Multiplying Matrices Without Multiplying

Davis Blalock [1 2] **John Guttag** [2]

Abstract

Multiplying matrices is among the most fundamental and compute-intensive operations in machine learning. Consequently, there has been significant work on efficiently approximating matrix multiplies. We introduce a learning-based algorithm for this task that greatly outperforms existing methods. Experiments using hundreds of matrices from diverse domains show that it often runs $100\times$ faster than exact matrix products and $10\times$ faster than current approximate methods. In the common case that one matrix is known ahead of time, our method also has the interesting property that it requires zero multiply-adds. These results suggest that a mixture of hashing, averaging, and byte shuffling–the core operations of our method–could be a more promising building block for machine learning than the sparsified, factorized, and/or scalar quantized matrix products that have recently been the focus of substantial research and hardware investment.

1. Introduction

Matrix multiplication is among the most fundamental subroutines used in machine learning and scientific computing. As a result, there has been a great deal of work on implementing high-speed matrix multiplication libraries (Paszke et al., 2017; Guennebaud et al., 2010; Abadi et al., 2016), designing custom hardware to accelerate multiplication of certain classes of matrices (Han et al., 2016; Chen et al., 2016; Parashar et al., 2017; Jouppi et al., 2017), speeding up distributed matrix multiplication (Yu et al., 2017; Dutta et al., 2016; Yu et al., 2020; Irony et al., 2004), and designing efficient Approximate Matrix Multiplication (AMM) algorithms under various assumptions.

We focus on the AMM task under the assumptions that the matrices are tall, relatively dense, and resident in a single machine's memory. In this setting, the primary challenge is minimizing the amount of compute time required to approximate linear operations with a given level of fidelity. This setting arises naturally in machine learning and data mining when one has a data matrix $\boldsymbol{A}$ whose rows are samples and a linear operator $\boldsymbol{B}$ one wishes to apply to these samples. $\boldsymbol{B}$ could be a linear classifier, linear regressor, or an embedding matrix, among other possibilities.

As a concrete example, consider the task of approximating a softmax classifier trained to predict image labels given embeddings derived from a neural network. Here, the rows of $\boldsymbol{A}$ are the embeddings for each image, and the columns of $\boldsymbol{B}$ are the weight vectors for each class. Classification is performed by computing the product $\boldsymbol{AB}$ and taking the argmax within each row of the result. In Figure 1, we see the results of approximating $\boldsymbol{AB}$ using our method and its best-performing rivals (Dasgupta et al., 2010; Mairal et al., 2009) on the CIFAR-10 and CIFAR-100 datasets.

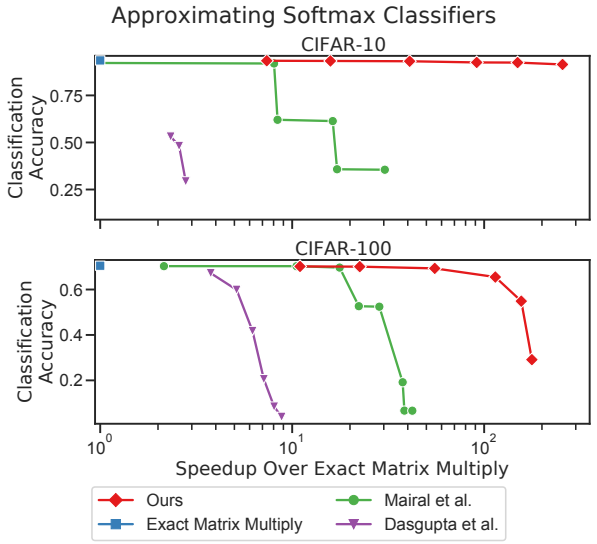

Figure 1: Our method achieves a dramatically better speed-accuracy tradeoff than the best existing methods when approximating two linear classifiers.

Our method represents a significant methodological departure from most traditional approaches to this problem. Traditional AMM methods construct matrices $\boldsymbol{V}_A, \boldsymbol{V}_B \in \mathbb{R}^{D\times d}, d \ll D$ such that

$$\boldsymbol{AB} \approx (\boldsymbol{AV}_A)(\boldsymbol{V}_B^\top \boldsymbol{B}). \tag{1}$$

Often, $\boldsymbol{V}_A$ and $\boldsymbol{V}_B$ are sparse, embody some sort of sampling scheme, or have other structure such that these pro-

[1]MosaicML, San Francisco, CA, USA [2]MIT CSAIL, Cambridge, MA, USA. Correspondence to: Davis Blalock <davis@mosaicml.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

jection operations are faster than a dense matrix multiply. In short, these methods use linear functions to preprocess $\boldsymbol{A}$ and $\boldsymbol{B}$ and reduce the problem to exact matrix multiplication in a lower-dimensional space.

Our proposed method, MADDNESS[1], instead employs a *nonlinear* preprocessing function and reduces the problem to table lookups. Moreover, in the case that $\boldsymbol{B}$ is known ahead of time—which happens when applying a trained linear model to new data, among other situations—MADDNESS does not require any multiply-add operations. Our method is most closely related to vector quantization methods used for similarity search (e.g., (Blalock & Guttag, 2017; André et al., 2017; 2019; Jegou et al., 2011; Ge et al., 2014)). However, instead of using an expensive quantization function that requires many multiply-adds, we introduce a family of quantization functions that require no multiply-adds.

Our contributions can be summarized as follows:

- An efficient family of learned vector quantization functions that can encode over 100GB of data per second in a single CPU thread.
- A high-speed summation algorithm for low-bitwidth integers that avoids upcasting, saturation, and overflow.
- An algorithm based on these functions for approximate matrix multiplication. Experiments across hundreds of diverse matrices demonstrate that this algorithm significantly outperforms existing alternatives. It also features theoretical quality guarantees.

1.1. Problem Formulation

Let $\boldsymbol{A} \in \mathbb{R}^{N \times D}$ and $\boldsymbol{B} \in \mathbb{R}^{D \times M}$ be two matrices, with $N \gg D \geq M$. Given a computation time budget τ, our task is to construct three functions $g(\cdot)$, $h(\cdot)$, and $f(\cdot, \cdot)$, along with constants α and $\boldsymbol{\beta}$, such that

$$\|\alpha f(g(\boldsymbol{A}), h(\boldsymbol{B})) + \boldsymbol{\beta} - \boldsymbol{A}\boldsymbol{B}\|_F < \varepsilon(\tau)\|\boldsymbol{A}\boldsymbol{B}\|_F \quad (2)$$

for as small an error $\varepsilon(\tau)$ possible. The constants α and $\boldsymbol{\beta}$ are separated from $f(\cdot, \cdot)$ so that $f(\cdot, \cdot)$ can produce low-bitwidth outputs (e.g., in the range $[0, 255]$) even when the entries of $\boldsymbol{A}\boldsymbol{B}$ do not fall in this range.

We assume the existence of a training set $\tilde{\boldsymbol{A}}$, whose rows are drawn from the same distribution as the rows of $\boldsymbol{A}$. This is a natural assumption in the case that rows of $\boldsymbol{A}$ represent examples in training data, or structured subsets thereof (such as patches of images).

2. Related Work

Because our work draws on ideas from randomized algorithms, approximate matrix multiplication, vector quantization, and other fields, the body of work related to our own is vast. Here, we provide only a high-level overview, and refer the interested reader to (Wang et al., 2016a; 2014a; Desai et al., 2016) for more detailed surveys. We also defer discussion of related vector quantization methods to the following sections.

