

Binary Code based Hash Embedding for Web-scale Applications

Bencheng Yan*, Pengjie Wang*, Jinquan Liu, Wei Lin, Kuang-Chih Lee, Jian Xu and Bo Zheng†

Alibaba Group

{bencheng.ybc,pengjie.wpj,vjinquan.ljq,kuang-chih.lee,xiyu.xj,bozheng}@alibaba-inc.com,lwsaviola@163.com

ABSTRACT

Nowadays, deep learning models are widely adopted in web-scale applications such as recommender systems, and online advertising. In these applications, embedding learning of categorical features is crucial to the success of deep learning models. In these models, a standard method is that each categorical feature value is assigned a unique embedding vector which can be learned and optimized. Although this method can well capture the characteristics of the categorical features and promise good performance, it can incur a huge memory cost to store the embedding table, especially for those web-scale applications. Such a huge memory cost significantly holds back the effectiveness and usability of EDRMs. In this paper, we propose a binary code based hash embedding method which allows the size of the embedding table to be reduced in arbitrary scale without compromising too much performance. Experimental evaluation results show that one can still achieve 99% performance even if the embedding table size is reduced 1000× smaller than the original one with our proposed method.

CCS CONCEPTS

- Information systems → Recommender systems.

KEYWORDS

Embedding Learning, Web-scale Application, Hash Embedding

1 INTRODUCTION

Embedding learning for categorical features plays an important role in embedding-based deep recommendation models (EDRMs) [3, 7, 23]. A standard method, often referred to as *full embedding*, for embedding learning is to learn the representation of each feature value [19]. Specifically, let F be a categorical feature and $|F|$ be its vocabulary size, each feature value $f_i \in F$ is assigned an embedding index k_i so that the k_i -th row of the embedding table $W \in \mathbb{R}^{N \times D}$ is the embedding vector of f_i , where $N = |F|$ in the full embedding method and D is the embedding dimensionality (see Fig 1 (a)).

However, such full embedding learning suffers from severe memory cost problems. Actually, the memory cost of the embedding

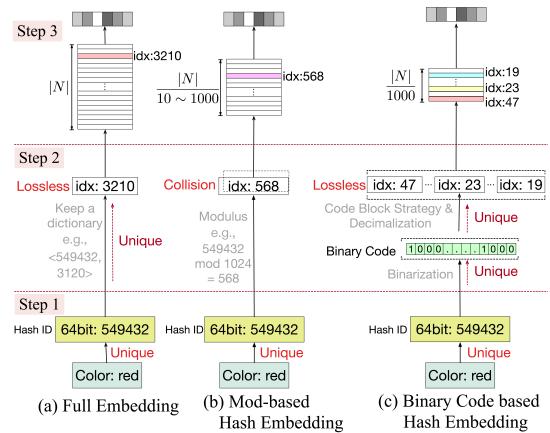


Figure 1: Comparisons of different embedding methods. Step 1, 2, and 3 refer to feature hashing, embedding index generation, and embedding generation respectively.

table is $O(|F|D)$ which grows linearly with $|F|$. For web-scale applications, one may need to store a huge embedding table since the vocabulary size may be millions or even billions. For example, suppose $|F| = 500$ million and $D = 256$, the corresponding memory cost will be 475GB. In practice, such a huge cost becomes a bottleneck in deploying EDRMs in memory-sensitive scenarios.

Therefore, it is crucial to reduce the size of the embedding table [2, 19]. In this paper, we highlight two challenges: (1) **Challenge one: Flexibility**. The memory constraint varies with different scenarios (from distributed servers to mobile devices). The embedding reduction methods need to be flexible enough to meet different memory requirements. Especially for mobile devices, a tiny EDRM is needed to meet the limited memory requirement. (2) **Challenge two: Performance Retention**. Since a big model usually has a better capacity and hence a better performance, embedding reduction may bring a performance gap due to the fewer parameters used in the reduced model. Hence, how to keep high performance when the memory size is reduced is a big challenge, especially for the memory-sensitive scenarios (e.g., in mobile devices).

