Mixed Precision Strategies

Mixed Precision Strategy

Achieving a balance between performance and efficiency by using different precision levels for each layer

Strategy Examples

Input Layer

INT8/FP16
Fast processing enabled

Hidden Layers

INT8
Most computations

Output Layer

FP16/FP32
Accurate probability calculation

Layer Sensitivity Analysis

Sensitive Layers → **Maintain FP16/FP32**

Less Sensitive Layers → **Apply INT8/INT4**

Tools

PyTorch Quantization, TensorRT, ONNX Runtime