2.1. Linear Approximation

Most AMM methods work by projecting $\boldsymbol{A}$ and $\boldsymbol{B}$ into lower-dimensional spaces and then performing an exact matrix multiply. One simple option for choosing the projection matrices is to use matrix sketching algorithms. The most prominent deterministic matrix sketching methods are the Frequent Directions algorithm (Liberty, 2012; Ghashami et al., 2016) and its many variations (Teng & Chu, 2019; Francis & Raimond, 2018b; Ye et al., 2016; Huang, 2019; Luo et al., 2019; Francis & Raimond, 2018a). There are also many randomized sketching methods (Sarlos, 2006; Kyrillidis et al., 2014; Pagh, 2013; Dasgupta et al., 2010; Nelson & Nguyên, 2013) and sampling methods (Drineas et al., 2006b;c).

A weakness of matrix sketching methods in the context of matrix multiplication is that they consider each matrix in isolation. To exploit information about both matrices simultaneously, Drineas et al. (2006a) sample columns of $\boldsymbol{A}$ and rows of $\boldsymbol{B}$ according to a sampling distribution dependent upon both matrices. Later work by Manne & Pal (2014) reduces approximation of the matrices to an optimization problem, which is solved by steepest descent. Mroueh et al. (2016), Ye et al. (2016), and Francis & Raimond (2018a) introduce variations of the Frequent Directions algorithm that take into account both matrices.

All of the above methods differ from our own not merely in specifics, but also in problem formulation. These methods all assume that there is no training set $\tilde{\boldsymbol{A}}$ and nearly all focus on large matrices, where provably reduced asymptotic complexity for a given level of error is the goal.

2.2. Hashing to Avoid Linear Operations

In the neural network acceleration literature, there have been several efforts to accelerate dense linear layers using some form of hashing (Spring & Shrivastava, 2017; Chen et al., 2019; Bakhtiary et al., 2015; Dean et al., 2013; Chen et al., 2015). These methods differ from our own in the hash functions chosen, in not exploiting a training set, and in the overall goal of the algorithm. While we seek to approximate the entire output matrix, these methods seek to either sample outputs (Spring & Shrivastava, 2017; Chen et al., 2019), approximate only the largest outputs (Bakhtiary et al., 2015; Dean et al., 2013), or implement a fixed, sparse linear operator (Chen et al., 2015).

[1] Multiply-ADDitioN-lESS

3. Background - Product Quantization

To lay the groundwork for our own method, we begin by reviewing Product Quantization (PQ) (Jegou et al., 2011). PQ is a classic vector quantization algorithm for approximating inner products and Euclidean distances and serves as the basis for nearly all vector quantization methods similar to our own.

The basic intuition behind PQ is that $\boldsymbol{a}^\top \boldsymbol{b} \approx \hat{\boldsymbol{a}}^\top \boldsymbol{b}$, where $\|\hat{\boldsymbol{a}} - \boldsymbol{a}\|$ is small but $\hat{\boldsymbol{a}}$ has special structure allowing the product to be computed quickly. This structure consists of $\hat{\boldsymbol{a}}$ being formed by concatenating learned prototypes in disjoint subspaces; one obtains a speedup by precomputing the dot products between $\boldsymbol{b}$ and the prototypes once, and then reusing these values across many $\boldsymbol{a}$ vectors. The $\boldsymbol{a}$ vectors here are the (transposed) rows of $\boldsymbol{A}$ and the $\boldsymbol{b}$ vectors are the columns of $\boldsymbol{B}$.

In somewhat more detail, PQ consists of the following:

1. **Prototype Learning** - In an initial, offline training phase, cluster the rows of $\boldsymbol{A}$ (or a training set $\tilde{\boldsymbol{A}}$) using K-means to create prototypes. A separate K-means is run in each of C disjoint subspaces to produce C sets of K prototypes.
2. **Encoding Function**, $g(\boldsymbol{a})$ - Determine the most similar prototype to $\boldsymbol{a}$ in each subspace. Store these assignments as integer indices using $C \log_2(K)$ bits.
3. **Table Construction**, $h(\boldsymbol{B})$ - Precompute the dot products between $\boldsymbol{b}$ and each prototype in each subspace. Store these partial dot products in C lookup tables of size K.
4. **Aggregation**, $f(\cdot, \cdot)$ - Use the indices and tables to *lookup* the estimated partial $\boldsymbol{a}^\top \boldsymbol{b}$ in each subspace, then sum the results across all C subspaces.

PQ is depicted for a single pair of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ in Figure 2. We elaborate upon each of these steps below.

Prototype Learning: Let $\tilde{A} \in \mathbb{R}^{N \times D}$ be a training set, K be a number of prototypes per subspace, C be a number of subspaces, and $\{\mathcal{J}^{(c)}\}_{c=1}^{C}$ be the mutually exclusive and collectively exhaustive sets of indices associated with each subspace. The training-time task of PQ is to learn C sets of prototypes $\boldsymbol{P}^{(c)} \in \mathbb{R}^{K \times |\mathcal{J}^{(c)}|}$ and assignments $\boldsymbol{z}^{(c)} \in \mathbb{R}^N$ such that:

$$\sum_{i=1}^{N} \sum_{c=1}^{C} \sum_{j \in \mathcal{J}^{(c)}} \left(\tilde{\boldsymbol{A}}_{ij} - \boldsymbol{P}^{(c)}_{\boldsymbol{z}^{(c)}, j} \right)^2 \tag{3}$$

is minimized. It does this by running K-means separately in each subspace $\mathcal{J}^{(c)}$ and using the resulting centroids and assignments to populate $\boldsymbol{P}^{(c)}$ and $\boldsymbol{z}^{(c)}$.

Encoding Function, $g(\boldsymbol{A})$: Given the learned prototypes, PQ replaces each row $\boldsymbol{a}$ of $\boldsymbol{A}$ with the concatenation

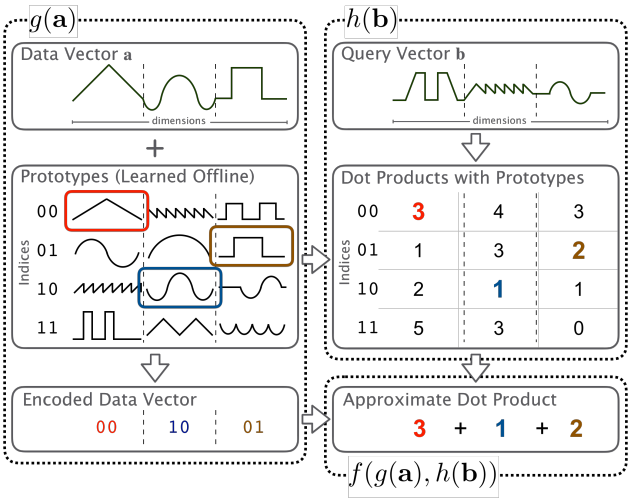

Figure 2: Product Quantization. The $g(\cdot)$ function returns the index of the most similar prototype to the data vector a in each subspace. The $h(\cdot)$ function computes a lookup table of dot products between the query vector b and each protoype in each subspace. The aggregation function $f(\cdot, \cdot)$ sums the table entries corresponding to each index.

of its C K-means centroid assignments in each of the C subspaces. Formally:

$$g^{(c)}(\boldsymbol{a}) \triangleq \operatorname*{argmin}_k \sum_{j \in \mathcal{J}^{(c)}} \left(\boldsymbol{a}_j - \boldsymbol{P}^{(c)}_{k,j} \right)^2. \tag{4}$$

We will refer to the resulting sequence of indices as the *encoding* of $\boldsymbol{a}$ and the set of K centroids as a *codebook*. For convenience, we will also refer to the vector $\boldsymbol{a}^{(c)} \triangleq \langle a_j \rangle, j \in \mathcal{J}^{(c)}$ as the *subvector* of $\boldsymbol{a}$ in subspace c.