In general, there are two directions to reduce the embedding table size, i.e., reducing the size of each embedding vector and reducing the number (i.e., N) of the embedding vectors in an embedding table. The embedding table size of the former methods (e.g., product quantization [5, 9], K-D method [2, 11, 13, 15, 20], and AutoDim [6, 10, 16, 17, 28, 29]) is still linearly increased with $|F|$, failing to tackle the memory problem caused by a large vocabulary size in web-scale applications [19]. Hence these methods are not considered in our paper. **For the latter methods, they typically apply a mod-based hash embedding to reduce N . The key idea of them is to apply modulo operation on the unique Hash ID of each feature value, i.e., focusing on Step 2 in Fig 1 (b).** For example, Hash embedding [24]

* These authors contributed equally to this work and are co-first authors.
† Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-8446-9/21/11...\$15.00
https://doi.org/10.1145/3459637.3482065

Table 1: Comparison about embedding methods.

	Full	Hash	MH	Q-R	Ours
Flexibility	Bad	Good	Good	Fair	Good
Performance Retention	Good	Bad	Fair	Good	Good

takes the remainder of the Hash ID divided by M as the embedding index, reducing the embedding size from $O(|F|D)$ to $O(MD)$. The problem of this method is that different feature values may have the same embedding index and hence the same embedding vector, leading to poor performance. Multi-Hash (MH) [22] adopts multiple embedding indices for one feature value, reducing the collision rate among feature values. But different feature values may still be indistinguishable especially for a tiny model, failing to the challenge two. Q-R trick [19] uses both the remainder and the quotient as embedding indices to identify a feature value. However, Q-R trick fails to the challenge one since its minimal reduced size is related to $\sqrt{|F|}$ rather than any scales. Although the generalized Q-R tries to address this problem, it needs a lot of effort to design the divisor [19]. The comparisons are summarized in Table 1.

In this paper, unlike the existing methods which adopt a modulo (collision) operation, we bring the idea of binary code (e.g., the binary code of integer 13 is 1101_2) which is unique for different Hash ID and propose a *binary code based hash embedding* method to tackle this reduction problem (see Fig 1 (c)). Specifically, we first binarize the Hash ID into a binary code. Then, to address the challenge one, we propose a code block strategy and reduce the embedding table size by adjusting the code block length flexibly. To address the challenge two, the generated embedding index is designed to be unique for different feature values at any reduction ratios. The uniqueness at any reducing ratios allows EDRMs to distinguish different feature values, leading to a good performance even for a tiny model. Furthermore, Step 2 of our method is a deterministic and non-parametric process and can be computed on-the-fly. This property is friendly for EDRMs both on the convenient application and handling new (out-of-vocabulary) feature values.

We also note that we are aware of some recent works using similar terms such as *learning binary embedding* [8, 14, 26]. We want to point out that they are in totally different contexts. In these works, binary refers to that each element in an embedding vector is a binary number for a fast similar embedding search. While in our work, binary refers to binarize the integer ID into a binary code.

To summarize, the main contributions are listed as follows: (1) We propose binary code based hash embedding, a simple but effective embedding method, to reduce the embedding table size and keep a high performance at the same time. (2) A code block strategy is presented to adjust the embedding table size flexibly and a lossless embedding index generation process is elaborately designed to allow the model to distinguish different feature values and achieve better performance. (3) Experimental results on large-scale real-world datasets show that with the help of the proposed method, the model size can be $1000\times$ smaller than the original model, and keep 99% performance as the original model achieves at the same time.

2 BINARY CODE BASED HASH EMBEDDING

In this section, we introduce the framework (see Fig 2) of our methods. In general, we also adopt three steps as introduced in Fig 1.

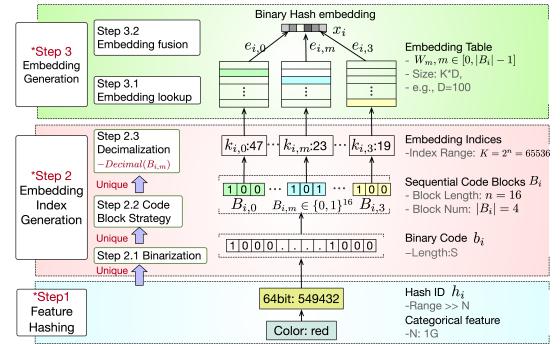


Figure 2: The framework of the proposed method.

2.1 Feature Hashing

In practice, the raw categorical feature values may be represented as various types, such as String and Integer values. To handle different types of categorical feature values, in practice, a feature hashing [19, 22, 24] is firstly applied to map these raw feature values into a uniformed integer number, called Hash ID (h_i) (see Fig 2). Formally, the feature hashing process can be expressed as $h_i = \mathcal{H}(f_i)$ where \mathcal{H} refers to a hash function (e.g., Murmur Hash [25]) and h_i is an integer number, called the Hash ID of f_i . In practice, the output length of \mathcal{H} is always a large value (e.g., h_i is a 64-bits integer) to make the collision among h_i as small as possible. In this case, h_i can be basically taken as a unique ID for different f_i [12, 24].

