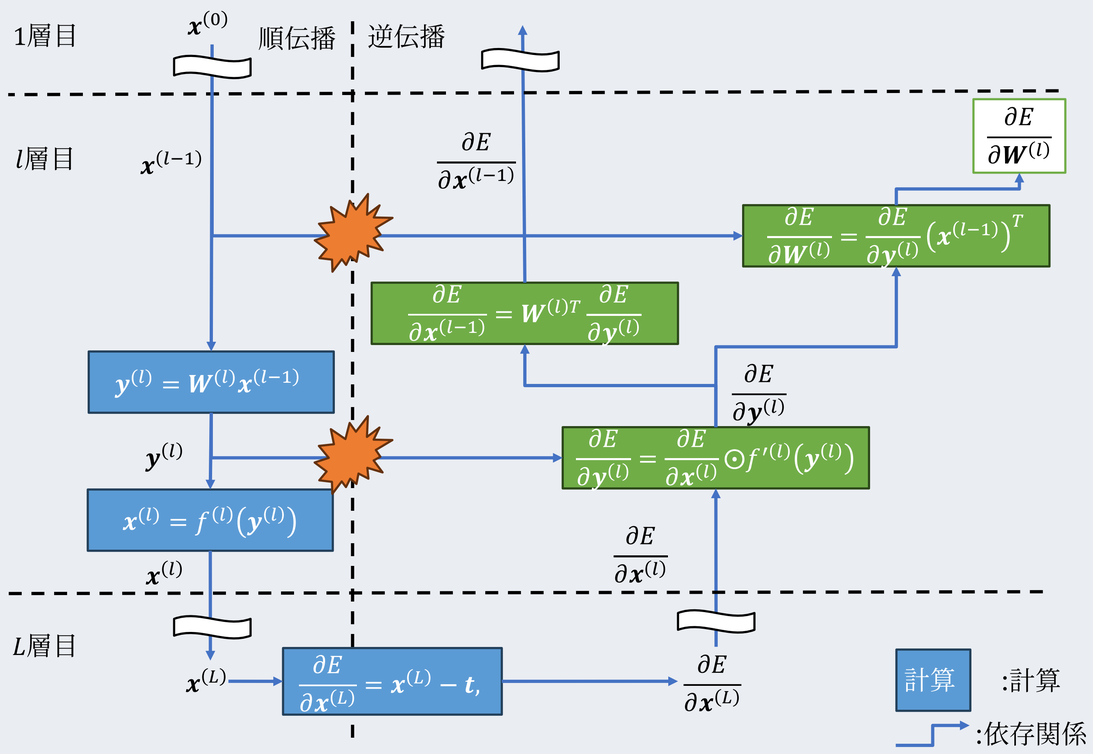
誤差逆伝播法（Backpropagation，BP）

- 誤差勾配を逆伝播し，連鎖律を用いて各層の勾配を効率的に計算

🥰高い計算効率，汎用性，

🥵中間activations $\boldsymbol{x}^{(l)}$,$\boldsymbol{y}^{(l)}$がメモリを圧迫，学習にはメモリが必要

🥵パイプライン並列性が制限される



研究目的

- 膨大なメモリ容量と帯域幅を必要とするBPに代わる学習アルゴリズムの提案，およびそのアーキテクチャ探索

準備，Reversible Networks

- 各層の変換を可逆なものに限定するネットワーク
- 情報圧縮せず解離表現（disentangled representations）を学習
- 可逆性を利用して，逆伝播時に中間activationsを再構築

🥰中間Activationsを保持しないでよいので，省メモリ

🥵順伝播に匹敵する逆計算コスト，計算誤差の累積

🥵表現力の制約（e.g., 線形層は$\boldsymbol{W}$が正方行列かつdet$\boldsymbol{W} \neq \boldsymbol{0}$が必要）

準備，Additive Coupling [1,2]

- 入力$\boldsymbol{x}$をm個に分割（$\boldsymbol{x} = [\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]$）
- m個とばしの残差接続，順方向/逆計算は以下　：

$$\boldsymbol{x}_t = \boldsymbol{F}_t\left(\boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_{t-m+1}\right) + \gamma \boldsymbol{x}_{t-m}$$

$$\boldsymbol{x}_{t-m} = \gamma^{-1}\left(\boldsymbol{x}_t - \boldsymbol{F}_t\left(\boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_{t-m+1}\right)\right)$$

