

Bringing UMAP Closer to the Speed of Light with GPU Acceleration

Corey J. Nolet^{1,2}, Victor Lafargue¹, Edward Raff^{2,3}, Thejaswi Nanditale¹, Tim Oates²,
John Zedlewski¹, Joshua Patterson¹

¹Nvidia

²University of Maryland Baltimore County

³Booz Allen Hamilton

Abstract

The Uniform Manifold Approximation and Projection (UMAP) algorithm has become widely popular for its ease of use, quality of results, and support for exploratory, unsupervised, supervised, and semi-supervised learning. While many algorithms can be ported to a GPU in a simple and direct fashion, such efforts have resulted in inefficient and inaccurate versions of UMAP. We show a number of techniques that can be used to make a faster and more faithful GPU version of UMAP, and obtain speedups of up to 100x in practice. Many of these design choices/lessons are general purpose and may inform the conversion of other graph and manifold learning algorithms to use GPUs. Our implementation has been made publicly available as part of the open source RAPIDS cuML library (<https://github.com/rapidsai/cuml>).

1 Introduction

Like other manifold learning algorithms, the Uniform Manifold Approximation and Projection algorithm (UMAP) (McInnes, Healy, and Melville 2018) relies upon the manifold hypothesis (Fefferman, Mitter, and Narayanan 2016) to preserve local neighborhood structure by modeling high-dimensional data in a low-dimensional space. This is in contrast to linear dimensionality reduction techniques like PCA, which aim only to preserve global Euclidean structure (He et al. 2005). UMAP produces low-dimensional embeddings that are useful for both visual analytics and downstream machine learning tasks. Unlike other manifold learning algorithms, such as IsoMap (Tenenbaum, De Silva, and Langford 2000), Locally Linear Embeddings (LLE) (Roweis and Saul 2000), Laplacian Eigenmaps (Belkin and Niyogi 2002), and t-Distributed Stochastic Neighbor Embeddings (T-SNE) (Maaten and Hinton 2008), UMAP has native support for supervised, unsupervised, and semi-supervised metric learning. Since its introduction in 2018, it has found use in exploratory data analysis applications (Ordun, Purushotham, and Raff 2020; Wander et al. 2020; Obermayer et al. 2020; Oden 2020), as well as bioinformatics, cancer research (Andor et al. 2018), single-cell genomics (Travaglini et al. 2019; Becht et al. 2018; Clara-Parabricks 2020), and the interpretation of highly non-linear models like deep neural networks

(Carter et al. 2019). This combination of features and quality of results has made UMAP a widely used and popular tool.

The wide array of applications and use of UMAP makes it desirable to produce faster versions of the algorithm. This is compounded by an increasing demand for exploratory and interactive visualization (Carter et al. 2019; Pezzotti et al. 2018; Chatzimparmpas, Martins, and Kerren 2020; Obermayer et al. 2020), that necessitates a lower latency in results. GPUs are a strong candidate for achieving faster implementation by trading per-core clock speeds for significantly more cores, provided that they can be utilized effectively in parallel computation. A direct conversion of UMAP to the GPU already exists in the GPUMAP (Pachev and Lupo 2017) project but, due to technical details, is not always faithful in reproducing the same quality of results. In this work, we show that applying a few general techniques can produce a version that is both faster and faithful in its results.

This paper contributes three components to the growing ecosystem (Okuta et al. 2017; Raschka, Patterson, and Nolet 2020; Johnson, Douze, and Jégou 2019; Paszke et al. 2019) of GPU-accelerated tools for data science in Python. First, we contribute a near drop-in replacement of the UMAP-learn Python library, which has been GPU-accelerated end-to-end in CUDA/C++. The speedup of this new implementation is evaluated against the current state-of-the-art, including UMAP-learn on the CPU and an existing GPU port of UMAP. Second, we contribute a GPU-accelerated near-drop-in replacement of the trustworthiness score, which is often used to evaluate the extent to which manifold learning algorithms preserve local neighborhood structure. Finally, we contribute a distributed version of UMAP and provide empirical evidence of its effectiveness.

The remainder of our paper is organized as follows. We will discuss related work in Section 2, with a brief review of UMAP in Section 3. Our approach to implementing UMAP for the GPU is detailed in Section 4, with code available as part of the RAPIDS cuML library (<https://github.com/rapidsai/cuml>). Our results show up to 100x speedups in Section 5, followed by our conclusions in Section 6.

2 Related Work

AI and ML research often strikes a balance in scientific exploration and software engineering, with well-engineered

software having a dramatic impact on both researchers and practitioners. Well implemented single-purpose packages that contain a single method or limited scope have proven to have a large influence on both of these factors. Many such works have limited theoretical contribution, but instead detail the work that goes into a careful and thoroughly efficient version of the algorithm. LIBLINEAR (Fan et al. 2008) and LIBSVM (Chang and Lin 2011) became tools for users and software to modify for years of research. Similar has happened for Optuna (Akiba et al. 2019) with bayesian hyperparameter optimization, and XGBoost (Chen and Guestrin 2016) has played a role in re-kindling decision tree & boosting research, its details on coding optimizations impacting other influential tools like LightGBM (Ke et al. 2017) and CatBoost (Dorogush, Ershov, and Gulin 2017). Likewise the original Cover-Tree implementation (Beygelzimer, Kakade, and Langford 2006) has influenced over a decade of nearest-neighbor search. The UMAP algorithm has quickly fallen into this group of single, influential implementations, but lacks GPU acceleration. We resolve this shortcoming in our work while simultaneously improving UMAP’s qualitative results through better implementation design, providing an empirical “existence proof” that these issues are not fundamental limitations of the original algorithm.

UMAP’s ability to shortcut the need for storing n^2 pairwise distances by defining local neighborhoods with the k -nearest neighbors around each data point is like other manifold learning algorithms. T-SNE predates UMAP and has found popularity in many of the same communities that UMAP has now become considered the state of the art (Oden 2020). T-SNE models point distances as probability distributions, constructing a students-t kernel from training data and minimizing the Kullback-Liebler divergence against the low-dimensional representation. While originally intractable for datasets containing more than a few thousand points, GPU-accelerated variants have recently breathed new life into the algorithm (Chan et al. 2018). Still, T-SNE has not been shown to work well for downstream machine learning tasks and lacks support for supervised learning.

