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

April 30, 2024

Development of the Large Language Model “LLM-jp-13B v2.0”

-- The NII-hosted LLM Research Group (LLM-jp) releases the successor model of “LLM-jp-13B,” and makes all resources used for development open for the public --

The National Institute of Informatics (NII, Director-General: Sadao Kurohashi, Tokyo, Japan) has been supervising the LLM Research Group (LLM-jp), consisting of over 1200 participants (As of the end of April 2024) including researchers in natural language processing and computer systems from universities and corporations, since May of last year. In October 2023, we released the Large Language Model “LLM-jp-13B” with 13 billion parameters(*1) as an initial product.

Based on the experience, we developed the “LLM-jp-13B v2.0” as the successor, utilizing the platform for building a data-empowered society “mdx(*2)” as a computational resource, improving the corpora(*3) and model structure, and introducing fine-tuning that takes safety into account. Today, we are pleased to announce that the model has been released.

As the performance of LLM improves and its use in society comes into general, it is essential to ensure their transparency and reliability. We will advance research using this and future models and contribute to the promotion of LLM research and development.

1. Overview of the Newly Built LLM "LLM-jp-13B v2.0"

(1) Major changes from "LLM-jp-13B v1.0"

- Improving the Japanese web corpus: A new corpus called “Japanese Common Crawl” was constructed and used for pre-training. We used “Uzushio” to extract and filter Japanese text from the entire large-scale web archive Common Crawl. The quality has been significantly improved compared to the Japanese web corpus "Japanese mC4" used for "LLM-jp-13B v1.0".
- Improving model architecture: We adopted a modern model architecture with various improvements. Extended the maximum token length from 2,048 to 4,096 to handle longer contexts.
- Fine-tuning with considering safety: A new dataset based on a safety perspective is constructed and used for fine-tuning.

National Institute of Informatics

Research Organization of Information and Systems
National Institute of Informatics
Publicity Team
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo
101-8430 JAPAN
Direct: +81(0)3-4212-2164 FAX : +81(0)3-4212-2150
E-Mail: media@nii.ac.jp

Web: https://www.nii.ac.jp
X(旧 Twitter): @jouhouken
facebook: https://www.facebook.com/jouhouken

(2) Computational Resources Used

- mdx (a platform for building a data-empowered society): Utilized 16 nodes, NVIDIA A100 GPU 128 pieces.
- Funded by: NII, RIKEN Center for Advanced Intelligence Project (AIP), and Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures (JHPCN).
- Model Construction: NVIDIA's Learning framework "Megatron-LM" was employed.
- Monitoring and Logging: The experiment tracking platform "Weights & Biases" was used for monitoring evaluation indices and logging during model construction.

(3) Corpora Used for Model Training

- Training Data Volume: Approximately 260 billion tokens
  - Japanese: Approximately 130 billion tokens (Japanese Common Crawl, Japanese Wikipedia)
  - English: Approximately 120 billion tokens (English Pile, English Wikipedia)
  - Program code: Approximately 10 billion tokens

(4) Model

- Number of Model Parameters: 13 billion (13B)
- Model Architecture: Followed LLaMA Base
- Maximum token length: 4,096

(5) Fine-Tuning

- Conducted fine-tuning experiments using 8 types of Japanese instruction data and English instruction data translated into Japanese.

(6) Evaluation

- Used "llm-jp-eval v1.3.0," which enables a multi-perspective evaluation by using 22 types of evaluation data from existing Japanese language resources.
- Used "Japanese Vicuna QA" and "Japanese MT Bench," GPT-4 based automatic evaluation frameworks for generated text.
- Evaluated the safety of generated text manually.
- In all evaluations, significant performance improvements were confirmed compared to "LLM-jp-13B v1.0."

(7) URL for the Released Models, Tools, and Corpora

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

Note : Though the model released this time has been fine-tuned from a safety perspective, it is still in the initial stages of research and development. Therefore, it is not intended to be provided as is for practical service.

2. Overview of LLM Research Group (LLM-jp)

1. NII hosts the LLM-jp, where over 1200 participants (As of the end of April 2024), including researchers in natural language processing and computer systems from universities and corporations, participate. LLM-jp utilizes hybrid meetings, online conferences, and Slack and other tools, to share information on LLM research and development and collaboratively work on building LLMs. Specifically, activities are conducted for the following purposes:

- Construction of an open LLM that is proficient in Japanese and the promotion of related research and development
- Regular information exchange on model building insights and recent developments in research for researchers interested in natural language processing and related fields
- Promotion of cross-organizational collaborations among researchers, predicated on the sharing of data and computational resources
- Public release of outcomes, including models, tools, and technical materials

2. For LLM construction, LLM-jp has established working groups such as "Corpus Construction WG," "Model Construction WG," "Fine-tuning & Evaluation WG," “Safety WG” and “Multi-modal WG.” Each group, led respectively by Professor Daisuke Kawahara of Waseda University, Professor Jun Suzuki of Tohoku University, Professor Yusuke Miyao of the University of Tokyo, Project Professor Satoshi Sekine of NII, and Professor Naoaki Okazaki of Tokyo Tech, is actively engaged in research and development activities. Additionally, the initiative is propelled by the contributions of many others, including Professor Kenjiro Taura of the University of Tokyo, Associate Professor Yohei Kuga of the University of Tokyo (for the utilization of computational resource mdx), and Professor Rio Yokota of Tokyo Tech (parallel computation methods).

3. For more details, please refer to the website

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

3. Future Plans

To utilize LLMs in society, it is essential to ensure their transparency and reliability. Therefore, NII has established the Research and Development Center for Large Language Models in April 2024, supported by the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology (P7, https://www.mext.go.jp/content/20240118-ope_dev03-000033586-11.pdf).
We will advance research using this and future models, such as 175 billion parameters scale LLM supported by the GENIAC by the Ministry of Economy, Trade and Industry, contributing to the promotion of LLM research and development.

<Media Contact>
National Institute for Informatics Research Organization of Information and Systems
Publicity Team, Planning Division, General Affairs Department
TEL：+81-3-4212-2164　E-mail：media@nii.ac.jp

(*1) **Number of Parameters**: Large language models are massive neural networks trained on language data, and the number of parameters is one of the indicators of the network's size. It is generally believed that more parameters indicate higher performance.
(*2) **mdx (a platform for building a data-empowered society)**: A high-performance virtual environment focused on data utilization, jointly operated by a consortium of 9 universities and 2 research institutes. It is a platform for data collection, accumulation, and analysis that allows users to build, expand, and integrate research environments on-demand in a short amount of time, tailored to specific needs.
(*3) **Corpus**: A database that stores large amounts of natural language texts in a structural manner.