A Preliminary Study

We show the accuracy-rank trade-offs on MRPC, RTE, and CoLA in Figure 1 (CoLA is additionally included compared to the main body of the paper). The observation on CoLA is similar to MRPC/RTE: first-order unstructured pruning can extract subnetworks that are most accurate while having the lowest average matrix rank, which lays the crucial foundation of later factorization.

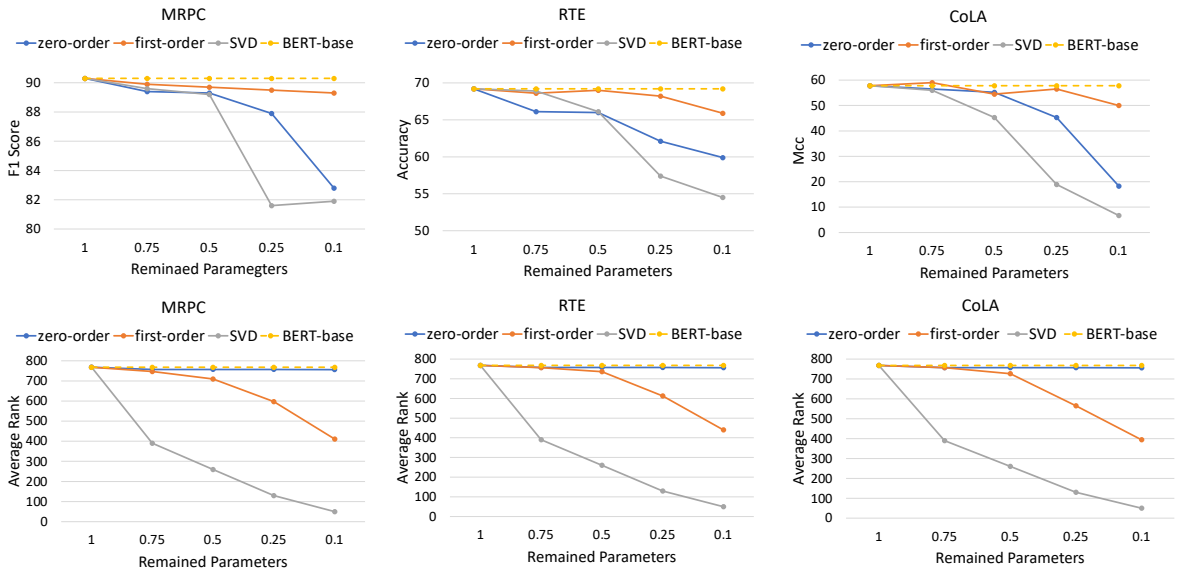

Figure 1: Task accuracy (top half) and average matrix rank (bottom half) v.s. percentage of original parameters retained. The dashed line indicates the performance/rank upper bound by fine-tuning the full-scale BERT-base model.