The reference implementation of UMAP is built on top of the Numba(Lam, Pitrou, and Seibert 2015) library and uses just-in-time (JIT) compilation to make use of parallel low-level CPU optimizations. The GPUMAP library (Pachev and Lupo 2017) is a direct port of the reference library to the GPU, using Numba’s `cuda.jit` feature, along with the CuPy library, to directly replace many SciPy library invocations with CUDA-backed implementations. Like other libraries that require fast nearest neighbors search on GPUs (Chan et al. 2018), GPUMAP uses the FAISS library (Johnson, Douze, and Jégou 2019). Our implementation also uses the FAISS library. GPUMAP invokes FAISS through the Python API, missing opportunities for zero-copy exchanges of memory pointers on device (Raschka, Patterson, and Nolett 2020) that our implementation leverages.

Manifold learning algorithms typically use the trustworthiness (Venna and Kaski 2006) score to evaluate a trained model’s preservation of local neighborhood structure. The trustworthiness score penalizes divergences in the nearest neighbors between the algorithm’s input and output, rank-

ing the similarities of the neighborhoods. Scikit-learn (Pedregosa et al. 2011) provides an implementation of trustworthiness, but the computational costs and memory footprint associated with computing the entire n^2 pairwise distance matrix makes it prohibitively slow to evaluate datasets greater than a couple thousand samples. To the best of our knowledge, there are no existing ports of the trustworthiness score to the GPU. We fill this gap with our new batchable implementation, which we demonstrate can scale well over 100s of thousands of samples on a single GPU with reasonable performance.

3 Uniform Manifold Approximation and Projection

Like many manifold learning algorithms, the UMAP algorithm can be decomposed into three major stages, which we briefly describe in this section. For explanations, derivations, and further details, we refer the reader to the official UMAP paper (McInnes, Healy, and Melville 2018).

In the first stage, a k -nearest neighbors (k -NN) graph is constructed using a distance metric, $d(x, y)$. The second stage weights the closest neighbors around each vertex in the nearest neighbors graph, converting them to fuzzy sets and combining them into a fuzzy union. The fuzzy set membership function learns a locally adaptive exponential kernel that smoothes the distances in each local neighborhood of the k -NN graph by finding a smoothing normalizer σ_i such that Equation 1 is satisfied. ρ in this equation contains the distances to the closest non-zero neighbor around each vertex. The triangular conorm (Klement, Mesiar, and Pap 1997; Dubois and Prade 1982) in Equation 2 combines the matrix of individual fuzzy sets, A , into a fuzzy union by symmetrizing the graph and adding the element-wise (Hadamard) product.

$$\sum_{j=i}^k \exp(-\max(0, d(x_i, x_{i_j}) - \rho_i)\sigma_i^{-1}) = \log_2(k) \quad (1)$$

$$B = (A + A^T) + (A \circ A^T) \quad (2)$$

In the third and final stage, the embeddings are laid out in the topological space using stochastic gradient descent. An initial layout is performed either by sampling embeddings from a uniform distribution or computing a spectral embedding over the fuzzy union. The cross-entropy objective function $-\sum_{a,b \in B} (\log(\Phi(a, b)) + \sum_{c \in B}^m \log(1 - \Phi(a, c)))$ is minimized over the edges of the fuzzy union, B , from Equation 2. This is done with negative sampling where m is the number of negative samples per edge. Φ (Equation 3) is the current membership strength in the newly embedded space and min_dist controls the minimum separation distance between points in the embedded space. We use the approximate form of Φ in this paper for simplicity. The $\log(\Phi)$ term in the objective is computed using the source and destination vertices on each edge and $\log(1 - \Phi)$ is computed using the source vertex with negative sampling.

$$\Phi(x, y) \approx \begin{cases} 1 & \|x - y\|_2 \leq min_dist \\ \exp(-\|x - y\|_2 - min_dist) & otherwise \end{cases} \quad (3)$$

When training labels are provided, an additional step in the neighborhood weighting stage adjusts the membership strengths of the fuzzy sets based on their labels. In addition to its learned parameters, the trained UMAP model keeps a reference to the k -NN index computed on the training data. This is used to create a mapping of the trained embeddings to a new set of vertices during inference.

4 GPU-Accelerating UMAP

Our implementation is primarily written in C++, which is wrapped in a Python API through the Cython library. Data can be passed into our Python API using common formats like Numpy (Harris et al. 2020) or Pandas (McKinney et al. 2011), as well as GPU array libraries such as CuPy (Okuta et al. 2017), Numba (Lam, Pitrou, and Seibert 2015), or RAPIDS cuDF (Raschka, Patterson, and Nolet 2020). When necessary, the data is automatically copied onto the device (e.g., when a Numpy array is passed in). Columnar memory layouts, such as those used in Apache Arrow (LeDem 2017), tend to exploit the optimized memory access patterns on GPUs, such as coalesced accesses (Davidson and Jinturkar 1994). Like UMAP-learn, our Python API maintains compatibility with the Scikit-learn (Pedregosa et al. 2011) API. We used the RAPIDS memory manager (RMM) to create a single memory pool for each process to avoid device synchronization from allocations and deallocations of temporary device memory. When available, we made use of existing libraries with optimized CUDA primitives, such as Thrust (Bell and Hoberock 2012), cuSparse (Naumov et al. 2010; Li et al. 2015), cuGraph, and cuML (Ocsa 2019; Raschka, Patterson, and Nolet 2020).

Our implementation begins with a straightforward port of UMAP-learn to CUDA/C++, diverging from the design of UMAP-learn only where we found a significant benefit to performance or the memory footprint. The prior GPUMAP implementation attempted a direct conversion of the code design, using Numba CUDA-JIT (Oden 2020) functions and CuPy, without any significant diversions. While hypothetically easier to maintain, we will show this produces results that do not always match the original implementation, and does not deliver meaningful speedups. In each section below, we will detail some of the major design choices that make our GPU implementation faster and more faithful to the original results.

