MATFYZ
CONNECTIONS

Contribution ID: **128** Type: **Zamestnanci informatika**

Addition is almost all you need: Compressing neural networks with double binary factorization

Wednesday, November 26, 2025 2:45 PM (30 minutes)

Binary quantization approaches, which replace weight matrices with binary matrices and substitute costly multiplications with cheaper additions, offer a computationally efficient approach to address the increasing computational and storage requirements of Large Language Models (LLMs). However, the severe quantization constraint (± 1) can lead to significant accuracy degradation.

In this paper, we propose Double Binary Factorization (DBF), a novel method that factorizes dense weight matrices into products of two binary (sign) matrices, each accompanied by scaling vectors. DBF preserves the efficiency advantages of binary representations while achieving compression rates that are competitive with or superior to state-of-the-art methods.

Specifically, in a 1-bit per weight range, DBF is better than existing binarization approaches. In a 2-bit per weight range, DBF is competitive with the best quantization methods like QuIP\# and QTIP. Unlike most existing compression techniques, which offer limited compression level choices, DBF allows fine-grained control over compression ratios by adjusting the factorization's intermediate dimension. Based on this advantage, we further introduce an algorithm for estimating non-uniform layer-wise compression ratios for DBF, based on previously developed channel pruning criteria.

Pracovisko fakulty (katedra)/ Department of Faculty

Katedra Aplikovanej Informatiky

Tlač postru/ Print poster

Budem požadovať tlač /I hereby required to print the poster in faculty

Author: BOZA, Vladimir (Comenius University)

Co-author: MACKO, Vladimir

Presenter: BOZA, Vladimir (Comenius University)

Session Classification: Významné výsledky z matematiky, fyziky, informatiky a didaktiky

Track Classification: Významné výsledky z matematiky, fyziky, informatiky a didaktiky