Table Construction, $h(\boldsymbol{B})$: Using these same prototypes, PQ constructs a lookup table $h^{(c)}(\boldsymbol{b}) \in \mathbb{R}^K$ in each of the C subspaces for each column $\boldsymbol{b}$ of $\boldsymbol{B}$, where

$$h^{(c)}(\boldsymbol{b})_k \triangleq \sum_{j \in \mathcal{J}^{(c)}} \boldsymbol{b}_j \boldsymbol{P}^{(c)}_{k,j}. \tag{5}$$

Existing work has shown that setting $K = 16$ and quantizing the lookup tables to 8 bits can offer enormous speedups compared to larger K and/or floating-point tables (Blalock & Guttag, 2017; André et al., 2017; 2019). This is because 16 1-byte entries can be stored in a SIMD register, allowing 16 or more table lookups to be performed in parallel using a byte shuffle instruction. Since the table entries naturally occupy more than 8 bits even for 8-bit data, some means of quantizing these entries is necessary. This can easily be done by subtracting off the minimum entry in each table and linearly rescaling such that the maximum entry in any table is at most 255. Ignoring rounding error, this affine transform is invertible, and is reflected by the constants α and $\boldsymbol{\beta}$ in equation 2. See Appendix A for further details.

Aggregation, $f(\cdot,\cdot)$: Given the encoding of $\boldsymbol{a}$ and the lookup tables for $\boldsymbol{b}$, the product can be approximated as

$$\boldsymbol{a}^\top \boldsymbol{b} = \sum_{c=1}^{C} \boldsymbol{a}^{(c)\top} \boldsymbol{b}^{(c)} \approx \sum_{c=1}^{C} h^{(c)}(\boldsymbol{b})_k,\ k = g^{(c)}(\boldsymbol{a}). \quad (6)$$

4. Our Method

Product Quantization and its variants yield a large speedup with $N, M \gg D$. However, we require an algorithm that only needs $N \gg M, D$, a more relaxed scenario common when using linear models and transforms. In this setting, the preprocessing time $g(\boldsymbol{A})$ can be significant, since ND may be similar to (or even larger than) NM.

To address this case, we introduce a new $g(\boldsymbol{A})$ function that yields large speedups even on much smaller matrices. The main idea behind our function is to determine the "most similar" prototype through locality-sensitive hashing (Indyk & Motwani, 1998); i.e., rather than compute the Euclidean distance between a subvector $\boldsymbol{a}^{(c)}$ and each prototype, we hash $\boldsymbol{a}^{(c)}$ to one of K buckets where similar subvectors tend to hash to the same bucket. The prototypes are set to the means of the subvectors hashing to each bucket.

4.1. Hash Function Family, $g(\cdot)$

Because we seek to exploit a training set while also doing far less work than even a single linear transform, we found that existing hash functions did not meet our requirements. Consequently, we designed our own family of trainable hash functions. The family of hash functions we choose is balanced binary regression trees, with each leaf of the tree acting as one hash bucket. The leaf for a vector $\boldsymbol{x}$ is chosen by traversing the tree from the root and moving to the left child if the value x_j at some index j is below a node-specific threshold v, and to the right child otherwise. To enable the use of SIMD instructions, the tree is limited to 16 leaves and all nodes at a given level of the tree are required to split on the same index j. The number 16 holds across many processor architectures, and we refer the reader to Appendix B for further vectorization details.

Formally, consider a set of four indices $j^1, \ldots, j^4$ and four arrays of split thresholds $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^4$, with v_t having length 2^{t-1}. A vector $\boldsymbol{x}$ is mapped to an index using Algorithm 1. This function is simple, only depends on a constant number of indices in the input vector, and can easily be vectorized provided that the matrix whose rows are being encoded is stored in column-major order.

4.2. Learning the Hash Function Parameters

The split indices $j^1, \ldots, j^4$ and split thresholds $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^4$ are optimized on the training matrix $\tilde{\boldsymbol{A}}$ using a greedy tree construction algorithm. To describe this algorithm, we introduce the notion of a *bucket* $\mathcal{B}_i^t$, which is the set of vectors mapped to node i in level t of the tree. The root of the tree is level 0 and $\mathcal{B}_1^0$ contains all the vectors. It will also be helpful to define the sum of squared errors (SSE) loss associated with a bucket, or a specific (index, bucket) pair:

$$\mathcal{L}(j, \mathcal{B}) \triangleq \sum_{\boldsymbol{x} \in \mathcal{B}} \left(x_j - \frac{1}{|\mathcal{B}|} \sum_{\boldsymbol{x}' \in \mathcal{B}} x'_j \right)^2$$
$$\mathcal{L}(\mathcal{B}) \triangleq \sum_j \mathcal{L}(j, \mathcal{B}). \quad (7)$$

Using this notation, it suffices to characterize the learning algorithm by describing the construction of level t of the tree given the buckets $\mathcal{B}_1^{t-1}, \ldots, \mathcal{B}_{2^{t-1}}^{t-1}$ from the previous level. This procedure is given in Algorithm 2.

In line 2, we select a fixed number of indices to evaluate. Several heuristics are possible, including evaluating all indices. We found that simply selecting the top n indices that contributed the most loss summed across all buckets was difficult to beat. In preliminary experiments, we found that using $n > 4$ indices offered no clear benefit (and even choosing $n = 1$ was nearly as good), so we fix $n = 4$.

In lines 4-15, we find the minimal loss obtainable by splitting all buckets along that index, but with bucket-specific cutoffs. This loss is minimal not in the sense that it leads to the globally optimal tree, but in that it minimizes the sum of the losses in the buckets produced in this iteration. To do this, we invoke the subroutine `optimal_split_threshold`, which takes in a bucket $\mathcal{B}$ and an index j and tests all possible thresholds to find one minimizing $\mathcal{L}(j, \mathcal{B})$. This can be done in time $O(|\mathcal{J}^{(c)}||\mathcal{B}| \log(|\mathcal{B}|))$. The pseudocode for this subroutine is given in Algorithms 3 and 4 in Appendix C.

Once a split index j and an array of split thresholds $\boldsymbol{v}$ are chosen, all that remains is to split the buckets to form the next level of the tree (lines 16-21). This entails forming two child buckets from each current bucket by grouping vectors whose jth entries are above or below the bucket's split threshold.

Algorithm 1 MADDNESSHASH

1: **Input:** vector $\boldsymbol{x}$, split indices $j^1, \ldots, j^4$, split thresholds $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^4$
2: $i \leftarrow 1$ // node index within level of tree
3: **for** $t \leftarrow 1$ **to** 4 **do**
4: $\quad v \leftarrow \boldsymbol{v}_i^t$ // lookup split threshold for node i at level t
5: $\quad b \leftarrow x_{j^t} \geq v \ ? \ 1 : 0$ // above split threshold?
6: $\quad i \leftarrow 2i - 1 + b$ // assign to left or right child
7: **end for**
8: **return** i

Algorithm 2 Adding The Next Level to the Hashing Tree

Input: buckets $\mathcal{B}_1^{t-1}, \ldots, \mathcal{B}_{2^{t-1}}^{t-1}$, training matrix $\tilde{\boldsymbol{A}}$
// greedily choose next split index and thresholds
$\hat{\mathcal{J}} \leftarrow$ `heuristic_select_idxs`$(\mathcal{B}_1^{t-1}, \ldots, \mathcal{B}_{2^{t-1}}^{t-1})$
$l^{min}, j^{min}, \boldsymbol{v}^{min} \leftarrow \infty$, NaN, NaN
for $j \in \hat{\mathcal{J}}$ **do**
  $l \leftarrow 0$ // initialize loss for this index to 0
  $\boldsymbol{v} \leftarrow [\,]$ // empty list of split thresholds
  for $i \leftarrow 1$ **to** 2^{t-1} **do**
    $v_i, l_i \leftarrow$ `optimal_split_threshold`$(j, \mathcal{B}_i^{t-1})$
    `append`$(\boldsymbol{v}, v_i)$ // append threshold for bucket i
    $l \leftarrow l + l_i$ // accumulate loss from bucket i
  end for
  if $l < l^{min}$ **then**
    $l^{min} \leftarrow l$, $j^{min} \leftarrow j$, $\boldsymbol{v}^{min} \leftarrow \boldsymbol{v}$ // new best split
  end if
end for
// create new buckets using chosen split
$\mathcal{B} \leftarrow [\,]$
for $i \leftarrow 1$ **to** 2^{t-1} **do**
  $\mathcal{B}_{below}, \mathcal{B}_{above} \leftarrow$ `apply_split`$(v_i^{min}, \mathcal{B}_i^{t-1})$
  `append`$(\mathcal{B}, \mathcal{B}_{below})$
  `append`$(\mathcal{B}, \mathcal{B}_{above})$
end for
return $\mathcal{B}, l^{min}, j^{min}, v^{min}$

4.3. Optimizing the Prototypes

At this point, we have a complete algorithm. We could simply drop our hash-based encoding function into PQ and approximate matrix products. However, we contribute two additional enhancements: a means of optimizing the prototypes with no runtime overhead, and a means of quickly summing low-bitwidth integers.