2.2 Embedding Index Generation

In this section, we introduce the embedding index generation process including binarization, code block strategy, and decimalization.

2.2.1 Binarization. After Step 1, each feature value f_i is mapped to h_i , which is basically regarded as a non-collision mapping due to the large output space [12, 24]. Then the binary code $b_i \in \{0, 1\}^S$ (where S refers the binary code length) of f_i can be generated by transforming this unique h_i to a binary form (e.g., the binary code of integer 13 is 1101_2). Note b_i is also unique for different f_i .

2.2.2 Code Block Strategy. To allow the model can flexibly reduce memory, we propose a novel strategy called code block strategy. Generally speaking, the code block strategy divides each 0-1 value in b_i to different blocks. Then, the ordered 0-1 values (i.e., 0-1 code) in each block can represent $K = 2^n$ unique integers where n is the number of 0-1 values in this block (see Step 2.2 in Fig 2).

If we take the decimal form of 0-1 code in each block as an embedding index and map each index to an embedding table $W \in \mathbb{R}^{K \times D}$, the size of the embedding table can be flexibly adjusted by n . For example, when $n = 1$ (i.e., the number of 0-1 values in each block is 2), the embedding table size is $O(2D)$. When all 0-1 values in b_i are arranged into one block, the embedding table size is $O(|F|D)$ (i.e., full embedding). In other words, by controlling the value of n , we can adjust the embedding table size to meet various scenarios (from distributed services to mobile devices).

Formally, we define $B_i = [B_{i,0}; B_{i,1}; \dots; B_{i,m}; \dots]$ as the sequential code blocks produced by a code block strategy on b_i , and $|B_i|$ refers to the number of blocks. Then the m -th code block $B_{i,m} \in \{0, 1\}^n$ can be represented as

$$B_{i,m} = \text{Order}(\{b_{i,j} | \text{Alloc}(b_{i,j}) = m\}) \quad (1)$$

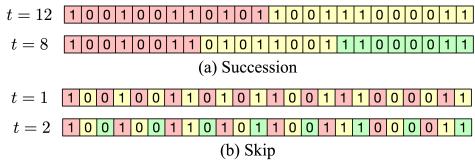


Figure 3: Code block strategy examples. The binary code $b_i = 100100110101100111000011$. The 0-1 values plotted with the same color in each case are divided into the same block.

where the function Alloc is the allocation function which allocates each 0-1 value to different blocks. Order is a function which gives an order for the 0-1 value in each block and generates a 0-1 code for each code block. Here we give two code block strategies (including Succession and Skip) as examples to show how it works (other possible strategies can also be allowed).

Succession. As shown in Fig 3 (a), the succession strategy puts the t successive 0-1 values in a binary code into the same block. The Order function keeps 0-1 values in the same relative position in b_i . Note if the number of the last 0-1 values in b_i is less than t , all of the left values are divided into a new code block.

Skip. As shown in Fig 3 (b), if the number of interval values of two 0-1 values in a binary code is t , they will be divided into the same block. The Order function is the same as that in Succession.

Note given one of the above code block strategies, we can obtain a unique sequence of code blocks B_i for the binary code b_i . This property guarantees the process of code block strategy is lossless.

2.2.3 Decimalization. The embedding index of each block can be obtained by decimalizing $B_{i,m}$ (e.g., $\text{Decimalize}(1101_2) = 13$), i.e., $k_{i,m} = \text{Decimalize}(B_{i,m})$ where $k_{i,m}$ is the embedding index of $B_{i,m}$.

2.3 Embedding Generation

When obtaining multiple indices for f_i , to get its embedding, two steps are proposed, i.e., embedding lookup and embedding fusion.

2.3.1 Embedding Lookup. As introduced above, each code block $B_{i,m}$ in B_i can obtain an embedding index $k_{i,m}$. The number of blocks is $|B_i|$, leading to a total of $|B_i|$ embedding indices. Then we can map each embedding index into an embedding vector, i.e., $e_{i,m} = \mathcal{E}(W_m, k_{i,m})$ where W_m is a embedding table, $e_{i,m}$ refers to the embedding of $B_{i,m}$ and \mathcal{E} is a embedding lookup function which usually returns the $k_{i,m}$ -th row of W_m . In practice, keeping $|B_i|$ embedding tables for different $B_{i,m}$ may also cost a lot memory consumption. Therefore, it is common to keep a single embedding table and share this table among all $B_{i,m}$ [22].

