プレスリリース
2021 年 3 月 31 日

国立研究開発法人情報通信研究機構
国立大学法人東京大学

自動並列化深層学習ミドルウェア RaNNC（ランク）をオープンソースで公開

～超大規模ニューラルネットワークの学習が飛躍的に簡単に～

【ポイント】

- 自動並列化深層学習ミドルウェア RaNNC を開発、オープンソースで公開開始
- 高度な知識と大きな作業コストを要する大規模ニューラルネットワークの学習が飛躍的に簡単に
- 深層学習の大規模化を容易にし、多様な AI システムの更なる性能向上が期待される

国立研究開発法人情報通信研究機構（NICT、理事長: 徳田 英幸）と国立大学法人東京大学（総長: 五神 真）は、自動並列化深層学習ミドルウェア RaNNC（Rapid Neural Net Connector）を開発し、2021 年 3 月 31 日に公開を開始しました。近年、大規模化が進んだ深層学習におけるニューラルネットワークの学習では、複雑なネットワークの定義を書き換え、GPU[*1] のメモリに収まるように人手で分割する必要がありました。今回公開する RaNNC は、ニューラルネットワークを自動的に分割することにより、複数の GPU を用いた並列学習を容易に実現します。大規模ニューラルネットワークの定義を書き換えずに分割を自動化できるソフトウェアは、世界にも例がありません。

RaNNC のソースコードは GitHub[*2] に公開されます（URL: https://github.com/nict-wisdom/rannc）。ライセンスは MIT ライセンスであり、ダウンロードしていただければ、商用目的を含め、無償でご利用いただけます。

【背景】

近年の研究で、深層学習において、ニューラルネットワークの大規模化が大幅な性能向上をもたらすことが知られるようになり、これまでにない大規模なニューラルネットワークが次々に提案されています。深層学習では、計算の高速化のため GPU を用いることが一般的ですが、2020 年に提案された GPT-3[*3] 規模のニューラルネットワークの学習には、数千枚の GPU が必要とされています。このような大規模ニューラルネットワークによる深層学習の高性能化には、大量の GPU を効率よく使用できる、大規模ニューラルネットワークの並列計算の技術が重要になります。

従来の大規模ニューラルネットワークの学習では、GPU のメモリに収まるようにニューラルネットワークを分割するため、複雑なニューラルネットワークの定義を大幅に書き換える必要がありました。しかし、これには高度な知識と大きな作業コストを要するため、これまで大規模ニューラルネットワークの学習を実施できる組織はごく限られていました。

【今回の成果】

今回公開する RaNNC は、大規模ニューラルネットワークの学習を容易にするためのミドルウェアで、NICT データ駆動知能システム研究センターと東京大学情報基盤センターとの共同研究によって開発されました。

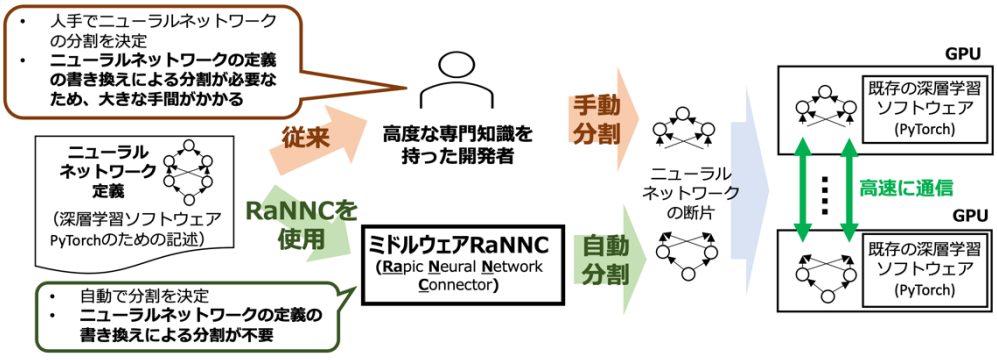

図 1　大規模ニューラルネットワークの並列学習

RaNNC は、代表的な深層学習ソフトウェアである PyTorch[*4] のために記述された既存のニューラルネットワークの定義を与えられると、実行速度を最適化しつつ、各々の GPU のメモリに収まるよ

