TSUBAME 共同利用　令和元年度　学術利用　成果報告書

利用課題名　HPC を利用した自然言語処理技術の研究
英文：High Performance Computing for Natural Language Processing Technology Research

利用課題責任者
鳥澤 健太郎

所属
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
http://www2.nict.go.jp/direct/

邦文抄録（300 字程度）
情報通信研究機構では、対話システム WEKDA を初めとする複数の自然言語処理アプリケーションで、深層学習による様々な分析を行っている。近年発表されている大規模なニューラルネットワークは、一枚の GPU に学習パラメータを収容することが困難であることから、研究代表者らはニューラルネットワークを分割し複数の GPU に格納して分散計算を行う、モデルパラレルと呼ばれる分散学習を行うフレームワーク RaNNC を開発してきた。本課題では RaNNC を拡張し、より複雑なネットワークの学習を可能とすると共に、データパラレルとのハイブリッド学習を実現した。BERT-Large を 5 倍以上の規模に拡大したネットワークを 256 枚の GPU を用いて事前学習し、BERT-Large より優れた学習性能が得られることを確認した。

英文抄録（100 words 程度）
NICT has been developing language processing applications including WEKDA, which intensively use deep learning to analyze a large-scale Web archive. Since some modern networks have billion-scale parameters and do not fit to single GPU, we have been developing RaNNC, which is a deep learning framework for model parallelism. RaNNC divides a neural network into smaller parts and distributes them onto multiple GPUs. In this work, we extended RaNNC to learn more complicated networks with the hybrid data/model parallelism. We trained a BERT network that has five times more parameters than the original BERT using RaNNC. Our pre-training with 256 GPUs showed that the scaled-up BERT results in a better training loss than BERT-Large.

Keywords: 自然言語処理, 大規模情報分析, テキスト分析, モデルパラレル

背景と目的

近年、深層学習で用いられるニューラルネットワークの大規模化が進んでいる。深層学習では GPU 等のアクセラレータを用いるのが一般的であるが、メモリサイズの制約から、巨大なネットワークの学習は容易ではない。そこで計算の並列化と、GPU1 枚あたりの必要メモリ削減の目的のため、データパラレルと呼ばれる、ミニバッチをさらに分割して学習する方式が広く用いられている。しかし、データパラレルではニューラルネットワーク自体が各 GPU に複製されるため、極めて多数のパラメータを持つネットワークは、GPU のメモリに収めることができず、データパラレルでは学習できない。そこで研究代表者らは、ネットワークを分割するモデルパラレルを自動的に行うフレームワーク RaNNC（Rapid Neural Network Connector）を開発してきた。

本課題では、RaNNC を拡張し、より複雑なネットワークへの対応やスケーラビリティの強化に取り組み、BERT-Large[Devlin 2018]の 5 倍以上のパラメータ数を持つネットワークを、256 枚の GPU で学習することに成功した。

概要

本課題では、モデルパレルのための深層学習フレームワーク RaNNC の機能強化を行い、TSUBAME 上で動作検証や性能測定を行った。RaNNC は、既存の深層学習フレームワーク（現在は PyTorch[1]に対応）が出力する計算グラフを複数の部分グラフに分割し、それぞれの部分グラフを異なる GPU に配置して計算する。各 GPU 上では、既存の深層学習フレームワーク

[1] https://pytorch.org/

の計算エンジンを用いて部分グラフを計算するが、入出力を MPI によって通信することにより、部分グラフを結合する。課題開始時点で、AlexNet などの基本的かつ小規模なニューラルネットワークを、数枚程度の GPU 上での学習できる段階まで実現済みであった。課題開始後の進捗は、以下の通りである。