Copying data between host memory and a GPU device comes at a high cost and can quickly become a bottleneck. Transferring 1MB of data can take several hundreds of microseconds. Even when memory is copied asynchronously between device and host, the underlying CUDA stream needs to be synchronized before the data can be used on the host, further increasing this latency. We reduce the need to transfer between host and device as much as possible, even if that means running code on the GPU that has little to no speedup over the CPU (Harris 2012). Standards like the `__cuda_array_interface__` (Raschka, Patterson, and Nolet 2020) and `dlpack` (Chen 2020) enable Python libraries to share CUDA memory pointers directly, without the need for copies (Raschka, Patterson, and Nolet 2020). This further reduces the need for transfers to host, and like the standard

`__array_interface__` (Harris et al. 2020), enables implicit conversion between types even as data is passed between different libraries.

4.1 GPU Architecture

The NVIDIA General-purpose GPU computing architecture (Owens et al. 2008; Luebke 2008) enables parallelism through the single-instruction multiple data design paradigm (Raschka, Patterson, and Nolet 2020). A single GPU device contains several banks of global memory that are accessible from a grid containing thousands of instruction processing cores called symmetric-multiprocessors, or SMs, which execute blocks of threads, called thread-blocks, in parallel.

Each SM has its own faster but smaller bank of memory called shared memory which is partitioned across the thread-blocks it executes, enabling a series of concurrently executing threads within each thread-block to share data. Groupings of 32 threads, known as warps, are executed on the physical hardware in each SM. A scheduler logically groups some number of warps into blocks based on user configuration. Each thread has a set of registers available for local variables and performing warp-level collective operations. The number of concurrently executing warps and blocks, as well as the total number of registers and amount of shared memory available to each, are limited by the SM.

GPUs provide the most performance gains when memory access patterns are able to take advantage of features like collective warp-level operations on registers, shared memory, uniform conditional branching, and coalesced memory accesses. Though core-for-core typically not as fast as a CPU, parallelizing operations over GPU threads can still provide significant performance gains even when memory access patterns and the uniformity of computations across threads are not efficient (Harris 2012).

4.2 Constructing the World k -NN Graph

The UMAP-learn library utilizes nearest neighbors descent (Dong, Moses, and Li 2011) for construction of an approximate nearest neighbors graph, however no known GPU-accelerated versions of this algorithm exist at the time of writing. Tree-based approximate variants, such as the algorithms available in Scikit-learn, also don't have a straightforward port to the GPU (Wieschollek et al. 2016). This results, in part, from the iterative nature of traversal, as well as the storage and representation requirements for the trees after they are constructed.

Our implementation of UMAP makes use of the FAISS library (Johnson, Douze, and Jégou 2019) for fast nearest neighbors search on GPUs. Other GPU-accelerated manifold implementations have used this same approach (e.g., t-SNE (Chan et al. 2018)). FAISS provides both exact and approximate methods to nearest neighbors search, the former being used by default in our implementation. We use the exact search provided by FAISS since it is performant and doesn't require underlying device memory be copied during the hand-off.

For smaller datasets of a few hundred thousand samples and a few hundred features, we found the quadratic scale of the exact k -NN graph computation to comprise 26% of the

total time UMAP spends in compute, making it the second largest performance bottleneck next to the optimization of the embeddings. However, as we demonstrate in Figure 2, the k -NN graph can quickly become the largest bottleneck as the number of samples and features increase to larger sizes, while the optimization stage consistently maintains high performance. This is not a surprising find, since the brute force approach requires exhaustive distances to be computed along with a heap, which is required to maintain the sorted order of closest neighbors. An additional cause of significant performance degradation during this stage is FAISS’ incompatibility with outside memory managers, causing unavoidable and expensive synchronous device memory allocations and deallocations for temporary scratch space. We extended the API to accept a k -NN graph that has already been computed. This modification provides a strategy to avoid these expensive synchronizations in exploratory environments where a model might be trained several times on the same data.

4.3 Handling Sparse Data

Once computed, many operations are performed over the sparse k -NN graph. This is common in many modern manifold learning approaches as well as network analysis problems, where these performance optimizations may be reused. The edges of the k -NN graph are unidirectional and the index and distance arrays represent the column and data arrays of a sparse format. The fixed degree makes the row array implicit and allows the use of dense matrix operations until the construction of the fuzzy union, where the degree is no longer fixed.

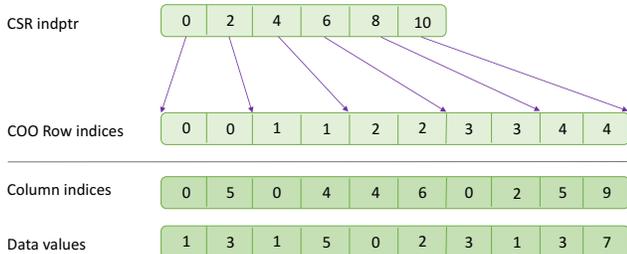


Figure 1: Example of our CSR index being used to index into a sorted COO index.

We use the COOrdinate (COO), or edge list format for efficient out-of-order parallel construction and subsequent element-wise operations. We have found sorting an out-of-order COO can take up to 3% of the total time spent in compute. When it is efficient to do so, we sort the COO arrays by row and create a Compressed Sparse Row (CSR) index into the column and data arrays, enabling both efficient row- and element-wise parallelism so long as the sorted order is maintained. See Figure 1 for a diagram.

While our implementation makes use of libraries like cuSparse and cuGraph for common operations on sparse data, we built several reusable primitives for operations such as reductions, sparse L_1 and L_∞ normalization, removal of nonzeros, and the symmetrization required for computing the triangular conorm, where existing implementations of these oper-

ations were not available. Aside from custom kernels that don’t have much potential for reuse outside of UMAP, such as those described in the following three sections, reusable primitives comprise a large portion of the algorithm.

