Topic: EdgeML: Enhancing On-Device ML with Quantised Models and Out-of-Distribution detection

Authors- Aditya Bansal, Eshani Agrawal

Introduction

Incorporating Out-of-distribution (OOD) detection ML models into autonomous vehicles and medical systems has the potential to revolutionize these domains. However, deploying these models on edge devices presents several challenges:

Limited Resources: On-device applications often run on hardware with restricted computational capabilities, necessitating efficient model deployment.

AI Safety: The safety of ML models in critical applications is paramount. OOD detection is crucial to identify situations where the model's predictions may be unreliable.

Sustainability: Sustainable ML aims to reduce the environmental impact of model deployment. Efficient model quantization contributes to sustainability goals.

Objectives: Optimized OOD-Detection

The primary objective of this solution is to optimize OOD detection for improving AI safety in autonomous vehicles and medical systems:

Robust OOD Detection: Develop and implement robust methodologies to detect OOD data points, enabling ML models to identify scenarios beyond their training distribution with high reliability.

Efficient System: The model is small enough to be deployed on small devices and doesn’t add lot of compute for the out of detection task. Hence, we applied performed neural network compression technique of quantization.

Implications

Enhanced AI Safety

This solution profoundly enhances AI safety by improving OOD detection. Robust methods for identifying OOD data points are integrated into ML models used in autonomous vehicles and medical systems. This ensures real-time identification of scenarios where the model's predictions may be unreliable, mitigating the risk of accidents and errors. Ultimately, this translates into improved safety for users and the public.

Sustainable ML

Sustainable ML, an essential aspect of our solution, embodies our commitment to minimizing the environmental footprint of machine learning practices. By incorporating quantization techniques into our approach, we not only optimize model size and computational efficiency but also significantly reduce the energy consumption during both model training and deployment.

Experiment Setup

Model: ResNet18
Metrics: Number of Parameters, Model Size
Devices: CPU, GPU
In Distribution Data - CIFAR10
Out-of-Distribution Data - ImageNet

Results

	FP32 ResNet18	**INT8 ResNet18**
Number of Parameters	11.1 M	11.1 M
Model Size	44.8 MB	12.4 MB
CPU Inference Latency (ms/sample)	7.91	2.87
GPU Inference Latency (ms/sample)	5.66	1.56

TABLE 1: Performance Metrics

	Baseline OOD Detector	**Quantized Model**
Detection error	10.0%	12.5%
AUROC	95.3%	89.1%
AUPR In	96.4%	93.0%
AUPR Out	93.8%	81.4%

TABLE 2: OOD Metrics

Conclusion

The solution focusing on OOD detection using quantized models for on-device applications addresses a critical aspect of AI safety, ensuring that ML models can operate reliably in high-stakes scenarios. As the use of AI continues to grow in importance, initiatives like this will shape the future of AI technology, making it safer and more dependable in mission-critical applications.