Sparse benchmarks

These sets of benchmarks are for the sparse matrix functionality using a popular real dataset collection called the Deep Learning Matrix Collection (DLMC), which were used in recent studies [1, 2].

Performance benchmarks scripts for matrix-matrix and matrix-vector ops (dense-sparse, sparse-sparse, and compare to dense-dense) are implemented here.

- `matmul_bench.py` with `--operation sparse@sparse|sparse@dense` is for Sparse matrix-matrix multiplication (SPMM) performance test. It can run in forward and backward mode with `--backward_test`, on CPU or CUDA with `--with_cuda`, using different datasets from the dataset collection DLMC. For more details see `test.sh` file.
- `matmul_bench.py` with `--operation sparse@vector` is for Sparse matrix-vector multiplication (SPMV) performance test.

References:

1. Trevor Gale, Matei Zaharia, Cliff Young, Erich Elsen. Sparse GPU Kernels for Deep Learning. Proceedings of the International Conference for High Performance Computing, 2020. https://github.com/google-research/google-research/tree/master/sgk
2. Trevor Gale, Erich Elsen, Sara Hooker. The State of Sparsity in Deep Neural Networks. https://github.com/google-research/google-research/tree/master/state_of_sparsity