2.3.2 Embedding Fusion. To generate the final embedding vector x_i of f_i , an embedding fusion function g is applied,

$$x_i = g(e_{i,0}, e_{i,1}, e_{i,2}, \dots, e_{i,|B_i|-1}) \quad (2)$$

The design of the fusion function can be various, such as pooling, LSTM, concatenation and so on. In this paper, by default, we adopt sum pooling as the fusion function (others can also considered).

2.4 Discussion

2.4.1 Desiderata. There are several key desiderata of our method, which EDRMs can be benefited. (1) **Determinacy.** The indices

generation is a deterministic and non-parametric process. It is computed on the fly, making it simple to practical implementations and friendly to new feature values. (2) **Flexibility.** The size of embedding table $W \in \mathbb{R}^{K \times D}$ is mainly determined by n (i.e., the number of 0-1 values in each code block). It means the memory reduction ratio can be flexibly adjusted from $2/|F|$ to 1 (assuming adopting embedding table sharing strategy). This benefits EDRMs can be developed on memory insensitive scenarios to sensitive scenarios. (3) **Uniqueness.** No matter what the reduction ratio is, B_i is unique for each feature value. This enables the model to distinguish different feature values and further improve the model performance.

2.4.2 Sub-collision Problem. We should point out although B_i is unique, there may exist sub-collision among two feature values (e.g., $B_i \neq B_j$ but $B_{i,1} = B_{j,1}$), called sub-collision problem. Actually, it is an open problem which exists in most mod-based hash methods [19, 22, 24]. We leave it as one of the future work. In practice, a hash function (e.g., Murmur hash [25]) which can randomly map values to a large space is used to relieve this problem.

2.4.3 The Relation with Existing Methods. Here, we discuss the relation between ours and other methods. (1) **Full Embedding.** Both of full embedding and ours can distinguish different feature values. Besides, our method has the ability to reduce memory flexibly. (2) **Hash Embedding.** It is a simplified form of ours, where the code block strategy is Succession, and only the first top t 0-1 values are used as the embedding index. (3) **Multi-Hash Embedding.** Both of them create multiple embedding indices. But our method goes further, i.e., keeping a uniqueness constraint for these indices. (4) **Q-R Trick.** Q-R trick is a special case of our method. When we utilize Succession and the block number is set to 2. The first top t 0-1 code and the left 0-1 code can be taken as the quotient and the remainder in Q-R trick respectively.

3 EXPERIMENTS

Datasets. (1) *Alibaba* is an industrial dataset which is obtained from Taobao. There are a total 4 billion samples, 100 million users. (2) *Amazon*¹ is collected from the Electronics category on Amazon. There are total 1,292,954 samples, 1,157,633 users. (3) *MovieLens*² is a reviews dataset and is collected from the MovieLens web site. There are total 1,000,209 samples, 6,040 users.

Baselines. (1) *Full Embedding (Full)* is a standard embedding learning method. (2) *Hash Embedding (Hash)*[24] applies the modulo operation on the Hash ID to obtain an embedding index. (3) *Multi-Hash Embedding (MH)* [22] applies multiple hash functions to the feature value to obtain multiple indices. (4) *Q-R Trick (Q-R)* [19] take both the remainder and the quotient as indices.

Training Details. All methods have the same EDRM architecture. The embedding dimensionality is also set the same for all methods. The methods (i.e., MH, Q-R trick, and ours) employ embedding table sharing strategy in different indices for memory reduction purpose and take sum pooling as the fusion function. For MH, we use 2 hash functions as suggested by authors [22]. For our method, the code block strategy is Succession. We use the Adagrad optimizer with a learning rate of 0.005. The batch size is 1024 for all datasets.

¹<https://www.amazon.com/>

²<https://grouplens.org/datasets/movielens/>

Table 2: The results for CTR tasks.

