QuaRot Extended

Outlier-Free 4-Bit Inference in Rotated LLMs

Machine Learning in Practice
Professor Gary Kazantsev
May 6, 2025

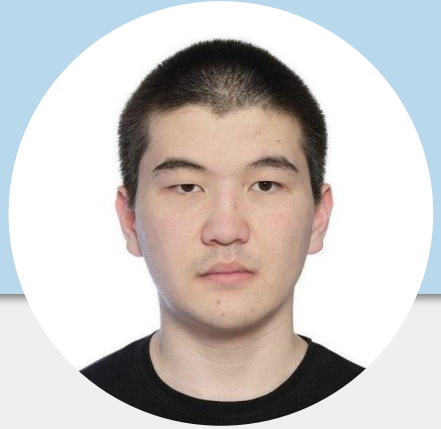
Abhinaya Anil Menon (AA5536)
Ali Bauyrzhan (AB5867)
Dacia John (DMJ2149)

COLUMBIA ENGINEERING
The Fu Foundation School
of Engineering and Applied Science

Instructor and the Team

Professor Gary Kazantsev (Instructor)

Abhinaya Menon



Ali Bauyrzhan

Daisy John

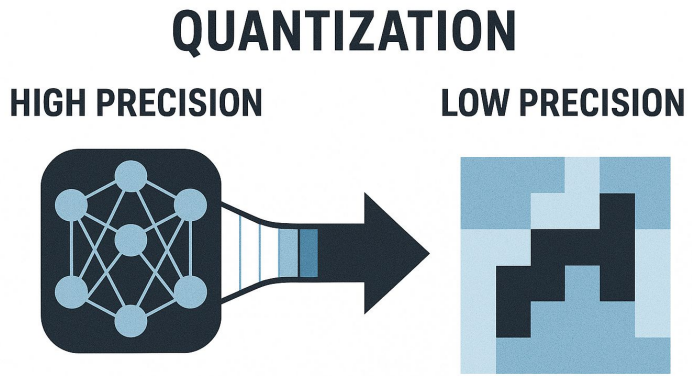
Topics

- Background
- QuaRot Paper's Purpose
- QuaRot's Contribution
- Challenges
- Three Main Extensions
- Matrix Optimization
- Dataset Integration
- Skew based quantization mode
- Challenges
- Future Work
- Conclusion
- References

Introduction and Background

Why did we choose this paper?

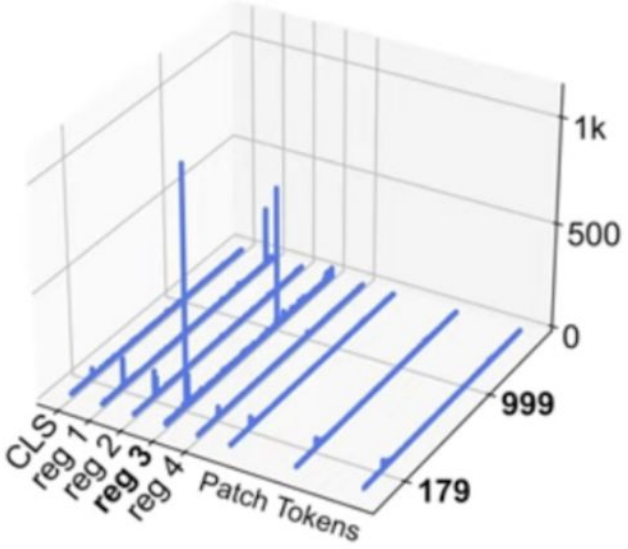
- LLMs are widely used, but hard to deploy privately due to high GPU/memory needs.
- Quantization is crucial for efficient, private LLM deployment.
- This 2024 paper proposes an innovative method (QuaRot) using rotations for 4-bit quantization.
- Curious how authors addressed all major quantization challenges using rotational methods.
- Achieved only 0.47 perplexity loss and retained 99% accuracy on LLAMA2-70B.

