

自動並列化深層学習ミドルウェア RaNNC が PyTorch Annual Hackathon 2021 において First Place(第 1 位)を受賞

国立研究開発法人情報通信研究機構(NICT、理事長: 徳田 英幸)データ駆動知能システム研究センター(DIRECT)と国立大学法人東京大学(総長: 藤井 輝夫)情報基盤センターが共同で開発した自動並列化深層学習ミドルウェア RaNNC(ランク、Rapid Neural Network Connector)が、PyTorch Annual Hackathon 2021 において、First Place(第 1 位、PyTorch Developer Tools & Libraries 部門)を受賞しました。

PyTorch Annual Hackathon は、世界中で広く使用される深層学習の代表的ソフトウェア PyTorch に関する成果を競うイベントとして、PyTorch の開発元である Facebook が公式に開催する唯一のイベントであり、世界中から数多くの参加者を集めています。今回受賞したミドルウェア RaNNC は、PyTorch の従来機能では困難であった大規模ニューラルネットワークの学習を、飛躍的に簡単化するものです。RaNNC はオープンソースで一般公開されており、ダウンロードしていただくことで、商用目的を含め無償でご利用いただけます。



PyTorch Annual Hackathon(URL: <https://pytorch2021.devpost.com/>)は、深層学習で使うニューラルネットワークを記述するためのフレームワークである PyTorch を使用したソフトウェアや機械学習モデルを開発するイベントです。PyTorch に関する成果を競うイベントとして、PyTorch の開発元である Facebook が公式に開催する唯一のものであり、2019 年から年に 1 度開催されています。今年は世界各国から 1,947 人が参加し、65 件の応募がありました。

RaNNC は、大規模なニューラルネットワークを自動で分割し、複数の GPU^{*1} を用いた並列処理によって高速に学習するためのミドルウェアです。大規模ニューラルネットワークの学習では、学習パラメータが GPU メモリに収まらないため、ニューラルネットワークを分割して複数の GPU 上で並列処理する必要があります。そのため、従来は、ニューラルネットワークの分割や、並列処理の速度を考慮したニューラルネットワークの記述を手で行う必要があり、多くの手間と専門知識が必要とされました。一方、RaNNC は、単一の GPU を使用することを想定した、並列化を考慮しないニューラルネットワーク記述を与えると、GPU メモリに収まり、かつ並列処理によって高い学習速度が得られるようにニューラルネットワークを自動で分割するため、大規模ニューラルネットワークの学習を大幅に容易化できます。

RaNNC は 2021 年 3 月から一般公開されていますが、PyTorch Annual Hackathon 2021 への応募に当たって、新機能を追加しました。具体的には、大規模ニューラルネットワーク学習に必要な GPU メモリを削減するため、学習パラメータの多くをより大きなメインメモリに配置しておき、GPU を用いた計算が必要となるタイミングで、必要なデータのみを GPU のメモリに移す機能を実現しました。これにより、より小さな GPU メモリで、大規模ニューラルネットワークの学習が可能となりました。

RaNNC は情報通信研究機構と東京大学の共同研究として開発されました。受賞の対象者は以下のとおりです。

田中 正弘	情報通信研究機構 データ駆動知能システム研究センター 主任研究員
田浦 健次郎	東京大学大学院情報理工学系研究科 電子情報学専攻 教授 / 東京大学情報基盤センター センター長
塙 敏博	東京大学 情報基盤センター スーパーコンピューティング研究部門 教授
鳥澤 健太郎	情報通信研究機構 フェロー / ユニバーサルコミュニケーション研究所 副所長 / データ駆動知能システム研究センター 主管研究員

これまでに RaNNC を用いて、1,000 億パラメータ規模のニューラルネットワークの学習の自動並列化が実現できています。従来、こうした規模のニューラルネットワークの学習は、元々は並列化を考慮せずに記述されたニューラルネットワークの記述を、専門の技術者が大幅に改変することで実現されており、その改変のコストは大きなものでした。一方 RaNNC は、並列化を一切考慮せず、単一の GPU を使用する想定で記述されたニューラルネットワークの学習を、自動的に並列化し、高速な学習を実現できています。これによって、並列処理向けのニューラルネットワークの記述の改変が不要となり、より低コスト・短時間で大規模なニューラルネットワークの並列での学習を可能にします。

また、大規模ニューラルネットワーク学習のための主要な既存ソフトウェアは、Transformer^{*2} 等、特定の種類のニューラルネットワークにしか適用できない一方で、RaNNC は基本的にニューラルネットワークの種類を選ばず適用可能であるという特長があります。

RaNNC は、ソースコードや利用例と共に、GitHub^{*3} で公開されています (URL: <https://github.com/nict-wisdom/rannnc>)。ライセンスは MIT ライセンスとしており、商用目的を含め、無償で利用できます。

<関連するプレスリリース>

- ・2021 年 3 月 31 日 自動並列化深層学習ミドルウェア RaNNC 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～
<https://www.nict.go.jp/press/2021/03/31-2.html>

<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 Transformer

2017 年に発表された、主に自然言語処理分野で用いられるニューラルネットワーク。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 BERT で利用されるなど、その後の深層学習研究に大きな影響を与えた。

*3 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。

< 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 広報担当 (大林)
Tel: 080-9422-7780
E-mail: itc-press@itc.u-tokyo.ac.jp