4.4 Neighborhood Weighting

The neighborhood weighting step begins with constructing the ρ and σ arrays, with one element for each vertex of the k -NN graph. ρ contains the distance to the closest neighbor of each vertex and σ contains the smoothing approximator to the fuzzy set membership function for the local neighborhoods of each source vertex in the k -NN graph. The operations for computing these two arrays are fused into a single kernel, which maps each source vertex of the k -NN graph in CSR format to a separate CUDA thread. The computations in this kernel are largely similar to corresponding Python code in the reference implementation and comprise less than 0.1% of the total time spent in compute.

The k -NN distances are weighted by applying the fuzzy set membership function from the previous step to the COO matrix containing the edges of each source vertex in the k -NN graph. Since this computation requires no dependencies between the edges in the neighborhood graph, the CUDA kernel maps each neighbor to their own thread individually.

As described in Section 3, the final step of the neighborhood weighting stage combines all the fuzzy sets, using the triangular conorm to build a fuzzy union. We implemented this step by fusing both symmetrization sum and product steps together into a single kernel, using the CSR indptr we introduced in Section 4.3 as a Compressed Sparse Column (CSC) indptr to look up the the transposed value and apply the triangular conorm to each element in parallel. This step comprises less than 0.2% of the total time spent in compute.

Larger kernels composed of smaller fused operations, such as computing the mean, min, and iterating for the adaptive smoothing parameters, allowed us to make use of registers where the alternative required intermediate and more expensive storage. We found a 12-15 \times speedup for the adaptive smoothing operations when compared to separate kernels that require intermediate results to be stored in global memory and accessed without memory coalescing. The end-to-end neighborhood weighting stage exploits parallelism at the expense of potential thread divergence from non-uniform conditional branching, and help the kernels to stay compute-bound.

4.5 Embedding Updates

The first step of the embeddings optimization stage initializes the array of output embeddings. We provide both random and spectral initialization strategies. While the reference implementation uses a spectral embedding of the fuzzy union through the nearest-neighbors variant of the Laplacian eigenmaps (Belkin and Niyogi 2002) algorithm, we use the spectral clustering implementation from cuGraph (Fender 2017), setting the number of clusters to 1 and removing the lowest eigenpairs. We have found spectral clustering to be sufficient for maintaining comparable trustworthiness in our experiments while comprising less than 0.1% of the total time spent in compute.

The optimization step performs stochastic gradient descent over the edges of the fuzzy union, minimizing the cross entropy described in Section 3. The gradient computation and update operations have been fused into a single kernel and parallelized so that each thread processes one edge of the fuzzy union. The CUDA kernel is scheduled iteratively for n_epochs to compute and apply the gradient updates to the embeddings in each epoch. The dependencies between the vertices in the updating of the gradients makes this step non-trivial to parallelize efficiently, which decreases potential for coalesced memory access and creates the need for atomic operations when applying gradient updates. As a result, we have seen this kernel take up to 30% of the total time spent in compute for datasets of a few hundred thousand samples with a few hundred features. When the k -NN graph is pre-computed, this step can comprise up to 50% of the remaining time spent in compute. The dependencies between vertices also create challenges to reproducibility, which we describe in Section 4.6.

Both the source and destination vertices are updated for each edge during training. Since the trained embeddings should remain unchanged, only the destination vertex is updated during inference. In addition, both training and inference require the source vertex be updated for some number of randomly sampled vertices. Each source vertex will perform $n_components * (n_negative_samples + 1)$ atomic writes in each thread plus an additional write for the destination vertex during training.

When $n_components$ is small enough, such as a few hundred, we use shared memory to create a small local cache per compute thread, accumulating the updates for each source vertex from multiple negative samples before writing the results atomically to global memory. When shared memory can be used, this reduces atomic updates per thread by a factor of $n_components * n_negative_samples$. We have measured performance gains of 10% for this stage when $n_components = 2$ to 56% when $n_components = 16$ and expect the performance benefits to continue increasing in proportion to $n_components$. For these cases where $n_components$ is very small, such as $n_components = 2$, these updates can be accumulated right in the registers, providing a speedup of 49% for this stage. We suspect these strategies, and any future optimizations, will be useful broadly given the many algorithms (e.g., word2vec (Mikolov et al. 2013b)) that make use of negative sampling.

4.6 Reproducibility

Following the original implementation of UMAP, the user can provide a seed to control the random initialization of weights to increase the reproducibility. This does not eliminate all inconsistency when working with parallel updates made from multiple threads. When using a limited number of CPU cores (≤ 40 in most circumstances), this effect is minimal. However, with a GPU that has thousands of parallel threads, even subtle timing differences between the thread-blocks can have a large impact on the consistency of results. In addition, large numbers of updates can become queued waiting to be performed atomically. A similar issue is observed with the Hogwild algorithm even when atomic up-

dates are used (Zhang, Hsieh, and Akella 2016; Recht et al. 2011; Hsieh, Yu, and Dhillon 2015; Raff and Sylvester 2018; Tran et al. 2015; Chin et al. 2015), but at a larger scale. This problem is further exacerbated by small divergences in the processing of instructions that results from non-uniform conditional branching across threads.

Our use of a local cache to accumulate updates as described in Section 4.5 alleviates this by decreasing the number of global atomic writes, helping to reduce the potential for thread divergence and resulting in higher quality solutions. While this minimizes the writes significantly, we still found the potential for inconsistencies to increase in proportion to the number of vertices in the dataset, the number of edges in the fuzzy union, and the number of components being trained.

The results are made fully repeatable with exact precision by optionally using a 64-bit float array to accumulate the updates to the embeddings and applying the updates at the end of each epoch. The additional precision avoids the numerical instabilities created by repeatedly summing small values in a finite range while the single application of the updates removes the potential for race conditions between reads and writes (Villa et al. 2009). We found the performance impact to increase with the number of components, from an end-to-end slowdown of 11 \times with $n_components = 2$ to 20 \times with $n_components = 16$ on a Volta GV100 GPU.