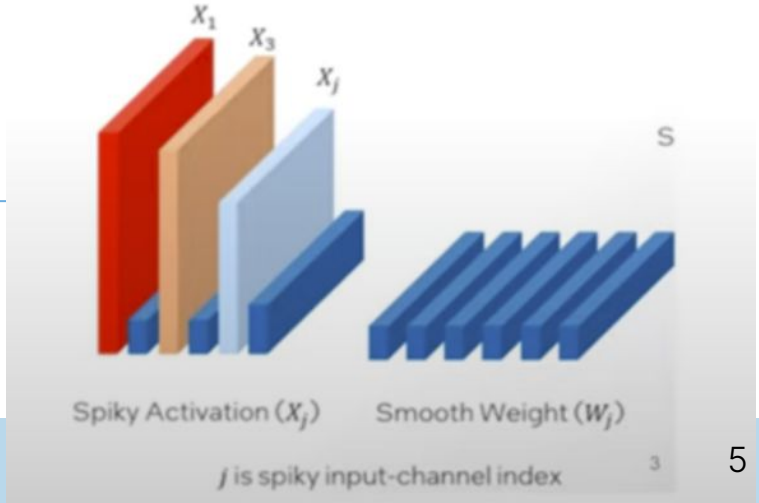


Purpose of QuaRot

What were the tasks?

- Quantize all parts of LLMs (weights, activations, and KV cache) to 4 bits for efficient inference and reduced memory usage.
- Remove outlier features in activations and cached values, which make low-bit quantization difficult.
- Apply a unified approach (randomized Hadamard rotations) to enable end-to-end quantization without significant accuracy loss.





Purpose of QuaRot

What were the paper's contributions?

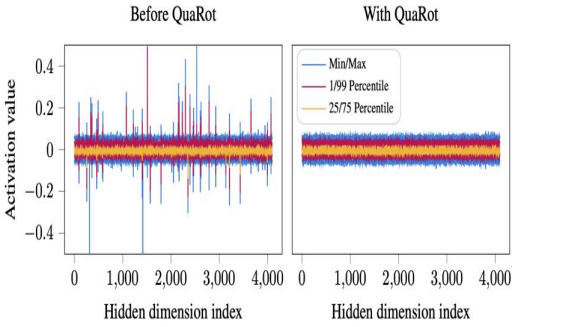


Figure 1: The distributions of activations at the input to the FFN block in LLAMA2-7B model, in the tenth layer. Left: using the default configuration as downloaded from Hugging Face. Right: after processing using QuaRot. The processed distribution has no outliers, leading to superior quantization.

- Rotates weights and activations with randomized Hadamard matrices, spreading outliers and making quantization easier.
- Fuses rotations into model weights using computational invariance, so model outputs remain unchanged.
- Applies online Hadamard transforms to the attention KV cache, enabling 4-bit quantization of all model parts.
- All quantization and modifications are done in preprocessing, so inference is fast and requires no extra computation.

Paper Extensions

We extended the results from the paper in the following ways:

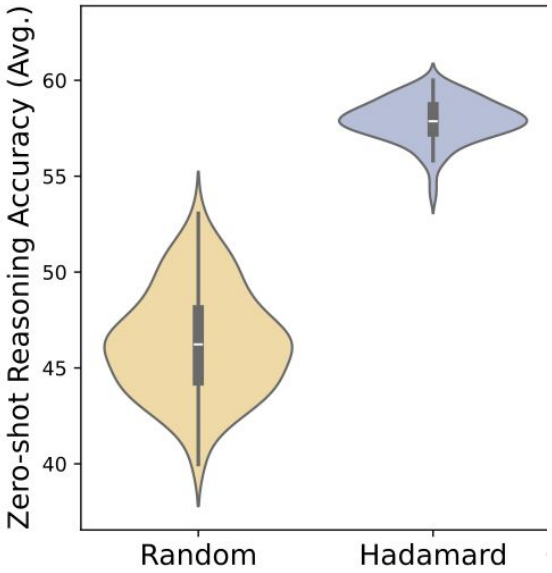
Alternative Rotation Matrix Optimization