Dataset	Alibaba					Amazon					MovieLens				
Reduction Ratio	0.1%	0.75%	1.5%	5%	37.5%	0.1%	0.75%	1.5%	5%	37.5%	0.1%	0.75%	1.5%	5%	37.5%
Full	70.57	70.57	70.57	70.57	70.57	68.56	68.56	68.56	68.56	68.56	80.23	80.23	80.23	80.23	80.23
Hash	69.06	69.35	69.63	69.86	70.24	64.66	66.27	66.66	67.32	67.67	73.62	75.50	76.42	77.68	79.12
MH	69.45	69.66	69.73	70.00	70.28	66.15	67.53	67.58	67.79	68.03	74.97	78.06	78.23	78.85	79.85
Q-R	69.47	69.58	69.82	70.10	70.28	66.67	67.43	67.62	67.73	68.17	75.92	78.05	78.30	78.90	79.78
BH	69.90	69.95	70.02	70.26	70.38	67.59	67.73	67.92	68.38	68.38	78.20	78.38	78.54	79.14	80.02

Table 3: The results of memory size when all the methods archive 99 % performance as the full embedding method achieves in AUC score.

	Full	Hash	Q-R	MH	BH
Model size (G)	843.2	44.1	13.0	12.5	0.8
User ID	288.94	10.16	3.11	2.88	0.15
Item ID	242.71	11.07	3.01	2.95	0.16
Query ID	165.24	11.07	3.17	3.15	0.16

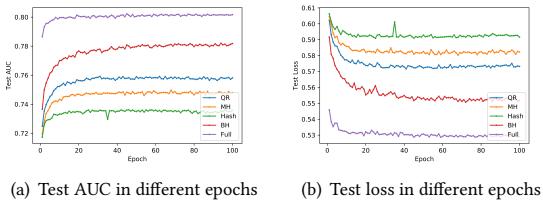


Figure 4: Convergence of different methods.

3.1 Click-Through Rate (CTR) Prediction Task

We conduct experiments on CTR prediction tasks and compare the performance with different memory reduction ratios of the full embedding. AUC (%) [4] score is reported as the metric. Note 0.1% absolute AUC gain is regarded as significant for the CTR task [3, 21, 30]. Results are shown in Table 2. BH refers to our method.

Comparison with Mod-based Hash Embedding Methods. In general, BH performs best on all cases, and the gain gap between BH and baselines is increased for a smaller model. For example, compared with Q-R on MovieLens, BH can achieve 0.24% gains when the reduction ratio is 37.5 %. While when the ratio becomes 0.1 %, the gain gap is increased to 2.28 %. It indicates that due to the nice properties of BH (see Section 2.4.1), BH can better represent each categorical feature value, especially for a tiny model.

Comparison with the Full Embedding. We can observe that: (1) Since the full embedding method contains significant parameters, it gets better performance. (2) In most cases, BH can obtain competitive performance with the full embedding method.

3.2 Memory Reduction Comparison

In this section, we conduct experiments to evaluate the model size of all methods when achieving similar performance. Specifically, we take the dataset Alibaba as an example due to its closeness with the web-scale application. Then we report the model size of different embedding methods when they achieve 99% performance as the full embedding method achieves in AUC score. Besides, to further evaluate the reduction ratio, we also report the size of the embedding table of the top 3 largest for each method. The results are shown in Table 3. Some observations are summarized as follows: (1)

Table 4: The results of different code block strategies.

	Succession	Skip	Q-R
Alibaba	69.90	69.87	69.47
Amazon	67.59	67.60	66.67
MovieLens	78.20	78.25	75.92

Compared with other embedding methods, when all of them archive similar scores, BH can cost the smallest memory size. (2) Compared with the full embedding method, BH can adopt an extremely tiny model (i.e., 1000× smaller) to achieve 99 % performance. Such a small model with high performance is urgently needed to develop EDRMs on a memory-sensitive scenarios (e.g., mobile devices).

3.3 The Effect of the Code Block Strategy

In this section, we evaluate the performance of the proposed two code block strategies, i.e., Succession and Skip. To have a fair comparison, we keep the same reduction ratio for these two strategies. Table 4 shows the results. Note we also provide the best performance (Q-R) among baselines for comparison. We can observe that Succession and Skip achieve similar performance overall datasets, and perform better than the best baselines. It indicates the uniqueness of code block strategy is helpful to improve the embedding performance no matter what kind of code block strategies we choose.

