T0-QAT: ξ-Aware Quantization-Aware Training

Experimental Validation of Noise-Resilient AI Training

Based on T0 Time-Mass Duality Theory

Zusammenfassung

This document presents experimental validation of ξ-aware quantization-aware training, where $\xi = \frac{4}{3} \times 10^{-4}$ is derived from fundamental physical principles in the T0-Theory (Time-Mass Duality). Our preliminary results demonstrate improved robustness to quantization noise compared to standard approaches, providing a physics-informed method for enhancing AI efficiency through principled noise regularization.

Inhaltsverzeichnis

0.1 Einleitung

Quantization-aware training (QAT) hat sich als entscheidende Technik für das Deployment von neuronalen Netzen auf ressourcenbeschränkten Geräten etabliert. Allerdings basieren aktuelle Ansätze oft auf empirischen Rausch-Injektionsstrategien ohne theoretische Grundlage. Diese Arbeit führt ξ-aware QAT ein, basierend auf der T0 Zeit-Masse-Dualitätstheorie, die eine fundamentale physikalische Konstante ξ bereitstellt, die numerische Präzisionsgrenzen natürlich regularisiert.

0.2 Theoretische Grundlagen

0.2.1 T0 Zeit-Masse-Dualitätstheorie

Der Parameter $\xi = \frac{4}{3} \times 10^{-4}$ ist keine empirische Optimierung, sondern leitet sich aus ersten Prinzipien der Fundamentale Fraktalgeometrische Feldtheorie (FFGFT, früher T0-Theorie) der Zeit-Masse-Dualität ab. Diese fundamentale Konstante repräsentiert den minimalen Rauschpegel, der physikalischen Systemen inhärent ist, und bietet eine natürliche Regularisierungsgrenze für numerische Präzisionslimits.

Die vollständige theoretische Herleitung ist im T0 Theory GitHub Repository verfügbar[1], einschließlich:

- Mathematische Formulierung der Zeit-Masse-Dualität
- Herleitung fundamentaler Konstanten
- Physikalische Interpretation von ξ als Quantenrauschgrenze

0.2.2 Implikationen für AI Quantization

Im Kontext der Neural Network Quantization repräsentiert ξ die fundamentale Präzisionsgrenze, unterhalb derer weitere Bit-Reduzierung aufgrund physikalischer Rauschbeschränkungen abnehmende Erträge liefert. Durch die Einbeziehung dieser physikalischen Konstante während des Trainings lernen Modelle, optimal innerhalb dieser natürlichen Präzisionsgrenzen zu operieren.

[1] https://github.com/jpascher/T0-Time-Mass-Duality/releases/tag/v3.2

0.3 Experimenteller Aufbau

0.3.1 Methodik

Wir entwickelten ein vergleichendes Framework zur Evaluierung von ξ-aware Training gegenüber standard Quantization-aware Ansätzen. Das experimentelle Design besteht aus:

- **Baseline:** Standard QAT mit empirischer Rausch-Injektion
- **T0-QAT:** ξ-aware Training mit physikalisch-informiertem Rauschen
- **Evaluation:** Quantisierungsrobustheit unter simulierter Präzisionsreduktion

0.3.2 Datensatz und Architektur

Für die initiale Validierung verwendeten wir eine synthetische Regressionsaufgabe mit einer einfachen neuronalen Architektur:

- **Datensatz:** 1000 Samples, 10 Features, synthetisches Regressionsziel
- **Architektur:** Einzelne lineare Schicht mit Bias
- **Training:** 300 Epochen, Adam Optimizer, MSE Loss

0.4 Ergebnisse und Analyse

0.4.1 Quantitative Ergebnisse

Methode	Volle Präzision	Quantisiert	Drop
Standard QAT	0.318700	3.254614	2.935914
T0-QAT (ξ-aware)	9.501066	10.936824	1.435758