学際大規模情報基盤共同利用・共同研究拠点 公募型共同研究 平成31年度採択課題
jh190023-ISH
11th Symposium
Joint Usage / Research Center for Interdisciplinary Large-scale Information Infrastructures

田仲正弘（情報通信研究機構）

超巨大ニューラルネットワークのための分散深層学習フレームワークの開発とスケーラビリティの評価

実施体制

田仲 正弘（研究代表）
国立研究開発法人 情報通信研究機構
ユニバーサルコミュニケーション研究所
データ駆動知能システム研究センター

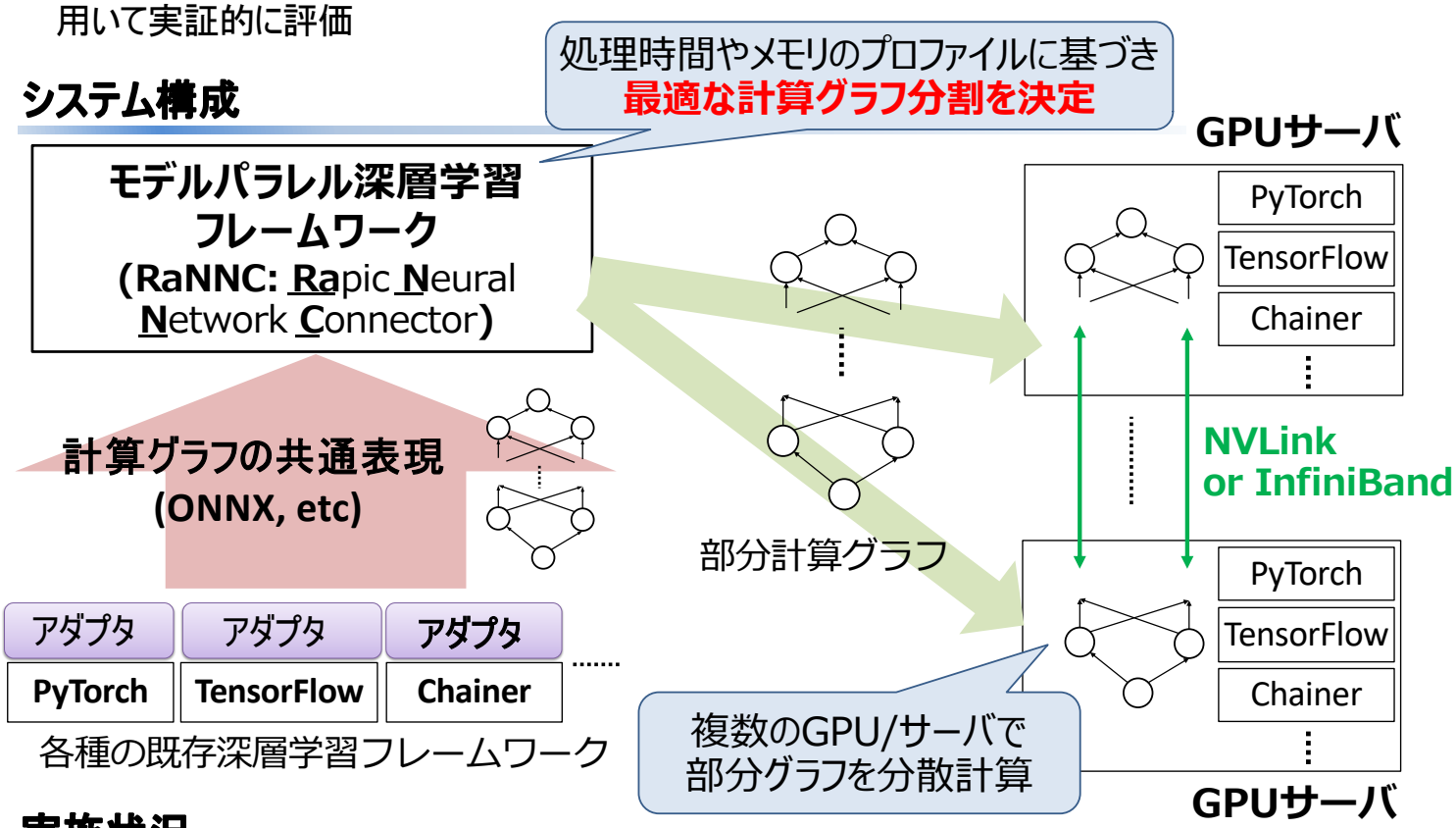
田浦 健次朗（副代表）
東京大学大学院情報理工学系研究科

塙 敏博
東京大学情報基盤センター

概要

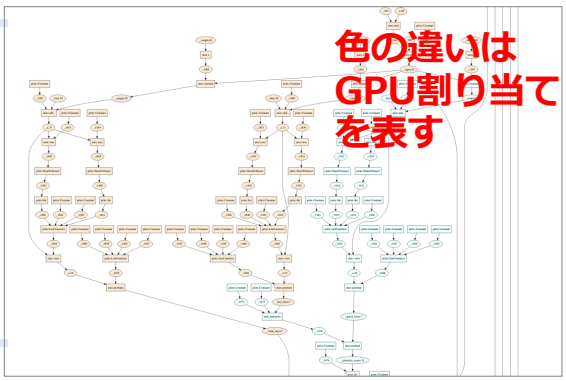
◆ 言語処理分野でブレークスルーをもたらしたBERT等、データパラレルのみでは学習困難な数億～数十億規模のパラメータを持つ超巨大ニューラルネットワークを、モデルパラレルにより学習
◆ ニューラルネットワークの自動分割アルゴリズムを提案し、東京大学情報基盤センターReedbushを用いて実証的に評価

システム構成



実施状況

◆ PyTorchをバックエンドのエンジンとして実装
◆ BERTの学習に成功 (タスク: SQuAD)
  ◆ 使用メモリがおよそ均等になるように分割
  ◆ バッチサイズ1でもGPU1枚に載らない大規模モデルBERT-large(系列長384)を学習可能



BERTネットワークの分割例（一部）

今後の予定

◆ モデルパラレル・データパラレルのハイブリッド機構の導入
◆ ニューラルネットワーク分割アルゴリズムの改善
◆ ソフトウェアはオープンソースで公開予定