Front. Comput. Sci., 2025, 19(3): 193501
https://doi.org/10.1007/s11704-024-40465-z

LETTER

DeepSwarm: towards swarm deep learning with bi-directional optimization of data acquisition and processing

Sicong LIU[1], Bin GUO (✉)[1], Ziqi WANG[1], Lehao WANG[1], Zimu ZHOU[2], Xiaochen LI[1], Zhiwen YU[1,3]

1 School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
2 School of Data Science, City University of Hong Kong, Hong Kong 999077, China
3 College Of Computer Science And Technology, Harbin Engineering University, Harbin 150001, China

© Higher Education Press 2025

1 Introduction

On-device deep learning (DL) on mobile and embedded IoT devices drives various applications [1] like robotics image recognition [2] and drone swarm classification [3]. Efficient local data processing preserves privacy, enhances responsiveness, and saves bandwidth. However, current on-device DL relies on *predefined patterns*, leading to accuracy and efficiency bottlenecks. It is difficult to provide feedback on *data processing* performance during the *data acquisition* stage, as processing typically occurs after data acquisition.

Harnessing the potential of swarms formed by physically adjacent mobile and embedded devices in IoT scenarios, we advocate a paradigm shift towards *Swarm DL* by drawing inspiration from *swarm intelligence* [4]. This shift entails moving from *reactive* on-device DL, which responds to given input data, to *proactive* systems known as swarm deep learning (Swarm DL). Conceptually, *swarm intelligence* involves the collective intelligent behavior of multiple agents, each acting *proactively* based on simple patterns in a self-organized, self-adaptive, self-evolved manner, leading to enhanced global performance. Building upon on-device DL as the foundation, Swarm DL *proactively* scales data acquisition and processing and provides *bi-directional optimization feedback* for them, forming a closed system loop. This paradigm aims to *maximize implicit complementarity* and *minimize redundancy* in both data acquisition and processing within the swarm, fostering a more efficient and scalable IoT system.

Specifically, Swarm DL achieves swarm intelligence by utilizing IoT devices equipped with advanced sensors and DL computing capabilities as *active agents*, with two major visions: *Proactive Data Acquisition for DL and Proactive Data Processing by DL*. To achieve these two visions, it is necessary to expand current research, facing two challenges: (i) from independent data collection and processing optimization to bidirectional optimization. (ii) from reactive data processing to active data collection and processing.

To handle these challenges in practical IoT environments, we propose a generic system framework, named *DeepSwarm*. We define modular design for DeepSwarm and identify optimization opportunities and techniques to deploy Swarm DL on resource-limited and decentralized mobile and embedded platforms. DeepSwarm pinpoints a set of *proactive* strategies inter- and intra-devices that contribute to a self-organized, self-adaptive, and self-evolving Swarm DL system.

Received May 8, 2024; accepted July 18, 2024

E-mail: guob@nwpu.edu.cn

2 Scope and framework

In this section, we introduce Swarm DL and a general framework, DeepSwarm, comprising two functional modules.

2.1 The concept of Swarm DL

As mentioned above, *Swarm DL* extends *reactive on-device DL*, which focuses on resource-efficient DL given input data on individual IoT devices, to a distributed setting inspired by *proactive swarm intelligence*. As shown in Fig. 1, compared with related concepts, Swarm DL exhibits unique characteristics of self-organization, *self-adaptive*, *and self-evolving*.

- **Self-organized** Swarm DL emphasizes the bottom-up emergence of collective behaviors among devices. Each agent achieves dynamic optimization of global system performance and resource efficiency by *proactively* triggering local operations for data acquisition and processing.
- **Self-adaptive** Each agent in Swarm DL exhibits greater pro-activity, participating in both DL inference on the device and adaptation, with a higher proportion of agents dedicated to adaptation.
- **Self-evolving** Swarm DL is real-time and human independent, consisting of mobile agents that actively learn from data through bidirectional feedback, optimizing data collection (supply) and processing (demand) for each device.

