UMEÅ UNIVERSITET

Towards Safe and Efficient Application of Deep Neural Networks in Resource-Constrained Real-Time Embedded Systems

Siyu Luan

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i Triple Helix, Samverkanshuset, Umeå universitet, Måndagen den 09 oktober 2023, kl. 13:00.
Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Universitetslektor, Lei Feng,
Institutionen för maskinkonstruktion, Kungliga tekniska högskolan, Stockholm, Sverige.

Department of Applied Physics and Electronics

Organization
Umeå University
Department of Applied Physics
and Electronics

Document type
Doctoral thesis

Date of publication
11 09 2023

Author
Siyu Luan

Title
Towards Safe and Efficient Application of Deep Neural Networks in Resource-Constrained Real-Time Embedded Systems

Abstract
We consider real-time safety-critical systems that feature closed-loop interactions between the embedded computing system and the physical environment with a sense-compute-actuate feedback loop. Deep Learning (DL) with Deep Neural Networks (DNNs) has achieved success in many application domains, but there are still significant challenges in its application in real-time safety-critical systems that require high levels of safety certification under significant hardware resource constraints. This thesis considers the following overarching research goal: How to achieve safe and efficient application of DNNs in resource-constrained Real-Time Embedded (RTE) systems in the context of safety-critical application domains such as Autonomous Driving? Towards reaching that goal, this thesis presents a set of algorithms and techniques that aim to address three Research Questions (RQs): RQ1: How to achieve accurate and efficient Out-of-Distribution (OOD) detection for DNNs in RTE systems? RQ2: How to predict the performance of DNNs under continuous distribution shifts? RQ3: How to achieve efficient inference of Deep Reinforcement Learning (DRL) agents in RTE systems?

For RQ1, we present a framework for OOD detection based on outlier detection in one or more hidden layers of a DNN with either Isolation Forest (IF) or Local Outlier Factor (LOF). We also perform a comprehensive and systematic benchmark study of multiple well-known OOD detection algorithms in terms of both accuracy and execution time on different hardware platforms, in order to provide a useful reference for the practical deployment of OOD detection algorithms in RTE systems. For RQ2, we present a framework for predicting the performance of DNNs for end-to-end Autonomous Driving under continuous distribution shifts with two approaches: using an Autoencoder that attempts to reconstruct the input image; and applying Anomaly Detection algorithms to the hidden layer(s) of the DNN. For RQ3, we present a framework for model compression of the policy network of a DRL agent for deployment in RTE systems by leveraging the relevance scores computed by Layer-wise Relevance Propagation (LRP) to rank and prune the convolutional filters, combined with fine-tuning using policy distillation.

The algorithms and techniques developed in this thesis have been evaluated on standard datasets and benchmarks. To summarize our findings, we have developed novel OOD detection algorithms with high accuracy and efficiency; identified OOD detection algorithms with relatively high accuracy and low execution times through benchmarking; developed a framework for DNN performance prediction under continuous distribution shifts, and identified most effective Anomaly Detection algorithms for use in the framework; developed a framework for model compression of DRL agents that is effective in reducing model size and inference time for deployment in RTE systems. The research results are expected to assist system designers in the task of safe and efficient application of DNNs in resource-constrained RTE systems.

Keywords
Machine Learning/Deep Learning, Real-Time Embedded systems, Out-of-Distribution Detection, Distribution Shifts, Deep Reinforcement Learning, Model Compression, Policy Distillation.

Language
English

ISBN
print: 978-91-8070-160-0
PDF: 978-91-8070-161-7

Number of pages
60 + 4 papers