うに、自動的にニューラルネットワークを分割した上で、データ並列・モデル並列[*5] のハイブリッドによって、複数のGPU を使い並列に学習を行います（図 1 参照）。開発者がニューラルネットワークの定義を書き換えて分割できるようにする必要がないため、大規模ニューラルネットワークの学習が飛躍的に容易になります。また、既存ソフトウェアのMegatron-LM[*6] や Mesh-TensorFlow[*7] は、BERT[*8][1]等の特定のタイプのニューラルネットワークにしか使用できないという制限がありますが、RaNNC は適用できるニューラルネットワークの種類に基本的に制限がないという点で優れています。

NICT の計算機環境における比較実験では、RaNNC は Megatron-LM より約 5 倍の規模のニューラルネットワークの学習が可能な一方、同一の規模のニューラルネットワークでは、ほぼ同等の学習速度を実現しています。なお、この成果は、並列分散処理分野におけるトップレベルの国際会議である IPDPS（IEEE International Parallel and Distributed Processing Symposium）に採択されています[2]。また、RaNNC の概要は、GPU テクノロジカンファレンス（GTC）（2021 年 4 月 12-16 日開催）で発表予定です。

NICT データ駆動知能システム研究センターでは、これまで収集してきた高品質な日本語テキスト約 350 GB を学習データとし、RaNNC を用いて、BERT を約 50 億パラメータ（原論文の 15 倍）に大規模化したニューラルネットワークを学習しています。こうした規模のニューラルネットワークを、ネットワークの定義を書き換えて分割できるようにする手間なしに、自動分割し、並列で学習させるソフトウェアは、我々の知る限り世界にも例がありません。

RaNNC のソースコードは、GitHub に公開されます。ライセンスは MIT ライセンスのため、ダウンロードしていただければ、商用目的を含め、無償でご利用いただけます。（URL: https://github.com/nict-wisdom/rannc）

【今後の展望】

NICT データ駆動知能システム研究センターでは、大規模 Web 情報分析システム WISDOM X[*9]、高齢者介護用マルチモーダル音声対話システム MICSUS[*10]、次世代音声対話システム WEKDA[*11]、対災害情報分析システム DISAANA[*12]、災害状況要約システム D-SUMM[*13] など多数のシステムを開発し、一般公開や民間企業へのライセンスを行っています。これらのシステムでは様々なニューラルネットワークが使用されていますが、RaNNC を用いて学習された、より大規模ニューラルネットワークをこれらのシステムで利用することで、更なる性能向上が期待されます。

また、RaNNC はオープンソースソフトウェアとして MIT ライセンスで一般公開されるため、深層学習を用いたシステムを開発する多くの組織で、大規模なニューラルネットワークの学習が可能になり、様々な技術や製品、サービスの開発が幅広く促進されることが期待されます。

＜各機関の役割分担＞

・NICT: RaNNC 全体の開発、動作検証、評価

・東京大学: RaNNC の高速化

＜関連するプレスリリース＞

・2021 年 3 月 31 日 大規模 Web 情報分析システム WISDOM X「深層学習版」の試験公開を開始
https://www.nict.go.jp/press/2021/03/31-3.html

＜ 本件に関する問合せ先 ＞
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
田仲 正弘
E-mail: wisdom-contact@ml.nict.go.jp

＜ 広報（取材受付）＞
国立研究開発法人情報通信研究機構
広報部 報道室
Tel: 042-327-6923
E-mail: publicity@nict.go.jp

国立大学法人東京大学
情報基盤センター 総務チーム
Tel: 03-5841-2710
E-mail: jouhousoumu.adm@gs.mail.u-tokyo.ac.jp

＜用語解説＞

***1 GPU（Graphics Processing Unit）**

元々は、画像処理に用いられる計算に特化した演算器を備えた装置であったが、近年は、並列処理により高い演算性能を得られるため、広く汎用演算に使われるようになった。特に、深層学習では膨大な演算を効率よく並列処理できるため、GPU が多く使われている。一方、CPU に比べて計算データを格納するメモリが小さいため、特に、深層学習においては、巨大なニューラルネットワークのデータを格納できないことが課題となる。