Several works propose prototype or table optimizations based on knowledge of $\boldsymbol{B}$ (Babenko et al., 2016; Wang et al., 2014b), and others optimize them at the expense of slowing down the function $g(\cdot)$ (Zhang et al., 2014; 2015). In contrast, we introduce a method that does not do either of these. The idea is to choose prototypes such that $\tilde{\boldsymbol{A}}$ can be reconstructed from its prototypes with as little squared error as possible—this improves results since less error means that less information about $\tilde{\boldsymbol{A}}$ is being lost.

Let $\boldsymbol{P} \in \mathbb{R}^{KC \times D}$ be a matrix whose diagonal blocks of size $K \times |\mathcal{J}^{(c)}|$ consist of the K learned prototypes in each subspace c. The training matrix $\tilde{\boldsymbol{A}}$ can be approximately reconstructed as $\tilde{\boldsymbol{A}} \approx \boldsymbol{G}\boldsymbol{P}$, where $\boldsymbol{G}$ serves to select the appropriate prototype in each subspace. Rows of $\boldsymbol{G}$ are formed by concatenating the one-hot encoded representations of each assignment for the corresponding row of $\tilde{\boldsymbol{A}}$. For example, if a row were assigned prototypes $\langle$`3 1 2`$\rangle$ with $K = 4$, $C = 3$, its row in $\boldsymbol{G}$ would be $\langle$`0010 1000 0100`$\rangle \in \mathbb{R}^{12}$. Our idea is to optimize $\boldsymbol{P}$ conditioned on $\boldsymbol{G}$ and $\tilde{\boldsymbol{A}}$. This is an ordinary least squares problem, and we solve it with ridge regression:

$$\boldsymbol{P} \triangleq (\boldsymbol{G}^\top \boldsymbol{G} + \lambda \boldsymbol{I})^{-1} \boldsymbol{G}^\top \tilde{\boldsymbol{A}}. \tag{8}$$

One could obtain better performance by cross-validating to find λ, but for simplicity, we fix $\lambda = 1$.

This procedure allows the prototypes to be nonzero outside of their original subspaces. Because of our hashing procedure, we avoid the dramatically increased overhead faced by other methods with non-orthogonal prototypes (c.f. (Babenko & Lempitsky, 2015; 2014; Zhang et al., 2014; Liu et al., 2016; Martinez et al., 2016; 2014)).

4.4. Fast 8-bit Aggregation, $f(\cdot, \cdot)$

Let $\boldsymbol{T} \in \mathbb{R}^{M \times C \times K}$ be the tensor of lookup tables for all M columns of $\boldsymbol{B}$. Given the encodings $\boldsymbol{G}$, the function $f(\cdot, \cdot)$ is defined as

$$f(g(\boldsymbol{A}), h(\boldsymbol{B}))_{n,m} \triangleq \sum_{c=1}^{C} \boldsymbol{T}_{m,c,k}, \; k = g^{(c)}(\boldsymbol{a}_n). \tag{9}$$

Because the entries of $\boldsymbol{T}$ are stored as 8-bit values, exact summation requires immediately upcasting each looked-up entry to 16 bits before performing any addition instructions (Blalock & Guttag, 2017). This not only imposes overhead directly, but also means that one must perform 16-bit additions, which have half the throughput of 8-bit additions.

We propose an alternative that sacrifices a small amount of accuracy for a significant increase in speed. Instead of using *addition* instructions, we use *averaging* instructions, such as `vpavgb` on x86 or `vrhadd` on ARM. While non-saturating additions compute $(a + b)$ % 256, these instructions compute $(a + b + 1)/2$. This means that they lose information about the low bit instead of the high bit of the sum. We estimate the overall mean by averaging pairs of values, then pairs of pairs, and so on. We refer the reader to Appendix D for details.

The challenging part of this approach is computing the bias in the estimated sum in order to correct for it. We prove in Appendix D that this bias has a closed-form solution under the realistic assumption that the low bits are equally likely to be 0 or 1.

4.5. Theoretical Guarantees

Our central theoretical result is a generalization guarantee for the overall approximation error of MADDNESS, stated below. See Appendix F for a proof and additional analysis, including a discussion of algorithmic complexity. Besides this main guarantee, we also inherit all of the guarantees for Bolt (Blalock & Guttag, 2017), modulo a small amount

of additional error from averaging integers rather than summing exactly. This follows from Bolt's guarantees depending only on the quantization errors, rather than the method used to obtain them.

Theorem 4.1 (Generalization Error of MADDNESS)**.** *Let $\mathcal{D}$ be a probability distribution over $\mathbb{R}^D$ and suppose that* MADDNESS *is trained on a matrix $\tilde{\boldsymbol{A}} \in \mathbb{R}^{N \times D}$ whose rows are drawn independently from $\mathcal{D}$ and with maximum singular value bounded by σ_A. Let C be the number of codebooks used by* MADDNESS *and $\lambda > 0$ be the regularization parameter used in the ridge regression step. Then for any $\boldsymbol{b} \in \mathbb{R}^D$, any $\boldsymbol{a} \sim \mathcal{D}$, and any $0 < \delta < 1$, we have with probability at least $1 - \delta$ that*

$$\mathbb{E}_{\mathcal{D}}[\mathcal{L}(\boldsymbol{a}, \boldsymbol{b})] \leq \mathbb{E}_{\tilde{\boldsymbol{A}}}[\mathcal{L}(\boldsymbol{a}, \boldsymbol{b})] + \frac{C \sigma_A \|\boldsymbol{b}\|_2}{2\sqrt{\lambda}} \left(\frac{1}{256} + \frac{8 + \sqrt{\nu(C, D, \delta)}}{\sqrt{2n}} \right) \tag{10}$$

where $\mathcal{L}(\boldsymbol{a}, \boldsymbol{b}) \triangleq |\boldsymbol{a}^\top \boldsymbol{b} - \alpha f(g(\boldsymbol{a}), h(\boldsymbol{b})) - \boldsymbol{\beta}|$, α is the scale used for quantizing the lookup tables, $\boldsymbol{\beta}$ is the constants used in quantizing the lookup tables plus the debiasing constant of Section 4.4, and

$$\nu(C, D, \delta) \triangleq C(4 \lceil \log_2(D) \rceil + 256) \log 2 - \log \delta. \tag{11}$$

5. Experiments

To assess MADDNESS's effectiveness, we implemented both it and existing algorithms in C++ and Python. All of our code and raw numerical results are publicly available at `https://smarturl.it/Maddness`. All experiments use a single thread on a Macbook Pro with a 2.6GHz Intel Core i7-4960HQ processor. Unless stated otherwise, all timing results use five trials, with each trial reporting the fastest among 20 executions. We use the best, rather than the average, since this is standard practice in performance benchmarking and is robust to the purely additive noise introduced by competing CPU tasks. Standard deviations are shown for all curves as shaded areas. Since training can be performed offline and all methods except SparsePCA (Mairal et al., 2009) train in at most a few minutes, we omit profiling of training times. We also do not profile the time to preprocess $\boldsymbol{B}$, since 1) this time is inconsequential in most cases, and 2) $\boldsymbol{B}$ is fixed and could be processed offline in all the problems we consider. In order to avoid implementation bias, we build upon the source code provided by Blalock & Guttag (2017)[2], which includes highly tuned implementations of many algorithms to which we compare.

