ON-DEVICE SPLIT INFERENCE FOR EDGE DEVICES: A LITERATURE REVIEW

Bora Kozan - b.kozan@student.tudelft.nl Responsible Professor: Qing Wang Supervisors: Mingkun Yang & Ran Zhu

TU Delft

01 INTRODUCTION

As the technology around us continuously improves, the need for computing increases. Nowadays, there are embedded devices all around us. The idea of on-device split inference comes from the idea of using embedded devices not just for simple tasks like data collection but also running more complex algorithms in order to achieve faster run times and keep the sensitive data local [1]. Usually, many algorithms are too demanding to fit inside single embedded devices. In that case, splitting the algorithm into multiple devices can be used to run the complex algorithm on the edge [1].

02 RESEARCH QUESTION

The research focus is to survey the key split inference technologies on edge devices.

Sub-questions:

1. Which machine learning models and algorithms are used for distributed inference on embedded devices?
2. What are the benefits and the limitations of the currently used on-device inference methods?
3. Are there any distributed computing and parallelization algorithms that are used for or can be applied to split inference on embedded devices?
4. What are the applications of on-device split inference?

03 METHODOLOGY

- The search engines used: ACM Digital Library, IEEE Xplore, Scopus, Google Scholar (Used for an initial search)
- Papers that focus on distributed inference on edge devices only were considered. The edge devices must be used for inference, not just for collecting data, etc. In the end, 50 papers were included in the review.
- Some of the keywords used for queries:

split	distributed	collaborative
inference	machine learning	neural network
TinyML	embedded	edge

04 ANALYSIS

To categorize and analyze papers, the following questions are asked:

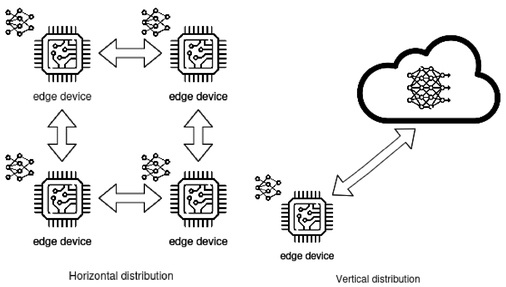
- What device is used?
- What machine learning or artificial algorithm is the paper based on?
- What optimization and preprocessing methods are they using?
- How are they achieving the distribution of the inference task?
- Are there any performance gains? What are the benefits and the shortcomings of the method?
- What is the purpose or the use case of the method?

05 RESULTS/ FINDINGS

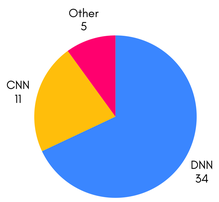
Inference Distribution

There are two main ways that a neural network can be distributed: horizontally and vertically [2]. Horizontal distribution means that the inference task is split between devices at the same level in an edge-cloud architecture. Vertical distribution is when the inference is distributed among various layers of an edge-cloud architecture [2].

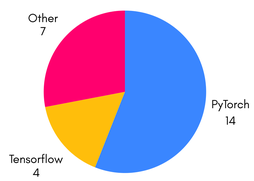
25/50 papers contained a method with **horizontal distribution**.
21/50 papers focused on **vertical distribution**.
4/50 papers contained **both** distribution types.



Which artificial intelligence or machine learning technique is the base of the methods?



Which programming libraries are being used?



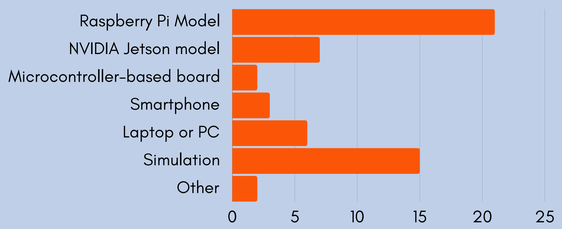
What are the optimization and preprocessing methods that stand out?

- Converting a neural network to a directed acyclic graph (DAG) representation. This makes it possible to run various algorithms on the neural network model.
- Pruning. It is removing some parts of the neural network that don't contribute much, therefore decreasing the size of the model without sacrificing much from accuracy [3]

Are the methods static or adaptive during runtime?

In some papers, the system is static. The optimizations were done during the design. Once it is deployed, it takes predetermined actions. However, several other systems are adaptive. They make decisions during run-time and adjust some system parameters to further increase the performance of the system.
20/50 papers talk about **static** systems.
30/50 papers talk about **adaptive** systems.

Which devices are being used?



Use Cases and Real World Applications

- Healthcare - The fact that IoT and wearable devices are popular can be leveraged. Automatic speech recognition applications with multiple IoT devices
- Sensor networks for monitoring large areas
- Industrial applications - Detecting product defects and equipment failure
- Video processing

06 CONCLUSION/ FUTURE WORK

Conclusion

- Inference distribution type is the main parameter that shapes the architecture of an on-device split inference system.
- There are not a lot of papers that focus on distributed inference systems for microcontrollers.

Future Work

- A systematic review can be done to find and review almost all of the papers in the area.
- Research can be conducted based on findings from distributed inference on edge devices to design a system for distributed inference on microcontrollers and similar embedded systems.

Related literature

[1] R. Sahu, R. Toepfer, M. D. Sinclair, and H. Duwe, "DENNI: Distributed Neural Network Inference on Severely Resource Constrained Edge Devices," in 2021 IEEE International Performance, Computing, and Communications Conference (IPCCC), Oct. 2021, pp. 1–10. doi: 10.1109/IPCCC51483.2021.9679448.
[2] M. G. Sarwar Murshed et al. "Machine Learning at the Network Edge: A Survey". In: ACM Comput. Surv. 54.8 (Oct. 2021). Place: New York, NY, USA Publisher: Association for Computing Machinery. ISSN: 0360-0300. DOI: 10.1145/3469029. URL: https://doi.org/10.1145/3469029.
[3] Hengyuan Hu et al. Network Trimming: A Data-Driven Neuron Pruning Approach towards Efficient Deep Architectures. July 12, 2016. DOI: 10.48550/arXiv.1607.03250. arXiv: 1607.03250[cs]. URL: http://arxiv.org/abs/1607.03250 (visited on 06/23/2024).