(1)　複雑なモデルへの対応

各 GPU に配置された部分グラフは、PyTorch の計算エンジンを用いて計算を行うが、本来は部分グラフの計算が想定されておらず、グラフの構造によっては、様々な問題が生じる。一例として、計算グラフへの入力値が、複数の計算オペレータから共有される場合の勾配計算が挙げられる。逆伝播において、共有される入力値にはそれぞれの計算オペレータ由来の勾配が加算されるが、PyTorch の計算エンジンはこのようなケースを想定しないため、加算の操作が競合し、正しく計算できない。そのため RaNNC では独自に排他処理を実装している。その他、通信タイミングの制御や複雑なデータ型の通信などを強化し、BERT において、PyTorch のみを用いる場合と全く等価な結果を得られることを確認した。

(2) データパラレルとのハイブリッド化

本課題では、極めて大規模なネットワークの学習を目的としているため、モデルパラレルのみでは、学習に要する時間が極めて長くなり、実用性に乏しい。そのため、モデルパラレル・データパラレルのハイブリッドで、GPU 数百枚規模の学習を可能とするよう拡張した。

本課題では、こうした機能の検証及び速度測定に TSUBAME を使用した。

結果および考察

以上に述べた成果を適用し、BERT の学習を試みた。異なる作成者による PyTorch を用いた BERT モデル[2][3]を対象に、ニューラルネットワークを定義する部分の実装を改変することなく、モデルパラレル・データパラレルのハイブリッドでの学習が実現できた。モデルパラレルを可能とする既存のフレームワークとして、Mesh-TensorFlow[Shazeer 2018], Megatron-LM [Shoeybi 2019]などがあるが、これらを適用するには、既存のモデルの大幅な改修が必要となる。無改変でのモデルパラレルを実現した研究やフレームワークは、研究代表者らの知る限り他にない。

また、BERT-Large の隠れ層サイズを 1920（原論文では 1024）、総数を 54 層（同 24 層）に大規模化したネットワーク（パラメータ総数にして BERT-Large の約 5 倍以上）を、GPU256 枚を用いて事前学習を行った（本実験は課題代表者らの計算設備で実行した）。計算資源確保の問題から、学習途中で中断しているが、図 1 に示すように、BERT-Large と比べて顕著に学習誤差が減少していることが確認できた。

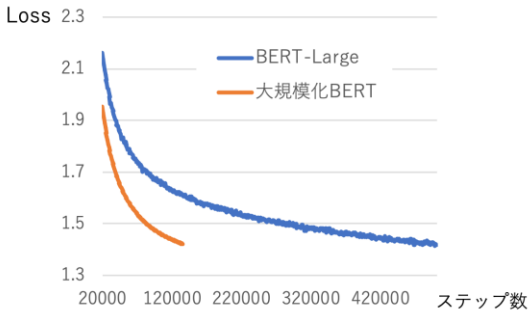

図 1: 大規模化 BERT の事前学習

まとめ、今後の課題

独自に開発したモデルパラレル深層学習フレームワーク RaNNC の強化を行い、BERT-Large の 5 倍以上の規模のネットワークを学習可能とした。今後、GPU 利用効率を向上するなどの更なる改善を進めつつ、T5[Raffel 2019]を初めとする、より巨大なネットワークの学習を進める。

参考文献

[Devlin 2018] Jacob Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019)*, pp. 4171–4186, 2018.

[Shazeer 2018] N. Shazeer, Y. Cheng, N. Parmar, D. Tran, A. Vaswani, P. Koanantakool, P. Hawkins, H. Lee, M. Hong, C. Young, R. Sepassi, and B. Hechtman. Mesh-TensorFlow: Deep learning for supercomputers. *In Neural Information Processing Systems*, 2018.

[Shoeybi 2019] Mohammad Shoeybi and Mostofa Patwary and Raul Puri and Patrick LeGresley and Jared Casper and Bryan Catanzaro. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. arXiv: 1909.08053, 2019.

[Raffel 2019] Colin Raffel et al., Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, arXiv: 1910.10683, 2019.

[2] https://github.com/NVIDIA/DeepLearningExamples/

[3] https://github.com/huggingface/transformers