We do not need to tune any hyperparameters for MADDNESS, but we do take steps to ensure that other methods are not hindered by insufficient hyperparameter tuning. Concretely, we sweep a wide range of hyperparameter values and allow them to cherry-pick their best hyperparameters on each test matrix. Further details regarding our experimental setup and choices (e.g., use of a single thread) can be found in Appendix E.

[2]https://github.com/dblalock/bolt

5.1. Methods Tested

Recall that most baselines take the form of selecting a matrix $\boldsymbol{V} \in \mathbb{R}^{D \times d}, d < D$ such that $\boldsymbol{AB} \approx (\boldsymbol{AV})(\boldsymbol{V}^\top \boldsymbol{B})$. Here d is a free parameter that adjusts the quality versus speed tradeoff. We therefore characterize most of these methods by how they set $\boldsymbol{V}$.

- **PCA**. Set $\boldsymbol{V}$ equal to the top principal components of $\tilde{\boldsymbol{A}}$.
- **SparsePCA** (Mairal et al., 2009). Set $\boldsymbol{V} = \text{argmin}_{\boldsymbol{V}} \min_{\boldsymbol{U}} \frac{1}{2N_{train}} \|\tilde{\boldsymbol{A}} - \boldsymbol{U}\boldsymbol{V}^\top\|_F^2 + \lambda \|\boldsymbol{V}\|_1$, where $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$. This is not the only dictionary learning formulation referred to as SparsePCA (Zou & Xue, 2018; Camacho et al., 2020), but it is a good representative and is the only one with support in a major Python library.
- **FastJL** (Ailon & Chazelle, 2009). $\boldsymbol{V}$ is set to Rademacher random variables composed with a Fast Hadamard Transform (FHT). For simplicity, we exclude the FHT in the timing.
- **HashJL** (Dasgupta et al., 2010). $\boldsymbol{V}$ is zero except for a ± 1 in each row, with both sign and position chosen uniformly at random.
- **ScalarQuantize**. The matrices are not projected, but instead linearly quantized to eight bits such that the smallest and largest entries map to either 0 and 255 or -128 and 127, as appropriate. We use FBGEMM (Khudia et al., 2018) to perform the quantized matrix multiplies. We neglect the time required to convert from other formats to eight bits, reflecting the optimistic scenario in which the matrices are already of the appropriate types.
- **Bolt** (Blalock & Guttag, 2017). Bolt is the most similar method to our own, differing only in the encoding function, the use of averaging instead of upcasting, and the optimization of centroids.
- **Exact Multiplication**. We simply compute the matrix product $\boldsymbol{AB}$ using a modern BLAS implementation.
- **MADDNESS-PQ**. A handicapped version of MADDNESS without the prototype optimization step. The gap between MADDNESS and MADDNESS-PQ is the gain from optimizing the prototypes.

We also compared to many additional methods (see Appendix E), but omit their results since they were not competitive with those listed here.

5.2. How Fast is MADDNESS?

We begin by profiling the raw speed of our method. In Figure 3, we time the $g(\boldsymbol{A})$ functions for various vector quantization methods. The $\boldsymbol{A}$ matrices have 2^{14} rows and varying numbers of columns D. Following Blalock & Guttag (2017), we also vary the number of codebooks C, profiling 8-, 16-, and 32-byte encodings. We measure in bytes rather than codebooks since PQ and OPQ use eight bits per codebook while Bolt and MADDNESS use four.

MADDNESS is up to two orders of magnitude faster than existing methods, and its throughput increases with row length. This latter property is because its encoding cost per row is $O(C)$ rather than $O(D)$.

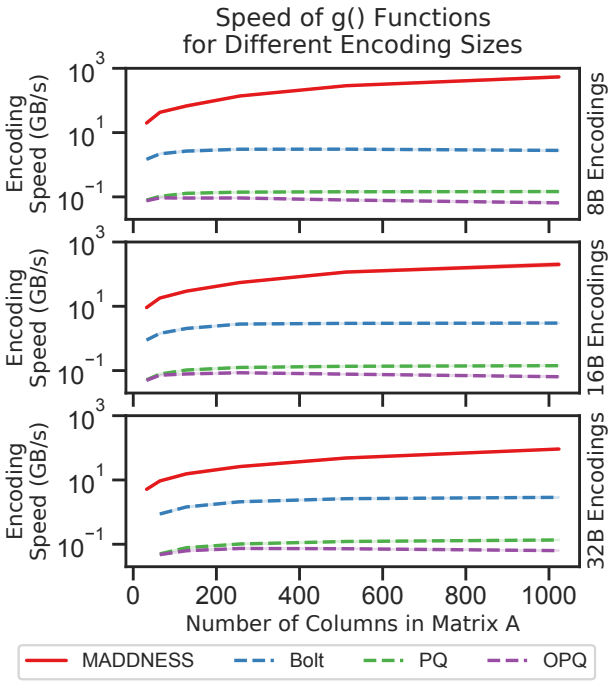

Figure 3: MADDNESS encodes the $\boldsymbol{A}$ matrix orders of magnitude more quickly than existing vector quantization methods.

We also profile the speed of our aggregation function $f(\cdot, \cdot)$ using the same baselines as Blalock & Guttag (2017). As Figure 4 shows, our average-based, matrix-aware aggregation is significantly faster than the upcasting-based method of Bolt, its nearest rival.

5.3. Softmax Classifier

As described in Section 1, we approximated linear classifiers on the widely used CIFAR-10 and CIFAR-100 datasets (Krizhevsky et al., 2009). The classifiers use as input features the 512-dimensional activations of open-source, VGG-like neural networks trained on each dataset (Geifman, 2018). The matrices $\boldsymbol{A}$ are the 10000×512-dimensional floating point activations for the full test sets, and the matrices $\boldsymbol{B}$ are each network's final dense layer. The 50000×512-dimensional activations from the training set serve as the training matrices $\tilde{\boldsymbol{A}}$. As shown in Figure 5, MADDNESS significantly outperforms all existing methods, achieving virtually the same accuracy as exact multiplication more than an order of magnitude faster.

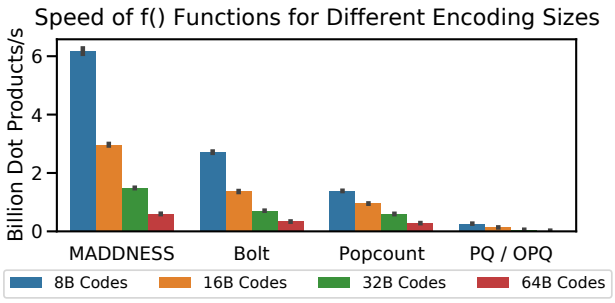

Figure 4: Given the preprocessed matrices, MADDNESS computes the approximate output twice as fast as the fastest existing method.

