RANDOM OFFSET BLOCK EMBEDDING (ROBE) FOR COMPRESSED EMBEDDING TABLES IN DEEP LEARNING RECOMMENDATION SYSTEMS

Aditya Desai [1] **Li Chou** [2] **Anshumali Shrivastava** [1 3]

ABSTRACT

Deep learning for recommendation data is one of the most pervasive and challenging AI workload in recent times. State-of-the-art recommendation models are one of the largest models matching the likes of GPT-3 and Switch Transformer. Challenges in deep learning recommendation models (DLRM) stem from learning dense embeddings for each of the categorical tokens. These embedding tables in industrial scale models can be as large as hundreds of terabytes. Such large models lead to a plethora of engineering challenges, not to mention prohibitive communication overheads, and slower training and inference times. Of these, slower inference time directly impacts user experience. Model compression for DLRM is gaining traction and the community has recently shown impressive compression results. In this paper, we present Random Offset Block Embedding Array (ROBE) as a low memory alternative to embedding tables which provide orders of magnitude reduction in memory usage while maintaining accuracy and boosting execution speed. ROBE is a simple fundamental approach in improving both cache performance and the variance of randomized hashing, which could be of independent interest in itself. We demonstrate that we can successfully train DLRM models with same accuracy while using $1000\times$ less memory. A $1000\times$ compressed model directly results in faster inference without any engineering effort. In particular, we show that we can train DLRM model using ROBE Array of size 100MB on a single GPU to achieve AUC of 0.8025 or higher as required by official MLPerf CriteoTB benchmark DLRM model of 100GB while achieving about $3.1\times$ (209%) improvement in inference throughput.

1 INTRODUCTION

Recommendation systems are one of the top applications of machine learning. For example, Facebook reports that recommendation inference accounts for over 79% of AI inference cycles (Gupta et al., 2020). Therefore, considerable efforts have been and continue to be expended to develop systems that help users make more personalized and well-informed choices in various application domains. Recent approaches utilize deep learning-based models to achieve state-of-the-art performance. However, a key challenge is the need to handle millions of categorical features that dominate the recommendation data (Naumov et al., 2019; Cheng et al., 2016). Following the work in natural language processing (Mikolov et al., 2013; Vaswani et al., 2017), current approaches (Naumov et al., 2019; Wang et al., 2017; Song et al., 2019; Guo et al., 2017; Lian et al., 2018; Huang et al., 2019) utilize a real-valued feature vector (i.e., embedding) to represent each categorical token. These categorical representations are learned, end-to-end, and organized in a memory structure called *embedding* tables.

Production scale models have large storage cost. If the set of all categories is S and the embedding dimension is d, then the embedding table size is $|S|\times d$. With the number of categorical tokens per feature as large as tens of millions, embedding tables consume over 99.9% of total model memory. Specifically, memory footprint for models that utilize embedding tables can easily surpass hundreds of terabytes (TB) (Mudigere et al., 2021; Naumov et al., 2019; Ginart et al., 2019; Shi et al., 2019; Gupta et al., 2020). For example, Facebook recently showcased training of a 50TB sized model distributed over 128 GPUs (Mudigere et al., 2021).

Inference with deep learning recommendation models is memory-bound. Access of embedding tables do not follow any recognizable pattern. Namely, the access is highly irregular. The large size of embedding tables coupled with irregular and sparse access causes high cache miss rates (Gupta et al., 2020). In fact, production scale systems spend 80% of inference cycles in embedding lookups (Gupta et al., 2020). Hence, these models are memory bound.

Training deep learning recommendation models suffers

[1]Department of Computer Science, Rice University, Houston, Texas [2]College of Engineering, West Texas A&M University, Canyon, Texas [3]ThirdAI Corp. Houston, Texas. Correspondence to: Aditya Desai <Aditya.P.Desai@rice.edu>.

Proceedings of the 5^{th} MLSys Conference, Santa Clara, CA, USA, 2022.

from high communication cost. Purely data parallel model training has a high communication cost equal to the size of the model. In case of deep learning recommendation models, data parallel training is infeasible due to large embedding tables. In fact, embedding tables have to be stored in a distributed manner across multiple nodes/GPUs. Thus the model has to be trained in a model parallel fashion introducing communication costs in both the forward and the backward passes. This makes training as well as inference slower for recommendation models.

Training of deep learning recommendation models is not accessible to a general user. Training models with large number of parameters, and on terabytes of data, comes with significant engineering challenges. In addition, such a task requires expensive hardware. Deep learning recommendation models have to be trained in a mixed model and data parallel setting on clusters of nodes or GPUs, which is cost prohibitive. Thus, these models are out of the reach for machine learning users without such access. This also severely restricts the possibility of fast research in this area.

The deep learning recommendation model (DLRM) architecture (Naumov et al., 2019) gave rise to an increased interest in constructing more memory-efficient embeddings. Recent state-of-the-art efforts in this direction include increasing expressive power of embeddings by using additional computing over smaller memory such as compositional embedding (Shi et al., 2019); learning different sized embeddings for different values to leverage the inherent power law in frequencies (Ginart et al., 2019; Joglekar et al., 2020; Liu et al., 2020; 2021; Cheng et al., 2020; Zhao et al., 2020), low rank decomposition of embedding tables (Yin et al., 2021). These approaches show a single ($\approx 10\times$, (Shi et al., 2019; Ginart et al., 2019)) or double order ($\approx 100\times$, (Yin et al., 2021)) of magnitude reduction in embedding table size with no (or minimal) loss of accuracy. In our empirical evaluation, we show that with ROBE Array for DLRM model, we can obtain as much as $1000\times$ compression with similar (or even improved) accuracy, at the same time giving a multi-fold increase in the inference throughput performance. Specifically, we can train $1000\times$ compressed DLRM MLPerf model for CriteoTB dataset which reaches the same MLPerf AUC value 0.8025 or higher with a inference throughput boost of $3.1\times$. Also, similar observations can be made on Criteo Kaggle dataset where $1000\times$ compressed model can achieve similar or better accuracy as original model over variety of state of the art deep learning recommendation models.

What are the implications of 1000× compression of embedding tables?

(1) Eliminate the need of model parallel training. For models as large as 50TBs, a $1000\times$ compression can reduce the model size to 50 gigabytes (GB), which can easily fit on a single high-end GPU (e.g., Nvidia A100). Hence, we can simply run a pure data-parallel model optimization.

(2) 1000× lower communication cost. With pure data-parallel model optimization, we would achieve a $1000\times$ reduction in communication cost at each step of model update. Therefore, this leads to significant savings in communication cost.

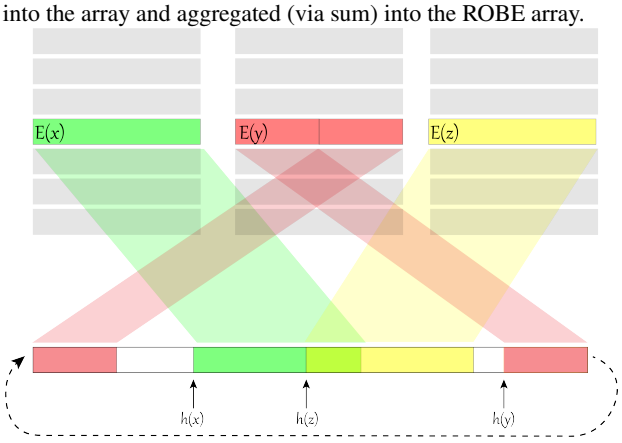
(3) Lower memory latency. In their paper, (Gupta et al., 2020), authors reveal that their production scale recommendation show cache miss rate of 8 MPKI (misses per 1000-instructions) as compared to 0.5 MPKI in RNN, 0.2 MPKI in FC and 0.06 MPKI in CNN. This high cache miss rate is the main cause of higher memory latency. With smaller memory footprint, we can potentially store the embedding tables or large parts of them on chip memory thus thwarting the problem at its root. In our experiments, we show a $3.1\times$ speedup in inference with $1000\times$ compression.

(4) Faster inference times and potentially faster training time Overall, we have the potential to construct compact models that have faster inference and training time. In our experiments, we will show $3.1\times$ improvement in inference throughput. Our proof-of-concept code does not show any improvement in training time per iteration. We leave optimizing training time for future work.

(5) Faster refresh cycle for industrial models. With changing interests, recommendation data suffers from frequent concept shift (Gama et al., 2014). Faster training would imply a better refresh rate for models and thus better service to users of recommendation system.

Our approach: Weight sharing is a widely used idea in machine learning to reduce memory required for the model. Some examples include feature hashing (Weinberger et al., 2009) to reduce input dimension, HashedNet (Chen et al., 2015) to compress fully connected multi-layered neural networks, usage of filters in convolution neural networks, and recently demonstrated some success with LSH based weight sharing in recommendation models (Desai et al., 2021). In this paper, we introduce a memory sharing technique – Random Offset Block Embedding Array (ROBE). We use universal hash functions on chunks/blocks of the embeddings in the embedding table to locate it in a small circular array of memory. We refer to this form of hashing as ROBE hashing. In a standard feature hashing scenario, where we project a vector in higher dimension to lower dimension, ROBE hashing outperforms the usual feature hashing as defined in (Weinberger et al., 2009). We discuss the theoretical results in Section 4. In addition to being theoretically superior, ROBE also leads to better cache performance due to coalesced array access ($3.1\times$ boost in inference throughput). Our results shed new light on how to make randomized hashing algorithms cache friendly, and at the same time,

Figure 1. ROBE-D : Working of ROBE-D (A) **Forward pass:** embedding for a token x is extracted from the location specified by the hash function, $h(x)$. The ROBE array is circular, so the embeddings that overflow are continued from beginning. (B) **Backward Pass:** the gradients of each of the embeddings are mapped back into the array and aggregated (via sum) into the ROBE array.



have superior variance. We also provide precise quantification of various trade-offs involved. The results could be of independent interest to algorithms community working on randomized hashing algorithms.

