論文 / 著書情報
Article / Book Information

題目(和文)	高効率・高精度なエッジAIアンサンブル手法の研究
Title(English)	A Study on Efficient and Accurate Ensemble Methods for Edge AI
著者(和文)	熊澤峻悟
Author(English)	Shungo Kumazawa
出典(和文)	学位:博士(学術), 学位授与機関:東京科学大学, 報告番号:甲第382号, 授与年月日:2025年3月26日, 学位の種別:課程博士, 審査員:本村 真人,髙橋 篤司,CHU VAN THIEM,ISLAM A K M MAHFUZUL,佐々木 広
Citation(English)	Degree:Doctor (Academic), Conferring organization: Institute of Science Tokyo, Report number:甲第382号, Conferred date:2025/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報通信 情報通信	系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of	（ 学術 ）
学生氏名： Student's Name	熊澤 峻悟		審査員主査： Chief Examiner	本村 真人	

要旨（英文800語程度）
Thesis Summary (approx.800 English Words)

Advancements in low-cost, high-performance device technologies have driven the rapid proliferation of IoT devices, expanding their applications across various domains such as healthcare, urban infrastructure, and industrial automation. This surge has brought attention to edge AI, which enhances responsiveness and energy efficiency by enabling distributed data processing at the network edge. As the number of IoT devices continues to grow in edge AI environments, efficiently utilizing these expanding device networks has become increasingly important. In particular, collaborative inference methods, which leverage multiple devices to overcome the limitations of individual device performance, are gaining significance.

One promising approach is Edge Ensembles, which achieve higher accuracy than a single device by integrating predictions across multiple devices based on an ensemble approach. They offer key advantages such as scalability to accommodate increasing IoT devices, efficient resource utilization through dynamic load balancing, and reduced reliance on centralized servers. Additionally, their adaptability to dynamic environments allows seamless handling of device mobility and availability changes, ensuring robustness in real-world scenarios. However, increasing the number of collaborating devices in cross-device ensemble inference can lead to higher communication overhead and latency. Additionally, fluctuations in device availability may cause variations in inference accuracy. To mitigate these challenges, this study addresses two key issues to improve accuracy and efficiency, enabling high-precision inference even with a reduced number of available devices. The first is the need for lightweight machine learning (ML) methods for both training and inference that can efficiently run in highly resource-constrained environments while achieving high accuracy through ensemble techniques. As IoT devices proliferate, reducing server dependency is essential not only for inference but also for model updates, ensuring the adaptability of Edge Ensembles. Second, using computationally intensive models like deep learning (DL) for complex tasks in Edge Ensembles significantly increases the inference cost compared to traditional ML models. This leads to prolonged inference times per device, reducing the overall availability of devices within the ensemble cluster and potentially undermining the ensemble's benefits.

To overcome these limitations, this study introduces two novel proposals. The first is "Extra Ferns," a lightweight fern-based ensemble training method designed for resource- constrained edge environments. Extra Ferns significantly enhances inference accuracy compared to existing fern-based methods while consuming low energy consumption during training, making it particularly well-suited for edge device deployment. At its core, Extra Ferns represents a unique subset of decision tree ensembles, combining the strengths of existing methods while addressing their limitations. Specifically, it significantly decreases memory access and optimizes each tree in parallel, benefiting from the advantages of both Extra Trees and Random Ferns. As Extra Trees does, it generates nodes by randomly selecting attributes and generating thresholds. Then, as Random Ferns does, it builds ferns, which are decision trees that share identical nodes at each depth. In contrast to other ensemble methods using greedy optimization, Extra Ferns attempts global optimization of each fern. Experimental results show that Extra Ferns requires only 4.3% and 4.1% memory access for training models with 3.0% and 1.2% accuracy drops compared with Random Forest and Extra Trees, respectively. This study also proposes applying lightweight random projection to Extra Ferns as a preprocessing step, which improved further accuracy by up to 2.0% for image datasets.

The second proposal focuses on improving the efficiency and accuracy of model integration for DL-based Edge Ensembles. Enhancing model integration reduces the number of devices required to achieve the desired accuracy in collaborative inference tasks. This reduction in device usage leads to lower overall resource consumption across the cluster, enabling more efficient collaboration. To explore this improvement, this study investigates the effectiveness of conventional model integration techniques, such as cascade, weighted averaging, and test-time augmentation (TTA), when applied to Edge Ensembles to enhance their performance. Furthermore, this study proposes enhancements of these techniques tailored for Edge Ensembles. The cascade reduces the number of models required for inference but worsens latency by sequential inference processing. An increase in latency deteriorates real-time performance, posing significant challenges for its application in scenarios that demand immediate predictions. To address this issue, this study proposes m-parallel cascade, which reduces latency by adjusting the number of models processed simultaneously to m. This study also proposes learning TTA policies and weights for weighted averaging using ensemble prediction labels instead of ground truth labels. In the experiments, this study verified the effectiveness of each technique for the context of Edge Ensembles. The proposed m-parallel cascade achieved a 2.8 times reduction in latency compared to the conventional cascade, even with a 1.06 times increase in computational costs. Additionally, the ensemble label-based learning demonstrated comparable effectiveness to the approach using ground truth labels.

This study advances collaborative inference methods through two key contributions. The first contribution enhances the practicality of fern-based Edge Ensembles in environments with extremely limited computational resources. Additionally, it enables these methods to perform training directly within edge environments, facilitating more adaptive and environment-specific inference. The second contribution strengthens the collaborative capabilities of DL-based Edge Ensembles, improving the overall performance of Edge Ensemble systems and demonstrating the feasibility of optimizing model integration in edge environments using only inference data. These contributions not only enhance the accuracy and efficiency of Edge Ensemble inference but also provide model training methods and analyses on collaborative frameworks, serving as valuable guidelines for designing practical cooperative inference systems.