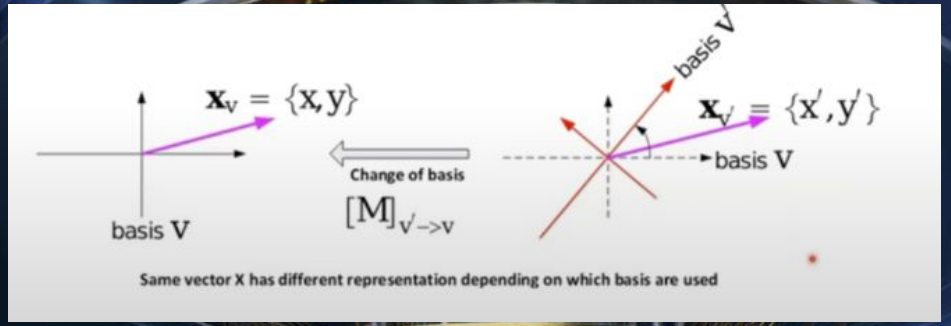
Skew Based Quantization mode selection

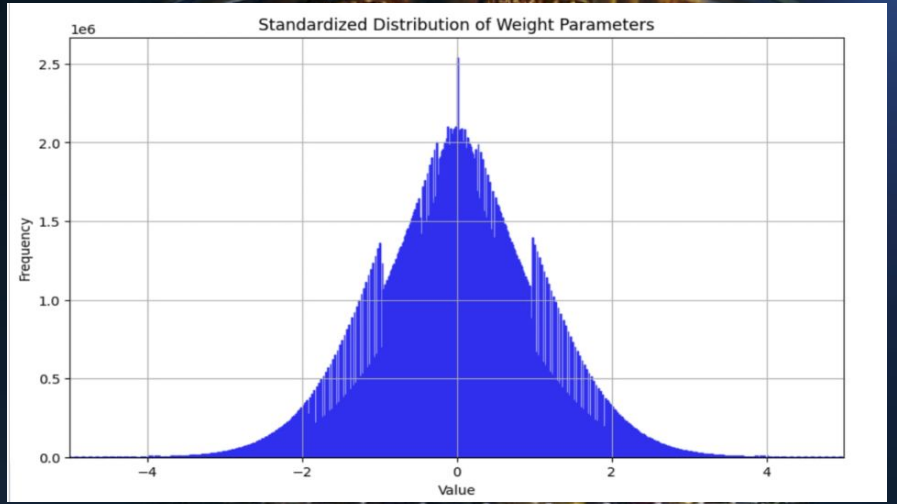
Integrated datasets

Paper Extensions

Extensions 1
Matrix Optimization







Extensions 2 : Integrated Datasets

WIKITEXT2

collection of over 100 million tokens extracted from verified articles on Wikipedia

C4

C4 is a colossal, cleaned version of Common Crawl's web crawl corpus.

PTB

Diagnostic ECG Database obtained using non-commercial, PTB - prototype recorder

WIKITEXT103

collection of over 100 million tokens extracted from verified articles on Wikipedia

PILE

consists of 22 smaller, high-quality datasets combined together

Paper Extensions

Extensions 3 : Skew Based Quantization mode selection

How can we **decrease performance loss**?

Symmetric Quantization:

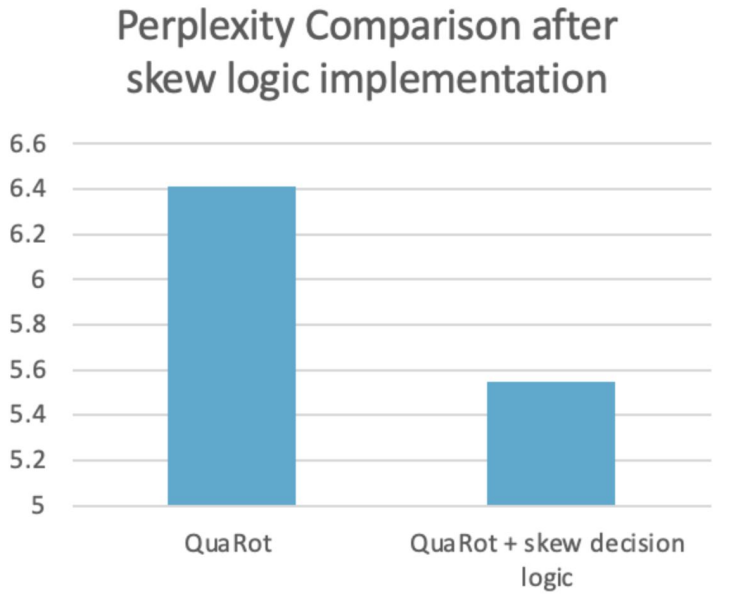
If distribution skewed, this will waste range reserving space

