Contribution ID: **66** Type: **Študenti informatika**

MACKO: Sparse Matrix-Vector Multiplication for Low Sparsity

Wednesday, November 26, 2025 3:28 PM (1 minute)

Sparse Matrix-Vector Multiplication (SpMV) is a fundamental operation in the inference of sparse Large Language Models (LLMs). Because existing SpMV methods perform poorly under the low and unstructured sparsity (30–90%) commonly observed in pruned LLMs, unstructured pruning struggled to deliver real memory reduction or speedup. We propose **MACKO-SpMV**, a GPU-optimized format and kernel co-designed to reduce storage overhead while preserving compatibility with the GPU's execution model. This enables efficient SpMV for unstructured sparsity without specialized hardware units (e.g., tensor cores) or format-specific precomputation.

Empirical results show that at sparsity 50%, MACKO is the first approach with significant 1.5× memory reduction and 1.2
$textup--1.5\times$ speedup over dense representation. Speedups over other SpMV baselines: 2.8
$textup--13.0\times$ over cuSPARSE, 1.9
$textup--2.6\times$ over Sputnik, and 2.2
$textup--2.5\times$ over DASP. Applied to Llama2-7B pruned with Wanda to sparsity 50%, it delivers 1.5× memory reduction and 1.5× faster inference at fp16 precision. Thanks to MACKO, unstructured pruning at 50% sparsity is now justified for practical deployment in real-world LLM workloads.

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

Authors: BOZA, Vladimir (Comenius University); MACKO, Vladimir

Session Classification: Poster session + káva: prezentácie vedeckých výsledkov FMFI UK Zamestnanci Informatika

Track Classification: Poster session + káva: prezentácie vedeckých výsledkov FMFI UK Zamestnanci: Poster session + káva: prezentácie vedeckých výsledkov FMFI UK Zamestnanci Informatika