EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 111 – 117

Proceedings of Emerging Tech Conference: Edge Intelligence 2023

Transforming the Path Towards Automation of Monitoring and Management for Edge Computing

Georgios Samaras[1], Marinela Mertiri[1], Maria-Evgenia Xezonaki[1], Nikolaos Psaromanolakis[1], Vasileios Theodorou[1], and Theodoros Bozios[1]
Intracom Telecom, Greece
{gsamaras, marmert, maxez, nikpsarom, theovas, tmpo}@intracom-telecom.com

Abstract

The breakthrough of Artificial Intelligence (AI) has revolutionized Machine Learning (ML), particularly in the form of Transformer models such as Chat Generative Pre-training Transformer (ChatGPT), achieving state-of-the-art (SoTA) performance in various do mains, including Edge Computing. This paper introduces the TANDEM Smart Resource Monitoring and Management approach, which addresses the challenges of resource monitoring and management at the edge. TANDEM leverages AI/ML mechanisms to enable efficient service distribution, user-centered operation, and diverse edge applications support. It proposes novel forms of dynamic resource monitoring and management, utilizing Trans former models for accurate system usage prediction. Furthermore, TANDEM provides an AutoML framework and its AutoTinyML extension, which enhances IoT applications with powerful ML services. The proposed architecture and approach contribute to the intersection of AutoML and Transformer models with Edge Computing. Extensive experimental evaluations and analysis demonstrate the effectiveness and potential of our approach.

1 Introduction

The breakthrough of AI has sparked a surge in ML and Deep Learning (DL) applications and services, driving remarkable growth. Transformer ML models have gained immense popularity, mainly because of the Chat Generative Pre-training Transformer (ChatGPT), attracting significant attention to Automated ML (AutoML). These advanced models achieve SoTA performance across various domains, e.g. edge computing [4]. The widespread adoption of AI/ML will greatly benefit the Edge Computing domain.

Edge network infrastructure capacity poses limitations compared to centralized cloud re sources, requiring resource monitoring and management mechanisms for efficient Life Cycle Management (LCM) and service distribution. Recent advances in processing and AI/ML capabilities of smart devices offer exploitation opportunities, but challenges remain due to varying technology support and ongoing standardization efforts (e.g. ETSI Multi-access Edge Computing (MEC)). TANDEM addresses these challenges with a novel edge platform architecture, user-centered operations, and support mechanisms. It enables various users to develop, exe cute, and manage edge services without infrastructure complexities, leveraging reusable services. TANDEM explores innovative aspects of edge computing, including service and application com

position across edge nodes, incorporating AI/ML mechanisms and online monitoring functions. It automates service chain execution, adheres to standards, and encompasses business aspects such as pricing, Service Level Agreements (SLAs), and service/product life cycles.

Developing real-time prediction models for resource monitoring & management in edge network infrastructures is a complex challenge. Existing systems rely on simplistic heuristics, limiting their effectiveness in diverse applications and environments [6]. However, recent advancements in AI/ML offer the potential for improved models that outperform generic heuristics. By leveraging data-driven techniques and extracting relevant input data, customized ML models can be created. This approach addresses the need for efficient resource allocation and improved performance in modern applications while capitalizing on the growing volume of IoT generated Big Data. The proposed transformer models, incorporating a monitoring module for enhanced extraction of time dependencies, provide a solution for accurate system usage prediction.

This paper proposes novel forms of dynamic resource monitoring & management for edge network infrastructures to enable the automated selection of distributed and diverse resources. Additionally, we propose an AutoML framework and its AutoTinyML extension to enable automated online monitoring at the edge. The intention of this work is to provide an in-depth analysis and extensive experimentation on the intersection of AutoML and Transformer DL models with Edge AI- an area not yet adequately explored. In this sense, the contribution of this work is two-fold:

- A novel architectural element is presented, named Smart Resource Monitoring & Management, which is equipped with intelligent and data-driven prediction capabilities, able to provide adequate resource monitoring and management, to enhance coordination and delivery across stakeholders on the edge continuum;
- Transformer ML models are designed to draw patterns across resource types and determine suitable edge network infrastructure resources;

The rest of the article is organized as follows. Section 2 presents the related work. Section 3 presents the TANDEM and Smart Resource Monitoring & Management architecture. Section 4 discusses the evaluation results, while concluding remarks are presented in Section 5.