- 🥰任意の非可逆変換$\boldsymbol{F}_t$を利用可能（e.g., CNN, ViT）

準備，RevCol[3]

- 割と大規模なReversible Networks (ICLR 2023)
- サブネットワークをAdditive Couplingで接続
- 可逆性（情報損失無），多様なタスクへ適応
  - ImageNet-22K+プライベートデータ事前学習
  - ImageNet-1K 画像分類: 90.0%
  - COCO 物体検出: APBox 63.8 %
  - ADE20K Semantic segmentation: mloU 61.0%

🥰大規模モデル，データでも省メモリで学習可

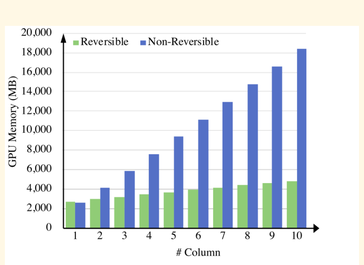
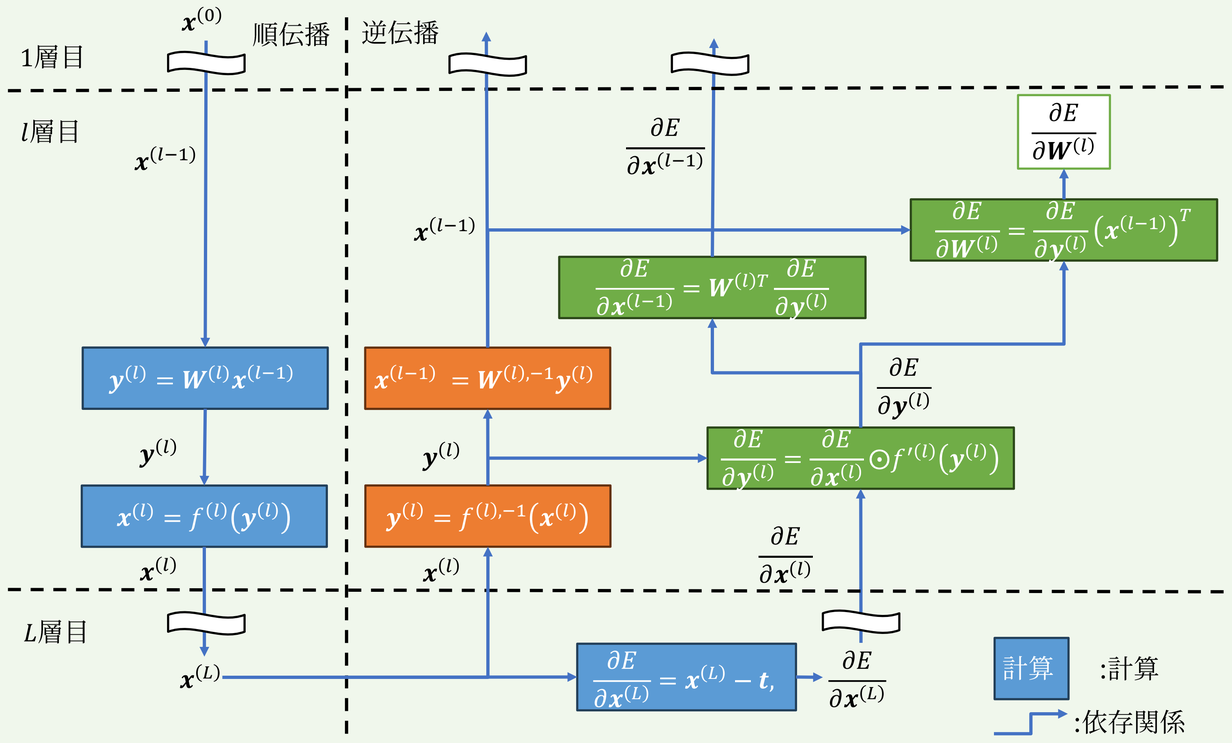


