NEWS RELEASE

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

2023 年（令和 5 年）10 月 20 日

LLM勉強会
LLM-jp

130 億パラメータの大規模言語モデル「LLM-jp-13B」を構築 ～NII 主宰 LLM 勉強会（LLM-jp）の初期の成果を アカデミアや産業界の研究開発に資するために公開～

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所（ＮＩＩ、所長：黒橋 禎夫、東京都千代田区）は、本年 5 月から、自然言語処理及び計算機システムの研究者を中心として、大学・企業等から 500 名以上が参加する LLM 勉強会（LLM-jp）を主宰しています。7 月から、計算資源としてデータ活用社会創成プラットフォーム mdx[*1] を活用し、パラメータ数[*2]130 億の大規模言語モデル（LLM）の構築を開始しました。このたび、同 LLM の事前学習及びチューニングが終了し、モデルを公開しましたので、お知らせします。

同モデルは LLM 研究開発としては初期段階のものであり、モデルの性能を示す評価値はこれまでに国内で公開されたモデルと比較して同程度ではありますが、今後のアカデミアや産業界の研究開発に資するため、コーパス[*3]等を含めてすべてオープンにするものです。

今後、国立研究開発法人 産業技術総合研究所（産総研）や国立大学法人 東京工業大学（東工大）とも協力して世界トップレベルの性能を持つ LLM の構築を目標に研究開発を進めます。その第一歩として、産総研の計算資源である AI 橋渡しクラウド（ABCI）[*4]の利用により、より本格的な 1750 億パラメータ（GPT-3 級）の LLM の構築に着手しました。

また、LLM を社会で利活用していく上では、LLM の透明性・信頼性の確保が必要であり、モデルの高度化に伴い、安全性の配慮がより重要となります。今回のモデルや今後構築するモデルを活用してそれらの研究を進め、LLM 研究開発の促進に貢献します。

(*1) データ活用社会創成プラットフォーム mdx：9 大学 2 研究所が連合して共同運営する、データ活用にフォーカスした高性能仮想化環境。研究環境を用途に合わせてオンデマンドで短時間に構築・拡張・融合できる、データ収集・集積・解析のためのプラットフォーム。
(*2) パラメータ数：大規模言語モデルは言語を学習した大規模なニューラルネットワークで、パラメータはニューラルネットワークの規模を示す指標のひとつ。パラメータ数が多いほど高い性能であるといわれている。
(*3) コーパス：自然言語の文章を構造化し大規模に集積したデータベース。
(*4) AI 橋渡しクラウド（ABCI）：産業技術総合研究所（AIST）が提供する AI 向け計算用で現状国内最大の計算資源。

National Institute of Informatics

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
総務部企画課 広報チーム
〒101-8430 東京都千代田区一ツ橋 2-1-2
直通：03-4212-2164 FAX：03-4212-2150
E-Mail: media@nii.ac.jp

Web: https://www.nii.ac.jp
Twitter: @jouhouken
facebook: https://www.facebook.com/jouhouken

1．　LLM 勉強会（LLM-jp）の概要

（1）　NII が主宰する LLM-jp では、自然言語処理及び計算機システムの研究者を中心として、大学・企業等から 500 名以上が集まり、ハイブリッド会議、オンライン会議、Slack 等を活用して LLM の研究開発について情報共有を行うとともに、共同で LLM 構築等の研究開発を行っています。具体的には、以下の目的で活動しています。

- オープンかつ日本語に強い LLM の構築とそれに関連する研究開発の推進
- 上記に関心のある自然言語処理および関連分野の研究者によるモデル構築の知見や最近の研究の発展についての定期的な情報交換
- データ・計算資源等の共有を前提とした組織横断的な研究者間の連携の促進
- モデル・ツール・技術資料等の成果物の公開

（2）　LLM 構築に当たっては、「コーパス構築 WG」「モデル構築 WG」「チューニング・評価 WG」等を設置し、それぞれ、早稲田大学 河原大輔教授、東北大学 鈴木潤教授、東京大学 宮尾祐介教授を中心に研究開発活動に取り組んでいます。また、このほか、東京大学情報基盤センター 田浦健次朗センター長／教授、空閑洋平准教授（計算資源 mdx の利用）、東工大 横田理央教授（並列計算手法等）等、多数の方々の貢献により、活動を進めています。

（3）　詳細については、ホームページ https://llm-jp.nii.ac.jp/ をご参照ください。

2．　今回構築した LLM「LLM-jp-13B」の概要

（1）　利用計算資源等

- データ活用社会創成プラットフォーム mdx　12 ノード（A100 96 枚）
- NII、理化学研究所革新知能統合研究センター（AIP）、学術大規模情報基盤共同利用・共同研究拠点（JHPCN）で経費を負担
- モデル構築に当たっては、Microsoft の DeepSpeed の技術を利用
- モデル構築時の監視やログの保存には Weights & Biases を利用

（2）　モデル学習用コーパス

- トークナイザー、ウェブコーパスのフィルタリングツールを整備し、コーパスを構築
- 学習データ量　約 3000 億トークン
  （日本語 約 1450 億トークン：日本語 mC4、日本語 Wikipedia、
  英語 約 1450 億トークン：英語 Pile、英語 Wikipedia、
  プログラムコード 約 100 億トークン）

（3） モデルのパラメータ数

- 130 億個（13B）

（4） チューニング・評価

- チューニング：日本語インストラクションデータおよび英語インストラクションデータの和訳データ 12 種類を用いてチューニング実験を実施、評価結果を分析した上でチューニング済みモデルとして今回公開。
- 評価：既存の日本語言語資源を利用した 9 種類の評価データを整備、横断的に評価を行うフレームワークを構築（現在開発中のデータを含めると 50 種類程度の評価データが構築される予定）。

（5） 開発モデル・ツール・コーパスの公開 URL

https://llm-jp.nii.ac.jp/release

注：今回公開するモデルは、まだ研究開発の初期段階のものであり、安全性の観点など人間の意図に沿った出力を行うようにチューニングを行ったものではないため、そのまま実用的なサービスに供することを想定しているものではありません。

3． 今後の予定

- モデルの挙動を分析する上で重要なモデル学習用コーパスの検索ツールも整備中で、近日公開します。
- 今後、産総研や東工大とも協力して世界トップレベルの性能を持つ LLM の構築を目標に研究開発を進めます。その第一歩として、2023 年 9 月、産業技術総合研究所 ABCI の第 2 回大規模言語モデル構築支援プログラム（https://abci.ai/ja/link/llm_support_program2023.html）に採択されましたので、今回構築・公開した 130 億パラメータの LLM 構築の知見に基づき、国内の他の組織が構築、公開していない 1750 億パラメータの LLM の構築に 10 月より着手しました。
- LLM を社会で利活用していく上では、LLM の透明性・信頼性の確保が必要であり、モデルの高度化に伴い、安全性の配慮がより重要となります。今回のモデルや今後構築するモデルを活用してそれらの研究を進め、LLM 研究開発の促進に貢献します。

〈メディアの皆様からのお問い合わせ先〉
大学共同利用機関法人 情報・システム研究機構 国立情報学研究所
総務部企画課 広報チーム
TEL:03-4212-2164 E-mail：media@nii.ac.jp