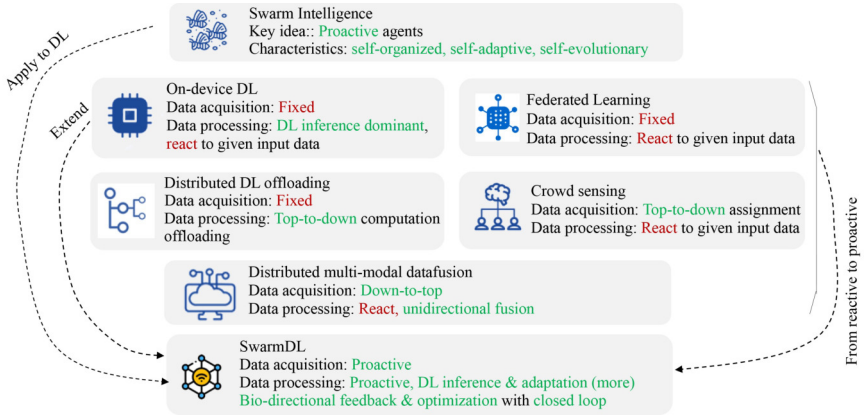

Fig. 1 Comparison of Swarm DL and related concepts

2.2 The DeepSwarm framework

To realize the vision of Swarm DL, we present a generic system framework, DeepSwarm (see Fig. 2). DeepSwarm takes a modular design and decomposes the requirements of the Swarm DL into two modules: ***proactive swarm data acquisition for DL*** and ***proactive swarm data processing by DL***. These two modules work in synergy with bi-directional feedback to optimize the system performance (e.g., accuracy, latency) and resource efficiency (energy efficiency, memory fragmentation). It functions with heterogeneous AIoT hardware (e.g., CPU, GPU, or MCU-equipped embedded devices), adapts to dynamic application contexts (i.e., data distribution drifts and runtime resource availability), and generalizes to various AIoT applications (e.g., mobile health, smart homes, autonomous vehicles, industrial automation).

2.3 Proactive swarm data acquisition module

This module coordinates the *self-organization of distributed agents and sensors*, drawing inspiration from swarm intelligence. Each agent actively engages in data resampling, sensing parameter adjustment, and association with other agents by analyzing information extracted from cross-modal, cross-task, and cross-clock sensor data, aiming to *maximize fusion quality* and *minimize redundancy* of agent data. Additionally, it aims to achieve **complementary enhancement as early as possible**, addressing challenges such as modal information loss, clock asynchrony, and data shifts. This mitigates subsequent resource costs (e.g., sampling, computing, transmission bandwidth) as early as possible. Specifically, we emphasize the simultaneous assessment of the *explicit* and *implicit* importance of data on the performance of subsequent data processing tasks at runtime. The explicit data importance estimation is data-driven and has abundant existing works. While implicit data complementarity profiling is system-driven and non-trivial, requiring a comprehensive consideration of dynamic system factors.

2.4 Proactive swarm data processing module

This module encompasses data processing tasks, including DL inference and adaptation. *Traditional* on-device DL systems primarily *focus on inference*, especially in task-specific embedded devices. In contrast, Swarm DL agents exhibit **higher proactivity**, engaging in **both inference** and **adaptation**, **even with a *greater emphasis on adaptation*.** Balancing resource allocation for DL inference and adaptation presents a novel challenge in this context. Specifically, the primary objective in Swarm DL is to perform data processing tasks in a *self-adaptive* and *self-evolutionary* manner, akin to general swarm intelligence. Technically, unlike approaches solely optimizing DL algorithms, DeepSwarm also addresses system asynchrony in resource competition and varied resource availability among agents, managing peak memory usage, and optimizing system delay and accuracy tradeoffs.

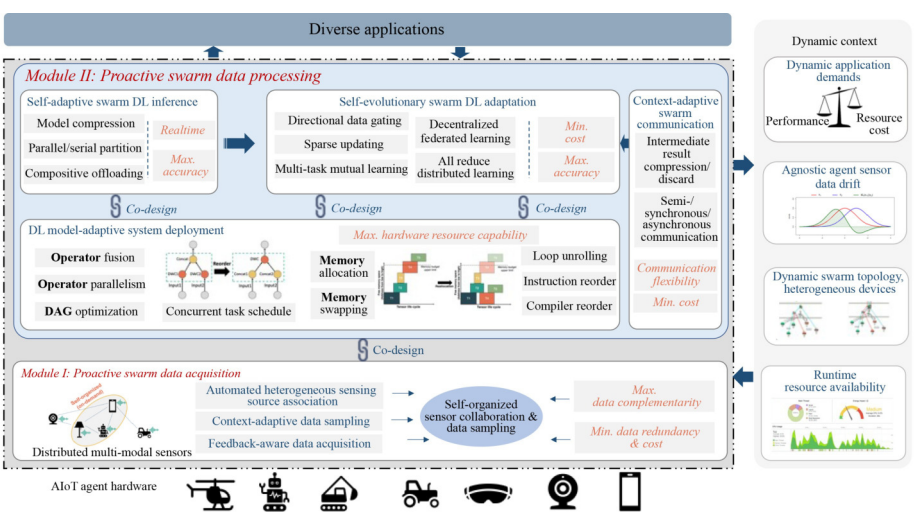

Fig. 2 Illustration of DeepSwarm, a generic system framework to realize bio-directional optimization of Swarm DL

Table 1 Performance comparison of different DL model adaptation methods

Method	Accuracy gain of global model	Accuracy of mobile model after adaptation				Accuracy gain of mobile model			
	IoU = 0.50	IoU = 0.50				IoU = 0.50			
		Mobile model A	Mobile model B	Mobile model C	Average	Mobile model A	Mobile model B	Mobile model C	Average
Domain adaptation	None	0.504	0.469	0.497	0.49	14.3%	48.9%	13.7%	23.4%
NestEvo without data generation	1.3%	0.504	0.475	0.501	0.505	14.3%	50.8%	15.5%	27.2%
Original mobile model	None	0.441	0.315	0.437	0.397	None			
Only mobile model adaptation	0	0.501	0.478	0.493	0.491	13.6%	51.7%	12.8%	23.7%
NestEvo	9.13%	0.571	0.543	0.584	0.566	29.5%	72.4%	33.6%	42.6%

