Seminar Cloud Computing

From Concept to Production: Deploying TinyML in Industry

Trung Nguyen
Technische Universität München

November 2024

Abstract

In recent years, Artificial Intelligence (AI) and Machine Learning (ML) have received tremendous amount of attention in both industry and research world. However, conventional Machine Learning demands high computing capability which limits its usage to only larger computing units. The pardigm shift to Tiny Machine Learning (TinyML) is revolutionizing industries by enabling the deployment of machine learning models on low-power, resource-constrained devices. Being one of the most rapid developing field of Machine Learning, TinyML promises to benifits multiple industries. However, building a production-ready tinyML system poses different unique challenges. In this paper, we explore the key obstacles faced when developing and deploying TinyML models in production environments, including model optimization, hardware limitations, software integration, and maintaining performance in real-world conditions. Additionally, we present real-world use cases of TinyML in industrial settings, showcasing its transformative impact. We also discuss practical approaches and strategies presented by recent researches [6] to overcome these challenges, providing insights into how TinyML systems can be successfully scaled and implemented in production.

1 Introduction

Traditional Machine Learning Models, especially Deep Learning Models typically require substantial amount of computing capability to operate effectively. These models are often trained on powerful Graphics Processing Units (GPUs) and produce large models ranging from tens or hundreds of gigabytes (GB) down to smaller models in the range of 10 to 100 megabytes (MB). However, the memory requirements during runtime for these models far exceed what microcontrollers (MCUs) can handle. The pardigm shift to TinyML is driven by the prevailing number of Microcontroller Units (MCU) currently circulating in the industry. According to a recent report [2, 1], as of 2021, around 31 billion MCUs were shipped worldwide annually. The MCU market size is projected to increase in upcoming years [2]. This creates a big incentive for researchers and industry players to put effort into developing the technology further. TinyML aims to enable the operation of ML model to run on energy-, and memory-constraint devices by limiting communication overhead with better suited architecture design and applying different compressing techniques such as: quantization and pruning.

In this paper, Chapter 2,

2 TinyML Overview

2.1 Definition, Key Concepts, and Techniques

2.2 TinyML pipeline

$$a^2 + b^2 = c^2 \tag{1}$$

Again, refering to this equation is easy (see Eq. 1). If you do not need numbering for equations, use the *displaymath* environment:

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

3 Production tinyML

> "I think there is a world market for maybe five computers." (T.J. Watson, IBM, 1943)

Figure 1: The caption explaining what can be seen in the image/figure. Readers often read captions first if they do not have much time. Thus, it is important to find a good short explanation.

The rest of the work (especially all the regular text) must be written/phrased by you. If you write about some results or fact stated in another paper, you should refer to it. The 'Analytical Engine" — a mechanical calculation machine — created by Charles Babbage in the year 1838 was based on the decimal system [4, 5, 3, 7, 8].

4 Bringing TinyML to industrial setting

Enumerations using bullet points:

- Agriculture
- Environmental Monitoring
- Industrial predictive maintenance
- Edge AI and Autonomous Systems

5 Challenges and Future of TinyML

6 Conclusion

References

[1] Microcontroller market research, 2030.

[2] Microcontroller Market Size, Share & Trends By 2034.

[3] Miguel de Prado, Manuele Rusci, Romain Donze, Alessandro Capotondi, Serge Monnerat, Luca Benini And, and Nuria Pazos. Robustifying the Deployment of tinyML Models for Autonomous mini-vehicles, July 2020.

[4] Dina Hussein, Dina Ibrahim, and Norah Alajlan. Original Research Article TinyML: Adopting tiny machine learning in smart cities. *Journal of Autonomous Intelligence*, 7:1–14, January 2024.

[5] Aditya Jyoti Paul, Puranjay Mohan, and Stuti Sehgal. Rethinking Generalization in American Sign Language Prediction for Edge Devices with Extremely Low Memory Footprint, February 2021. arXiv:2011.13741.

[6] Haoyu Ren, Darko Anicic, and Thomas Runkler. TinyOL: TinyML with Online-Learning on Microcontrollers. April 2021. arXiv:2103.08295.

[7] Haoyu Ren, Darko Anicic, and Thomas A. Runkler. The synergy of complex event processing and tiny machine learning in industrial IoT. In *Proceedings of the 15th ACM International Conference on Distributed and Event-based Systems*, DEBS '21, pages 126–135, New York, NY, USA, June 2021. Association for Computing Machinery.

[8] A. Navaas Roshan, B. Gokulapriyan, C. Siddarth, and Priyanka Kokil. Adaptive Traffic Control With TinyML. In *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 451–455, March 2021.