Adversarial Training in Stochastic Bitwise Neural Networks: Robustness Analysis Against FGSM Attacks on CIFAR-10

Assignee Research

June 11, 2026

Abstract

Recently published methods enable training of bitwise neural networks which allow reduced representation of down to a single bit per weight. We present a method that exploits ensemble decisions based on multiple stochastically sampled network models to increase performance figures of bitwise neural networks in terms of classification accuracy at inference. Our experiments with the CIFAR-10 and GTSRB datasets show that the performance of such network ensembles surpasses the performance of the high-precision base model. With this technique we achieve 5.81% best classification error on CIFAR-10 t

1 Introduction

This paper examines: Efficient Stochastic Inference of Bitwise Deep Neural Networks. Research question: How does adversarial training with stochastic sampling in bitwise neural networks impact robustness against FGSM attacks on CIFAR-10 compared to deterministic weight-based BNNs, measured by classification accuracy under varying perturbation magnitudes?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.7/10.

3 Results

16 papers retrieved. 10 claims extracted; 10 independently verified. Quality review score: 8.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The CIFAR-10 dataset contains 60,000 images in 32$\times$32 pixel RGB resolution and 10 different classes.	✓	0.26
The network structure used is 128C3–128C3–MP2–256C3–256C3–MP2–512C3–512C3–MP2–1024FC–1024FC–10SVM4.	✓	0.27
The model was trained for 500 epochs with hyperparameters from [4] and without any preprocessing or augmentations on the	✓	0.15
The high-precision model parameters with the lowest error on the validation set were used to generate multiple instances	✓	0.33
A stochastic BNN ensemble with at least four members always performs better than the floating-point reference model, whi	✓	0.27
The best classification error rate achieved with an ensemble of 23 networks was 9.41%.	✓	0.19
Better classification results can be achieved when the same network is trained with ReLU activation function, binary pro	✓	0.27
The best classification error rate achieved with an ensemble of 29 networks using ReLU activation was 5.81%.	✓	0.24
The high-precision reference model with ReLU activation achieves a classification error of 6.13%.	✓	0.18
The classification error rates for different accumulation lengths (numbers of ensemble members) are plotted in Figure 2a	✓	0.25

References

- http://arxiv.org/abs/2409.15190v1
- http://arxiv.org/abs/1611.06539v1
- http://arxiv.org/abs/2601.09933v1