***2 GitHub**

オープンソースの様々なソフトウェアのソースコードが登録されたサイト。登録されたソフトウェアのソースコードは誰でも取得できる。

***3 GPT-3**

人工知能研究を行う非営利団体 OpenAI が 2020 年に提案したニューラルネットワーク。人間と判別が困難なレベルの高品質な文章の生成が可能なことで知られている。現在は、マイクロソフトに独占ライセンスされている。

***4 PyTorch**

Facebook が開発した、深層学習のためのソフトウェア。Google が開発した TensorFlow と並び、最も広く使われている深層学習のためのソフトウェアの一つである。

***5 データ並列・モデル並列**

いずれも深層学習の並列化方式。データ並列では、ニューラルネットワークの全体を複製し、異なる GPU 等で並列に計算する。実現が容易なため、多くの既存の深層学習ソフトウェアが、データ並列のための機能を標準で備えている。一方、モデル並列は、ニューラルネットワークを分割し、分割によって得られた断片を異なる GPU 等で並列に計算する。データ並列と異なり、高い並列化の効果を得られるようにニューラルネットワークを分割することが難しいため、既存の多くの深層学習ソフトウェアは、モデル並列のための機能をほとんど持たない。

***6 Megatron-LM, *7 Mesh-TensorFlow**

それぞれ NVIDIA と Google が開発した、モデル並列によって大規模ニューラルネットワーク学習を行うソフトウェア。データ並列も併用する。BERT 及び同種の構造を持つニューラルネットワークの学習に特化しており、その他のニューラルネットワークに適用するのは困難である。

***8 BERT**

2018 年に Google から発表されたニューラルネットワーク。質問応答など、言語処理分野における様々なタスクで、従来の最高性能を更新した。その後、BERT を拡張あるいは参考にして多くのニューラルネットワークが提案されるなど、言語処理分野の深層学習研究に極めて大きな影響を与えた。

***9 大規模 Web 情報分析システム WISDOM X（ウィズダム エックス）**

数十億件の Web ページを深く意味まで分析し、「なに？」「なぜ？」「どうなる？」といったタイプの様々な質問に回答できる質問応答システム。2015 年から https://wisdom-nict.jp で試験公開中。どのような質問を入力すべきか分からない場合には、キーワードを入力すると回答可能な質問を提案するほか、質問の回答から更なる質問を提案し、情報の更なる深堀りを行ったり、Web 上に書かれていない仮説を生成したりすることも可能。2021 年 3 月から、深層学習を導入するとともに、新たに「どうやって/どうしたら」型の質問に回答できるように機能強化。約 350 GB という大量の Web テキストや NICT で構築した高品質かつ大量の学習データで BERT を学習し、さらに、独自技術と組み合わせるなどして、より広範な質問へのより高い精度での回答を実現した。

***10 高齢者介護用マルチモーダル音声対話システム MICSUS（ミクサス）**

要支援等の認定を受けている在宅高齢者に対して、介護モニタリングと呼ばれる、健康状態や生活習慣のチェックを音声での対話で行い、ケアプランの作成等に役立てるために開発中の対話システム。本来介護モニタリングを実施するケアマネジャーと呼ばれる職種の介護職の負担を軽減するとともに、現在月一回とされている介護モニタリングの頻度を増やし、より高品質なケアにつなげることを目的とする。また、次世代音声対話システム WEKDA や WISDOM X の技術を使って Web 情報を用いた雑談的対話も行い、高齢者に飽きられることなく普段使いをしてもらうことも狙っており、さらには高齢者の健康状態を阻害する要因となるコミュニケーション不足の解消も狙う。内閣府の戦略的イノベーション創造プログラム（SIP）（第 2 期）の支援の下、KDDI 株式会社、NEC ソリューションイノベータ株式会社、株式会社日本総合研究所と NICT が共同開発を行っている。更なる詳細は、https://www.youtube.com/watch?v=gCUrC3f9-Go を参照のこと。（YouTube で“MICSUS”と検索しても閲覧可能）