Better when mean centered around 0

Asymmetric Quantization:

Shift range to match actual data

Thresholds: skewness, value of the mean

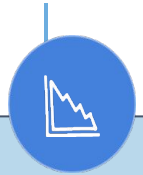


Challenges

Challenges in reproducing the results

System and environment requirements were not up to date

Hardware limitations



No documentation and limited comments

Hugging Face access (takes a while)

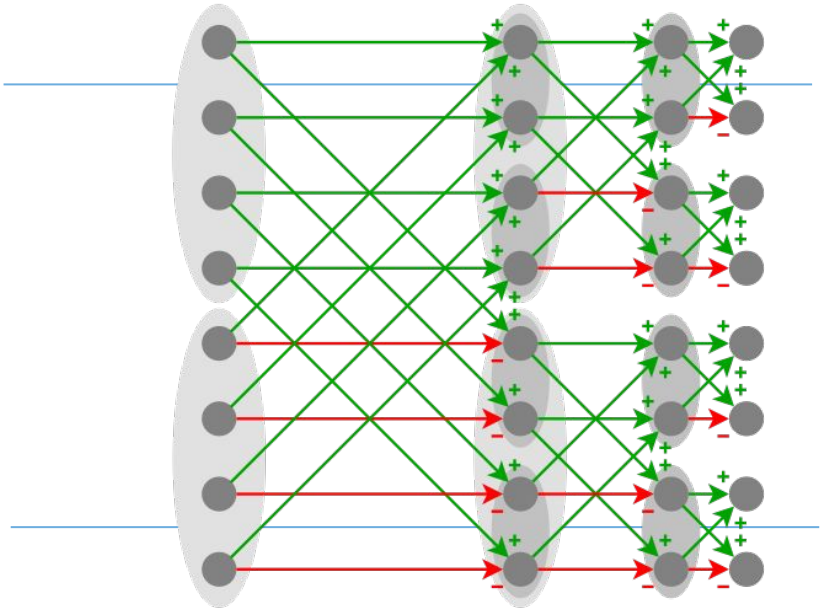
Unable to reproduce the results for large models

Conclusion

Future developments:

Combine with spinquant

Further testing in different resources



References

Research Paper Chosen

- QuaRot: Outlier-Free 4-Bit Inference in Rotated LLMs (https://arxiv.org/pdf/2404.00456)

Research Paper code link

- https://github.com/spcl/QuaRot

Other Sources

- https://arxiv.org/abs/2405.16406
- https://huggingface.co/meta-llama/Llama-2-7b
- Wikitext2, C4, PTB, Wikitext103 and Pile datasets
- Chat GPT used in debugging