Figure 4: GPU Memory Consumption *vs.* Model size

提案手法，データフローBP

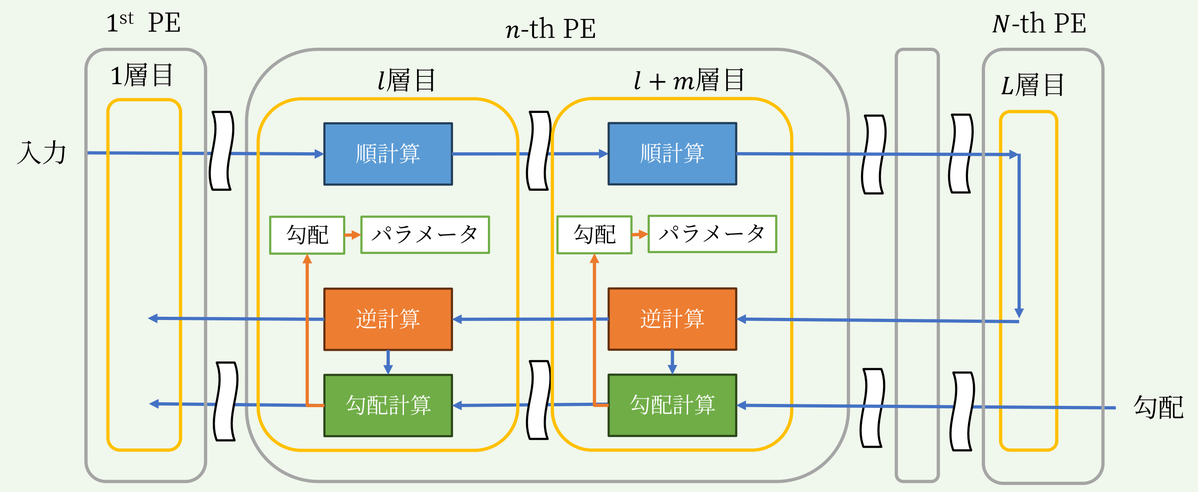
- Reversible Networksは順伝播と逆伝播が完全に分離できる



データフローBP

- モデルを深さ方向に分割し，並列計算資源（PE）に割り当て
- バッチサイズ1の双方向パイプライン，通信遅延は隠蔽される
- 勾配はある程度累積してから更新

Note: 現状でもこのような分割によってパイプライン並列性を利用することはできるが，中間activationsがメモリを圧迫するため，並列性はメモリ容量によってボトルネックされる[4]



🥰パイプライン並列性がもたらす高いスループット

🥰省メモリ（ある程度のバッファは必要）

🥰省電力性（メモリが電力消費の大部分を占めるため）

🥰モデルの深さ方向へのスケーラビリティ

今後の展望

- BPとのトレードオフをどの程度正当化できるのか
  - メモリ使用量のピークがどの程度抑えられるのか
  - 逆計算コストによって計算時間はどの程度遅くなるのか
  - 計算誤差が累積することによって性能はどの程度低下するのか
  - 性能とメモリ使用量のモデリング
  - 本当にスケーラブル…？
- FPGAでやってみたい
  - 推論自体はよくある，学習は新しいはず
  - FPGAクラスタでのスケーラビリティ解析
  - 消費電力やスループットでGPUと勝負できないか？

謝辞

本研究は，JST CREST (JPMJCR21D2) およびAIPチャレンジプログラム, JST SPRING (JPMJSP2108), JSPS 科研費 (24KJ0578) の助成を受けたものです.

参考文献

[1] Dinh, Laurent, David Krueger, and Yoshua Bengio. "Nice: Non-linear independent components estimation." arXiv preprint arXiv:1410.8516 (2014).
[2] Aidan N Gomez, Mengye Ren, Raquel Urtasun, and Roger B Grosse. “The reversible residual network: Backpropagation without storing activations.” NeurIPS 2017.
[3] Yuxuan Cai, Yizhuang Zhou, Qi Han, Jianjian Sun, Xiangwen Kong, Jun Li, and Xiangyu Zhang. Reversible column networks. ICLR 2023.
[4] Penghui Qi, Xinyi Wan, Guangxing Huang, and Min Lin. “Zero Bubble (Almost) Pipeline Parallelism.” ICLR 2024.