***11 次世代音声対話システム WEKDA（ウェクダ）**

多様な話題に関して、ユーザとブレインストーミング的な雑談を行うことを最終目標として狙った、NICT が開発している次世代音声対話システム。WISDOM X とほぼ同じ仕組みで様々な質問に回答するほか、「対話システムを作っています。」のような平叙文の音声入力に対しても、Web の情報を用いて「対話システムを用いて回想法を行い、認知症の予防、改善をしよう」といった応答を行う。

***12 対災害情報分析システム DISAANA（ディサーナ）**

Twitter で発信された災害関連情報をリアルタイムに分析し、「熊本県で土砂災害が起きているのはどこか？」「熊本県で何が不足しているか？」といった質問への回答を地図上に可視化し、災害の被災状況の把握を容易にするシステム。NICT が開発し、2015 年から https://disaana.jp/にて試験公開中。さらに、日本電気株式会社が、この技術の商用ライセンスを受け、2020 年から商用サービスの販売を行っている。

***13 災害状況要約システム D-SUMM（ディーサム）**

DISAANA と同様に、Twitter で発信された災害関連情報をリアルタイムに分析するが、質問に回答するのではなく、自治体名を指定すると関連する被災報告等を簡潔に要約し、被災状況の全体像の把握を容易にするシステム。内閣府の戦略的イノベーション創造プログラム（SIP）（第 1 期）による支援の下、NICT が開発し、2016 年から https://disaana.jp/d-summ/にて試験公開中。自治体等においても実際に活用され、豪雨による鉄橋流失を鉄道会社に先駆けて発見することに成功する等の事例もある。DISAANA と同様に、日本電気株式会社が、この技術の商用ライセンスを受け、2020 年から商用サービスの販売を行っている。

補足資料

深層学習のためのニューラルネットワークの大規模化

近年の研究で、深層学習のためのニューラルネットワークが、大規模化によって大幅な性能向上を示すことが知られるのに伴って、従来になかった大規模ニューラルネットワークが次々に提案されています。例えば、2018 年に発表され言語処理分野のブレークスルーとなった BERT は、発表当時として最大規模の 3.4 億の学習パラメータを持ちますが、2020 年に提案された GPT-3 は、BERT の 500 倍以上の学習パラメータ（1,750 億）を持ちます。深層学習では、計算の高速化のため GPU を用いることが一般的ですが、GPT-3 規模のニューラルネットワークの学習には、数千枚の GPU が必要とされています。このような大規模ニューラルネットワークによる深層学習の高性能化には、複数の GPU を効果的に使用できる、大規模ニューラルネットワークの並列計算の技術が重要になります。

大規模ニューラルネットワークの並列計算技術

従来の深層学習のためのソフトウェアの多くは、データ並列と呼ばれる並列化方式のみに対応していました。しかし、データ並列はニューラルネットワークの全体を複数の GPU 上に複製するため、極めて多数の学習パラメータを持つニューラルネットワークで GPU のメモリに収まらないものは、やはり学習できないという問題がありました。そこで、特に巨大なニューラルネットワークの学習には、モデル並列と呼ばれる、ニューラルネットワークを分割して異なる GPU で並列に計算する方法が用いられます。しかし、モデル並列を用いる既存のソフトウェアでは、利用者がニューラルネットワークを GPU のメモリに収まるように分割する作業が必要となります。具体的には、既存の著名な大規模ニューラルネットワーク学習のソフトウェアである Megatron-LM や Mesh-TensorFlow を用いる場合、しばしば数千行にも及ぶプログラムとして定義されるニューラルネットワークを、計算サーバや GPU、それらをつなぐネットワークの性能等を考慮しながら、全体にわたって書き換える必要があります。これには高度な知識と大きな作業コストを要するため、これまで大規模ニューラルネットワークの学習を実施できる組織はごく限られていました。

参考文献

[1] 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.

[2] Masahiro Tanaka, Kenjiro Taura, Toshihiro Hanawa and Kentaro Torisawa, Automatic Graph Partitioning for Very Large-scale Deep Learning, In the Proceedings of 35th IEEE International Parallel and Distributed Processing Symposium (IPDPS 2021), May, 2021.（発表予定）