

# 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

田中 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

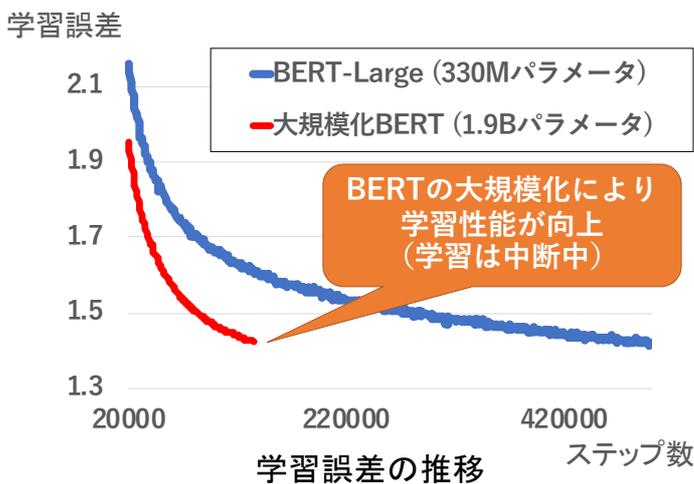
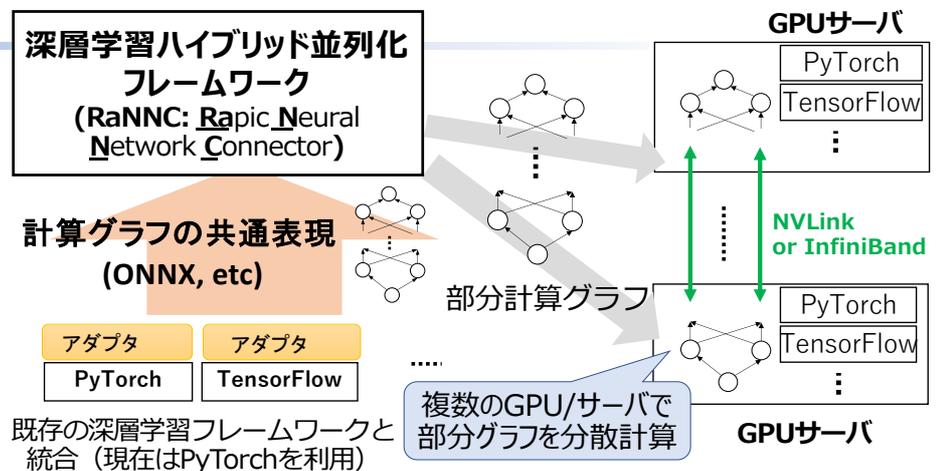


## 概要

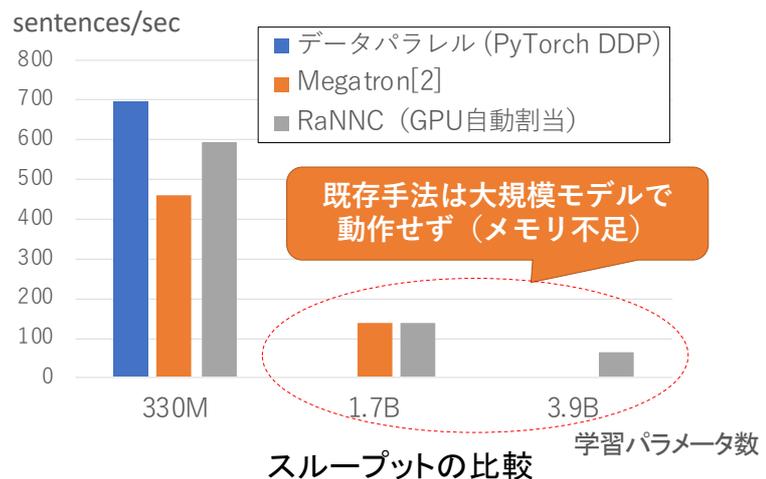
- ◆ 深層学習の並列化において一般的なデータパラレルは、ニューラルネットワークの全体を複製するため、数億・数十億規模のパラメータを持つ超巨大ニューラルネットワークは学習困難
- ◆ ニューラルネットワークを分割するモデルパラレルを自動化するフレームワーク RaNNC (Rapid Neural Net Connector) を開発し、BERT等の巨大ネットワークに適用

## 実施状況

- ◆ モデルパラレル・データパラレルのハイブリッド機構を実現、最大480枚のGPUで分散学習
- ◆ BERT-Largeの5倍以上のパラメータを持つ巨大ニューラルネットワークを学習
- ◆ パイプライン並列の実行効率を最適化するGPU自動割り当てアルゴリズムを導入



系列長128でのpretrainingで評価, NVIDIA V100 256枚を使用, gradient accumulationによりバッチサイズ4kに設定



系列長512でのBERTのpretrainingで評価, バッチサイズを512に設定, NVIDIA V100 32枚 (8枚\*4ノード)を使用, ノード間通信は100Gbps

## 今後の予定

- ◆ より多様なニューラルネットワークでの検証
- ◆ オープンソースでの公開

[1] J. Devlin, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL-HLT'2019, pp. 4171–4186. (2019).  
[2] M. Shoenybi, et al. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism, arXiv:cs.CL/1909.08053 (2019).