Per-Axis Weight Deltas for Frequent Model Updates

Kuyumdzhiev, Stefan (School: High School of Mathematics and Natural Sciences "Vasil Drumev")

Using numerous task-specialized large language models in deployment is often constrained by the storage footprint of fine-tuned checkpoints and significant cold-start latency during loading. Although fine-tuned weights are close to their base model, they are typically stored in full precision, resulting in redundant storage and inefficient loading. This work introduces a compact 1-bit delta scheme for efficient representation and deployment of fine-tuned LLM variants. Instead of storing full residual weights, we use only the sign of each weight difference and learn lightweight per-axis scaling vector from a small calibration set. This design uses the advantages of compression benefits of 1-bit deltas, substantially improving reconstruction fidelity compared to scalar scaling approaches. The resulting components are several times smaller than full FP16 checkpoints. The method is fully compatible with existing fine-tuning pipelines, requires minimal calibration data, and enables scalable multi-variant model serving with reduced storage and loading time. Experimental results demonstrate improved reconstruction quality over previous quantized methods with negligible runtime overhead. The approach can be integrated in standard fine-tuning pipelines and requires only a lightweight calibration set of 150 examples. It supports efficient multi-variant model serving while substantially reducing storage requirements and accelerating model loading. Importantly, the experimental results shows that despite the agrasive 1-bit compression, it has better accuracy compared to the original version. Our implementation and experiments are publicly available.

Awards Won: