CS 5463: Survey-based Term Project

Topic: Energy Efficiency of Machine Learning on Edge Devices

Scope of the Survey: With the rising use of Machine Learning (ML) on edge devices, optimizing energy efficiency has become a critical research area. Traditional Deep Neural Networks (DNNs) are computationally intensive and require significant power and memory. This makes them unsuitable for resource-constrained environments such as embedded systems, IoT devices, and smartphones. This survey aims to explore emerging models and frameworks that improve the energy efficiency of ML models on edge devices while maintaining accuracy and low latency.

The focus of this study includes:

- **Model Optimization Techniques**: Investigating approaches such as knowledge distillation, pruning, quantization, and autoencoders to reduce model size and inference time.
- **Autoencoder-based Methods**: Analyzing frameworks such as EncodeNet, CAE-Net, and converting autoencoders that transform complex images into an easy-to-classify image of its class.
- **Speeding Up Model Decisions**: Reviewing techniques such as early-exit DNNs, entropy-based clustering and others to minimize computational overhead.
- **Testing on Real Devices**: Evaluating performance metrics such as energy consumption, latency, and accuracy trade-offs on edge devices like Raspberry Pi and Nvidia Jetson Nano.

By reviewing recent research on energy-efficient Machine Learning, including EncodeNet, CAE-Net, and other edge-optimized DNN frameworks, this survey aims to understand the best strategies for deploying ML models efficiently on edge devices.

Course Website: https://github.com/asifulhoque23/utsa-cs5463-course-website