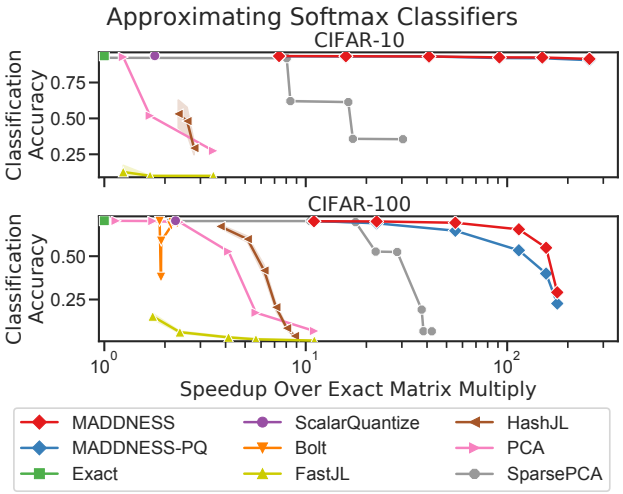

Figure 5: MADDNESS achieves a far better speed-accuracy tradeoff than any existing method when approximating two softmax classifiers.

Moreover, our method achieves this performance despite having worse support from the hardware. More precisely, it obtains speedups much smaller than the level of compression it provides. For example, the third points from the right in both plots correspond to speedups of roughly $100\times$. However, they are compressing each $512 \times 4\text{B} = 2048\text{B}$ row of the input down to a mere 4B, a savings of $512\times$ (sizes not shown in figure). *If the hardware could lookup-accumulate as many bytes per cycle as it can multiply-accumulate, our method could be over $4\times$ faster.* Combined with the fact that multiplexers require many fewer transistors than multipliers, this suggests that a hardware implementation of our method might offer large efficiency gains compared to existing accelerators.

5.4. Kernel-Based Classification

To assess the efficacy of our method on a larger and more diverse set of datasets than CIFAR-10 and CIFAR-100, we trained kernel classifiers on the datasets from the UCR Time Series Archive (Dau et al., 2018). To enable meaningful speed comparison across datasets, we resampled the time series in all datasets to the median length and obtained the matrix $\boldsymbol{B}$ for each dataset by running Stochastic Neighbor Compression (Kusner et al., 2014) on the training set with an RBF kernel of bandwidth one. We approximate the Euclidean distances used by the kernel via the identity $\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 = \|\boldsymbol{x}\|_2^2 - 2\boldsymbol{x}^\top \boldsymbol{y} + \|\boldsymbol{y}\|_2^2$, which consists only of dot products. This is not the state-of-the-art means of classifying time series, but it does yield fixed-sized matrices and is representative of several modern techniques for constructing highly efficient classifiers (Kusner et al., 2014; Wang et al., 2016b; Zhong et al., 2017; Gupta et al., 2017). Because Stochastic Neighbor Compression optimizes the classifiers to avoid redundancy, this task is quite difficult.

As shown in Figure 6, MADDNESS is significantly faster than alternatives at a given level of accuracy. A counterintuitive result, however, is that optimization of the prototypes occasionally reduces accuracy—see the red line dipping below the blue one in the lowest subplot. Since the optimization strictly increases the expressive power, we believe that this is a product of overfitting and could be corrected by not fixing $\lambda = 1$ in the ridge regression.

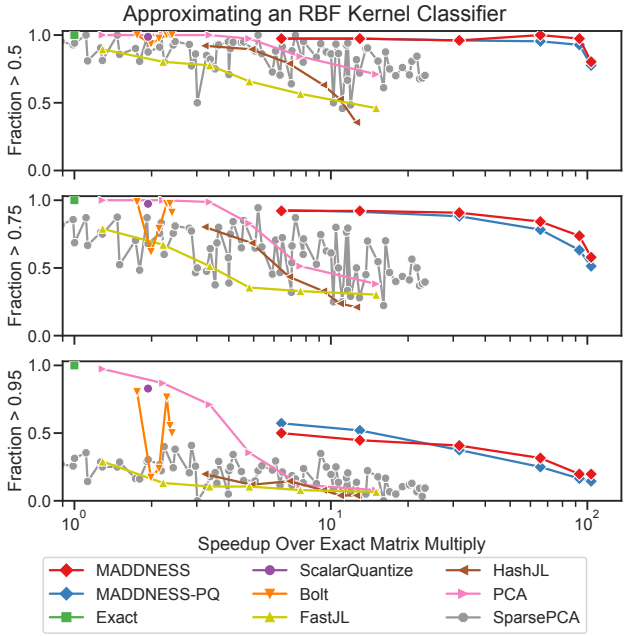

Figure 6: Fraction of UCR datasets for which each method preserves a given fraction of the original accuracy. MADDNESS enables much greater speedups for a given level of accuracy degradation.

5.5. Image Filtering

To test the extreme limits of MADDNESS, we benchmarked the various techniques' ability to apply small filters to images (after an offline im2row transform to reduce the task to matrix multiplication). This task is extreme in that D and M are tiny, affording almost no opportunity to amortize preprocessing costs. As representative example filters, we chose 3×3 Sobel kernels and 5×5 Gaussian kernels. These are common high-pass and low-pass filters, respectively. We took the first 10 images from the first 50 classes of the Caltech101 dataset (Fei-Fei et al., 2004) as a single training set, and the first 10 images from the remaining 51 classes as 510 test sets. We constructed the $\boldsymbol{A}$ matrices by extracting each patch of each image as one row. The $\boldsymbol{B}$ matrices have two columns, corresponding to one pair of Sobel or Gaussian filters (since using these filters in pairs is common). We report the normalized mean-squared error (NMSE), defined as $\|\hat{\boldsymbol{C}}_{i,j} - \boldsymbol{AB}\|_F^2 / \|\boldsymbol{AB}\|_F^2$, where $\hat{\boldsymbol{C}}$ is the method's estimate of $\boldsymbol{AB}$. An NMSE of 0 is perfect and an NMSE of 1 corresponds to always predicting 0.

In Figure 7, we see that it is only MADDNESS that offers any advantage over exact matrix products. This is likely because two columns afford almost no time to preprocess $\boldsymbol{A}$; indeed, rival vector quantization methods cannot logically do less work than brute force in this setting, and dense linear methods can only save work by embedding rows of $\boldsymbol{A}$ in one-dimensional space. MADDNESS performs much worse on the high-pass filters (top) than the low-pass filters (bottom). This is likely because the former produce outputs with variance that is orders of magnitude lower than that of the original image, making the NMSE denominator tiny.

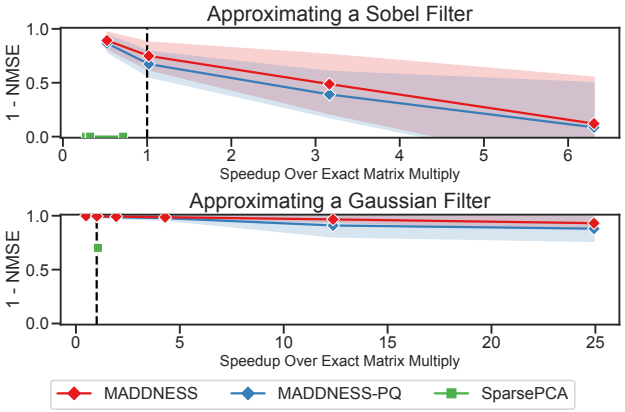

Figure 7: Despite there being only two columns in the matrix B, MADDNESS still achieves a significant speedup with reasonable accuracy. Methods that are Pareto dominated by exact matrix multiplication on both tasks are not shown; this includes all methods but MADDNESS and SparsePCA.

6. Discussion and Conclusion

Because our work draws on a number of different fields but does not fit cleanly into any of them, it is useful to discuss what we have and have not demonstrated, as well as possible implications and extensions of our work.

Our main empirical finding is that our proposed method, MADDNESS, achieves order-of-magnitude speedups compared to existing AMM methods and up to two-order-of-magnitude speedups compared to the dense baseline. It also compresses matrices by up to three orders of magnitude. These results are evaluated on a CPU, and are obtainable only when there is a training set for one matrix. We also claim superior performance only when one matrix is larger than the other, and both matrices are tall—the regime wherein our extremely fast (but less accurate) encoding function is beneficial. Our method also loses utility when the larger matrix is known ahead of time; this assumption is common in similarity search, and eliminates the need for a fast encoding function entirely. Our approximate integer summation and fused table lookups would likely be useful independent of any of these assumptions, but demonstrating this is future work.

