Optimasi Model Menggunakan Teknik Pruning dan Kuantisasi untuk Implementasi pada ESP32 dan ESP32-S3 = Optimization of Human Activity Recognition Model using Pruning and Quantization Techniques for ESP32 and ESP32-S3 Implementations

Azzam Muhammadi Rizqun Karima, author

Abstrak

Penelitian ini mengeksplorasi optimasi model pembelajaran mesin untuk implementasi pada perangkat edge, dengan studi kasus Human Activity Recognition (HAR). Fokus utama adalah teknik optimasi model untuk meningkatkan efisiensi komputasi dan penggunaan sumber daya pada perangkat dengan keterbatasan, khususnya ESP32 dan ESP32-S3.Lima arsitektur model diuji sebagai basis evaluasi: Shallow Network, Deep Network, Gated Recurrent Unit (GRU), 1D Convolutional Neural Network (1DCNN), dan Long Short-Term Memory-Convolutional Neural Network (LSTM-CNN). Penelitian menerapkan kombinasi teknik pruning dan kuantisasi untuk mengoptimalkan model-model tersebut.Hasil menunjukkan bahwa teknik optimasi dapat mengurangi ukuran model hingga 90,43% dengan penurunan akurasi minimal. Evaluasi pada ESP32 dan ESP32-S3 mengungkapkan peningkatan kinerja yang signifikan setelah optimasi, dengan peningkatan throughput mencapai 592,09% pada ESP32-S3 yang dilengkapi fitur SIMD. Analisis menggunakan Analytical Hierarchy Process (AHP) mengidentifikasi model 1DCNN teroptimasi sebagai solusi paling seimbang, dengan ukuran 0,18 MB, latency 17,06 ms, dan akurasi 89,56%. Penelitian ini memberikan kerangka kerja sistematis untuk optimasi model pembelajaran mesin pada perangkat edge, serta pemahaman mendalam tentang trade-off antara efisiensi komputasi dan akurasi model.
.....This research explores machine learning model optimization for edge device implementation, using Human Activity Recognition (HAR) as a case study. The primary focus is on model optimization techniques to enhance computational efficiency and resource utilization on constrained devices, specifically ESP32 and ESP32-S3.Five model architectures were tested as evaluation bases: Shallow Network, Deep Network, Gated Recurrent Unit (GRU), 1D Convolutional Neural Network (1DCNN), and Long Short-Term Memory-Convolutional Neural Network (LSTM-CNN). The research applies a combination of pruning and quantization techniques to optimize these models.Results show that optimization techniques can reduce model size by up to 90.43% with minimal accuracy loss. Evaluation on ESP32 and ESP32-S3 reveals significant performance improvements after optimization, with throughput increases reaching 592.09% on the SIMD-equipped ESP32-S3. Analysis using the Analytical Hierarchy Process (AHP) identifies the optimized 1DCNN model as the most balanced solution, with a size of 0.18 MB, latency of 17.06 ms, and accuracy of 89.56%. This research provides a systematic framework for machine learning model optimization on edge devices, as well as deep insights into the trade-offs between computational efficiency and model accuracy.