Edge Deployment Guide

Optimized for low-resource devices using quantization.

Features

- Dynamic quantization
- Small model size
- Efficient execution

Example

```
quantized_model = torch.quantization.quantize_dynamic(
trained_dml,
{nn.Linear, nn.LayerNorm},
dtype=torch.qint8
)
```