2023 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（東京工業大学 学術国際情報センター 教授）

研究成果の概要

The goal for this year is to make substantial progress on the Knowledge representation and transfer part. We continued collaborations among the core members on several project and published a few papers.

1. Khan group has 1 paper accepted at NeurIPS 2023 on Knowledge Representation by Memory Perturbation to relate models' sensitivity to uncertainty and predictability. Khan and Yokota groups have 1 paper accepted on Knowledge Transfer at TMLR (improving continual learning at ImageNet scale). Arbel group has 1 paper on accepted at ACML 2023. The paper is about cold-posterior effect in deep learning and its connections to PAC-Bayes.
2. We organized an ICML 2023 workshop on Duality Principles and also an internal meeting on Bayes-Duality.

We are now building on our Memory-Perturbation work and scale it to large models. We also applied some of our work to improve model merging for LLMs. We believe model merging also gives us a way to understand the memory of models. 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（2023 年 4 月～2024 年 3 月）の目標は、知識表現と知識伝達の部分を大幅に進歩させることである。いくつかのプロジェクトでコア・メンバー間のコラボレーションを継続し、いくつかの論文を発表した。

1. Khan グループは、不確実性と予測可能性に対するモデルの感度を関連付けるためのメモリ摂動による知識表現に関する論文 1 本が NeurIPS 2023 にアクセプトされた。Khan グループと Yokota グループの共著による ImageNet スケールでの継続的学習の改善の論文は TMLR にアクセプトされた。Arbel グループはディープラーニングにおける cold-posterior 効果と PAC-Bayes との関連に関する論文が ACML 2023 に採択された。
2. ICML 2023 にて、ベイズ双対原理に関するワークショップを開催し、ベイズ双対性原理に関するプロジェクト全体のワークショップも開催した。

今後は、記憶摂動に関する研究をさらに発展させ、これを大規模モデルにスケールアップしていく。また、LLM のモデル・マージを改善するために、ベイズ双対原理を適用した。モデル・マージは、モデルの記憶を理解する方法にもなると考えている。今年度も当初予定していた以上の進捗があり、これまでの成果は期待以上であった。

【代表的な原著論文情報】

1) Improving Continual Learning by Accurate Gradient Reconstructions of the Past, (TMLR) E. Daxberger, S. Swaroop, K. Osawa, R. Yokota, R. turner, J. M. Hernández-Lobato, M.E. Khan
2) The Memory Perturbation Equation: Understanding Model's Sensitivity to Data, (NeurIPS 2023) P. Nickl, L. Xu, D. Tailor, T. Möllenhoff, M.E. Khan
3) Memory-Based Dual Gaussian Processes for Sequential Learning, (ICML 2023) P. E. Chang, P.

Verma, S. T. John, A. Solin, M.E. Khan

4) Lie-Group Bayesian Learning Rule, (AISTATS 2023) E. M. Kiral, T. Möllenhoff, M.E. Khan
5) SAM as an Optimal Relaxation of Bayes, (ICLR 2023), T. Möllenhoff, M.E. Khan