2 Related work

In the context of cloud, fog, and IoT environments, Edge Intelligence (EI), or Edge AI, moves AI frontiers from the cloud to the network edge to fully unlock the potential of big data at its production source and facilitate efficient management in various domains. [2] offers insights into the challenges in resource management for edge computing environments and lists several re source management techniques, covering aspects such as resource allocation, task offloading and load balancing. [3] presents a taxonomy of different ML-based resource management techniques in edge computing and analyzes their strengths, limitations, and potential applications. [9] introduces a dynamic resource management approach that allocates resources across cloud and edge environments, ensuring efficient resources utilization and meeting SLAs.

Growing research interest has been recently devoted to the use of ML transformer models at the edge, although research work is still at an early stage. [1] studies various deployment methods for transformer models at the edge and explores relevant tooling options and practical approaches to illustrate the utilization of these models effectively. [7] examines the use of transformer models in resource-constrained

edge environments with a specific focus on ultra-low power devices. [5] introduces Transformers on the Edge (Edgeformers) and demonstrates their effectiveness in capturing contextual information in edge computing scenarios, showcasing their potential for enhancing representation learning tasks in distributed edge environments.

Based on the above, in this work, a mechanism building on transformer models is proposed to facilitate the management and pattern identification of edge computing infrastructure resources. For the monitoring realization, a proposed AutoML pipeline and its AutoTinyML extension are used, showing the improved capabilities of transformer models in edge computing environments.

3 Architecture

The TANDEM platform offers support for Platform-as-a-Service (PaaS) edge and serverless computing models, enabling users to create complex service/function chains by combining existing services with custom ones. It focuses on edge computing environments like smart domains, connected vehicles and Industry 4.0, addressing challenges such as multi-protocol and multi data type support, device interconnection and management. The platform employs reusable and custom services, utilizing edge resources, connected devices, and auto-scaling/auto-healing mechanisms to facilitate users with easy applications development through its API.

The TANDEM architecture, shown in Figure 1, incorporates elements from the European Telecommunications Standards Institute (ETSI) Network Functions Virtualisation (NFV) reference architecture and the ETSI MEC architecture while introducing modules to support various usage scenarios. Its key feature is the ability to create more complex services and applications by integrating services and devices from different edge hosts or the cloud. The architecture is based on microservices, with edge nodes organized in edge clouds forming clusters. The architecture consists of three levels: the System Level, the Edge Computing Level and the Device Level. The System Level coordinates resources, devices, and services across all edge nodes via the Edge Orchestrator, while also providing user management, pricing policies, and billing functions via the Service Orchestrator. The Edge Level encompasses the TANDEM Edge Platform, IoT support services and customer-installed services. Finally, the Device Level includes all devices connected to TANDEM, which can send data and receive commands or microservices.