Caveat: The caveat while using our compression technique in recommendation models is that while training the model, we require more iterations than those required for training the original model. For example, the original CriteoTB MLPerf model (100GB) takes 1 epoch to reach the target AUC of 0.8025, while the same model using 1000× less memory with ROBE Array (100MB) takes 2 epochs to reach the same AUC. We see similar trend in our experiments with the Criteo Kaggle dataset. While we see a clear improvement of inference throughput (3.1×), our current proof-of-concept code does not show training time benefits.

We believe that this 2x epochs might still be cheaper in terms of time, if we consider the post-processing efforts invested in model reduction techniques to make these models leaner. Even if we ignore post-processing costs, with memory optimizations leveraging on 1000× less memory footprint, we believe we should be able to get faster end-to-end training times even while requiring more number of iterations. We leave this aspect for future work.

2 Related Work

The related lines of research in model compression can be broadly classified into two groups: (i) learning a compressed representation (or compression-aware training), and (ii) compressing the learned model via post processing. ROBE-Z extends the former line of research, which has been widely applied to compressing DLRMs. Here, for models, we are referring to recommendation models. In addition, we focus our discussions on compressing embedding tables for these models.

2.1 Learning compressed representations

(1) Low rank decomposition: Low rank decomposition of large matrices is a well known technique to reduce the memory footprint of the model. This entails representing the matrix under consideration, say $A \in \mathbb{R}^{D_1 \times D_2}$, as product of two low rank matrices, $B \in \mathbb{R}^{D_1 \times d}$ and $C \in \mathbb{R}^{d \times D_2}$ where $d \ll D_1, D_2$. In previous research on compressing DLRMs, low rank decomposition was applied in MD Embeddings (Ginart et al., 2019) and TT-Rec (Yin et al., 2021). In both works, embedding tables are first split by grouping together tokens based on their frequency in the dataset. Next, different rank-decompositions are applied to matrices representing parts of the embedding tables belonging to a particular group. The key idea is to use lower rank for tokens that appear sparsely in the dataset and hence leads to more memory saving. TT-Rec uses tensor train decomposition instead of the standard low rank decomposition to optimize for GPU computations. MD Embeddings has sparked a lot of related research for automatically partitioning the table and choosing the ranks of low-rank decomposition for optimal performance (Joglekar et al., 2020; Liu et al., 2020; 2021; Cheng et al., 2020; Zhao et al., 2020). While MD Embeddings show compression of 16× without loss of quality, TT-Rec shows compression of 112× on Criteo Kaggle dataset.

(2) Feature hashing and compositional techniques : An embedding table that is a tall matrix creates the problem of a high dimension input space. This problem has been traditionally solved in machine learning by feature hashing (Weinberger et al., 2009; Shi et al., 2009) where each input value is hashed to a smaller range using a hash function. This is quite similar to low rank decomposition where the first matrix is a fixed sparse matrix defined by the hash function. The authors of QR-Trick (Shi et al., 2019) show that feature hashing does not work well for compressing embedding tables in recommendation systems. The reason is feature hashing forces the embedding of different tokens to be exactly same, thus causing a loss in the quality of model. QR-Trick ensures that each token gets a unique embedding by combining embeddings from multiple smaller embedding tables into a single embedding. The combining operation can be element-wise multiplication/ addition or even concatenation. QR-Trick gives 4× compression with a slight loss in model quality.

(3) HashedNet In their paper (Chen et al., 2015), authors introduced a technique to reduce memory usage of matrices in MLP networks. The weights of matrix are grouped randomly using a xxHash function and all the weights grouped together only use a single value from the underlying memory. This reduces the total memory foot print of the model.

While this scheme is good to compress memory and reduce its footprint, it has some serious issues when it comes to efficiency. HashedNet randomly distributes the elements of a matrix to varied locations. So, in order to access a vector of size, say d, we have to potentially fetch d cache lines utilizing only $1/B$ fraction of bandwidth B. This large wastage of bandwidth can be one of the reasons why the community has not evaluated HashedNet style compression with latency critical application such as recommendation. In this paper, we evaluate HashedNet style compression and propose ROBE-Z scheme which is better than HashedNet both in terms of quality and performance. Essentially, ROBE-Z achieve better quality of approximation than HashedNet while maintaining fraction of bandwidth usage of 1.

(4) Quantization for training (low precision models) : Research in reduced precision models for deep architectures has gained momentum recently (Courbariaux et al., 2016; Gupta et al., 2015; Han et al., 2015; Judd et al., 2016; Reagen et al., 2016). However, the challenges in recommendation models under consideration are unique and these techniques, developed primarily for CNN and RNN cannot naturally extend to recommendation. Recently, Facebook published their effort on using low precision models for DLRM in (Zhang et al., 2018) and shows up to $2\times$ memory savings and $1.2\times$ speed up.

2.2 Compressing learned models

These approaches require us to first train the baseline model and then compress them. Thus, making these approaches less attractive to compression in recommendation models.

(1) Quantization for inference: Quantization can be performed post model training with the goal of reducing the inference time. The idea behind this quantization is to convert the floating point values to smaller representations, (e.g., int16 and int8), and replace floating point operations to integer operations, which are known to be faster. This approach can be applied in conjunction with the earlier approaches (learning compressed representation based) and ROBE to improve performance further.

(2) Pruning: Pruning (Liu et al., 2018) compresses the model by removing edges from the computational graph of the model. It enables faster inference for models by reducing the computation. There is no straight forward way to apply pruning to compression of embedding tables and has not been explored in literature.

(3) Knowledge distillation: Knowledge distillation as proposed by (Hinton et al., 2015), is a way of training a smaller model (called student) from a larger model (called teacher). Generally, the student model trained in this way outperforms the same model when trained standalone. One can imagine training a smaller dimensional embedding table from a larger embedding table. This approach has not been evaluated in the literature and can be explored further in an independent manner.

3 Random Offset Block Embedding Array

The memory footprint of the model is determined by the memory used to store the parameters of the model. In the case when the number of parameters far exceed the total amount of memory we intend to use, there are approaches such as mixed-precision learning (Zhang et al., 2018), low rank decomposition (Yin et al., 2021; Ginart et al., 2019) or specialized methods (Shi et al., 2019) used to fit the parameters in the memory. To achieve order of magnitude more reduction in memory footprint of the model, we share memory among the elements of embeddings. Weight sharing scheme to compress MLP networks was proposed in HashedNet (Chen et al., 2015) but was never evaluated on embedding tables. We evaluate HashedNet in our experiments and propose a weight sharing scheme that is provably better than standard weight sharing defined by HashedNet, both in terms of quality and performance.

Instead of storing an embedding table, we maintain a single array for learned parameters which is a compressed representation of embedding table. All embedding tables share the same array of learned parameters. The embeddings are accessed in a blocked manner from the embedding array using GPU-friendly universal hashing. We call this scheme of embedding compression as Random Offset Block Embedding Array (ROBE). As we will see in Section 4, in learning a shared memory array via ROBE, we can expect to get good quality models even with very high compression. This is further supported by our experiments in Section 5. What's more, ROBE surmounts the memory bandwidth issues created by HashedNet style hashing by making coalesced access.

The section is organized as follows. We first describe the most useful form of ROBE i.e. ROBE-D where D is the dimension of the embedding in embedding table. We then generalize the approach to consider ROBE-Z for $Z \in \mathbf{N}$. In the next subsection, we contrast ROBE-Z with HashedNet style weight sharing. We end this section with a discussion on advantages of ROBE-Z.

3.1 ROBE-D : ROBE with block size equal to embedding size.

Consider that we are looking to build an embedding table of size $|S| \times D$ with embedding size D. We use a circular array $\mathcal{M}$ to store the learned parameters. Let $h : \mathbf{N} \rightarrow \{0, .., m-1\}$ be a hash function drawn uniformly randomly from a universal hash family. Similarly let $g : \mathbf{N} \times \mathbf{N} \rightarrow \{-1, 1\}$

be an independent hash function drawn from a different hash family with range 2. The working of ROBE-D is illustrated in the figure 1.

Forward Pass: The embedding for a given token x as a whole is located in the ROBE array using a universal hashing function, h. In case, the $h(x) + D \geq |\mathcal{M}|$, the embedding is continued in the first part of the ROBE array. The embedding can, optionally, be multiplied element-wise with a value from $\{+1, -1\}$ as obtained via the hash function $g(x, i)$. Let vector $\mathbf{G}(x) = \{g(x, 1), g(x, 2), ..., g(x, D)\}$. If P(x) is a primary embedding obtained from the ROBE array. Then the final embedding can be considered as $\mathbf{E}(x) = \mathbf{G}(x) \circ \mathbf{P}(x)$ where $\circ$ is element wise multiplication. Thus, we can write

$$\begin{aligned}
&\mathbf{P}(x) = \mathcal{M}[h(x) : h(x) + D] \quad \text{if} \quad h(x) + D < |\mathcal{M}| \\
&\mathbf{P}(x) = p_1.p_2 \quad \text{if} \quad h(x) + D \geq |\mathcal{M}| \\
&\qquad where \\
&\qquad\quad p_1 = \mathcal{M}[h(x) : |\mathcal{M}|] \\
&\qquad\quad p_2 = \mathcal{M}[0 : (D - (|\mathcal{M}| - h(x)))] \\
&\mathbf{E}(x) = \mathbf{G}(x) \circ \mathbf{P}(x)
\end{aligned} \tag{1}$$

where "." denotes concatenation.