We also have several theoretical findings, taking the form of guarantees regarding the errors introduced by our method and its constituent subroutines. While we do obtain an overall generalization guarantee, this guarantee is not tight. In particular, it should grow looser with the large matrix's Frobenius norm and tighter as its singular values become more concentrated; at present, however, it simply grows looser as the largest singular value grows. The missing step is a guarantee that our encoding function will yield lower quantization errors when the singular values are more concentrated, which is its behavior in practice.

We have not demonstrated results using GPUs or other accelerators. While such accelerators are a small minority of hardware, they are often used in machine learning. Our method is not inherently tied to CPUs, but the differing performance characteristics of accelerators mean that adapting our method to them would require both algorithmic and implementation work, with the details depending on the device. We also have not evaluated a multi-CPU-threaded extension of our algorithm, though this is because our method is intended to serve as the low-level, compute-bound, block matrix product routine called by individual threads.

Finally, we have not demonstrated results using convolutional layers in neural networks, or results accelerating full networks. The weight reuse in convolutional layers presents many opportunities for algorithmic optimizations, and we hope to exploit them using a specialized extension of our method in future work. Accelerating overall networks will require two significant undertakings: first, the engineering work of building and integrating custom operators, data layouts, etc., into existing frameworks and networks; and second, the research necessary to determine when, how, and to what extent to include approximate kernels inspired by our approach. A particular difficulty with the latter is that our hash function is not differentiable.

We believe that accelerating full networks with our ideas is a promising direction, particularly for inference. This is especially true at the hardware level—our method requires only multi*plexers*, not multi*pliers*, and can therefore be implemented easily and with far less power than current matrix product logic. Moreover, our encoded representation and lookup tables have contiguous and uniformly-sized elements, making our approximate GEMM inner loops nearly identical to their dense counterparts–i.e., there is no need for complex access patterns or sparsity handling.

In summary, we introduced MADDNESS, an algorithm that achieves up to a $10\times$ better speed-quality tradeoff than existing methods for the well-studied problem of approximate matrix multiplication (AMM), as measured on a large, diverse, and challenging set of real-world matrices. Our approach is a significant departure from existing AMM work in that it relies on hashing and table lookups rather than multiply-add operations. Our results suggest that future methods similar to our own might hold promise for accelerating convolution, deep learning, and other workloads bottlenecked by linear transforms.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pp. 265–283, 2016.

Achlioptas, D. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 274–281, 2001.

Ailon, N. and Chazelle, B. The Fast Johnson-Lindenstrauss Transform and Approximate Nearest Neighbors. *SIAM Journal on Computing (SICOMP)*, 39(1):302–322, 2009. doi: 10.1137/060673096.

André, F., Kermarrec, A.-M., and Le Scouarnec, N. Accelerated nearest neighbor search with quick adc. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pp. 159–166, 2017.

André, F., Kermarrec, A.-M., and Le Scouarnec, N. Quicker adc: Unlocking the hidden potential of product quantization with simd. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Babenko, A. and Lempitsky, V. Additive quantization for extreme vector compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 931–938, 2014.

Babenko, A. and Lempitsky, V. Tree Quantization for Large-Scale Similarity Search and Classification. *CVPR*, pp. 1–9, 2015. URL papers3://publication/uuid/F4762974-BB97-4208-B035-508945A90EFC.

Babenko, A., Arandjelović, R., and Lempitsky, V. Pairwise quantization. *arXiv preprint arXiv:1606.01550*, 2016.

Bakhtiary, A. H., Lapedriza, A., and Masip, D. Speeding up neural networks for large scale classification using wta hashing. *arXiv preprint arXiv:1504.07488*, 2015.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Blalock, D. W. and Guttag, J. V. Bolt: Accelerated data mining with fast vector compression. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 727–735. ACM, 2017.

Blalock, D. W., Ortiz, J. J. G., Frankle, J., and Guttag, J. V. What is the state of neural network pruning? In Dhillon, I. S., Papailiopoulos, D. S., and Sze, V. (eds.), *Proceedings of Machine Learning and Systems 2020, MLSys 2020, Austin, TX, USA, March 2-4, 2020*. mlsys.org, 2020. URL https://proceedings.mlsys.org/book/296.pdf.

Camacho, J., Smilde, A., Saccenti, E., and Westerhuis, J. All sparse pca models are wrong, but some are useful. part i: Computation of scores, residuals and explained variance. *Chemometrics and Intelligent Laboratory Systems*, 196:103907, 2020.

Chen, B., Medini, T., and Shrivastava, A. Slide: In defense of smart algorithms over hardware acceleration for large-scale deep learning systems. *arXiv preprint arXiv:1903.03129*, 2019.

Chen, W., Wilson, J. T., Tyree, S., Weinberger, K. Q., and Chen, Y. Compressing neural networks with the hashing trick. In *ICML*, pp. 2285–2294, 2015.

Chen, Y.-H., Emer, J., and Sze, V. Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. *ACM SIGARCH Computer Architecture News*, 44(3):367–379, 2016.

Dasgupta, A., Kumar, R., and Sarlós, T. A sparse johnson: Lindenstrauss transform. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 341–350, 2010.

Dau, H. A., Keogh, E., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., Yanping, Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G., and Hexagon-ML. The ucr time series classification archive, October 2018. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

Dean, T., Ruzon, M. A., Segal, M., Shlens, J., Vijayanarasimhan, S., and Yagnik, J. Fast, accurate detection of 100,000 object classes on a single machine. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1814–1821, 2013.

Desai, A., Ghashami, M., and Phillips, J. M. Improved practical matrix sketching with guarantees. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1678–1690, 2016.

Drineas, P., Kannan, R., and Mahoney, M. W. Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication. *SIAM Journal on Computing*, 36(1):132–157, January 2006a. ISSN 0097-5397, 1095-7111. doi: 10.1137/S0097539704442684. URL http://epubs.siam.org/doi/10.1137/S0097539704442684.

Drineas, P., Kannan, R., and Mahoney, M. W. Fast Monte Carlo Algorithms for Matrices II: Computing a Low-Rank Approximation to a Matrix. *SIAM Journal on Computing*, 36(1):158–183, January 2006b. ISSN 0097-5397, 1095-7111. doi: 10.1137/S0097539704442696. URL http://epubs.siam.org/doi/10.1137/S0097539704442696.

Drineas, P., Kannan, R., and Mahoney, M. W. Fast Monte Carlo Algorithms for Matrices III: Computing a Compressed Approximate Matrix Decomposition. *SIAM Journal on Computing*, 36(1):184–206, January 2006c. ISSN 0097-5397, 1095-7111. doi: 10.1137/S0097539704442702. URL http://epubs.siam.org/doi/10.1137/S0097539704442702.

Dutta, S., Cadambe, V., and Grover, P. Short-dot: Computing large linear transforms distributedly using coded short dot products. In *Advances In Neural Information Processing Systems*, pp. 2100–2108, 2016.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Francis, D. P. and Raimond, K. An improvement of the parameterized frequent directions algorithm. *Data Mining and Knowledge Discovery*, 32(2):453–482, March 2018a. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-017-0542-x. URL http://link.springer.com/10.1007/s10618-017-0542-x.

Francis, D. P. and Raimond, K. A practical streaming approximate matrix multiplication algorithm. *Journal of King Saud University - Computer and Information Sciences*, September 2018b. ISSN 13191578. doi: 10.1016/j.jksuci.2018.09.010. URL https://linkinghub.elsevier.com/retrieve/pii/S1319157818306396.