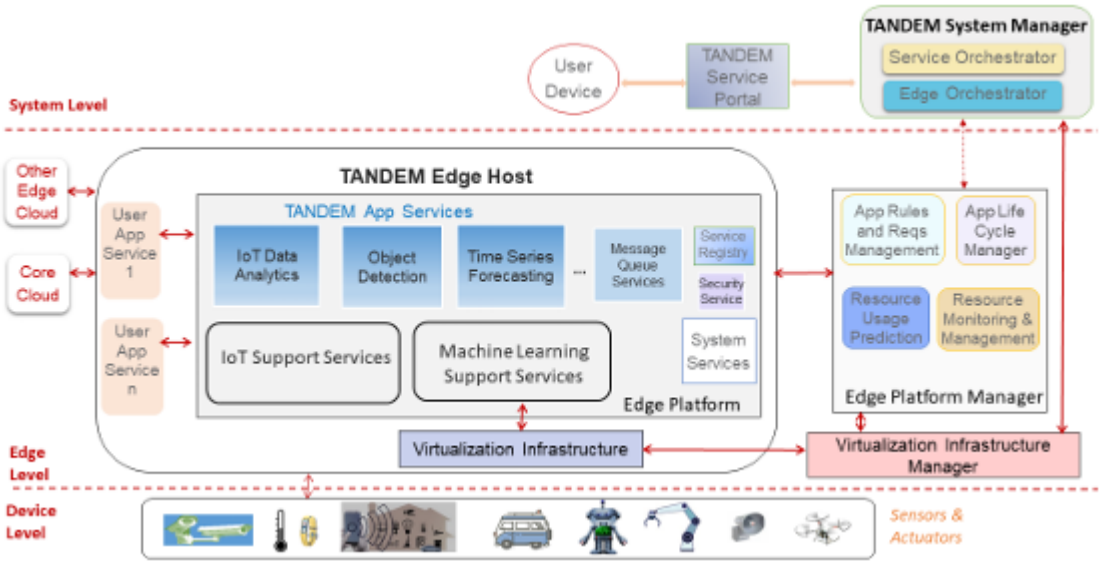

Figure 1: The architectural components and core technological stack of TANDEM

3.1. Smart Resource Monitoring & Management

In this work, we introduce a new subsystem into this architecture, named Smart Resource Monitoring & Management (SRMM), which performs zero-touch monitoring, management, and life cycle functions for one edge node or a cluster of edge hosts (edge cloud). It utilizes powerful AutoML mechanisms for accurate resource usage prediction, enabling efficient resource management and Quality of Service (QoS)

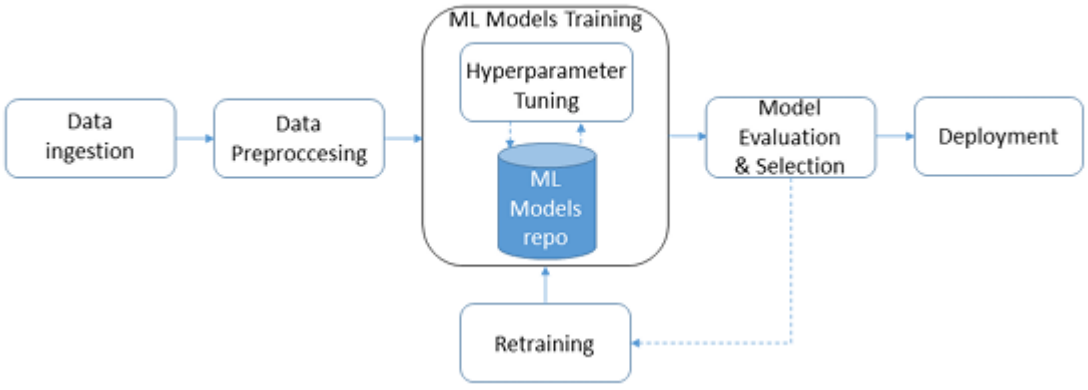

Figure 2: Smart Resource Monitoring & Management AutoML pipeline.

monitoring. Precise predictions are essential for optimal decision-making, which considers both existing and new services for intelligent resource allocation based on predicted resource usage, resource characteristics and user service requirements.

SRMM's AutoML pipeline ingests data and deploys a high-performing ML model for Time Series Forecasting to predict resource usage. The pipeline stages are data pre-processing, i.e. cleaning &

transformations, models training in parallel and hyperparameter tuning, model evaluation & selection and an optional retraining stage in case no model passes the evaluation phase, as shown in Figure 2. Moreover, AutoTinyML extends the SRMM's AutoML pipeline with model compression capabilities (e.g. post-training model pruning and quantization). The extension follows the same philosophy with the AutoML pipeline, but comprises an evaluation stage that checks for both model accuracy and size- particularly useful for Edge AI.

