Lookups are not (yet) all you need for deep learning inference

Calvin McCarter‡
calmcc@amazon.com

Nicholas Dronen‡
ndronen@amazon.com

‡Work done prior to employment at Amazon.

Background

MADDNESS (Blalock & Guttag, 2021) proposed matrix multiplication without multiplying.

- Accelerated $\boldsymbol{AB}$, given training set $\tilde{\boldsymbol{A}}$.
- Replaced dot-products with comparison-based hashing.
- Decomposed full dot-products into partial dot-products, each with its own codebook (Jegou et al., 2010).

Blalock, D., & Guttag, J. (2021). Multiplying matrices without multiplying. ICML.
Jegou, H., Douze, M., & Schmid, C. (2010). Product quantization for nearest neighbor search. IEEE TPAMI.

Motivation

Apply to deep learning inference:

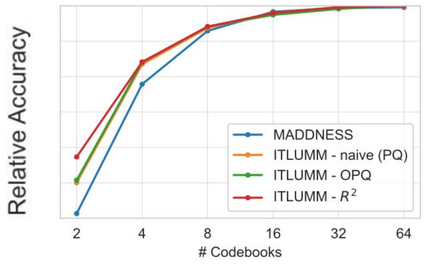
- Accelerate $\sigma(\boldsymbol{AB})$ given training set $\tilde{\boldsymbol{A}}$ and weights $\boldsymbol{B}$.
- Accelerate composition of layers $\sigma(\sigma(\sigma(\boldsymbol{AB_1})\boldsymbol{B_2})\boldsymbol{B_3})$

Our Approach

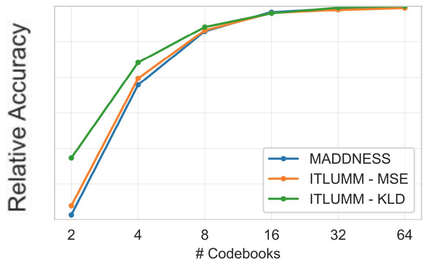
- Partition to partial products via
  - Learning rotation matrix and projecting onto space of permutation matrices.
  - Hierarchical clustering dimensions according to R^2.
- Utilize knowledge of fixed weights $\boldsymbol{B}$ and activation/loss σ.
- Replace layers incrementally with fine-tuning.

For a single layer, our method improves accuracy over MADNESS without any inference-time penalty.

Benefit of Intelligent Partitioning



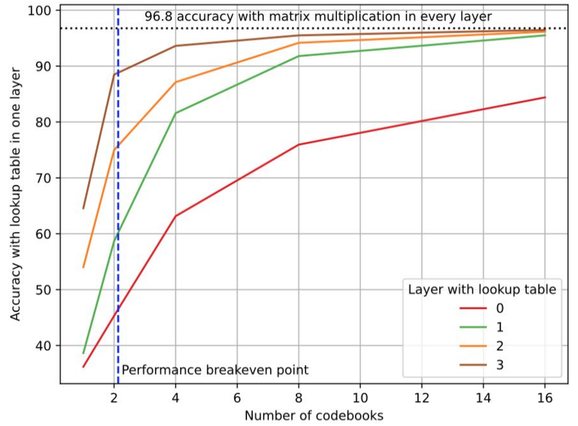
Benefit of Utilizing Weights, Activation, & Loss



CIFAR100 (reproducing *MADDNESS* result)

But a network loses too much accuracy when lookup tables are faster than matrix multiplies.

Effect of Individual Layer Replacement on Accuracy



Effect of All-Layers Replacement on Accuracy

# of Codebooks	1	2	4	8	16
Accuracy	36.1	36.3	52.0	70.5	84.9
Faster?	Yes	Yes	No	No	No

4-layer MLP for MNIST