Novelty Assessment Report

Paper: Tractability via Low Dimensionality: The Parameterized Complexity of Training Quantized Neural Networks
PDF URL: https://openreview.net/pdf?id=BAQNrsr987
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

The training of neural networks has been extensively studied from both algorithmic and complexity-theoretic perspectives, yet recent results in this direction almost exclusively concern real-valued networks. In contrast, advances in machine learning practice highlight the benefits of quantization, where network parameters and data are restricted to finite integer domains, yielding significant improvements in speed and energy efficiency. Motivated by this gap, we initiate a systematic complexity-theoretic study of ReLU Neural Network Training in the full quantization mode. We establish strong lower bounds by showing that hardness already arises in the binary setting and under highly restrictive structural assumptions on the architecture, thereby excluding parameterized tractability for natural measures such as depth and width. On the positive side, we identify nontrivial fixed-parameter tractable cases when parameterizing by input dimensionality in combination with width and either output dimensionality or error bound, and further strengthen these results by replacing width with the more general treewidth.

Core Task Landscape

This paper addresses: **Parameterized Complexity of Training Quantized Neural Networks**

A total of **8 papers** were analyzed and organized into a taxonomy with **7 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Complexity-Theoretic Foundations**
- **Algorithm Design and Training Methods**
- **Hardware-Oriented Inference Optimization**

Complete Taxonomy Tree

- Parameterized Complexity of Training Quantized Neural Networks Survey Taxonomy
- Complexity-Theoretic Foundations
  - Parameterized Complexity Analysis ★ (1 papers)
  - [0] Tractability via Low Dimensionality: The Parameterized Complexity of Training Quantized Neural Networks (Anon et al., 2026) View paper
  - General Computational Hardness (1 papers)
  - [1] On the hardness of training deep neural networks discretely (Ilan Doron-Arad, 2025) View paper
  - Discrete Neural Network Theory (2 papers)
  - [6] Complexity issues in discrete neurocomputing (Wiedermann, 1990) View paper
  - [7] Discrete mathematics of neural networks: selected topics (Martin Anthony, 2001) View paper
- Algorithm Design and Training Methods
  - Low-Precision Training Algorithms (2 papers)
  - [2] High-accuracy low-precision training (De Sa, 2018) View paper
  - [8] Progressive Learning of Low-Precision Networks (Zhou Zheng-guang, 2019) View paper
  - Gradient Estimation for Discrete Variables (1 papers)
  - [5] Bias-variance tradeoffs in single-sample binary gradient estimators (Shekhovtsov, 2021) View paper
  - Rule Extraction and Generalization (1 papers)
  - [3] Extraction of rules from discrete-time recurrent neural networks (C. Omlin, 1996) View paper
- Hardware-Oriented Inference Optimization (1 papers)
  - [4] Wrapnet: Neural net inference with ultra-low-precision arithmetic (Ni, 2021) View paper

Narrative

Core task: parameterized complexity of training quantized neural networks. The field structure suggested by the taxonomy divides into three main branches. Complexity-Theoretic Foundations investigates the fundamental computational hardness of training networks with discrete or low-precision parameters, often drawing on classical complexity theory and parameterized analysis to identify which problem features make training tractable or intractable. Algorithm Design and Training Methods focuses on practical techniques—such as gradient approximation schemes, progressive quantization strategies, and specialized optimizers—that enable effective learning despite discrete constraints. Hardware-Oriented Inference Optimization emphasizes deployment considerations, exploring ultra-low precision representations and efficient inference architectures that exploit quantization for speed and energy savings. Early theoretical work like Complexity Discrete Neurocomputing[6] and Discrete Mathematics Neural Networks[7] laid groundwork by characterizing discrete neural computation, while more recent efforts such as Hardness Training Deep Discretely[1] and High Accuracy Low Precision[2] bridge theory and practice by analyzing training difficulty and demonstrating that carefully designed low-bit schemes can preserve accuracy.

