R&D perspectives as an industry scientist in Japan

IJST-A*STAR Project Kick-off Meeting

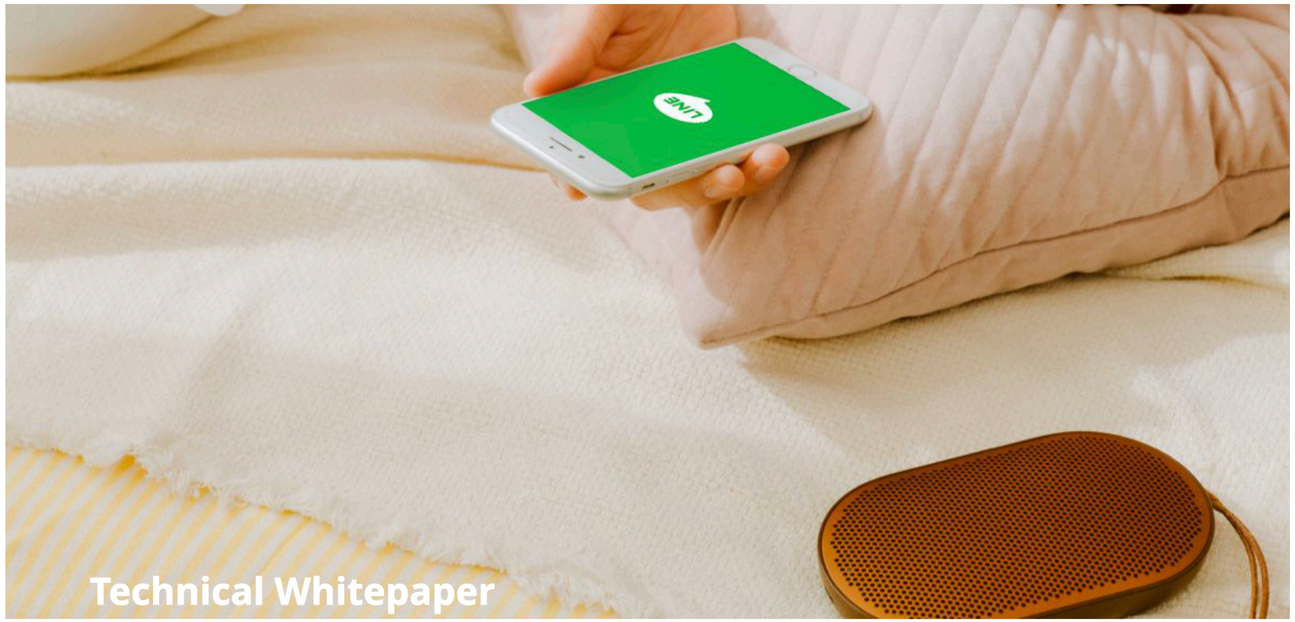
Seng Pei Liew, 2025-04-22

Seng Pei Liew

- Malaysia -> Japan -> PhD in Physics (Univ. of Tokyo) -> TU Munich (Germany) -> NEC Corporation (Japan) -> LINE/LY Corporation (Japan) -> LY Corporation and SB Intuitions (Japan)
- https://spliew.github.io/
- Collaboration with Prof. Cao (2021~)
  - SIGMOD 2022, VLDB 2022 etc.

LINE / LY

- LINE: Popular communication app like WhatsApp
- R&D on technologies that protect user privacy (especially differential privacy, DP)
  - (DP) Federated learning
  - Differential privacy of shuffle model
  - DP data synthesis



https://s.yimg.jp/images/cdo/privacycenter/2023renewal/pdf/ja/line-differential-privacy-whitepaper-ver1.0.pdf

SB Intuitions

A subsidiary of Softbank

- Large language models (LLMs) specialized for Japanese
- R&D on pretraining strategies of LLMs
  - Mixture-of-Experts
  - Scaling laws
  - Optimization

SoftBank Corp. Launches New Company to Develop Homegrown Large Language Models (LLM)

August 4, 2023
SoftBank Corp.

SoftBank Corp. (TOKYO: 9434, “SoftBank”) today announced it launched a new company to research and develop homegrown Large Language Models (LLM) specialized for the Japanese language. The fully operational company will also develop, market and provide generative AI services. First established as a preparatory company on March 27, 2023, the company changed its name to SB Intuitions Corp. (“SB Intuitions”) on August 1, 2023.

To realize its corporate philosophy, “Information Revolution — Happiness for everyone,” SoftBank is actively promoting the utilization of AI. Positioning generative AI as an important and innovative technology that accelerates its vision, SoftBank is implementing various generative AI initiatives. In the fiscal year ending March 31, 2024 (FY2023), SoftBank plans to build a computing platform with advanced data processing capabilities that can be utilized for developing generative AI and other applications. SoftBank also plans to rapidly offer services based on its computing platform to academic and research institutions, enterprises and other organizations.

