DOOHAE JUNG

Location: Seoul, Korea
Notion | LinkedIn | GitHub
Email: rick7213@gmail.com | Mobile: 010-2036-9712

Research Interest

- Large Language Models
- Natural Language / Code Generation
- Large Pre-training Data Processing / Alignment Data Curation

Experience

Research Engineer Aug 2023 – Present
kakaobrain *Seongnam, Gyeonggi*

- Development of pipeline for training and evaluating code generation / medical domain language model
- Construction of alignment data for building chat-based code generation agent model
- Searching for various methods for language models to interact with external tools
- Exploration of strategies for continual pre-training of large language model
- Development of quality filter models for pre-training corpus

Language Model Research Intern Sep 2022 – Jul 2023
kakaobrain *Seongnam, Gyeonggi*

- Development of a framework for easy pre-training of language models
- Development of large-scale data preprocessing pipelines compatible with various pre-training frameworks
- Development of Korean benchmark evaluation pipeline compatible with all model types based on fine-tuning
- Experiment for maximizing in-context learning performance of encoder-decoder models

Education

Sogang University Seoul, Korea
Bachelor of Science expected in Physics *Feb 2017 – Present*

Projects

KoGPT-2.0 for Code
2023-01 - present

* Large-scale code data collection and pre-processing
* Exploration and development of efficient tokenizers suitable for code models
* Constructon of large pre-training code dataset and chat-based code dataset
* Pre-training of billion-scale models and developing an execution-based text-to-code evaluation pipeline
* Adaptation of language models to behave as a code generative agent
* Exploration of strategies for continually training of pre-trained language models to efficiently injecting new features
* Development of quality filter models for pre-training data

Medical LLM
2023-11 - 2023-12

* Continually training large language models with large medical domain corpus
* Building few-shot chain-of-thought pipeline for evaluating medical knowledge of language model
* Curation of synthetic fine-tuning data for "Korean Medical Licensing Examination" and medical clinic
* Development of end-to-end pipeline for language model to pass the KMLE 2022 and 2023

Exploiting the Potential of Seq2Seq Models as Robust Few-Shot Learners
2022-12 - 2023-06

* Exploring new upper bounds for few-shot capabilities of encoder-decoder models
* Constructing a comprehensive baseline for experimenting with various IO formats tailored to the architecture and diverse prompts

LASSL (LAnguage Self-Supervised Learning) link
2022-06 - 2022-12

* Constructing a framework for convenient pre-training Transformer models based on HuggingFace
* Implementation of pre-training objectives for models such as BART, T5, UL2 and Electra based on papers
* Acceleration of training speed through improvements in collator operation

Presentations

SSL narratives Part 1 Link
2022-04

* Presentation for review of paper "CPC v1: Representation Learning with Contrastive Predictive Coding"

Jiphyeonjeon Season 3 Link
2023-04

* Presentation for review of paper "Transformer Memory as a Differentiable Search Index"

kakaobrain Mini Conference December 2022
2022-12

* Presentation for projects that the Language Model Team is working on for constructing reproducible model pipelines and discussing the future direction

Competitions

Medal Award at Korean Sentences Relationship Competition by Dacon *2022-03*

Medal Award at Jigsaw Rate Severity of Toxic Comments by Kaggle *2022-02*

Participation Award at Plant Growth Period Prediction Competition by KIST *2021-11*

Certifications

- BoostCamp AITech
- ModuLabs (completed a total of 5 sessions)

Technical Skills

Python : 4/5

PyTorch : 3/5

Google Cloud : 3/5