Particularly active lines of work contrast algorithmic innovation with rigorous complexity analysis. On one hand, methods like Progressive Low Precision Networks[8] and Wrapnet Ultra Low Precision[4] push the envelope of how few bits suffice for competitive performance, often relying on heuristic training tricks and empirical validation. On the other hand, studies such as Bias Variance Binary Gradients[5] and Hardness Training Deep Discretely[1] probe the statistical and computational trade-offs inherent in discrete optimization, revealing when and why quantized training becomes hard. The original paper, Tractability via Low Dimensionality[0], sits squarely within the Complexity-Theoretic Foundations branch, offering a parameterized lens that identifies low-dimensional structure as a key to tractability. Compared to Hardness Training Deep Discretely[1], which emphasizes worst-case intractability, Tractability via Low Dimensionality[0] highlights positive algorithmic results under restricted parameterizations, thereby complementing the broader landscape of hardness and approximation results in quantized network training.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses specifically on parameterized complexity analysis, examining how structural parameters (depth, width, treewidth, dimensionality) affect the tractability of training quantized neural networks. The sibling subtopics cover complementary aspects: Discrete Neural Network Theory addresses foundational theoretical properties of discrete networks, while General Computational Hardness establishes baseline complexity results without parameter-specific analysis. Together, these form a hierarchy from general hardness to parameter-specific tractability studies.

Similarities: - All three subtopics address theoretical complexity aspects of discrete/quantized neural networks rather than practical algorithms - Each examines computational difficulty of problems related to neural network training or learning - All exclude practical training algorithms and implementation methods, focusing on theoretical analysis

Differences: - Parameterized Complexity Analysis uniquely focuses on fixed-parameter tractability under specific structural parameters, while General Computational Hardness establishes unconditional hardness results - Discrete Neural Network Theory emphasizes foundational properties (threshold functions, Boolean representations) rather than complexity classification - The original leaf specifically targets quantized networks with parameter-dependent analysis, while General Computational Hardness provides parameter-free baseline results - Discrete Neural Network Theory is broader in scope, covering learning complexity and representational theory beyond just training hardness

Suggested Search Directions: - Investigate connections between structural parameters (treewidth, depth) and the transition from hardness to tractability - Explore whether results from Discrete Neural Network Theory (e.g., threshold function properties) inform parameter choices in parameterized analysis - Examine how general hardness results serve as starting points for identifying meaningful parameterizations

Sibling Subtopics

- **Discrete Neural Network Theory** (leaves: 1, papers: 2)
- Scope: Examines theoretical foundations of discrete neural networks including threshold functions, Boolean representations, and learning complexity.
- Exclude: Excludes training algorithms and quantization methods; those belong in Algorithm Design and Training Methods.
- **General Computational Hardness** (leaves: 1, papers: 1)
- Scope: Establishes fundamental complexity results for discrete neural network training without parameterized analysis.
- Exclude: Excludes parameterized complexity studies; those belong in Parameterized Complexity Analysis.

Contributions Analysis

Overall novelty summary. The paper establishes parameterized complexity results for training quantized ReLU networks, contributing both hardness and tractability findings. It resides in the Parameterized Complexity Analysis leaf under Complexity-Theoretic Foundations, where it is currently the sole paper in that specific leaf. This positioning reflects a relatively sparse research direction: while the broader taxonomy encompasses eight papers across multiple branches, the parameterized complexity perspective on quantized training remains underexplored. The taxonomy reveals that most related work either addresses general computational hardness without parameterization or focuses on algorithmic and hardware-oriented methods rather than fine-grained complexity analysis.

The taxonomy structure shows that neighboring work divides into three main branches. Complexity-Theoretic Foundations contains sibling leaves on General Computational Hardness and Discrete Neural Network Theory, which establish fundamental intractability results and theoretical properties of discrete networks but do not employ parameterized analysis. Algorithm Design and Training Methods develops practical low-precision training strategies and gradient estimation techniques, explicitly excluding complexity-theoretic analysis. Hardware-Oriented Inference Optimization targets deployment efficiency rather than training complexity. The paper's focus on parameterized tractability via structural parameters like treewidth and dimensionality distinguishes it from these neighboring directions, which either lack parameterization or emphasize empirical performance over theoretical boundaries.

Among the 22 candidates examined, the contribution on fixed-parameter tractability via input dimensionality combined with width and output/error parameters shows one refutable candidate out of ten examined, suggesting some overlap with prior parameterized approaches. The other two contributions—strong lower bounds for quantized training and tractability results using treewidth—each examined ten and two candidates respectively, with no clear refutations found. The limited search scope means these statistics reflect top semantic matches rather than exhaustive coverage. The lower bounds contribution appears more novel within this sample, while the dimensionality-based tractability result encounters at least one overlapping prior work among the candidates reviewed.