Multiple embedding tables: As discussed earlier, all embedding tables in the system, even with varying embedding dimensions, will share the ROBE array. In order to achieve independent locations for embeddings in each table, the hash function h and g is modified to include embedding table id as a parameter $h : N \times N \to \{0, ..., m - 1\}$, $g : N \times N \times N \to \{+1, -1\}$

Hashing functions h, g and idx **mapping**. We use universal hash function families to choose h and g. The significant aspect of universal hash functions is that they are cheap to compute, GPU-implementation friendly and provide decent guarantee-bound on collision probability. With universal hash functions for h and g the mapping, idx, for element i of the embedding for a token x from embedding table e can be written as,

$$\begin{aligned}
&h(e, x) = ((A_h e + B_h x + C_h) \mod P) \mod |\mathcal{M}| \\
&\text{idx}(e, x, i) = (h(e, x) + i) \mod |\mathcal{M}| \\
&g(e, x, i) = 2(\\
&\qquad ((A_g e + B_g x + C_g i + D_g) \mod P \mod 2 \\
&\qquad) - 1
\end{aligned} \tag{2}$$

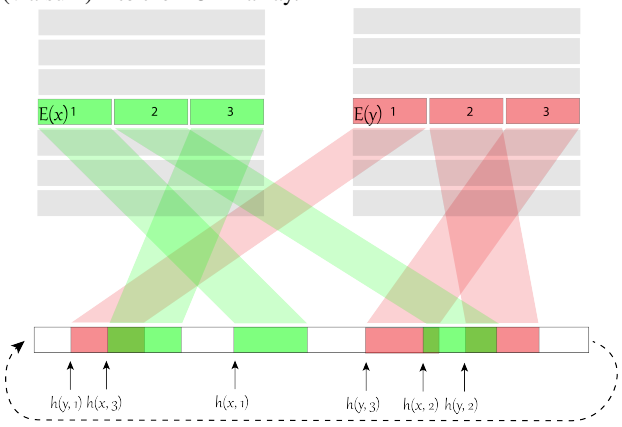
where P is a large prime and $A_h, B_h \in \{1, ...P - 1\}, C_h \in \{0, ..., P - 1\}$ are randomly chosen values. Thus, i^{th} element of the embedding of x from embedding table e can be written as $\mathcal{M}[\text{idx}(e, x, i)] * g(x, i)$

Backward Pass: The backward pass can also be illustrated with the same figure 1. Essentially, the gradients of the embeddings are mapped into the ROBE array according to the idx and undergo a signed aggregation.

$$\Delta(\mathcal{M}[j]) = \sum_{\text{idx}(e,x,i)=j} g(e, x, i)\Delta(\mathbf{E}_e(x)[i]) \tag{3}$$

where $\mathbf{E}_e$ represents the embedding table with table id e and $\Delta(p)$ is the gradient of the loss function w.r.t the parameter p.

Figure 2. ROBE-Z : Working of ROBE-Z (A) **Forward pass:** embedding for a token x is assembled by extracting chunks of memory from the locations specified by the hash function, $h(x, c_{id})$. The ROBE array is circular, so the embeddings that overflow are continued from beginning. (B) **Backward Pass:** the gradients of each of the embeddings are mapped back into the array and aggregated (via sum) into the ROBE array.



3.2 ROBE-Z : ROBE with arbitrary block size $Z \leq D$

We generalize ROBE-D to ROBE-Z with arbitrary block size. In this case, the embedding of a particular token is split into chunks of size Z. These chunks are independently mapped into the ROBE array using universal hash functions similar to ROBE-D. The procedure is illustrated in figure 2. Borrowing notation from the previous section,

$$\begin{aligned}
&\mathbf{P}_c(x) = \mathcal{M}[h(x, c) : h(x, c) + Z] \text{ if } h(x, c) + Z < |\mathcal{M}| \\
&\mathbf{P}_c(x) = p_1.p_2 \quad \text{if} \quad h(x) + Z \geq |\mathcal{M}| \\
&\qquad where \\
&\qquad\quad p_1 = \mathcal{M}[h(x) : |\mathcal{M}|] \\
&\qquad\quad p_2 = \mathcal{M}[0 : (Z + h(x) - |\mathcal{M}|)] \\
&\mathbf{P}(x) = \mathrm{P}_{c_1}.\mathrm{P}_{c_2}.\mathrm{P}_{c_3}...\mathrm{P}_{c_{D/Z}} \\
&\mathbf{E}(x) = \mathbf{G}(x) \circ \mathbf{P}(x)
\end{aligned} \tag{4}$$

where "." denotes concatenation. The idx mapping for a multiple table ROBE-Z can be written as,

$$\begin{aligned}
&\mathcal{Z}_{id}(i) = \lfloor i/Z \rfloor \quad \mathcal{O}_{id}(i) = i \mod Z \\
&h(e, x, c) = ((A_h e + B_h x + C_h c + D) \mod P \mod |\mathcal{M}| \\
&\text{idx}(e, x, i) = (h(e, x, \mathcal{Z}_{id}(i) + \mathcal{O}_{id}(i))) \mod |\mathcal{M}|
\end{aligned} \tag{5}$$

Table 1. Number of memory fetches on varying sizes of Z, where D is the embedding size, B is the bus size and Z is the block size. $B|D$ denotes B divides D.

	Condition	Max number of memory fetches	Comment
Original	$B\|D$	$D/B+1$	
HashedNet/ROBE-1		D	
ROBE-Z $Z < B < D$	$Z\|B\|D$	$2 \times D/Z$	With high probability as $\|M\| \gg d > Z$
ROBE-Z $B < Z < D$	$B\|Z\|D$	$D/B + D/Z$	With high probability as $\|M\| \gg d > Z$
ROBE-Z $Z \geq D$	$D\|Z$	$D/B+2$	

where P is a large prime and $A_h, B_h, C_h \in \{1, ...P-1\}, D_h \in \{0, ..., P-1\}$ are randomly chosen values. In the equations above, $\mathcal{Z}_{id}(i)$ represents the chunk id of the index i and $\mathcal{O}_{id}(i)$ computes the offset within the chunk of i.

The expression for g(x, e, i) is same as in ROBE-D. Also, the backward pass functions similarly to ROBE-D according to equation 3

3.3 ROBE-1 vs HashedNet

ROBE-1 hashing scheme is similar to the hashing scheme proposed by HashedNet (Chen et al., 2015) for compressing matrices in MLP network. There are some differences though. ROBE-1 uses light weight universal hashing as opposed to xxHash used by HashedNet. Thus, ROBE-1 compromises the collision guarantees for better performance. Using universal hashing makes implementing the computation on GPU very convenient and efficient. Additionally, HashedNet, demonstrated on MLP networks, keeps separate arrays for separate matrices, whereas ROBE-1 use a single array to map all the elements from all the embedding tables. What is most exciting about ROBE approach is its ROBE-Z (or ROBE-D) which is theoretically superior to hashing proposed by HashedNet and is cache efficient due to appropriate usage of cache-lines via coalesced access.

The setup can also be extended to $Z > D$ by clubbing multiple embeddings together in a chunk. The formulation follows the same scheme as shown in sections 3.1 and 3.2 This actually leads to better feature hashing quality as shown in section 4.

3.4 Advantages of ROBE-*Z*

Memory Latency and Issue of Irregular memory access As mentioned in (Gupta et al., 2020), recommendation models suffer a very high cache-miss rate due to large embedding tables and irregular access as compared to other architectures. ROBE-Z can partially solve this problem by potentially storing large part of the embedding table (or even entire embedding table) in a compressed format in LLC . For example, embedding tables with a collective size of 100GB, when allocated a memory of 100MB (i.e. $1000\times$ reduction), can be stored on last level cache. The original model, in this case, without any memory sharing has to be stored on RAM, or even worse on disk.

Better compute intensity In their paper, (Gupta et al., 2020), authors highlight low compute intensity as one of the unique challenges in embedding tables. With reusing a lot of memory locations, ROBE-Z improves the compute intensity of the embedding tables.

Memory Fetches: The number of memory fetches can potentially increase when using a ROBE-Z allocation scheme, especially worse during ROBE-1 (or HashedNet). The reason is wasting band-width of cache line. We present the number of cache-line fetches while using ROBE-Z and compare it against the memory fetches with using original embedding and HashedNet, which is shown in Table 1.

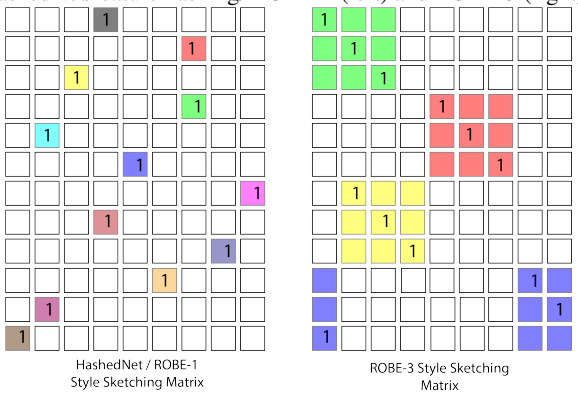
Consider the original embedding of size D (generally kept in multiples of cache-line size). Let the cache-line size be B. Thus, in order to fetch a single embedding from original embedding table, we would require a maximum of $(D/B + 1)$ (+1 for non-aligned access) memory fetches. As we can see from Table 1, as we increase the value of Z the number of cache-line fetches decrease from the $2D/Z$ to $D/B + 2$ due to the coalesced access pattern when Z is greater than D. Also, as we will see in section 4, the greater the value of Z, the better is dimensionality reduction. So it is advisable to choose a large value for Z.

