

2024 年度年次報告書

数学・数理科学と情報科学の連携・融合による情報活用基盤の創出と社会課題解決に向けた展開

2021 年度採択研究代表者

カーン エムティヤズ

理化学研究所 革新知能統合研究センター

チームリーダー

ベイズ双対性に基づく適合的・頑健・継続的な人工知能システム

主たる共同研究者:

坂内 健一 (慶應義塾大学 理工学部 教授)

横田 理央 (東京科学大学 総合研究院 教授)

## 研究成果の概要

The goal for this year is to start preparing for the finish of the knowledge representation and transfer part. We made substantial progress on several projects.

1. Khan group published a paper on Model Merging at ICLR 2024 where they proposed new methods by using uncertainty-based gradient matching.
2. Khan and Yokota groups have a paper accepted at ICML 2024 on a new optimizer called IVON which is derived from the Bayesian Learning Rule and gives state-of-the-art result on large models, such as, GPT-2. We also had a new version to do LoRA adaptation. Khan group published a paper at ICLR 2025 showing new connections between Bayes and Federated ADMM.
3. Khan group organized two workshops at ICLR 2025, namely “Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI” and “XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge”.
4. Emtiyaz Khan gave a keynote at the 3rd Conference on Lifelong Learning Agents (CoLLAs) 2024

We are continuing our work to make progress on continual and federated learning. We are also working more on model merging and improving understanding of LLMs, and their low-precision training. This year too we have made more progress than we originally planned for, and the results we have so far exceed our expectations.

本研究はベイズ双対性の理論を発展させ、それを応用することで適応的で頑健な継続学習が可能な AI システムを構築することを目的とする。今年度(2024年4月~2025年3月)の目標は、知識表現と転移に向けた作業を完了することであった。複数のプロジェクトで着実な進展を遂げることができた。

1. Khan グループは、不確実性に基づく勾配一致手法を用いた新しい方法を提案し「Model Merging」に関する論文を ICLR 2024 で発表した。
2. Khan と Yokota グループは、ICML 2024 で IVON と呼ばれる新しい最適化アルゴリズムに関する論文が採択された。このアルゴリズムはベイズ学習則から導出され、GPT-2 のような大規模モデルにおいて最先端の結果を達成しすることができた。また、LoRA 適応を行う新しいバージョンも開発した。Khan 研究グループは、ICLR 2025 でベイズ学習と連合学習の新たな関連性を示す論文を発表した。
3. Khan グループは、ICLR 2025 で「Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI」と「XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge」という2つのワークショップを主催した。
4. Emtiyaz Khan は、3rd Conference on Lifelong Learning Agents (CoLLAs) 2024 で基調講演を行った。

今後は、継続学習と連合学習の進展に向けた研究を継続していく。また、モデル統合や LLM

の理解の深化、低精度トレーニングの改善にも注力する。今年度も当初の計画を上回る進展を遂げ、現在の成果は期待を上回るものとなっている。

【代表的な原著論文情報】

- 1) Cong, B., Daheim, N., Shen, Y., Cremers, D., Yokota, R., Khan, M. E., & Möllenhoff, T. Variational Low-Rank Adaptation Using IVON. NeurIPS 2024 workshop on Fine-Tuning in Modern ML (FTML), 2024.
- 2) Shen, Y., Daheim, N., Cong, B., Nickl, P., Marconi, G.M., Bazan, C., Yokota, R., Gurevych, I., Cremers, D., Khan, M.E. and Möllenhoff, T. Variational Learning is Effective for Large Deep Networks, The 41st International Conference on Machine Learning (ICML), 2024
- 3) Swaroop, S., Khan, M. E., & Doshi-Velez, F. Connecting Federated ADMM to Bayes. International Conference on Learning Representation (ICLR). 2024.
- 4) Wolinski P. and Arbel J. Gaussian Pre-Activations in Neural Networks: Myth or Reality? Transactions of Machine Learning Research (TMLR). 2025.
- 5) Arbel J., Pitas K., Vladimirova M., and Fortuin V. A primer on Bayesian neural networks: review and debates. Statistical Science. 2024.