Given the limited search of 22 candidates and the sparse population of the Parameterized Complexity Analysis leaf, the paper appears to occupy a relatively underexplored niche. The taxonomy context suggests that while quantized network training has received attention from algorithmic and hardware perspectives, the systematic parameterized complexity treatment remains less developed. The analysis captures top semantic matches but does not claim exhaustive field coverage, leaving open the possibility of additional related work outside the examined candidate set.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Strong lower bounds for quantized ReLU neural network training

Description: The authors prove that training quantized ReLU neural networks remains NP-hard even in the binary (2-bit) case and under severe restrictions including constant depth, constant width, single output neurons, and zero error bounds. These results exclude fixed-parameter tractability for standard architectural parameters like depth and width.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Tight Hardness Results for Training Depth-2 ReLU Networks

URL: View paper

Brief Assessment

Tight Hardness Depth Two[30] focuses on training depth-2 ReLU networks with real-valued parameters and square loss minimization, not quantized networks with integer domains. The technical settings are fundamentally different.

2. Hardness of noise-free learning for two-hidden-layer neural networks

URL: View paper

Brief Assessment

Hardness Two Hidden Layer[29] focuses on statistical query lower bounds for learning two-hidden-layer ReLU networks in the noise-free setting with Gaussian inputs, not on training quantized networks with constant depth/width restrictions.

Contribution 2: Fixed-parameter tractability via input dimensionality combined with width and output/error parameters

Description: The authors develop algorithms showing that quantized neural network training becomes fixed-parameter tractable when parameterized by the input dimensionality combined with network width and either the output dimensionality or the error bound. These are the first non-trivial tractable parameterizations for this problem.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Complexity of Injectivity and Verification of ReLU Neural Networks

URL: View paper

Brief Assessment

Complexity Injectivity Verification ReLU[23] focuses on injectivity and verification problems for ReLU networks, not on training problems. The candidate studies different computational problems (deciding injectivity, network verification) rather than the training problem addressed in the original paper.

2. Nonlinear initialization methods for low-rank neural networks

URL: View paper

Brief Assessment

Nonlinear Initialization Low Rank[25] focuses on low-rank neural network initialization methods and provides a fixed-parameter tractable algorithm for the ReLU low-rank approximation problem (Algorithm 3, Corollary 3.4). However, this addresses a different problem domain (initialization and function approximation) rather than the quantized neural network training problem studied in the original paper.

3. Complexity of Training ReLU Neural Network

URL: View paper

Prior Art Analysis

Complexity Training ReLU[26] demonstrates that similar fixed-parameter tractability results for neural network training parameterized by input dimensionality combined with network width were established prior to the original paper. The candidate paper proves that training k-ReLU neural networks is fixed-parameter tractable when parameterized by input dimension (d) and the number of nodes in the first layer (k), which directly corresponds to the original paper's claim about input dimensionality and width. Both papers establish FPT algorithms for the same parameter combinations, though the candidate focuses on ReLU networks without quantization while the original addresses quantized networks.

Evidence

Evidence 1 - **Rationale**: Both papers establish fixed-parameter tractability for neural network training when parameterized by input dimensionality combined with network width. The candidate's 'k' (number of nodes in first layer) corresponds to the original's 'width', and both achieve polynomial-time algorithms under these parameterizations. - **Original**: as our second set of contributions , we show that parameterizing by α enables fixed-parameter neural network training in the quantized setting-but only when combined with additional restrictions. in particular, our results imply that for every fixed d, d-qnnt is fixed-parameter tractable w.r.t. the ... - **Candidate**: under the assumption that the dimension of input, dand the number of nodes in the first layer, k, are constant, then there exists a poly(n)-time solution to the training problem of k-relu neural network, where n is the number of data-points.

Evidence 2 - **Rationale**: Both papers use similar algorithmic approaches to achieve FPT results. The candidate explicitly describes enumerating combinations based on fixed input dimension and network structure, which is the same conceptual approach as the original paper's FPT algorithms for input dimensionality combined with width. - **Original**: we identify nontrivial fixed-parameter tractable cases when parameterizing by input dimensionality in combination with width and either output dimensionality or error bound, and further strengthen these results by replacing width with the more general treewidth. - **Candidate**: the high-level idea of the proof is the following: each data point "passes through" the three relu nodes and the activation function in these nodes is "turned on" or "turned off" (i.e., the output is 0 or not). we will enumerate all possible combinations of the data points being turned on or not, whi...