4.7 Distributed Inference

Because of its ability to embed out-of-sample data points (Bengio et al. 2004), we scaled the UMAP algorithm to support datasets larger than a single GPU by training a model on a random sample of the training dataset, sending the trained embeddings to a set of workers, each mapped to its own GPU, and performing inference in parallel on the remaining data samples. Our implementation minimizes the use of host memory during communication by using CUDA IPC (Potluri et al. 2012) to support fast communication over NVLink (Li et al. 2019) internal to a physical machine and GPUDirect RDMA (Venkatesh et al. 2014) to communicate across machine boundaries. We use the Dask library, which has been GPU-accelerated (Raschka, Patterson, and Nolet 2020) and optimized with the Unified-Communications-X library (UCX) (Shamis et al. 2015) to support CUDA IPC and GPUDirect transports automatically, without the need to invoke the aforementioned transports directly.

We have found our distributed implementation to scale linearly with the number of GPUs and find it can acceptably preserve structure for a small single-cell RNA dataset containing only 23 thousand cells when trained on as little as 3% of the data with less than a 1% drop in trustworthiness. Further, we find only a 0.05% drop in trustworthiness when we embed the remaining 97% of the dataset over 16 separate workers.

5 Experiments

We compare the execution time and correctness of GPUMAP and our implementation against the multi-core implementation of UMAP-learn on CPU. The datasets are summarized in

Dataset	UMAP-Learn		GPUMAP		cuML UMAP	
	$\mu \pm \sigma^2$	Trust%	$\mu \pm \sigma^2$	Trust%	$\mu \pm \sigma^2$	Trust%
digits	6.328 \pm 2.897	98.79	2.483 \pm 1.058	95.58	0.3583\pm0.0111	98.77
fashion mnist	45.87 \pm 10.23	97.81	4.158 \pm 1.800	97.50	0.455\pm0.006	97.73
mnist	52.575 \pm 1.1677	95.94	10.6071 \pm 0.45444	94.43	0.70781\pm0.0088	95.74
cifar100	105.85 \pm 2.482	84.72	6.186 \pm 1.770	84.01	1.009\pm0.0188	83.42
coil20	11.210 \pm 2.571	99.36	2.582 \pm 0.0050	95.67	0.757\pm0.5752	99.28
shuttle	38.88 \pm 8.039	100.0	9.064 \pm 3.431	97.78	0.5825\pm0.0252	100.0
scRNA	223.9 \pm 9.071	62.38	10.89 \pm 1.604	94.35	4.103\pm0.0601	97.81

Table 1: Each result shows mean \pm variance, followed by max trustworthiness score, of each implementation of UMAP for the unsupervised case with default parameters. Fastest result in bold.

Dataset	UMAP-Learn		GPUMAP		cuML UMAP	
	$\mu \pm \sigma^2$	Trust%	$\mu \pm \sigma^2$	Trust%	$\mu \pm \sigma^2$	Trust%
digits	6.756 \pm 0.1109	98.76	2.553 \pm 1.095	95.55	0.4063\pm0.0135	98.80
fashion mnist	53.09 \pm 6.183	97.81	6.477 \pm 0.0632	96.97	1.0370\pm0.0002	97.76
mnist	89.1877 \pm 6.4658	95.85	23.905 \pm 7.057	94.69	0.9175\pm0.00297	95.74
cifar100	98.42 \pm 2.273	84.91	5.954 \pm 0.0236	83.01	1.0816\pm0.0003	83.82
coil20	12.34 \pm 0.0217	98.68	8.210 \pm 0.0275	93.33	0.3695\pm0.0066	98.70
shuttle	50.17 \pm 17.55	100.0	17.15 \pm 23.92	96.67	0.5560\pm0.0111	100.0

Table 2: Each result shows mean \pm variance, followed by max trustworthiness score, of each implementation of UMAP for the supervised case with default parameters. Fastest result in bold.

Dataset	Size
Digits (Garris et al. 1994)	1.7K \times 64
Shuttle (Dua and Graff 2017)	58K \times 9
Fashion MNIST (Xiao et al. 2017)	60K \times 784
MNIST (Deng 2012)	60K \times 784
CIFAR-100 (Krizhevsky 2009)	60K \times 1K
COIL-20 (Nene et al. 1996)	1.4K \times 1.6K
scRNA (Travaglini et al. 2019)	64.5K \times 5K
GoogleNews Word2vec (Mikolov et al. 2013a)	3M \times 300

Table 3: Datasets used in experiments

Table 3, showing the number of rows, columns, and classes. We evaluated the execution times of unsupervised training on each dataset for all three implementations and recorded the resulting times, in seconds. Where classes were provided, we also evaluated the supervised training mode. All experiments were conducted on a single DGX1 containing 8 Nvidia GV100 GPUs with Dual Intel Xeon 20-core CPUs. UMAP-learn was configured to take advantage of all the available threads on the machine.

We use trustworthiness to rank the degree to which local neighborhood structure is preserved between input and embedded spaces. Scikit-learn provides an implementation of this score, but the execution time and memory requirement of computing the pairwise distance matrix make it prohibitive on some of the datasets used in this paper. We implemented a batched GPU-accelerated version of trustworthiness that

Samples	Scikit-learn	cuML UMAP
2k	0.33	0.13
5k	2.06	0.18
10k	8.64	0.24
20k	35.76	0.54
50k	303.08	2.07
100k	FAIL	5.74
1M	FAIL	446.26

Table 4: Execution times for computing the Trustworthiness score with UMAP’s default of $n_neighbors = 15$. The first column shows the number of samples used, and the right two columns present run time in seconds. The number of features was fixed to 1024. Best results are in bold. Done using isotropic blobs

provides reasonably low execution times for datasets up to 1M samples. Table 4 contains the execution times of computing the trustworthiness score on various different numbers of samples in both Scikit-learn and cuML UMAP. This contribution was necessary to perform our evaluations and was used to evaluate the correctness of each implementation. Because users often select the result with the highest trustworthiness score, we report the max score in results.