Dimensionality Reduction: As we will see in Section 4, ROBE-Z hashing is better than ROBE-1 in terms of dimensionality reduction. As the value of Z increases, while the estimate of inner products in projected space is unbiased, the variance decreases until Z reaches $|M|$.

4 Theoretical Considerations

The procedure described in ROBE-Z is closely related to the sketching literature, and in particular, the area of random projections. A *parameter vector* can be created by joining all the flattened embedding matrices. The ROBE-Z hashing, essentially, projects this parameter vector into a $\mathbb{R}^{|\mathcal{M}|}$ space. We know from Johnson-Lindenstrauss Lemma, that random projections can provide us with low-dimensional and

Figure 3. The mapping from parameter vector in R^n to smaller memory vector R^m can be described by a sketching matrix of size $m \times n$. Here, we show sketching matrix without sign for HashedNet/feature hashing/ ROBE-1 (left) and ROBE-3 (right).



low-distortion embeddings of vectors from high dimension. Feature hashing is an efficient form of random projection where the sketching matrix is sparse - i.e. each row of the matrix has exactly one non-zero (usually ± 1) and this location is determined randomly. We can visualize the sketching matrix for ROBE-Z as shown in Figure 3.

Using the mapping function, or alternatively the sketching matrix, we can recover the original embedding vector from the memory. In fact, in this paper, we directly learn the compressed representation of the parameter vector (hence embedding tables).

We provide two analysis. (1) The first analysis measures the quality of the dimensionality reduction while using ROBE-Z hashing. This is a standard analysis on the lines of that presented in (Weinberger et al., 2009). We show that ROBE-Z (with $Z > 1$) is better than ROBE-1 which is essentially the feature hashing described in (Weinberger et al., 2009).

(2) The previous papers such as HashedNet only evaluate their method on dimensionality reduction. However, depending on how the compressed memory is being used, we believe it is important to also measure the application specific effect of compression. Hence, we analyse the quality of embedding structure maintained by the ROBE-Z hashing in the projected space. In this analysis, we measure how the relation between two embeddings is maintained under this memory allocation scheme.

4.1 Dimensionality reduction : ROBE-Z beats feature hashing

In order to assess the quality of dimensionality reduction of the parameter vector in $\mathbb{R}^n$, we look at the estimation of the inner product of two vectors in the projected space. This is a standard way to measure the preservation of distances under projection.

Let x and y be two parameter vectors in $\mathbb{R}^n$. Note that these are not embedding vectors but two parameter vectors. Let the inner product between x and y be denoted as $\langle x, y \rangle$. Let ROBE-Z sketching matrix be a $n \times m$ matrix projecting the vectors in space of $\mathbb{R}^n$ to $\mathbb{R}^m$ such that $m < n$. Let the projected vectors be $\hat{x}$ and $\hat{y}$ respectively. The projections are obtained by using two hash functions h and g. h is the memory allocation function described in section above, which maps the block id into the range $\{0, ..., m-1\}$, and g assigns a value in $\{+1, -1\}$ to each index of the vector. The final mapping obtained is idx as described in section 3.2. Note that while hash function h applies to blocks, g gives independent values to each index within the block as well. We do not use g in practice, but here we use g to simplify the analysis. To begin with, the inner product estimator can be written as

$$\widehat{\langle x, y \rangle} = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} x_i g(i) \mathbb{1}(\text{idx}(i) == j) \right) \left(\sum_{i=1}^{n} y_i g(i) \mathbb{1}(\text{idx}(i) == j) \right)$$

where $\mathbb{1}(\cdot)$ is an indicator function. The expectation and variance of the inner product estimator can be written as below (Theorem 1). Here, we use the function $\mathcal{Z}_{id}$ directly on the index of parameter vector as opposed to how it is defined in Equation (5) when we are dealing with embedding tables. Both notations are equivalent and we use them interchangeably depending on context.

Theorem 1 *The inner product of two parameter vectors $x, y \in \mathbb{R}^n$ projected into the space $\mathbb{R}^m$ using the sketching matrix for ROBE-Z with block size Z, $m < n$ and $Z < m$ is a random variable with expectation and variance as noted below. Let $\mathcal{Z}_{id}(i)$ denote the block id of index i as computed in Equation (5).*

$$\mathbb{E}(\widehat{\langle x, y \rangle}) = \langle x, y \rangle$$

$$\mathbb{V}(\widehat{\langle x, y \rangle}) = \frac{1}{m}(\Sigma_{\mathcal{Z}_{id}(i) \neq \mathcal{Z}_{id}(j)} x_i^2 y_j^2 + \Sigma_{\mathcal{Z}_{id}(i) \neq \mathcal{Z}_{id}(j)} x_i y_i x_j y_j)$$

Let $\mathbb{V}_Z(x, y, n, m)$ be the variance of inner product between $x, y \in \mathbb{R}^n$ while using ROBE-Z on memory size $|M| = m$. Then the above equation can be rewritten as follows.

$$\mathbb{V}_Z(x, y, n, m) = \mathbb{V}_1(x, y, n, m) - \sum_{i=0}^{n/Z-1} \mathbb{V}_1(x[Zi : Z(i+1)], y[Zi : Z(i+1)], Z, m),$$

where $x[i : j]$ refers to slice of vector from index i to j.

Note that $\mathbb{V}_1(x, y, n, m)$ is exactly the variance observed with ROBE-1 or feature hashing matrix (Weinberger et al., 2009). As we can clearly see that ROBE-Z has lower variance than ROBE-1. Hence, ROBE-Z is better at dimensionality reduction than feature hashing.

Intuition: The results are not surprising and can explained by observing that once we hash blocks, we ensure that elements of parameter vector that lie within a particular block do not collide under random projection (benign correlations due to blocking). Also, the marginal probability of collision of two elements that lie in different blocks is same as that in ROBE-1. While there is additional constraint in ROBE-Z of the form "if i and j collide then $i+1$ and $j+1$ also collide if they lie in same blocks as i and j respectively," these relations do not affect the variance as can be seen in detailed analysis in the Appendix. The exciting part is that the improved variance also comes with improved cache performance.

The analysis can be extended to get concentration inequalities on the lines of the analysis provided by (Weinberger et al., 2009).

4.2 Embedding structure preservation under projection

Let a and b be two tokens and their corresponding embedding vectors be $\vec{\theta_a}, \vec{\theta_b} \in \mathbb{R}^D$. Let θ be the parameter vector, then $\theta_a = \theta[id_a : id_a + D]$ for some id_a. We assess the quality of embedding structure preservation under the ROBE-Z projection by measuring the inner product estimation between θ_a and θ_b under the projection. The inner product estimator can be written as

$$\widehat{\langle \vec{\theta_a}, \vec{\theta_b} \rangle} = \sum_{j=1}^{D} \left(\sum_{i=1}^{n} \theta_i g(id_a + j) g(i) \mathbb{1}(\text{idx}(i) == \text{idx}(id_a + j)) \right) \left(\sum_{i=1}^{n} \theta_i g(id_b + j) g(i) \mathbb{1}(\text{idx}(i) == \text{idx}(id_b + j)) \right).$$

Theorem 2 *The inner product for embeddings of two distinct values a and b, say $\theta_a, \theta_b \in \mathbb{R}^D$, when the parameter vector $\theta \in \mathbb{R}^n$ is projected onto a space R^m using the ROBE-Z hashing has an expected value as shown below.*

$$\mathbb{E}(\widehat{\langle \vec{\theta_a}, \vec{\theta_b} \rangle}) = \langle \theta_a, \theta_b \rangle,$$

when a and b have embeddings in same block.

$$\mathbb{E}(\widehat{\langle \vec{\theta_a}, \vec{\theta_b} \rangle}) = \langle \theta_a, \theta_b \rangle (1 + \frac{1}{m}),$$

when a and b have embeddings in different blocks or $Z < D$. The variance for the case when embeddings of a and b lie in separate blocks and $Z \geq D$ can be expressed as

$$\mathbb{V}(\langle \vec{\theta_a}, \vec{\theta_b} \rangle) \leq \mathbb{O}\left(\frac{D}{m^2} ||\theta||^4 + \frac{1}{m^2} (\sum_{i \neq j=1}^{D} \theta_a^{(i)} \theta_b^{(i)} \theta_a^{(j)} \theta_b^{(j)}) \right)$$

We provide variance of inner product estimator for a commonly occuring case. Detailed proof of Theorem 2 can be found in the Appendix. Variance for other cases can be computed similarly. We factor out the dependence on Z to simplify the expression. However , it can be noted that if $Z < D$ there will be additional interactions which will increase the variance.

Intuition: The expected value of inner product of two embeddings is unbiased only in the case when embeddings of a and b lie in the same block. This is expected as when embeddings of a and b lie in different blocks, i^{th} element of a and b can be potentially mapped to the same location in memory leading to biased estimates. Also, the variance depends on the $\ell 2$-norm of the parameter vector (equivalently frobenius norm of embedding tables). Again, one can expect this as every element can potentially collide with the elements of embeddings of a and b. It is interesting to note that while dimensionality reduction's variance was proportional to $\frac{1}{m}||\theta||^2$, the variance in this case is proportional to $(\frac{1}{m}||\theta||^2)^2$. One can expect this to happen as in this analysis, vectors share the same memory (as opposed to dimensionality reduction).

5 Experimental Results

We evaluate ROBE-Z for embedding tables in deep learning recommendation models in this section. Subsection 5.1 compiles various baselines from literature which try to compress embedding tables for various recommendation datasets. One can clearly see that ROBE-Z provides the best compression of embedding tables that is orders of magnitude better than previous state of the art. In the section 5.2 we show that quality of model holds for different values of Z on CriteoTB dataset for Facebook DLRM model. In the section 5.3, we show that the results of ROBE-Z holds on Criteo Kaggle dataset for a set of state-of-art recommendation models proving the generalized success of ROBE-Z. In the last two subsections 5.4 and 5.5, we discuss the effect of ROBE-Z on inference and training times respectively. Specifically, we show that using ROBE-Z gives $3.1\times$ more throughput during inference. The two datasets are described in sections 5.2 and 5.3.

