Conference Paper Title*

*Note: Sub-titles are not captured in Xplore and should not be used

Orestis Kanaris
Delft University of Technology
Delft, Netherlands
O.Kanaris@student.tudelft.nl

Johan Pouwelse (msc supervisor)
Delft University of Technology
Delft, Netherlands
J.A.Pouwelse@tudelft.nl

***Abstract*—**
***Index Terms*—**

I. Introduction

II. Problem Description

A. Background

In recent years, the proliferation of mobile devices has reached unprecedented levels, with smartphones becoming an integral part of everyday life. These devices have increasingly powerful hardware, making them suitable candidates for running complex machine-learning models [1], [2]. Machine learning on mobile devices holds great potential for many applications, from personalized recommendations to democratizing big tech. One can imagine a world where every smartphone (or personal computer) holder holds their own portion of "Google's" database (and computation), having all smartphones intercommunication and share information to complete a search result, leading to a democratized distributed peer-to-peer search engine, cleansed from the big tech influence and hidden agendas [].

However, deploying machine learning models on mobile devices presents numerous challenges, including limited computational resources, memory constraints, and the need for efficient communication between devices. The main struggle that this paper focuses on is the efficient communication between devices since the rest can be overcome with efficient implementation and time-assuming Moore's Lawk [].

Personal devices, specifically smartphones communicate through home Wi-Fi and mobile networks like 5G. By using these networks the devices usually end up behind home or Carrier-Grade NATs. The existence of these NATs makes it harder for the devices to communicate with each other since they lock their discoverability by hiding the devices behind the NAT's private network and forcing each device to initiate the connection first. The problem starts when there are two devices behind different NATs and both need to initiate the connection first, but none of them knows the address of the other since it is hidden [].

Identify applicable funding agency here. If none, delete this.

B. Research problem

The central research problem addressed in this thesis revolves around the distribution of machine learning workloads on Android mobile phones using an adapted version of TensorFlow LITE[1] [3] as the model of choice to demonstrate the capabilities of the system in terms of running distributed machine learning models on mobile devices that communicate through 5G. Specifically, the challenge is to design and implement a framework that enables the execution of machine learning tasks in a distributed manner across a network of Android devices that communicate through 5G. This framework should harness the computational power of multiple mobile devices while overcoming connectivity issues, such as network address translation (NAT) and the establishment of overlay networks.

C. Objectives

The primary objectives of this thesis are as follows:

1) Develop a distributed machine learning framework that leverages TensorFlow Lite for efficient model execution on Android mobile phones.
2) Implement a mechanism for establishing overlay networks among Android devices to facilitate communication and task distribution.
3) Address the NAT puncturing problem to enable seamless connectivity among devices, even when they are behind NATs or firewalls.
4) Evaluate the performance, scalability, and resource utilization of the proposed framework through experimental validation and benchmarking on Android devices obtained from the Tribler lab[2].

III. System Design

IV. Implementation

V. Evaluation

VI. Discussion and Future Work

VII. Conclusion

References

[1] M. S. Louis, Z. Azad, L. Delshadtehrani, S. Gupta, P. Warden, V. J. Reddi, and A. Joshi, "Towards deep learning using tensorflow lite on risc-

[1] https://github.com/Tribler/tribler/issues/7254#issuecomment-1733143687
[2] https://www.tribler.org/about.html

v," in *Third Workshop on Computer Architecture Research with RISC-V (CARRV)*, vol. 1, 2019, p. 6.

[2] J. Dai, "Real-time and accurate object detection on edge device with tensorflow lite," in *Journal of Physics: Conference Series*, vol. 1651, no. 1. IOP Publishing, 2020, p. 012114.

[3] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/