We begin by demonstrating the speedups obtained by our new cuML UMAP implementation of UMAP in the standard unsupervised scenario. The timing results with standard deviation from 4 runs can be found in Table 1, with the trustwor-

thiness score on the right. cuML UMAP dominates all other implementations in speed, with $17\times$ speedups compared to UMAP-learn on the smallest datasets, and increasing to up to $104.9\times$ on moderate scale datasets like MNIST. Similar results can be seen in the supervised case in Table 2. cuML UMAP is also $2.65 - 15.6\times$ faster than the prior GPUMAP, with an average $7.29\times$ advantage. This is biased towards GPUMAP’s favor by the fact that it has regressions in solution quality, as measured by trustworthiness, on 6/7 datasets.

The trustworthiness and speedups show the value of our contributions in Section 4, and we, in addition, note that the CPU-based UMAP-learn has its own regression on the scRNA dataset. This is a known issue caused by the lack of synchronized updates in its implementation¹, following the hog-wild style update of parameters (Recht et al. 2011). This dataset’s large number of features and datapoints combine to create race conditions that are too significant for correct results. This shows the importance of our register accumulation strategy introduced in Section 4.6, allowing us to obtain better quality results in these extreme cases.

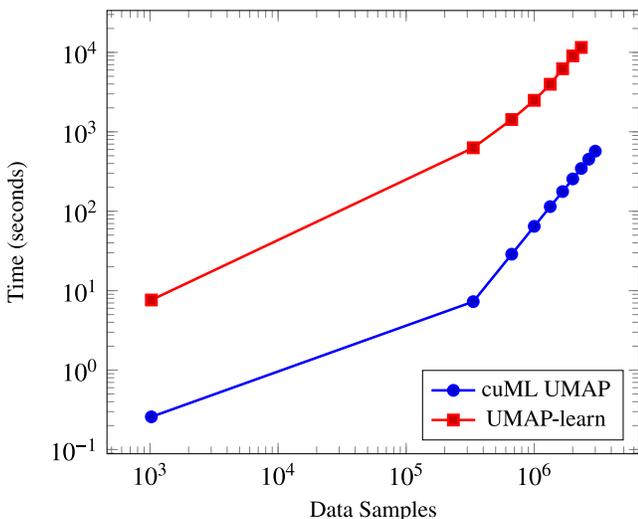


Figure 2: Google-News Results showing runtime (y-axis) as more of the dataset is sampled (x-axis).

The original GPUMAP implementation has, at times, had runtime failures where no results are produced, and significant time was spent attempting to re-compile/fix these issues without success. This prevented its use on our largest dataset, Google-News word2vec embeddings. We use the Google-News corpus in particular as a large-scale experiment to show the value of our results, compared to UMAP-learn up to a time limit of 3 hours. The runtime comparing many cores with UMAP-learn to our cuML UMAP on a single GPU is shown in Figure 2. We can clearly see that cuML UMAP continues to dominate runtime with no loss in quality, obtaining $\geq 30\times$ speedups across $n = 1,024$ samples all the way up to the full 3M samples while UMAP-learn reached its time limit after 2.3M samples.

¹see <https://umap-learn.readthedocs.io/en/latest/reproducibility.html>

Our cuML UMAP’s 9.5 minutes to process all 3 million datapoints of Google-News is already a significant advantage in runtime. In addition, we note that all non k -NN work of the UMAP algorithm took only 9.3 seconds of that total time. This is important for the interactive and hyper-parameter tuning scenarios. Our implementation allows computing the k -NN once, and then other hyperparameter settings can be adjusted with results obtained in seconds. Section 4.7 briefly discusses our distributed UMAP inference algorithm, with preliminary results demonstrating an ability to embed 10M points (including k -NN) across 8 GPUs in just under 5 seconds with only a marginal impact to trustworthiness. These results can speedup tuning and visualization by orders of magnitude, and is enabled by the optimizations we have contributed.

We also note that the consistent speedups on both small and large datasets is important for showing the isoefficiency (Grama, Gupta, and Kumar 1993) of our method. A poorly implemented parallel method may exhibit speedups if sufficiently large amounts of data/work are fed to counter-balance the overheads of communication primitives used. Since we obtain speedups on small datasets, our results are demonstrating good isoefficiency and users should be able to regularly obtain a better runtime by using our method — rather than having to judge if the dataset is “big enough” to make our implementation worth while. These speedups on small datasets are also important for practitioners where the difference between 1 minute and 1 second are noticeable and additionally enables the computer-human interaction usecases we hope to enable with this work.

6 Conclusion

The UMAP algorithm is becoming a widely popular tool, which increases the demand and utility of a faster implementation. We have detailed a number of techniques that are easy to apply in code, and allow us to obtain a solution that is faster and more accurate, even at times compared to the original CPU-based implementation. This obtains up to $100\times$ speedups, and by eliminating all non k -NN calculations to $\leq 2\%$ of runtime, we enable interactive exploration and parameter tuning use-cases that were previously untenable.

Acknowledgements

We extend our sincerest gratitude to all of those who helped enable our research, especially Philip Hynsu Cho and Dante Gama Dessavre from the RAPIDS cuML team as well as Brad Rees, Alex Fender, and Joe Eaton from the RAPIDS cuGraph team. We would also like to thank the Clara Genomics team at Nvidia, especially Avantika Lal, Johnny Israeli, Raghav Mani, and Neha Tadimeti. In addition, we thank Jeff Johnson & Matthijs Douze of the FAISS project for their continued support and Dmitri Kobak, whose single-cell RNA preprocessing scripts were helpful in our evaluations. Finally, we owe our deepest thanks to Leland McInnes & John Healy, because this research would not exist without their contributions.