Evidence 3 - **Rationale**: The candidate's Theorem 3.3 establishes FPT with respect to input dimension (d) and width (k), which is a subset of the original paper's claimed contributions. This shows that the core FPT result for input dimensionality combined with width was already known in prior work. - **Original**: corollary 1. d-qnnt is fpt with respect to $\alpha + \ell$ + width. [...] corollary 2. d-qnnt is fpt with respect to $\alpha + \omega$ + width. - **Candidate**: theorem 3.3 under the assumption that the dimension of input, dand the number of nodes in the first layer, k, are constant, then there exists a poly(n)-time solution to the training problem of k-relu neural network, where n is the number of data-points.

4. Beating the Perils of Non-Convexity: Guaranteed Training of Neural Networks using Tensor Methods

URL: View paper

Brief Assessment

Beating Non Convexity Tensor[28] addresses training neural networks using tensor methods for guaranteed convergence, not parameterized complexity theory or fixed-parameter tractability analysis of quantized networks.

5. Learning Deep ReLU Networks Is Fixed-Parameter Tractable

URL: View paper

Brief Assessment

Learning Deep ReLU Tractable[27] focuses on learning/inference algorithms for ReLU networks with Gaussian inputs, not the training problem with quantized parameters and data that the original paper addresses.

6. Training Fully Connected Neural Networks is -Complete

URL: View paper

Brief Assessment

Training Fully Connected Complete[22] focuses on ∃R-completeness of training fully connected networks, not on fixed-parameter tractability. The candidate establishes computational hardness rather than tractable parameterizations.

7. Training neural networks is np-hard in fixed dimension

URL: View paper

Brief Assessment

Training Fixed Dimension Hard[19] focuses on proving NP-hardness for fixed input dimension d=2, which is orthogonal to the original paper's positive algorithmic results showing FPT tractability when combining input dimensionality with width and output/error parameters.

8. SHAP Meets Tensor Networks: Provably Tractable Explanations with Parallelism

URL: View paper

Brief Assessment

SHAP Tensor Networks[21] focuses on computing SHAP explanations for tensor networks and neural networks, not on the parameterized complexity of training quantized neural networks. The technical domains are fundamentally different.

9. On the complexity of learning neural networks

URL: View paper

Brief Assessment

Complexity Learning Neural Networks[24] focuses on statistical query complexity lower bounds for learning neural networks, not fixed-parameter tractability of training algorithms. The paper establishes hardness results for learning rather than algorithmic tractability parameterized by structural measures.

10. The computational complexity of ReLU network training parameterized by data dimensionality

URL: View paper

Brief Assessment

ReLU Training Data Dimensionality[20] focuses on two-layer networks with a single output neuron and does not address the combined parameterization by input dimensionality, network width, and output dimensionality/error bounds that characterizes the original paper's contribution.

Contribution 3: Strengthened tractability results using treewidth instead of width

Description: The authors prove that their fixed-parameter tractability results can be strengthened by replacing the width parameter with treewidth, a more general graph-theoretic measure. This is achieved through a structural insight (Lemma 1) bounding the non-zero indegree of neurons in solutions, enabling dynamic programming over tree decompositions.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. A neural network approach for efficiently answering most probable explanation queries in probabilistic models

URL: View paper

Brief Assessment

Neural MPE Queries[10] focuses on neural network-based approximate inference for MPE queries in probabilistic models, not on parameterized complexity or tractability analysis of neural network training problems using treewidth.

2. Learning to Condition: A Neural Heuristic for Scalable MPE Inference

URL: View paper

Brief Assessment

Learning to Condition[13] focuses on learning heuristics for MPE inference in probabilistic graphical models, not on parameterized complexity or tractability results for neural network training problems.

3. Descriptive complexity of learning

URL: View paper

Brief Assessment

Descriptive Complexity Learning[18] focuses on learning logical formulas from relational structures using treewidth for graph classes, not neural network training with quantized parameters.