5.1 Comparison with baselines

We select the following baselines to compare our results against. More details on these methods can be found in the related work section of the paper. The results are aggregated in the table 2

- **Quantization with low precision models:** In (Zhang et al., 2018), authors note that low precision embedding tables gives upto $2\times$ reduction in embedding tables by using FP16 instead of FP32.
- **Hashing Trick** : This is the popular form of input size reduction technique as introduced in (Weinberger et al., 2009). In this technique, each category value is hashed

Table 2. Comparison of ROBE-Z and other baselines on Criteo datasets.

Method	Dataset	Memory Compression	Quality of model	Metric Used
Low-Precision(Zhang et al., 2018)	Criteo Kaggle	2×	similar to baseline	Logloss
Hashing Trick (Weinberger et al., 2009)	Criteo Kaggle	4×	Much worse than QR Trick	Logloss
QR Trick (Shi et al., 2019)	Criteo Kaggle	4×	Slightly worse than baseline	Logloss
MD Emeddings (Ginart et al., 2019)	Criteo Kaggle	16×	Better or Similar to baseline	Logloss
TT-Rec (Yin et al., 2021)	Criteo Kaggle	112×	Better or similar to baseline	Accuracy/logloss
TT-Rec	Criteo TB	117×	Better or similar to baseline	Accuracy/logloss
ROBE-Z	Criteo Kaggle	1000×	Better or similar to baseline	AUC/logloss
ROBE-Z	Criteo TB	1000×	Better or similar to baseline	AUC/logloss

Table 3. Criteo Kaggle dataset: Test AUC of 1000× reduced ROBE-Z model (2MB embedding) compared against original model (2GB). While all models overfit after a time, DLRM models do not and the training is cutoff at 11 epochs. All values of Z in ROBE-Z show similar std-deviation. So we only write one of them.

Model	Original #1	ROBE-1 (avg) (seed 1,2,3)	ROBE-1 (stdev)	ROBE-2 (seed 1)	ROBE-4 (seed 1)	ROBE-8 (seed 1)	ROBE-16 (seed 1)
DLRM	0.8031	0.8032	0.0001	**0.8050**	0.8049	0.8047	**0.8050**
DCN	0.7973	0.7991	0.0004	0.7994	**0.7995**	0.7994	0.7993
AUTOINT	0.7968	0.7987	0.0002	0.7984	0.7984	**0.7988**	0.7985
DEEPFM	**0.7957**	0.7951	0.0001	0.7949	0.7949	0.7947	0.795
XDEEPFM	**0.8007**	0.7989	0.0004	0.7987	0.7988	0.799	0.7991
FIBINET	**0.8016**	0.8011	0.0002	0.8011	0.8010	0.8013	0.8012

to a smaller range and all values that are mapped to a single value use the same embedding. As reported by the (Shi et al., 2019), hashing trick performs quite worse than baseline with a compression of 4×

- **Compositional Embeddings (QR Trick)** : This technique ensures that each value has a unique embedding by composing different chunks from shared pool of embeddings. As reported in (Shi et al., 2019), at 4× compression, we see slight drop in quality of the model as compared to original on Criteo Kaggle dataset.
- **Mixed Dimensional (MD) Embeddings** : We choose MD Embeddings (Ginart et al., 2019) as a representative technique of different techniques (Joglekar et al., 2020; Liu et al., 2020; 2021; Cheng et al., 2020; Zhao et al., 2020) to choose mixed dimensional embeddings for compression. The paper reports 16× compression with results similar to the baseline on criteo Kaggle dataset.
- **TT-Rec** (Yin et al., 2021) employs a tensor-train decomposition of the embedding matrix. As reported in (Yin et al., 2021), it shows little over 100× compression on both Criteo Kaggle and Criteo TB datasets. This is an impressive improvement in memory usage as compared to its predecessors.

As can be seen in table 2, ROBE-Z provides orders of magnitude improvement in memory compression while maintaining the quality of the model. Thus ROBE-Z beats the state of the art compression results on DLRM models.

Table 4. CriteoTB dataset: 1000× reduced memory with ROBE-Z for varying Z. AUC was reached in 1.89 epochs for all settings.

Model(100MB embedding)	0.8025 AUC reached?
ROBE-1	Yes
ROBE-8	Yes
ROBE-32	Yes
ROBE-128	Yes

5.2 1000× Compression of CriteoTB MLPerf Model with AUC 0.825 or higher with varying values of Z

Dataset: CriteoTB dataset has 13 integer features and 26 categorical features with around 800 million categorical tokens in total. This is the advertising data of 23 days published by criteo. We use exactly same setting as mentioned for official version by MLPerf for training.

Model: The official MLPerf model for DLRM on CriteoTB (see section A in appendix for code details) requires around **100GB** sized embedding tables and achieves the target MLPerf AUC of 0.8025 in 1 epoch. We will use the same quality metric of 0.8025 AUC as prescribed in MLPerf settings for CriteoTB dataset to evaluate ROBE-Z.

Results: With ROBE-Z using 1000× less memory, i.e. **only 100MB**, we achieve higher than 0.8025 AUC within 2 epochs with different settings of block sizes. The details are given in table 4 As we can see we can achieve the same target AUC, although with almost 2x time in terms of iter-

Table 5. Sample throughput: run with a batch size of 16384. The time includes the time taken to send the batch from RAM to GPU global memory and then the forward pass on the batch. The time also includes hash computation. There is 120% increase in throughput by using ROBE-1 which can be further improved using ROBE-32 to 209%. Original model is run on 4 QUADRO RTX 8000 GPUs while ROBE-Z models are run on a single GPU. All models have access to 120 CPUs. (CPUs are not involved in the measured computation though)

Model	samples/second	Improvement
Original(100GB)	341454	-
ROBE-1	755469	121%
ROBE-2	865757	153%
ROBE-8	913893	167%
ROBE-32	920183	170%
ROBE-128	1055470	209%

Table 6. Training iterations required for original embedding model and ROBE-Z 1000x compressed model. As can be seen, there is a consistent increase in number of iterations required for compressed model. (1) Matches the performance of the original model (2) Reaches its best performance. This is used when the model does not outperform the original model.

Model	**Dataset**	**Original epochs to reach best**	**ROBE-Z 1000× epochs to reach (1) or (2)**
DLRM	Criteo TB	1	1.94 (1)
DLRM	Criteo Kaggle	1.37	3.96 (1)
DCN	Criteo Kaggle	1	1.8 (1)
AutoInt	Criteo Kaggle	1	1.62 (1)
DeepFM	Criteo Kaggle	1	1.99 (2)
XDeepFM	Criteo Kaggle	1.625	3.93 (2)
FiBiNET	Criteo Kaggle	3	2.99 (2)

ations. We experiment with different block sizes and see that we can achieve the required quality with different block sizes. The results can be reproduced using our code. (see section A in appendix for code details)

5.3 1000× Compression of Embedding Tables on Criteo Kaggle Dataset

For more comprehensive study on different state-of-the-art recommendation models, we use criteo kaggle dataset.

Dataset: The Criteo Kaggle dataset (see section A in appendix) has 13 integer features and 26 categorical features with 33.7M total categorical values. It is similar to CriteoTB dataset with lesser number of days and different sampling strategy. We split the data randomly into partitions 9:1, the smaller partition being used for testing. The training partition is further divided into partitions 8:2, the smaller partition being used for validation. We use early stopping based on validation AUC to choose the model. **Models** We use six different embedding based models from the literature: DLRM (Naumov et al., 2019), DCN (Wang et al., 2017), AutoInt (Song et al., 2019), DeepFM (Guo et al., 2017), xDeepFM (Lian et al., 2018), and FiBiNET (Huang et al., 2019). The exact details of hyperparameters for the models and optimizer parameters, data split used for testing, and properties of the dataset used can be found in Appendix C. Specifically, we use embedding size of 16 for all the categorical values (around 33.7M). Hence, the original models have 540M parameters. We use Adam(Kingma & Ba, 2014) for all models except DLRM which uses SGD as provided in original code. In this experiment, we set the compressed memory size to 540K parameters for ROBE-Z (i.e. 1000× compression)

Results: Table 3 shows the results of AUC and along with the corresponding standard deviations for all the models in Table 3. The standard deviations of AUC of all settings are pretty similar and we exclude putting the results for other models (original and ROBE-Z for $Z > 1$) to save space.

We make the following observations from Table 3.

- Test AUC of ROBE-Z 1000 × compressed model is either better than original model (3/6 models) or similar (2/6 models). Only in case of XDeepFM , ROBE-Z performs worse than original model.
- The quality of model (i.e. AUC) reached is stable across different values of Z for ROBE-Z.

Our results can be reproduced using the DLRM code for DLRM model and deep-torch code for other models(see section A in appendix for code details)

5.4 Inference Time for ROBE-Z

With our proof-of-concept code (experimental and unoptimized), we measure the throughput of the samples during inference on CriteoTB dataset. We can see a phenomenal improvement in throughput during inference. While original 100GB model, run on 4 Quadro RTX-8000 (46GB) GPUs, can process around 341K samples per second, the ROBE-Z models which are only 100MB large, perform much faster. Using ROBE-Z we can process about 3.1× samples. Specifically, we see 120%(2.2×) improvement in throughput with $Z = 1$. As expected, increasing value of Z in ROBE-Z improves the throughput further upto 209% (3.1×) for ROBE-128.