Our ML model suite contains novel transformers (Temporal Fusion Transformer (TFT) & Informer), MLP-based (Neural basis expansion analysis for interpretable time series forecasting (N-BEATS) & Neural Hierarchical Interpolation for Time Series Forecasting (N-HiTS)), popular RNN-based models (Gated Recurrent Unit (GRU) & Long Short-Term Memory (LSTM)) and a lightweight regression technique; Support Vector Regression (SVR).

We incorporate MLOps methodologies for the LCM of ML models. Over time, ML model accuracy can degrade due to data drift [8]. Periodic evaluation compares the accuracy to predefined empirical thresholds, triggering SRMM's AutoML pipeline with the newly- unseen by the ML models during the training phase- stored data as input.

The proactive auto-scaling/auto-healing mechanisms of TANDEM rely on SRMM's monitored metrics, where an empirical threshold is defined for each metric. In case SRMM's prediction exceed the set threshold, its management service is notified and automatically performs the required service adaptation. SRMM decides what the adjustment will be based on the metric that is predicted to exceed the threshold and some predefined rules, e.g. a configurable scaling factor. For example, if the metric is service related, then SRMM may decide to scale-out the service itself by deploying a new instance.

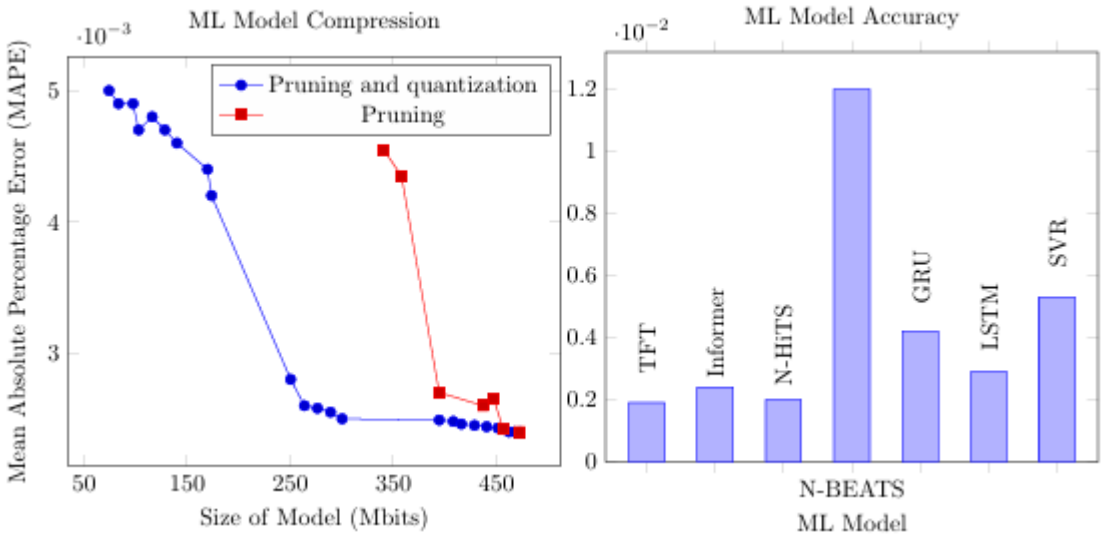

Figure 3: Left: Model accuracy after size pruning. Right: Resource usage prediction accuracy.

4 Performance Evaluation

This section details the thorough evaluation of the SRMM subsystem, focusing on the ML models accuracy performance and the effectiveness of its AutoTinyML extension.

4.1. Resource Usage Prediction Accuracy

Extensive experimentation shows the ML models', presented in section 3.1, robust accuracy. We collected data of four resource usage metrics, namely Memory Consumption, CPU Con sumption, Received Throughput and Transmitted Throughput. The resulted data set contains 8 days of 1 minute-based measurements.