Accelerating DP-FL

[2504.09850]

- Advanced federated optimization methods with DP guarantees are lacking
- We propose to modify the global LR adaptively while guaranteeing DP
- No client-side modification/additional hyperparameter is required
  - Least implementation intervention (important in industrial implementation)

$$\eta_g^{(t)} := \frac{\frac{1}{M}\sum_{i=1}^{M}\left\|\Delta_i^{(t)}\right\|^2}{\left\|\bar{\Delta}^{(t)}\right\|^2},$$

Sum of squares (of model update) divided by square of sum (of model update)

Privatize this under central/local DP

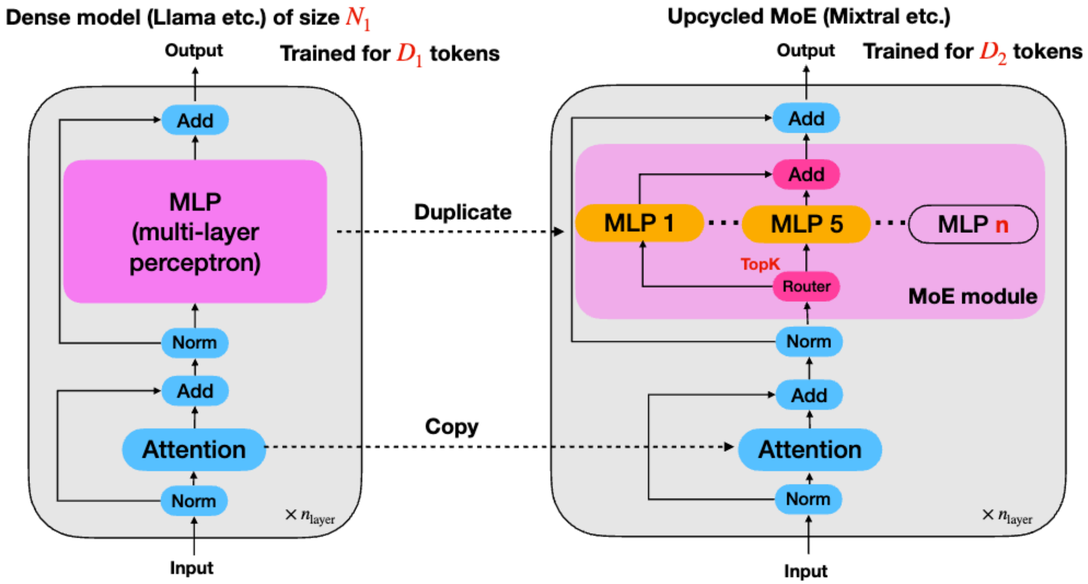
Training large models (MoE) efficiently

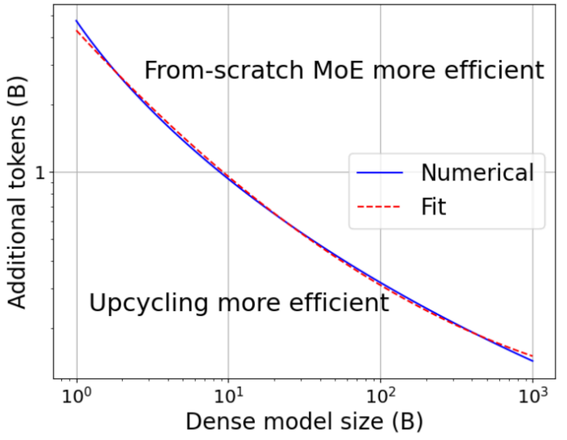
[2502.03009]

- Reusing (upcycling) dense models to train MoEs
- We find empirically that overtrained dense models are ineffective at continued training the MoEs, and this phenomenon can be captured by "scaling laws"

$$L = AD_1^{-\alpha_1} D_2^{-\alpha_2 + \alpha_3 \log D_1} + E$$

- "Scaling laws" usually lack theoretical justification, but important for making business decisions like total GPU hours to spend (which can be more than millions of USD).





Some other recent interests

- DP-SGD's privacy accounting for shuffled data is wrong (overestimated)!
  - https://arxiv.org/abs/2403.17673
  - Privacy amplification by random allocation/ balls-and-bins
    - https://arxiv.org/abs/2502.08202
  - What's its implication on LLM training/FL?
- New paradigm of optimization?
  - Muon (MomentUm Orthogonalized by Newton-Schulz)
    - https://kellerjordan.github.io/posts/muon/
  - Useful for LLM pretraining
  - What about (DP)FL, fine-tuning?