5.5 Training Time for ROBE-Z

Table 6 shows the running time of ROBE-Z models in terms of number of epochs needed to reach the best AUC or match the performance of original models. We note that training ROBE-Z models is 2-3× slower w.r.t number of epochs required to reach the same quality. This can potentially be

explained by the fact that having more parameters to tune can significantly speed up the learning. We see that with recommendation models with large embedding tables, we can achieve same quality with smaller compressed ROBE-Z given enough training time. A lot of research on recommendation models on click-through rate (CTR) data like Criteo, restrict themselves to 1 epoch of training. However, we want to stress that smaller models can potentially reach same quality and in these cases and it is just a matter of training more.

Our current proof-of-concept code does not show any training time per iteration improvement and thus as of now, training for ROBE-Z models is slower than original models. However, we believe that leveraging the smaller size of overall embeddings the training for these smaller ROBE-Z, we can improve the training time per iteration as well and potentially improve the wall-clock training times.

6 Conclusion

While industrial scale recommendation models are exploding due to large number of categorical features, ROBE Array is a perfect alternative to embedding tables and enable training models of $1000\times$ less memory to achieve same quality. ROBE Array also shows clear inference throughput benefit and can potentially be trained much faster than original models. Also, training models with ROBE Array is accessible to a average machine learning user who does not have access to high end hardware or engineering expertise required to train hundreds of TBs sized model. We believe DLRM with ROBE Array will serve as a new baseline for compression and expedite the research in recommendation models.

Acknowledgements

This work was supported by National Science Foundation IIS-1652131, BIGDATA-1838177, AFOSR-YIP FA9550-18-1-0152, ONR DURIP Grant, ONR BRC grant on Randomized Numerical Linear Algebra, and gift grants from Intel and VMware. We want to acknowledge Yanzhou Pan and Kuangyuan Sun for their initial code on which our code was built.

References

Chen, W., Wilson, J., Tyree, S., Weinberger, K., and Chen, Y. Compressing neural networks with the hashing trick. In *International conference on machine learning*, pp. 2285–2294. PMLR, 2015.

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pp. 7–10, 2016.

Cheng, W., Shen, Y., and Huang, L. Differentiable neural input search for recommender systems. *arXiv preprint arXiv:2006.04466*, 2020.

Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., and Bengio, Y. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

Desai, A., Pan, Y., Sun, K., Chou, L., and Shrivastava, A. Semantically constrained memory allocation (scma) for embedding in efficient recommendation systems. *arXiv preprint arXiv:2103.06124*, 2021.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

Ginart, A., Naumov, M., Mudigere, D., Yang, J., and Zou, J. Mixed dimension embeddings with application to memory-efficient recommendation systems. arXiv:1909.11810, 2019.

Guo, H., Tang, R., Ye, Y., Li, Z., and He, X. Deepfm: a factorization-machine based neural network for ctr prediction. *arXiv preprint arXiv:1703.04247*, 2017.

Gupta, S., Agrawal, A., Gopalakrishnan, K., and Narayanan, P. Deep learning with limited numerical precision. In *International conference on machine learning*, pp. 1737–1746. PMLR, 2015.

Gupta, U., Wu, C.-J., Wang, X., Naumov, M., Reagen, B., Brooks, D., Cottel, B., Hazelwood, K., Hempstead, M., Jia, B., et al. The architectural implications of facebook's dnn-based personalized recommendation. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 488–501. IEEE, 2020.

Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Huang, T., Zhang, Z., and Zhang, J. Fibinet: combining feature importance and bilinear feature interaction for click-through rate prediction. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 169–177, 2019.

Joglekar, M. R., Li, C., Chen, M., Xu, T., Wang, X., Adams, J. K., Khaitan, P., Liu, J., and Le, Q. V. Neural input search for large scale recommendation models. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2387–2397, 2020.

Judd, P., Albericio, J., Hetherington, T., Aamodt, T. M., and Moshovos, A. Stripes: Bit-serial deep neural network computing. In *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 1–12. IEEE, 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lian, J., Zhou, X., Zhang, F., Chen, Z., Xie, X., and Sun, G. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1754–1763, 2018.

Liu, H., Zhao, X., Wang, C., Liu, X., and Tang, J. Automated embedding size search in deep recommender systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2307–2316, 2020.

Liu, S., Gao, C., Chen, Y., Jin, D., and Li, Y. Learnable embedding sizes for recommender systems. *arXiv preprint arXiv:2101.07577*, 2021.

Liu, Z., Sun, M., Zhou, T., Huang, G., and Darrell, T. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*, 2018.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. arXiv:1301.3781, 2013.

Mudigere, D., Hao, Y., Huang, J., Tulloch, A., Sridharan, S., Liu, X., Ozdal, M., Nie, J., Park, J., Luo, L., et al. High-performance, distributed training of large-scale deep learning recommendation models. *arXiv preprint arXiv:2104.05158*, 2021.

Naumov, M., Mudigere, D., Shi, H. M., Huang, J., Sundaraman, N., Park, J., Wang, X., Gupta, U., Wu, C., Azzolini, A. G., Dzhulgakov, D., Mallevich, A., Cherniavskii, I., Lu, Y., Krishnamoorthi, R., Yu, A., Kondratenko, V., Pereira, S., Chen, X., Chen, W., Rao, V., Jia, B., Xiong, L., and Smelyanskiy, M. Deep learning recommendation model for personalization and recommendation systems. arXiv:1906.00091, 2019.

Reagen, B., Whatmough, P., Adolf, R., Rama, S., Lee, H., Lee, S. K., Hernández-Lobato, J. M., Wei, G.-Y., and Brooks, D. Minerva: Enabling low-power, highly-accurate deep neural network accelerators. In *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, pp. 267–278. IEEE, 2016.

Shi, H. M., Mudigere, D., Naumov, M., and Yang, J. Compositional embeddings using complementary partitions for memory-efficient recommendation systems. arXiv:1909.02107, 2019.

Shi, Q., Petterson, J., Dror, G., Langford, J., Smola, A., and Vishwanathan, S. Hash kernels for structured data. *J. Mach. Learn. Res.*, 10:2615–2637, December 2009. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1577069.1755873.

Song, W., Shi, C., Xiao, Z., Duan, Z., Xu, Y., Zhang, M., and Tang, J. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1161–1170, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. arXiv:1706.03762, 2017.

Wang, R., Fu, B., Fu, G., and Wang, M. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD*, 2017.

Weinberger, K., Dasgupta, A., Langford, J., Smola, A., and Attenberg, J. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1113–1120, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553516. URL https://doi.org/10.1145/1553374.1553516.

Yin, C., Acun, B., Wu, C.-J., and Liu, X. Tt-rec: Tensor train compression for deep learning recommendation models. *Proceedings of Machine Learning and Systems*, 3, 2021.

Zhang, J., Yang, J., and Yuen, H. Training with low-precision embedding tables. In *Systems for Machine Learning Workshop at NeurIPS*, volume 2018, 2018.

Zhao, X., Wang, C., Chen, M., Zheng, X., Liu, X., and Tang, J. Autoemb: Automated embedding dimensionality search in streaming recommendations. *arXiv preprint arXiv:2002.11252*, 2020.

A REPRODUCING RESULTS

We rely on the following repositories of code

- DLRM Patch(to run dlrm model on kaggle and criteotb dataset)[1]
- robe-z code[2]
- kaggle challenge data[3]
- deep-torch code (to run multiple state-of-the-art models on kaggle dataset) [4]
- Official Model for CriteoTB[5]
- Reproduce TB/kaggle results with DLRM model:
  - Install the robe-z code
  - apply dlrm patch on dlrm original code
  - setup data for criteoTB/kaggle
  - see the md file in dlrm file for commands
- Reproduce kaggle results on other models:
  - Install the robe-z code
  - use the deep-torch code
  - run the criteo_train file with appropriate config as provided in md file

B ROBE-Z

The parameter vector is constructed by flattening out the embedding table row-wise and concatenate all the embedding tables. let all the embeddings be of dimension d and let the chunk size be Z.

The ROBE-Z is performed as follows

- split the parameter vector into chunks of Z.
- hash each chunk to a particular location in the array of size m
- This chunk is added to the corresponding sub-array of the memory and in case we run outside of m we cycle through to add at the beginning of the array. So the array we are sketching the parameter vector is actually a circular array
- each element is actually multiplied by the sign which is obtained by using another hash function g() and this is applied at the element level. We do not use the sign in our experiments. However, it can be used and greatly simplifies the theory.

[1] https://github.com/apd10/dlrm
[2] https://github.com/apd10/universal_memory_allocation
[3] https://www.kaggle.com/c/criteo-display-ad-challenge
[4] https://github.com/apd10/criteo_deepctr
[5] https://github.com/facebookresearch/dlrm/tree/6d75c84d834380a365e2f03d4838bee464157516

B.1 Analysis 1: Analysis of quality of dimensionality reduction - feature hashing

Let the parameter vector be in $\mathbb{R}^n$. The projection maps this vector into $\mathbb{R}^m$.

Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ The estimator we want to analyse is that of inner product estimation . The estimator can be written in terms of the indicator functions $\mathbb{1}(\cdot)$ as follows:

$$\widehat{\langle x, y \rangle} = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} x_i g(i) \mathbb{1}(h(i) == j) \right) \left(\sum_{i=1}^{n} y_i g(i) \mathbb{1}(h(i) == j) \right).$$

We can simplify the above indicator as

$$\widehat{\langle x, y \rangle} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i y_j \mathbb{1}(h(i) == h(j)), \tag{6}$$

$$\widehat{\langle x, y \rangle} = \langle x, y \rangle + \Sigma_{i \neq j} x_i y_j \mathbb{1}(h(i)==h(j)) g(i) g(j). \tag{7}$$

Let $\mathcal{C}_i$ be the chunk-id that is assigned to to i, following the same notation used in Equations (5). Then, we know that $\mathbb{1}(h(i)==h(j))$ is 0 if $\mathcal{C}_i == \mathcal{C}_j$. Using this fact, we have

$$\widehat{\langle x, y \rangle} = \langle x, y \rangle + \Sigma_{\mathcal{C}_i \neq \mathcal{C}_j} x_i y_j \mathbb{1}(h(i)==h(j)) g(i) g(j). \tag{8}$$

We can easily see that this estimator of $\langle x, y \rangle$ is unbiased. Let us now look at

$$\mathbb{E}(\widehat{\langle x, y \rangle}) = \langle x, y \rangle. \tag{9}$$

The variance of the estimator can be computed as

$$\mathbb{V}(\widehat{\langle x, y \rangle}) = E((\widehat{\langle x, y \rangle} - \langle x, y \rangle)^2)$$
$$= \mathbb{E}((\Sigma_{\mathcal{C}_i \neq \mathcal{C}_j} x_i y_j \mathbb{1}(h(i)==h(j)) g(i) g(j))^2),$$
$$\mathbb{V}(\widehat{\langle x, y \rangle}) = \mathbb{E}((\sum_{\mathcal{C}_i \neq \mathcal{C}_j, \mathcal{C}_{i'} \neq \mathcal{C}_{j'}} x_i y_j$$
$$\mathbb{1}(h(i)==h(j)) g(i) g(j)) x_{i'} y_{j'} \mathbb{1}(h(i')==h(j')) g(i') g(j')$$
$$)).$$

The expected value of term in summation above is non-zero only if pairs are equal to eliminate the gs. As i cannot be equal to j , $i = j', i' = j$ or $i = i', j' = j$. This implies that

$$\mathbb{V}(\widehat{\langle x, y \rangle}) =$$
$$\mathbb{E}((\sum_{\mathcal{C}_i \neq \mathcal{C}_j} x_i^2 y_j^2 \mathbb{1}(h(i)==h(j)) + x_i y_i x_j y_j \mathbb{1}(h(i)==h(j)).$$

Note that although there are some constraints that appear when we do chunk hashing. like if i and j collide then and i and j lie within the chunk, then $i+1$ and $j+1$ will also collide. But you can see that this relation does not really appear in the equation for variance. Maybe this appears in higher moments of the estimator.

Using the fact that $\mathbb{E}(g(i)) = 0$, we can simplify the expression above as

$$\mathbb{V}(\widehat{\langle x, y \rangle}) = \frac{1}{m}(\Sigma_{\mathcal{C}_i \neq \mathcal{C}_j} x_i^2 y_j^2 + \Sigma_{\mathcal{C}_i \neq \mathcal{C}_j} x_i y_i x_j y_j). \quad (10)$$

Note that when $Z = 1$, the equation for variance is exactly the random projection that is used for "feature hashing" as proposed by (Weinberger et al., 2009).

Let us denote the variance of inner product of vectors x and y projected from the dimension n to m while using chunk size of Z to be $\mathbb{V}(x, y, Z, n, m)$. We will use this notation so that we are very precise in our statements.

Note that
$$\mathbb{V}(x, y, 1, n, m) =$$
$$\frac{1}{m}(\Sigma_{\mathcal{C}_i \neq \mathcal{C}_j} x_i^2 y_j^2 + \quad \Sigma_{\mathcal{C}_i \neq \mathcal{C}_j} x_i y_i x_j y_j)$$
$$+ \frac{1}{m} \sum_{c \in chunks} (\Sigma_{i_c \neq j_c} x_{i_C}^2 y_{j_c}^2 + \Sigma_{i_c \neq j_c} x_{i_c} y_{i_c} x_{j_c} y_{j_c}),$$

where x_{i_c} is an element of sub-vector x_c, which refers to the chunk of the parameter vector.

$$\mathbb{V}(x, y, 1, n, m) = \mathbb{V}(x, y, Z, n, m) + \sum_{c \in chunks} \mathbb{V}(x_c, y_c, 1, Z, m) \quad (11)$$

It is clear from the above equation that ROBE-Z has better variance w.r.t feature hashing .

B.2 Effect of ROBE-Z on inner product of embeddings of two values - i.e. parts of parameter vector that are identified as two separate embeddings.

The previous analysis was the analysis of the sketching matrix and how good it is in preserving distances in a space. However, another important aspect of this projection - pertinent to the discussion of this paper is how does this projection of entire parameter vector affect the inter- embedding relation between embeddings of two different values.

Consider how the parameter vector is constructed. We flatten out each embedding table row wise (so each embedding is contiguous) and we concatenate all flattened embedding tables together to get a single parameter vector which is projected down.

There are three cases that we need to check. We will assume that either Z divides d or d divides Z (if $Z > d$) also, both Z and d divided n. Let the embeddings of two values under consideration be x and y

- x and y lie in same chunk ($Z > d$)
- x and y lie in different chunks $Z > d$
- $Z < d$

CASE: $Z > d$, x and y lie in the same chunk $\mathcal{C}$

Let us first look at the product of two elements $x_1 y_1$.

$$\hat{x_1} = x_1 + \sum_{i=1, i \notin \mathcal{C}}^{n} \theta_i g(x_1) g(i) \mathbb{1}(h(i) == h(1)) \quad (12)$$

$$\hat{y_1} = y_1 + \sum_{i=1, i \notin \mathcal{C}}^{n} \theta_i g(y_1) g(i) \mathbb{1}(h(i) == h(y_1)) \quad (13)$$

$$\hat{x_1 y_1} = x_1 y_1 +$$
$$x_1 (\sum_{i=1, i \notin \mathcal{C}}^{n} \theta_i g(y_1) g(i) \mathbb{1}(h(i) == h(y_1))) +$$
$$y_1 (\sum_{i=1, i \notin \mathcal{C}}^{n} \theta_i g(x_1) g(i) \mathbb{1}(h(i) == h(1))) +$$
$$\sum_{i=1, \notin \mathcal{C}, j=1, j \notin \mathcal{C}} \theta_i, \theta_j g(x_1) g(y_1) g(i) g(j)$$
$$\mathbb{1}(h(i) = h(x_1)) \mathbb{1}(h(j) == h(y_1)) \quad (14)$$

$$\mathbb{E}(\widehat{x_1 y_1}) = xy \quad (15)$$

Hence,

$$\mathbb{E}(\widehat{\langle x, y \rangle}) = \langle x, y \rangle \quad (16)$$

CASE : $Z > d$ or $Z \leq d$, x and y lie in different chunks $\mathcal{C}_1$ and $\mathcal{C}_2$

Let us first look at the product of two elements $x_1 y_1$.

$$\hat{x_1} = x_1 + \sum_{i=1, i \notin \mathcal{C}_1}^{n} \theta_i g(x_1) g(i) \mathbb{1}(h(i) == h(1)) \quad (17)$$

$$\hat{y_1} = y_1 + \sum_{i=1, i \notin \mathcal{C}_2}^{n} \theta_i g(y_1) g(i) \mathbb{1}(h(i) == h(y_1)) \quad (18)$$

$$
\begin{aligned}
x_1\hat{y}_1 &= x_1y_1 + \\
&x_1(\sum_{i=1,i\notin\mathcal{C}_2}^{n}\theta_i g(y_1)g(i)\mathbb{1}(h(i)==h(y_1))) + \\
&y_1(\sum_{i=1,i\notin\mathcal{C}_1}^{n}\theta_i g(x_1)g(i)\mathbb{1}(h(i)==h(1))) + \\
&\sum_{i=1,\notin\mathcal{C}_1,j=1,j\notin\mathcal{C}_2}\theta_i,\theta_j g(x_1)g(y_1)g(i)g(j) \\
&\qquad \mathbb{1}(h(i)=h(x_1))\mathbb{1}(h(j)==h(y_1))
\end{aligned} \tag{19}
$$

$$
\mathbb{E}(\widehat{x_1y_1}) = xy(1+\frac{1}{m}) \tag{20}
$$

Hence,

$$
\mathbb{E}(\widehat{\langle x,y\rangle}) = \langle x,y\rangle(1+\frac{1}{m}) \tag{21}
$$

Variance:

We will analyse the variance for specific case of $Z > E = d$ and x and y values have embeddings in different blocks. Other cases can be computed similarly as

$$
\begin{aligned}
&\widehat{\langle\vec{\theta_x},\vec{\theta_y}\rangle} = \\
&\sum_{j=1}^{d}\left(\sum_{i=1}^{n}\theta_i g(idx_x+j)g(i)\mathbb{1}(h(i)==h(idx_x+j))\right) \\
&\quad\left(\sum_{i=1}^{n}\theta_i g(idx_y+j)g(i)\mathbb{1}(h(i)==h(idx_y+j))\right)
\end{aligned}
$$

$$
\begin{aligned}
(\widehat{\langle\vec{\theta_x},\vec{\theta_y}\rangle})^2 = &\sum_{j_1=1}^{d}\sum_{j_2=1}^{d}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{i_3=1}^{n}\sum_{i_4=1}^{n} \\
&\theta_{i_1}\theta_{i_2}\theta_{i_3}\theta_{i_4} \\
&g(i_1)g(i_2)g(i_3)g(i_4) \\
&g(idx_x+j_1)g(idx_y+j_1)g(idx_x+j_2)g(idx_y+j_2) \\
&\mathbb{1}(h(i_1)=h(idx_x+j_1))\mathbb{1}(h(i_2)=h(idx_y+j_1)) \\
&\mathbb{1}(h(i_3)=h(idx_x+j_2))\mathbb{1}(h(i_4)=h(idx_y+j_2))
\end{aligned} \tag{22}
$$

separating cases when ($i_1 = idx_x + j_1$, $i_2 = idx_y + j_1$, $i_3 = idx_x + j_2$, $i_4 = idx_y + j_2$) and others

$$
\begin{aligned}
(\widehat{\langle\vec{\theta_x},\vec{\theta_y}\rangle})^2 = &\sum_{k_1=1}^{d}\sum_{k_2=1}^{d}x_{k_1}y_{k_1}x_{k_2}y_{k_2} + \\
&\sum_{j_1=1}^{d}\sum_{j_2=1}^{d}\sum_{i_1=1,i_1\neq idx_x+j_1}^{n}\sum_{i_2=1,i_1\neq idx_y+j_1}^{n} \\
&\sum_{i_3=1,i_3\neq idx_x+j_2}^{n}\sum_{i_4=1,i_4\neq idx_y+j_2}^{n} \\
&\theta_{i_1}\theta_{i_2}\theta_{i_3}\theta_{i_4} \\
&g(i_1)g(i_2)g(i_3)g(i_4) \\
&g(idx_x+j_1)g(idx_y+j_1)g(idx_x+j_2)g(idx_y+j_2) \\
&\mathbb{1}(h(i_1)=h(idx_x+j_1))\mathbb{1}(h(i_2)=h(idx_y+j_1)) \\
&\mathbb{1}(h(i_3)=h(idx_x+j_2))\mathbb{1}(h(i_4)=h(idx_y+j_2))
\end{aligned} \tag{23}
$$

$$
\begin{aligned}
&\mathbb{V}(\langle\vec{\theta_x},\vec{\theta_y}\rangle) = \\
&\mathbb{E}(\sum_{j_1=1}^{d}\sum_{j_2=1}^{d}\sum_{i_1=1,i_1\neq idx_x+j_1}^{n}\sum_{i_2=1,i_1\neq idx_y+j_1}^{n} \\
&\sum_{i_3=1,i_3\neq idx_x+j_2}^{n}\sum_{i_4=1,i_4\neq idx_y+j_2}^{n} \\
&\theta_{i_1}\theta_{i_2}\theta_{i_3}\theta_{i_4} \\
&g(i_1)g(i_2)g(i_3)g(i_4) \\
&g(idx_x+j_1)g(idx_y+j_1)g(idx_x+j_2)g(idx_y+j_2) \\
&\mathbb{1}(h(i_1)=h(idx_x+j_1))\mathbb{1}(h(i_2)=h(idx_y+j_1)) \\
&\mathbb{1}(h(i_3)=h(idx_x+j_2))\mathbb{1}(h(i_4)=h(idx_y+j_2)))
\end{aligned} \tag{24}
$$

We will analyse the case of embeddings of x and y lie in separate blocks and $Z > d$.

For convenience we are using $x = \vec{\theta_x}$ and similarity for y

$$
\begin{aligned}
\mathbb{V}(\langle\vec{\theta_x},\vec{\theta_y}\rangle) \leq\ & d\frac{1}{m^2}\sum_{i=1}^{n}\theta_i^4 + 2d\frac{1}{m^4}\sum_{i\neq j}\theta_i^2\theta_j^2 \\
&+ d\frac{1}{m^2}\sum_{i\neq j}\theta_i^2\theta_j^2 + 2(\frac{1}{m^2}+\frac{1}{m^3})\sum_{i\neq j=1}^{d}x_iy_ix_jy_j)
\end{aligned} \tag{25}
$$

We use less than as not all elements from θ are present in actual summation.

$$
\mathbb{V}(\langle\vec{\theta_x},\vec{\theta_y}\rangle) \leq \mathbb{O}\left(\frac{d}{m^2}||\theta||^4 + \frac{1}{m^2}(\sum_{i\neq j=1}^{d}x_iy_ix_jy_j)\right) \tag{26}
$$

		term	factor	comment
j1 = j2	i1=i2=i3=i4	t^4	1/m^2	The value of i_1, i_2, i_3, i_4 should be equal
	i1 = i2 != i3=i4	t1^2 t2^2	1/m^4	in pairs or all of them. Also, they should not match their
	i1 = i3 != i2 = i4	t1^2 t2^2	1/m^2	own elements. (i.e. i_1 != x+j1, etc)
	i1 = i4 != i2 =i3	t1^2 t2^2	1/m^4	t = typical element of theta
j1!=j2	i1 = y + j1, i3 = y + j2 , i2 = x + j1, i4=x+j2	x1y1x2y2	1/m^2	
	i1 = y + j1, i3 = y + j2 , i2 = x + j2, i4=x+j1	x1y1x2y2	1/m^3	The value of i_1, i_2, i_3, i_4 should be equal to
	i1 = y + j2, i3 = y + j1 , i2 = x + j1, i4=x+j2	x1y1x2y2	1/m^3	the 4 distinct elements (2 of x and 2 of y) without
	i1 = y + j2, i3 = y + j1 , i2 = x + j2, i4=x+j1	x1y1x2y2	1/m^2	matching their own elements.
				Also, as the indicators should be non-zero
				So,
				i_1 cannot be x+j1 also it cannot be x+j2 as the value of
				indicator in that case will always be 0. Since Z >d h(i_1=x+j2)
				cannot be equal to h(x+j1)
				Hence i_1 can be equal to y+j1 or y+j2
				x,y are vectors

Table 7. For x and y lie in different blocks and Z >d . so each embedding lies completely inside the block.

C CRITEO KAGGLE EXPERIMENT

The experimental settings are described in detail below

Dataset: We choose the Criteo Kaggle dataset in order to demonstrate the compressive power of UMA. The original dataset of Criteo has 13 integer features and 26 categorical features. The counts of the feature values in total is around 33M. the breakup in each of the feature is as follows.

```
counts:
[1460, 583, 10131227, 2202608, 305, 24,
12517, 633, 3, 93145, 5683, 8351593, 3194,
27, 14992, 5461306, 10, 5652, 2173, 4,
7046547, 18, 15, 286181, 105, 142572]
```

We do not perform any rare feature filtering, which reduces the number of categorical values as is done in papers reporting SOTA values for the Criteo dataset. We want to demonstrate an algorithm that can deal with large embedding tables (e.g., terabytes-size of embedding tables as observed in industries) and choose the entire feature set of the Criteo dataset to make a reasonable dataset. Also, this is standard practice in papers which deal with "compression" or efficient embedding tables research (Ginart et al., 2019; Shi et al., 2019).

Hyperparameters chosen for different models. The hyperparameters for running different models were chosen as mentioned in the respective original papers. We did not do hyperparameter tuning for UMA as the result sufficiently supports our hypothesis that these memory allocation approaches are valuable to efficiently learn compressed embedding tables. We fix the batch size of each of training to 2048 and cutoff the training at 300K iterations for all models except DLRM which is run till 540K (due to SGD). All the models have embedding size = 16.

Train/Test data split. We use the following way to split the data into random training and testing data for all models except DLRM. For DLRM, the code provided has their own data loader and we do not make any changes there.

```
from sklearn.model_selection \\
    import train_test_split
train, test = train_test_split \\
 (data, test_size=0.1,\\
      random_state=2020)
```

Choosing models while training. We observe that most models (especially the original models) start to overfit after some iterations. Hence, we use early stopping based on validation AUC to select the final model.

C.1 FAQ on Kaggle Experiment:

1. **Why do we not perform rare feature filtering?** We do not perform any sort of rare feature filtering which reduces the number of categorical values as is done in papers reporting SOTA values for criteo kaggle dataset. It is important to note that such tricks that are optimized on smaller public benchmarks do not help in the real production scale datasets (Gupta et al., 2020) and hence are avoided. We want to demonstrate an algorithm that can deal with large embedding tables (e.g., terabytes-size of embedding tables as observed in industries) and choose the entire feature set of the Criteo dataet to make a reasonable dataset. Also, this is a standard in papers which deal with "compression" or efficient embedding tables research (Ginart et al., 2019; Shi et al., 2019).

2. **Why do we not achieve the SOTA results (as reported on original papers)?** Reproducing results in the original papers would require access to their codes and dataset preprocessing if any and dataset split they use to create training and testing datasets. Unfortu-

Table 8. Hyperparameters of different model chosen as per the specification in their papers. The code used for running DLRM model is : https://github.com/facebookresearch/dlrm . For other models, the code used is https://github.com/shenweichen/DeepCTR-Torch.

	architecture	dropout	l2_regularization	optimizer	learning rate
DLRM	bot: 13-512-256-64-16 top: 512-256-1	0	0	SGD	1.0
DCN	1024-1024-1024-1024	0	0	ADAM	0.001
AutoInt	400-400-400 attention_embedding_size: 32	0	0	ADAM	0.001
DeepFM	400-400-400	0.5	0	ADAM	0.001
XDeepFM	dnn: 400-400-400 cross interaction: 200-200-200	0.5	0.0001	ADAM	0.001
FiBiNET	400-400-400	0.5	0.0001	ADAM	0.001

nately, most papers do not have their own codes public and do not specify the random seeds used to split the data. In our experiments, we use the random seed = 2020 for all the models except DLRM which has its own random splitter in the code provided. Apart from this, we also do not perform rare feature filtering which might affect the results. However, our experiments on some models with rare feature filtering showed that it does not help with performance of original model.

3. **Why we use fixed embedding size for models like DCN, which tell us to use custom embedding sizes?** While it is true that DCN model gives a custom way to choose the size of the embedding based on the number of values in that category, using the formula leads to very large memory tables for the Criteo dataset using full features (no rare feature filtering). Hence, it is not possible to use the custom formula in our case. We uniformly set the embedding size to 16 across the different models.