4. Efficient rematerialization for deep networks

URL: View paper

Brief Assessment

Efficient Rematerialization[12] focuses on memory-efficient training schedules for neural networks using treewidth-based algorithms for rematerialization, not on tractability of neural network training problems in the quantized setting that the original paper addresses.

5. Fast Second-order Method for Neural Networks under Small Treewidth Setting

URL: View paper

Brief Assessment

Fast Second Order Treewidth[17] focuses on second-order optimization methods for over-parameterized neural networks with quadratic convergence, not on the parameterized complexity of quantized neural network training problems.

6. A review on deep learning classification techniques for gait recognition on humans

URL: View paper

Brief Assessment

Gait Recognition Review[14] is a survey on deep learning for gait recognition in humans, focusing on computer vision applications. It does not address parameterized complexity, neural network training tractability, or graph-theoretic measures like treewidth.

7. Inapproximability of graph width parameters under small set expansion hypothesis

URL: View paper

Brief Assessment

Inapproximability Graph Width[11] focuses on inapproximability of graph width parameters under complexity assumptions, not on tractability algorithms for neural network training using treewidth-based dynamic programming.

8. Learning Treewidth-Bounded Bayesian Networks with Thousands of Variables

URL: View paper

Brief Assessment

Learning Treewidth Thousands Variables[15] focuses on learning Bayesian network structures with bounded treewidth, not neural network training tractability. The problem domains are fundamentally different.

---

9. New complexity-theoretic frontiers of tractability for neural network training

URL: View paper

Brief Assessment

Complexity Frontiers Neural Training[9] focuses on treewidth-based algorithms for linear and ReLU network training in the non-quantized setting, while the original paper addresses quantized neural network training. The technical contexts and problem formulations differ fundamentally.

---

10. Learning Bounded Treewidth Bayesian Networks

URL: View paper

Brief Assessment

Learning Bounded Treewidth[16] focuses on learning Bayesian network structures with bounded treewidth, not neural network training tractability. These are fundamentally different problem domains with different computational challenges.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Tractability via Low Dimensionality: The Parameterized Complexity of Training Quantized Neural Networks View paper
- [1] On the hardness of training deep neural networks discretely View paper
- [2] High-accuracy low-precision training View paper
- [3] Extraction of rules from discrete-time recurrent neural networks View paper
- [4] Wrapnet: Neural net inference with ultra-low-precision arithmetic View paper
- [5] Bias-variance tradeoffs in single-sample binary gradient estimators View paper
- [6] Complexity issues in discrete neurocomputing View paper
- [7] Discrete mathematics of neural networks: selected topics View paper
- [8] Progressive Learning of Low-Precision Networks View paper
- [9] New complexity-theoretic frontiers of tractability for neural network training View paper
- [10] A neural network approach for efficiently answering most probable explanation queries in probabilistic models View paper
- [11] Inapproximability of graph width parameters under small set expansion hypothesis View paper
- [12] Efficient rematerialization for deep networks View paper
- [13] Learning to Condition: A Neural Heuristic for Scalable MPE Inference View paper
- [14] A review on deep learning classification techniques for gait recognition on humans View paper
- [15] Learning Treewidth-Bounded Bayesian Networks with Thousands of Variables View paper
- [16] Learning Bounded Treewidth Bayesian Networks View paper
- [17] Fast Second-order Method for Neural Networks under Small Treewidth Setting View paper
- [18] Descriptive complexity of learning View paper
- [19] Training neural networks is np-hard in fixed dimension View paper
- [20] The computational complexity of ReLU network training parameterized by data dimensionality View paper
- [21] SHAP Meets Tensor Networks: Provably Tractable Explanations with Parallelism View paper
- [22] Training Fully Connected Neural Networks is -Complete View paper
- [23] Complexity of Injectivity and Verification of ReLU Neural Networks View paper
- [24] On the complexity of learning neural networks View paper
- [25] Nonlinear initialization methods for low-rank neural networks View paper
- [26] Complexity of Training ReLU Neural Network View paper
- [27] Learning Deep ReLU Networks Is Fixed-Parameter Tractable View paper
- [28] Beating the Perils of Non-Convexity: Guaranteed Training of Neural Networks using Tensor Methods View paper
- [29] Hardness of noise-free learning for two-hidden-layer neural networks View paper
- [30] Tight Hardness Results for Training Depth-2 ReLU Networks View paper