Ge, T., He, K., Ke, Q., and Sun, J. Optimized product quantization. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):744–755, 2014.

Geifman, Y. cifar-vgg, 3 2018. https://github.com/geifmany/cifar-vgg.

Ghashami, M., Liberty, E., Phillips, J. M., and Woodruff, D. P. Frequent Directions: Simple and Deterministic Matrix Sketching. *SIAM Journal on Computing*, 45(5):1762–1792, January 2016. ISSN 0097-5397, 1095-7111. doi: 10.1137/15M1009718. URL http://epubs.siam.org/doi/10.1137/15M1009718.

Guennebaud, G., Jacob, B., et al. Eigen v3. http://eigen.tuxfamily.org, 2010.

Gupta, C., Suggala, A. S., Goyal, A., Simhadri, H. V., Paranjape, B., Kumar, A., Goyal, S., Udupa, R., Varma, M., and Jain, P. Protonn: Compressed and accurate knn for resource-scarce devices. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1331–1340. JMLR. org, 2017.

Han, S., Liu, X., Mao, H., Pu, J., Pedram, A., Horowitz, M. A., and Dally, W. J. Eie: efficient inference engine on compressed deep neural network. In *Proceedings of the 43rd International Symposium on Computer Architecture*, pp. 243–254. IEEE Press, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016b.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Huang, Z. Near Optimal Frequent Directions for Sketching Dense and Sparse Matrices. *Journal of Machine Learning Research*, 20(1):23, February 2019.

Indyk, P. and Motwani, R. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pp. 604–613, 1998.

Irony, D., Toledo, S., and Tiskin, A. Communication lower bounds for distributed-memory matrix multiplication. *Journal of Parallel and Distributed Computing*, 64 (9):1017–1026, 2004.

Jegou, H., Douze, M., and Schmid, C. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2011.

Ji, J., Li, J., Yan, S., Zhang, B., and Tian, Q. Super-bit locality-sensitive hashing. In *Advances in Neural Information Processing Systems*, pp. 108–116, 2012.

Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, pp. 1–12. ACM, 2017.

Kakade, S. M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, pp. 793–800, 2009.

Khudia, D., Basu, P., and Deng, S. Open-sourcing fbgemm for state-of-the-art server-side inference, 2018.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Kusner, M., Tyree, S., Weinberger, K., and Agrawal, K. Stochastic neighbor compression. In *International Conference on Machine Learning*, pp. 622–630, 2014.

Kyrillidis, A., Vlachos, M., and Zouzias, A. Approximate Matrix Multiplication with Application to Linear Embeddings. *arXiv:1403.7683 [cs, math, stat]*,

March 2014. URL http://arxiv.org/abs/1403.7683. arXiv: 1403.7683.

Liberty, E. Simple and Deterministic Matrix Sketching. *arXiv:1206.0594 [cs]*, June 2012. URL http://arxiv.org/abs/1206.0594. arXiv: 1206.0594.

Liu, S., Shao, J., and Lu, H. Generalized Residual Vector Quantization for Large Scale Data. *Proceedings - IEEE International Conference on Multimedia and Expo*, 2016-Augus, 2016. ISSN 1945788X. doi: 10.1109/ICME.2016.7552944.

Luo, L., Chen, C., Zhang, Z., Li, W.-J., and Zhang, T. Robust Frequent Directions with Application in Online Learning. *Journal of Machine Learning Research*, 20 (1):41, February 2019.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pp. 689–696, 2009.

Manne, S. and Pal, M. Fast Approximate Matrix Multiplication by Solving Linear Systems. *arXiv:1408.4230 [cs]*, August 2014. URL http://arxiv.org/abs/1408.4230. arXiv: 1408.4230.

Martinez, J., Hoos, H. H., and Little, J. J. Stacked quantizers for compositional vector compression. *arXiv preprint arXiv:1411.2173*, 2014.

Martinez, J., Clement, J., Hoos, H. H., and Little, J. J. Revisiting additive quantization. In *European Conference on Computer Vision*, pp. 137–153. Springer, 2016.

Mroueh, Y., Marcheret, E., and Goel, V. Co-Occuring Directions Sketching for Approximate Matrix Multiply. *arXiv:1610.07686 [cs]*, October 2016. URL http://arxiv.org/abs/1610.07686. arXiv: 1610.07686.

Nelson, J. and Nguyên, H. L. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 ieee 54th annual symposium on foundations of computer science*, pp. 117–126. IEEE, 2013.

Pagh, R. Compressed matrix multiplication. *ACM Transactions on Computation Theory*, 5(3):1–17, August 2013. ISSN 19423454. doi: 10.1145/2493252.2493254. URL http://dl.acm.org/citation.cfm?doid=2493252.2493254.

Parashar, A., Rhu, M., Mukkara, A., Puglielli, A., Venkatesan, R., Khailany, B., Emer, J., Keckler, S. W., and Dally, W. J. Scnn: An accelerator for compressed-sparse convolutional neural networks. *ACM SIGARCH Computer Architecture News*, 45(2):27–40, 2017.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Sarlos, T. Improved Approximation Algorithms for Large Matrices via Random Projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pp. 143–152, Berkeley, CA, October 2006. IEEE. ISBN 978-0-7695-2720-8. doi: 10.1109/FOCS.2006.37. URL https://ieeexplore.ieee.org/document/4031351/.

Spring, R. and Shrivastava, A. Scalable and sustainable deep learning via randomized hashing. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 445–454, 2017.

Teng, D. and Chu, D. A Fast Frequent Directions Algorithm for Low Rank Approximation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1279–1293, June 2019. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2018.2839198. URL https://ieeexplore.ieee.org/document/8362693/.

Wang, J., Shen, H. T., Song, J., and Ji, J. Hashing for similarity search: A survey. *arXiv preprint arXiv:1408.2927*, 2014a.

Wang, J., Shen, H. T., Yan, S., Yu, N., Li, S., and Wang, J. Optimized distances for binary code ranking. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 517–526, 2014b.

Wang, J., Liu, W., Kumar, S., and Chang, S.-F. Learning to hash for indexing big dataa survey. *Proceedings of the IEEE*, 104(1):34–57, 2016a.

Wang, W., Chen, C., Chen, W., Rai, P., and Carin, L. Deep metric learning with data summarization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 777–794. Springer, 2016b.

Wu, C.-J., Brooks, D., Chen, K., Chen, D., Choudhury, S., Dukhan, M., Hazelwood, K., Isaac, E., Jia, Y., Jia, B., et al. Machine learning at facebook: Understanding inference at the edge. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 331–344. IEEE, 2019.

Ye, Q., Luo, L., and Zhang, Z. Frequent Direction Algorithms for Approximate Matrix Multiplication with Applications in CCA. In *IJCAI*, pp. 7, 2016.

Yu, Q., Maddah-Ali, M., and Avestimehr, S. Polynomial codes: an optimal design for high-dimensional coded matrix multiplication. In *Advances in Neural Information Processing Systems*, pp. 4403–4413, 2017.

Yu, Q., Ali, M., and Avestimehr, A. S. Straggler mitigation in distributed matrix multiplication: Fundamental limits and optimal coding. *IEEE Transactions on Information Theory*, 2020.

Zhang, T., Du, C., and Wang, J. Composite Quantization for Approximate Nearest Neighbor Search. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 32:838–846, 2014.

Zhang, T., Qi, G.-J., Tang, J., and Wang, J. Sparse composite quantization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4548–4556, 2015.

Zhong, K., Guo, R., Kumar, S., Yan, B., Simcha, D., and Dhillon, I. Fast classification with binary prototypes. In *Artificial Intelligence and Statistics*, pp. 1255–1263, 2017.

Zou, H. and Xue, L. A selective overview of sparse principal component analysis. *Proceedings of the IEEE*, 106 (8):1311–1320, 2018.