References

- Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *KDD*, 2623–2631. ACM.
- Andor, N.; Lau, B. T.; Catalanotti, C.; Kumar, V.; Sathe, A.; Belhocine, K.; Wheeler, T. D.; Price, A. D.; Song, M.; Stafford, D.; et al. 2018. Joint single cell DNA-Seq and RNA-Seq of gastric cancer reveals subclonal signatures of genomic instability and gene expression. *bioRxiv* 445932.
- Becht, E.; Dutertre, C.-A.; Kwok, I. W.; Ng, L. G.; Ginhoux, F.; and Newell, E. W. 2018. Evaluation of UMAP as an alternative to t-SNE for single-cell data. *BioRxiv* 298430.
- Belkin, M.; and Niyogi, P. 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NeurIPS*, 585–591.
- Bell, N.; and Hoberock, J. 2012. Thrust: A productivity-oriented library for CUDA. In *GPU computing gems Jade edition*, 359–371. Elsevier.
- Bengio, Y.; François Paiement, J.; Vincent, P.; Delalleau, O.; Roux, N. L.; and Ouimet, M. 2004. Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. In *NeurIPS*, 177–184. MIT Press.
- Beygelzimer, A.; Kakade, S.; and Langford, J. 2006. Cover trees for nearest neighbor. In *ICML*, 97–104. ACM.
- Carter, S.; Armstrong, Z.; Schubert, L.; Johnson, I.; and Olah, C. 2019. Exploring neural networks with activation atlases. *Distill* doi:10.23915/distill.00015.
- Chan, D. M.; Rao, R.; Huang, F.; and Canny, J. F. 2018. t-SNE-CUDA: GPU-Accelerated t-SNE and its Applications to Modern Data. In *2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, 330–338. IEEE.
- Chang, C.-C.; and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3). doi:10.1145/1961189.1961199.
- Chatzimpampas, A.; Martins, R. M.; and Kerren, A. 2020. t-viSNE: Interactive Assessment and Interpretation of t-SNE Projections. *arXiv preprint arXiv:2002.06910*.
- Chen, T. 2020. dmlc/dlpack: RFC for common in-memory tensor structure and operator interface for deep learning system. <https://github.com/dmlc/dlpack>. (Accessed on 06/02/2020).
- Chen, T.; and Guestrin, C. 2016. XGBoost: Reliable Large-scale Tree Boosting System. In *KDD*.
- Chin, W.-S.; Zhuang, Y.; Juan, Y.-C.; and Lin, C.-J. 2015. A Fast Parallel Stochastic Gradient Method for Matrix Factorization in Shared Memory Systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* 6(1): 2:1—2:24. doi:10.1145/2668133.
- Clara-Parabricks, N. 2020. Examples of single-cell genomic analysis accelerated with RAPIDS. <https://github.com/clara-parabricks/rapids-single-cell-examples>. (Accessed on 06/03/2020).
- Davidson, J. W.; and Jinturkar, S. 1994. Memory access coalescing: a technique for eliminating redundant memory accesses. *Acm Sigplan Notices* 29(6): 186–195.
- Deng, L. 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine* 29(6): 141–142.
- Dong, W.; Moses, C.; and Li, K. 2011. Efficient k-nearest neighbor graph construction for generic similarity measures. In *WWW*, 577–586.
- Dorogush, A. V.; Ershov, V.; and Gulin, A. 2017. CatBoost: gradient boosting with categorical features support. In *Workshop on ML Systems at NeurIPS 2017*.
- Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. URL <http://archive.ics.uci.edu/ml>. (Accessed on 03/28/2021).
- Dubois, D.; and Prade, H. 1982. A class of fuzzy measures based on triangular norms a general framework for the combination of uncertain information. *International Journal of General Systems* 8(1): 43–61.
- Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. LIBLINEAR: A Library for Large Linear Classification. *JMLR* 9: 1871–1874.
- Fefferman, C.; Mitter, S.; and Narayanan, H. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society* 29(4): 983–1049.
- Fender, A. 2017. *Parallel solutions for large-scale eigenvalue problems arising in graph analytics*. Ph.D. thesis, Université Paris-Saclay.
- Garris, M. D.; Blue, J. L.; Candela, G. T.; et al. 1994. NIST form-based handprint recognition system. In *Technical Report NISTIR 5469 and CD-ROM, National Institute of Standards and Technology*. Citeseer.
- Grama, A. Y.; Gupta, A.; and Kumar, V. 1993. Isoefficiency: Measuring the Scalability of Parallel Algorithms and Architectures. *IEEE Parallel Distrib. Technol.* 1(3): 12–21. doi:10.1109/88.242438.
- Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; et al. 2020. Array programming with NumPy. *Nature* 585(7825): 357–362. doi:10.1038/s41586-020-2649-2.
- Harris, M. 2012. NVIDIA Developer Blog. URL <https://devblogs.nvidia.com/how-optimize-data-transfers-cuda-cc/>. (Accessed on 03/28/2021).
- He, X.; Cai, D.; Yan, S.; and Zhang, H.-J. 2005. Neighborhood preserving embedding. In *ICCV*, volume 2, 1208–1213.
- Hsieh, C.-J.; Yu, H.-F.; and Dhillon, I. S. 2015. PASSCoDe: Parallel Asynchronous Stochastic Dual Co-ordinate Descent. In *ICML*, 2370–2379.
- Johnson, J.; Douze, M.; and Jégou, H. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*.
- Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. volume 30, 3146–3154.
- Klement, E.; Mesiar, R.; and Pap, E. 1997. Triangular norms. *Tatra Mountains Math. Publ* 13: 169–193.
- Krizhevsky, A. 2009. *Learning multiple layers of features from tiny images*. Master’s thesis, Department of Computer Science, University of Toronto, 40 St. George Street, Room 4283 Toronto, ON M5S 2E4.
- Lam, S. K.; Pitrou, A.; and Seibert, S. 2015. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*.
- LeDem, J. 2017. Apache Arrow and Apache Parquet: Why We Needed Different Projects for Columnar Data, on Disk and In-Memory. www.kdnuggets.com/2017/02/apache-arrow-parquet-columnar-data.html. Accessed: 03/26/2021.
- Li, A.; Mazhar, H.; Serban, R.; and Negrut, D. 2015. Comparison of SPMV performance on matrices with different matrix format using CUSP, cuSPARSE and ViennaCL. Technical report, Technical Report TR-2015-02.