We downsampled the data set and then split it in train, validation and test data sets. The train set and validation set are used internally by the SRMM's AutoML pipeline to automatically train and fine-tune (via 5-fold cross-validation) the ML models. Notice that for evaluation purposes we configured the pipeline to output all the trained models, and not just the best performing model; in other words we disabled the model selection phase. The test set is used to perform time series forecasting. Our results are shown in Figure 3, where it is evident that the transformer models (TFT and Informer) are performing better than RNN-based models (GRU and LSTM), and that they are comparable to the N-HiTS MLP-based model. SVR achieves lower accuracy than the aforementioned models, but outperforms N-BEATS, the other MLP-based model, which has the lowest accuracy in this experimentation setting.

4.2. AutoTinyML Model Compression

The AutoTinyML extension of the SRMM's AutoML pipeline takes into account size and accuracy of the ML model. Resource-constrained environments advocate for smaller models and can tolerate some minimal loss of accuracy. ML model pruning reduces the size and complexity of an ML model by removing unnecessary parameters, connections, or structures, thereby improving efficiency and reducing resource requirements without significant loss in performance. ML model quantization involves the process of reducing the precision of numerical values, typically from floating-point to fixed-point representation, in order to reduce memory footprint, improve computational efficiency, and enable deployment on resource-constrained devices. Our AutoTinyML extension prunes a configurable amount of trainable layer weights and quantizes the ML model post-training, by reducing data type precision. In this evaluation setting, the data set described in 4.1 was used. We tuned the data type precision from float32 to int8, including the corresponding unsigned and quantized types, in combination with applying polynomial decay for tensor sparsity from 50% up to 80%, i.e. from 50% up to 80% zeros in weights. Figure 3 shows the trade-off between size and accuracy, and depicts that model compression managed to reduce size by 6 times while the accuracy difference is significantly lower.

5 Conclusion

Enabling automated monitoring and management of edge computing resources via AI/ML is a growing demand to realize the envisioned capabilities of edge continuum. Towards such a goal, this paper provides insights into the design and implementation of the Smart Resource Monitoring and Management subsystem for the TANDEM approach, which offers a comprehensive solution for addressing resource monitoring and management challenges in edge computing. By integrating AI/ML mechanisms, TANDEM optimizes resource allocation, improves service distribution, and supports a wide range of edge applications. Experimental evaluations confirm its effectiveness and potential for practical deployment.

6 Acknowledgement

This work is supported by the TANDEM project, co-financed by the European Union and Greek national

funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH CREATE- INNOVATE (project code: T2ΕΛΚ-02825).

7 References

[1] Cassie Breviu. How to operationalize transformer models on the edge. QCon Plus ’22, 2022.

[2] Cheol-Ho Hong and Blesson Varghese. Resource management in fog/edge computing: A survey on architectures, infrastructure, and algorithms. ACM Computing Surveys, pages 1–37, 2019.

[3] Sundas Iftikhar et al. AI-based fog and edge computing: A systematic review, taxonomy and future directions. Internet of Things, 2023.

[4] Juan J. et al. Performance evaluation of state-of-the-art edge computing devices for DNN inference. IECON ’20, 2020.

[5] Bowen Jin, Yu Zhang, Yu Meng, and Jiawei Han. Edgeformers: Graph-empowered transformers for representation learning on textual-edge networks, 2023.

[6] Anshul Makkar. Scope and performance of credit-2 scheduler. Xen-Summit, 06 2016.

[7] Francesco Bianco Morghet, Daniele Jahier Pagliari, Alessio Burrello, et al. Application of trans formers to edge-computing in ultra-low power devices, 2021.

[8] Georgios Samaras et al. Qmp: A cloud-native mlops automation platform for zero-touch service assurance in 5g systems. In MeditCom, 2022.

[9] Shashank Shekhar et al. Dynamic resource management across cloud-edge resources for performance-sensitive applications. CCGRID ’17, 2017.