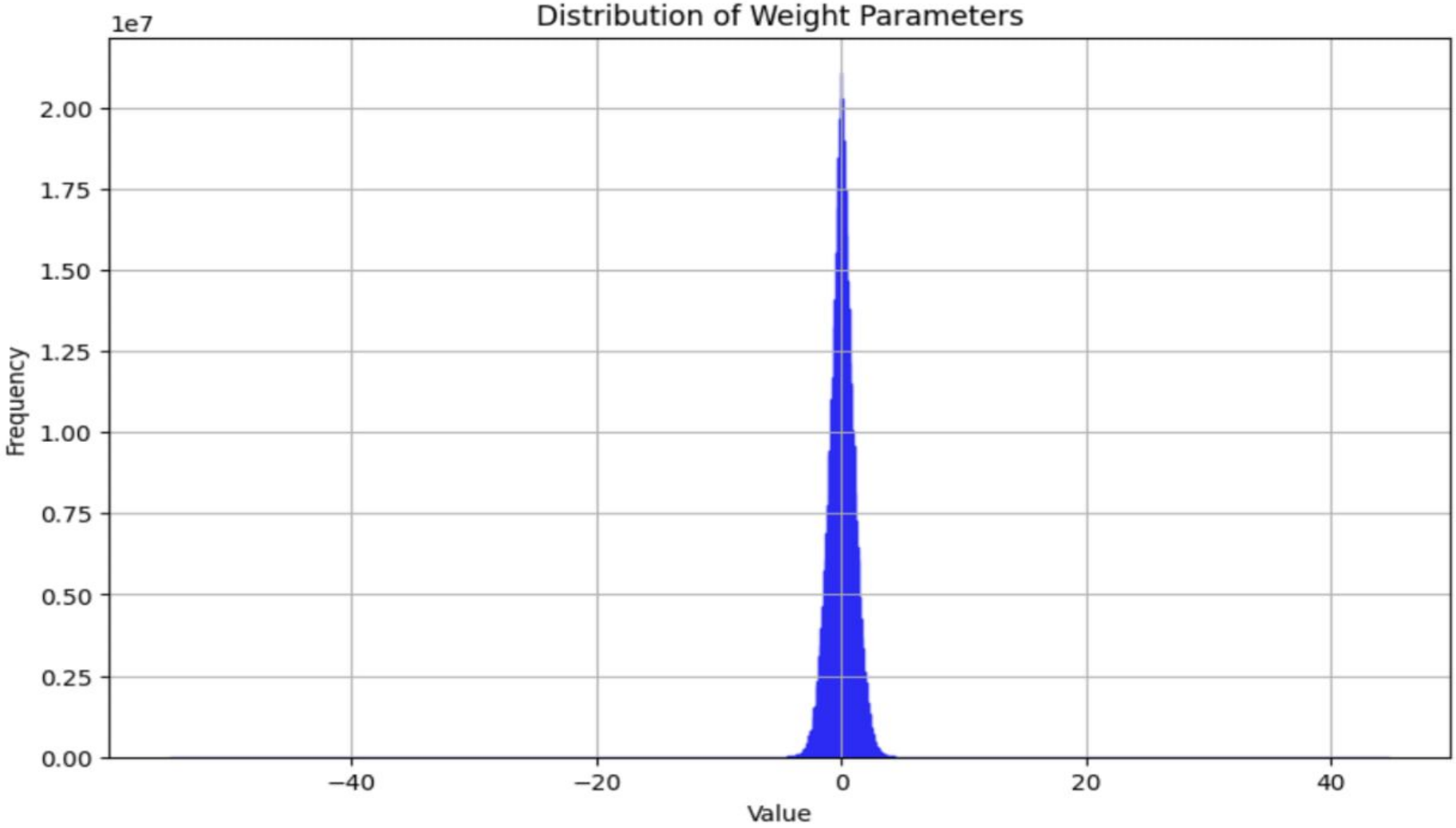
Background- QuaRot

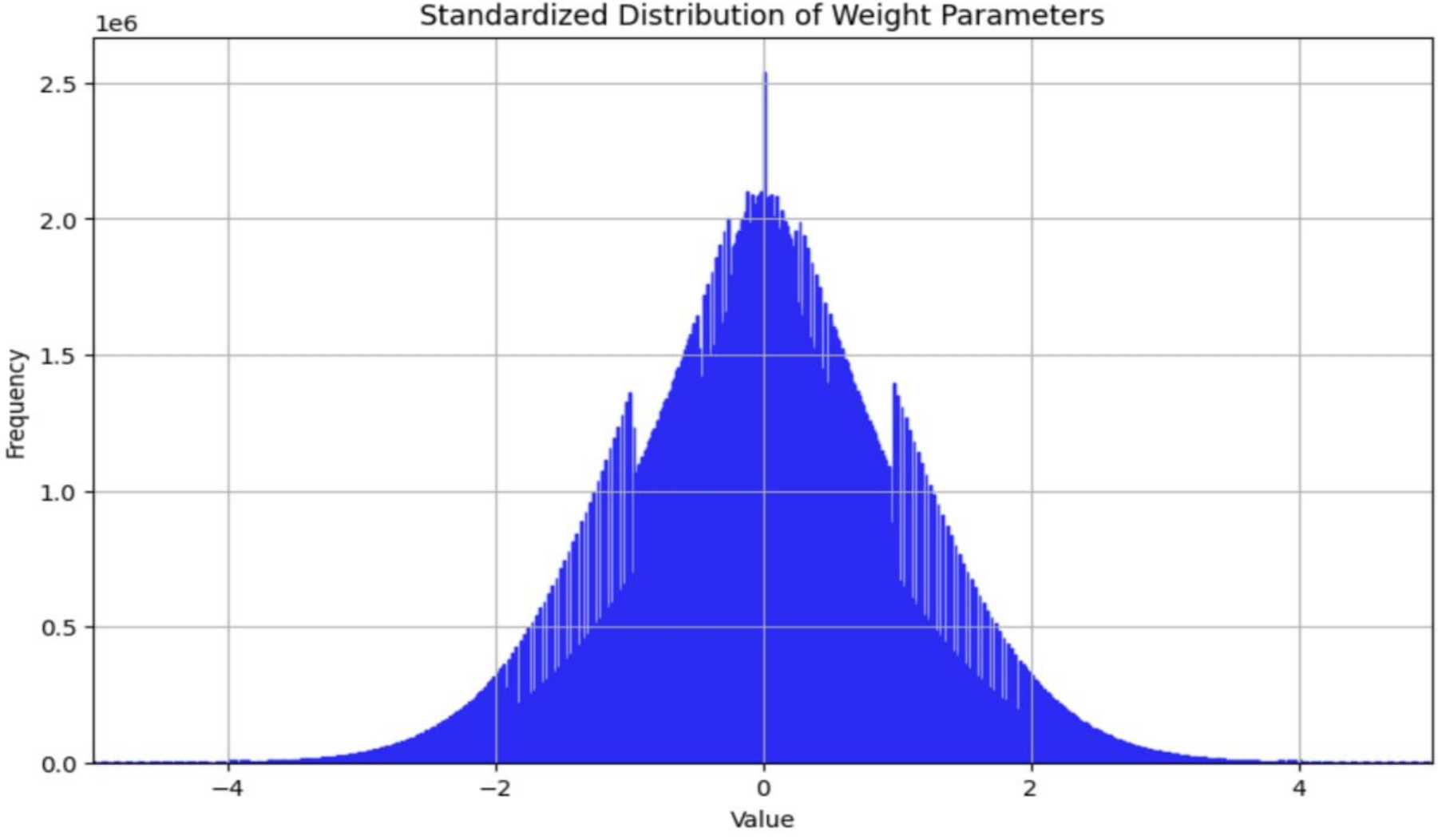
What Is the Hadamard Transform?

- The Hadamard transform is an orthogonal operation that rotates data into a new basis using only reflections and rotations.
- Unlike the Fourier transform, which uses sine waves and complex numbers, Hadamard uses simple square wave basis functions consisting of only +1 and -1.
- It is extremely efficient: requiring only additions and subtractions, with no complex math when working with real inputs.
- This makes it ideal for fast, hardware-friendly transformations in quantization workflows.

Fast Walsh–Hadamard transform

The fast Walsh–Hadamard transform applied to a vector of length 8


Distribution of Weight Parameters
1e7
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Frequency
−40
−20
0
20
40
Value


Standardized Distribution of Weight Parameters
1e6
2.5
2.0
1.5
1.0
0.5
0.0
Frequency
−4
−2
0
2
4
Value