- Li, A.; Song, S. L.; Chen, J.; Li, J.; Liu, X.; Tallent, N. R.; and Barker, K. J. 2019. Evaluating Modern GPU Interconnect: PCIe, NVLink, NV-SLI, NVSwitch and GPUDirect. *IEEE Transactions on Parallel and Distributed Systems* 31(1).
- Luebke, D. 2008. CUDA: Scalable parallel programming for high-performance scientific computing. In *2008 5th IEEE international symposium on biomedical imaging: from nano to macro*, 836–838. IEEE.
- Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR* 9(Nov): 2579–2605.
- McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- McKinney, W.; et al. 2011. pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing* 14(9).
- Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T.; Corrado, G.; Chen, K.; and Dean, J. 2013b. Efficient Estimation of Word Representations in Vector Space. *ICLR*.
- Naumov, M.; Chien, L.; VanderMersch, P.; and Kapasi, U. 2010. Cuspars library. In *GPU Technology Conference*.
- Nene, S. A.; Nayar, S. K.; Murase, H.; et al. 1996. Columbia object image library (coil-20).
- Obermayer, B.; Holtgrewe, M.; Nieminen, M.; Messerschmidt, C.; and Beule, D. 2020. SCellVis: exploratory single cell data analysis on the desktop and in the cloud. *PeerJ* 8: e8607.
- Oesa, A. 2019. SQL for GPU Data Frames in RAPIDS Accelerating end-to-end data science workflows using GPUs. LatinX in AI Research at ICML 2019. URL <https://hal.archives-ouvertes.fr/hal-02264776>. Poster.
- Oden, L. 2020. Lessons learned from comparing C-CUDA and Python-Numba for GPU-Computing. In *2020 28th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, 216–223. IEEE.
- Okuta, R.; Unno, Y.; Nishino, D.; Hido, S.; and Loomis, C. 2017. Cupy: A numpy-compatible library for nvidia gpu calculations. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in NeurIPS*.
- Ordun, C.; Purushotham, S.; and Raff, E. 2020. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*.
- Owens, J. D.; Houston, M.; Luebke, D.; Green, S.; Stone, J. E.; and Phillips, J. C. 2008. GPU computing. *Proceedings of the IEEE* 96(5): 879–899.
- Pachev, I.; and Lupo, C. 2017. GPUMap: A Transparently GPU-Accelerated Python Map Function. doi:10.1145/3149869.3149875.
- Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 8024–8035.
- Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *JMLR* 12: 2825–2830.
- Pezzotti, N.; Mordvintsev, A.; Hollt, T.; Lelieveldt, B. P.; Eisemann, E.; and Vilanova, A. 2018. Linear tsne optimization for the web. *arXiv preprint arXiv:1805.10817*.
- Potluri, S.; Wang, H.; Bureddy, D.; Singh, A. K.; Rosales, C.; and Panda, D. K. 2012. Optimizing MPI communication on multi-GPU systems using CUDA inter-process communication. In *2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum*.
- Raff, E.; and Sylvester, J. 2018. Linear Models with Many Cores and CPUs: A Stochastic Atomic Update Scheme. In *Big Data*, 65–73. doi:10.1109/BigData.2018.8622172.
- Raschka, S.; Patterson, J.; and Nolet, C. 2020. Machine Learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information* 11(4): 193.
- Recht, B.; Re, C.; Wright, S.; and Niu, F. 2011. Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent. In *NeurIPS*, 693–701.
- Roweis, S. T.; and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science* 290(5500).
- Shamis, P.; Venkata, M. G.; Lopez, M. G.; Baker, M. B.; Hernandez, O.; Itigin, Y.; Dubman, M.; Shainer, G.; Graham, R. L.; Liss, L.; et al. 2015. UCX: an open source framework for HPC network APIs and beyond. In *IEEE 23rd Annual Symposium on High-Performance Interconnects*, 40–43.
- Tenenbaum, J. B.; De Silva, V.; and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *science* 290(5500): 2319–2323.
- Tran, K.; Hosseini, S.; Xiao, L.; Finley, T.; and Bilenko, M. 2015. Scaling Up Stochastic Dual Coordinate Ascent. In *KDD*, 1185–1194. doi:10.1145/2783258.2783412.
- Travaglini, K. J.; Nabhan, A. N.; Penland, L.; Sinha, R.; Gillich, A.; Sit, R. V.; Chang, S.; Conley, S. D.; Mori, Y.; Seita, J.; et al. 2019. A molecular cell atlas of the human lung from single cell RNA sequencing. *bioRxiv* 742320.
- Venkatesh, A.; Subramoni, H.; Hamidouche, K.; and Panda, D. K. 2014. A high performance broadcast design with hardware multicast and GPUDirect RDMA for streaming applications on Infiniband clusters. In *21st International Conference on High Performance Computing (HiPC)*, 1–10.
- Venna, J.; and Kaski, S. 2006. Local multidimensional scaling. *Neural Networks* 19(6-7): 889–899.
- Villa, O.; Chavarria-Miranda, D.; Gurumoorthi, V.; Márquez, A.; and Krishnamoorthy, S. 2009. Effects of floating-point non-associativity on numerical computations on massively multi-threaded systems. In *Proceedings of Cray User Group Meeting (CUG)*, 3.
- Wander, L.; Vianello, A.; Vollertsen, J.; Westad, F.; Braun, U.; and Paul, A. 2020. Exploratory analysis of hyperspectral FTIR data obtained from environmental microplastics samples. *Analytical Methods* 12(6): 781–791.
- Wieschollek, P.; Wang, O.; Sorkine-Hornung, A.; and Lensch, H. 2016. Efficient large-scale approximate nearest neighbor search on the gpu. In *CVPR*, 2027–2035.
- Zhang, H.; Hsieh, C.-J.; and Akella, V. 2016. HogWild++: A New Mechanism for Decentralized Asynchronous Stochastic Gradient Descent. In *ICDM*.