3 Challengs and opportunities

Embracing the co-design principle, we identify the following challenges and opportunities:

- Self-adaptive Non-blocking DL Inference with misaligned, incomplete, and asynchrony data. Existing approaches relying on statistical analysis or divergence measures can not work well in real-time for local agents without access to global data. Challenges include identifying non-redundant data correlation, runtime sampling rate adaptation, and dynamically elastic DL model.
- Test-time Self-evolutionary DL Adaptation. The **asynchronism** of distributed multi-modal data streams **poses challenges**, leading to system delays (if waiting for slow devices) or a decrease in accuracy (if not waiting for slow data). Multi-task co-adaptation via data and computing reuse is an opportunity yet a challenge for agents lacking data and computational resources.
- Swarm DL-adaptive Adaptive Operator/System Resource Scheduling. Tailoring the system scheduling to characteristics of swarm data and tasks, such as **tensor/operator life cycles and dependencies**, enhances parallelism, increases data reuse, and reduces memory fragmentation during swarm data processing.

4 Experiment result and analysis

Mobile video applications today have attracted significant attention. The compressed DL model is widely used to enable on-device video inference, which, however, is vulnerable to the *non-stationary data drift* of the live video captured from dynamically changing mobile environments. To combat data drifts, present a **Swarm DL adaptation system** based on the DeepSwarm framework, which enables each agent to continuously update using newly collected sensor data from local and other agents The system consists of three components: *data drift-aware video frame sampling*, *feedback-aware DL adaption trigger*, and *adaptive DL adaptation & resource scheduling*. See more details in the supplementary material. We compare our system with four baselines, i.e., domain adaptation [5], DeepSwarm without data fusion from other agents, original agent DL model, and single-agent adaptation [6]. Table 1 demonstrates that DeepSwarm achieves the best overall performance in terms of accuracy, enhancing average accuracy by over 40% compared to the original models for the mobile model and by 9% for the global model.

5 Conclusion

Inspired by the collective intelligence observed in natural swarms, where individual *proactive* actions contribute to superior global performance, we advocate for a shift towards *Swarm DL*. By harnessing the potential of physically adjacent mobile devices in IoT scenarios, we present DeepSwarm, a closed-loop system framework architecture. DeepSwarm facilitates bidirectional optimization between data acquisition and processing, aiming to push the performance boundaries of on-device DL Specifically, DeepSwarm addresses the requirements of proactive Swarm DL by decomposing them into layers: self-organized swarm data acquisition and self-adaptive, self-evolutionary swarm data processing.

Acknowledgements The work was supported by the National Science Fund for Distinguished Young Scholars (62025205), the National Natural Science Foundation of China (Grant Nos. 62032020, 62102317), CityU APRC Grant (9610633).

Competing interests Bin Guo is an Editorial Board member of the journal and a co-author of this article. To minimize bias, he was excluded from all editorial decision-making related to the acceptance of this article for publication. The remaining authors declare no conflict of interest.

Electronic supplementary material Supplementary material is available in the online version of this article at journal.hep.com.cn and link.springer.com.

References

1. Liu S, Guo B, Fang C, Wang Z, Luo S, Zhou Z, Yu Z. Enabling resource-efficient AIoT system with cross-level optimization: a survey. IEEE Communications Surveys & Tutorials, 2024, 26(1): 389−427
2. Ma S, Pei J, Zhang W, Wang G, Feng D, Yu F, Song C, Qu H, Ma C, Lu M, Liu F, Zhou W, Wu Y, Lin Y, Li H, Wang T, Song J, Liu X, Li G, Zhao R, Shi L. Neuromorphic computing chip with spatiotemporal elasticity for multi-intelligent-tasking robots. Science Robotics, 2022, 7(67): eabk2948
3. Hu Y, Fang S, Lei Z, Zhang Y, Chen S. Where2comm: communication-efficient collaborative perception via spatial confidence maps. In: Proceedings of the 36th Conference on Neural Information Processing Systems. 2022, 4874−4886
4. Bonabeau E, Dorigo M, Theraulaz G. Swarm Intelligence: from Natural to Artificial Systems. Oxford: Oxford University Press, 1999
5. Mullapudi R T, Chen S, Zhang K, Ramanan D, Fatahalian K. Online model distillation for efficient video inference. In: Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. 2019. 3572−3581
6. Chen Y, Li W, Sakaridis C, Dai D, Van Gool L. Domain adaptive faster R-CNN for object detection in the wild. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. 3339−3348