3.4 Analysis of Convergence

We conduct experiments to analyze the convergence of different models. Specifically, we keep the same reduction ratio for all methods and report the AUC and the loss value of these methods on test data of MovieLens within 100 epochs (similar conclusions can be found in other datasets). As shown in Fig 4, we can find: (1) Compared the AUC and loss curves of mod-based hash embedding methods, BH converges faster than that of other baselines. Furthermore, BH can achieve a higher AUC score and a lower loss value. It demonstrates the effectiveness of our method when reducing EDRMs into a small-scale size. (2) Due to more parameters adopted in full embedding, it archives the best performance. But, BH can also achieve competitive performance compared with full embedding.

4 CONCLUSION

In this paper, to tackle the memory problem in embedding learning, we propose a binary code based hash embedding. A binary code is firstly generated to guarantee a unique index code. Then a code block strategy is designed to flexibly reduce the embedding table size. Finally, the feature embedding vector is obtained by combining the embedding vectors from different code blocks. Experimental results show that even if the model size is 1000× smaller, we can still obtain the 99% performance by binary code based hash embedding.

REFERENCES

- [1] John Anderson, Qingqing Huang, Walid Krichene, Steffen Rendle, and Li Zhang. 2020. Superbloom: Bloom filter meets Transformer. *arXiv preprint arXiv:2002.04723* (2020).
- [2] Ting Chen, Martin Renqiang Min, and Yizhou Sun. 2018. Learning k-way d-dimensional discrete codes for compact embedding representations. *ICML* (2018).
- [3] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.
- [4] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.
- [5] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. 2013. Optimized product quantization. *IEEE transactions on pattern analysis and machine intelligence* 36, 4 (2013), 744–755.
- [6] Antonio Ginart, Maxim Naumov, Dheevatsa Mudigere, Jiyan Yang, and James Zou. 2019. Mixed dimension embeddings with application to memory-efficient recommendation systems. *arXiv preprint arXiv:1909.11810* (2019).
- [7] Hufeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).
- [8] Weixiang Hong, Junsong Yuan, and Sreyasee Das Bhattacharjee. 2017. Fried binary embedding for high-dimensional visual features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2749–2757.
- [9] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2010. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence* 33, 1 (2010), 117–128.
- [10] Manas R Joglekar, Cong Li, Mei Chen, Taibai Xu, Xiaoming Wang, Jay K Adams, Pranav Khaitan, Jiahui Liu, and Quoc V Le. 2020. Neural input search for large scale recommendation models. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2387–2397.
- [11] Wang-Cheng Kang, Derek Zhiyuan Cheng, Ting Chen, Xinyang Yi, Dong Lin, Lichan Hong, and Ed H Chi. 2020. Learning Multi-granular Quantized Embeddings for Large-Vocab Categorical Features in Recommender Systems. In *Companion Proceedings of the Web Conference 2020*. 562–566.
- [12] Wang-Cheng Kang, Derek Zhiyuan Cheng, Tiansheng Yao, Xinyang Yi, Ting Chen, Lichan Hong, and Ed H Chi. 2020. Deep Hash Embedding for Large-Vocab Categorical Feature Representations. *arXiv preprint arXiv:2010.10784* (2020).
- [13] Valentin Khrulkov, Oleksii Hrinchuk, Leyla Mirvakhova, and Ivan Oseledets. 2019. Tensorized embedding layers for efficient model compression. *arXiv preprint arXiv:1901.10787* (2019).
- [14] Brian Kulis and Trevor Darrell. 2009. Learning to Hash with Binary Reconstructive Embeddings.. In *NIPS*, Vol. 22. Citeseer, 1042–1050.
- [15] Chaozhuo Li, Lei Zheng, Senzhang Wang, Feiran Huang, Philip S Yu, and Zhoujun Li. 2019. Multi-Hot Compact Network Embedding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 459–468.
- [16] Haochen Liu, Xiangyu Zhao, Chong Wang, Xiaobing Liu, and Jiliang Tang. 2020. Automated Embedding Size Search in Deep Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2307–2316.
- [17] Siyi Liu, Chen Gao, Yihong Chen, Depeng Jin, and Yong Li. 2021. Learnable Embedding Sizes for Recommender Systems. *arXiv preprint arXiv:2101.07577* (2021).
- [18] Joan Serra and Alexandros Karatzoglou. 2017. Getting deep recommenders fit: Bloom embeddings for sparse binary input/output networks. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 279–287.
- [19] Hao-Jun Michael Shi, Dheevatsa Mudigere, Maxim Naumov, and Jiyan Yang. 2020. Compositional embeddings using complementary partitions for memory-efficient recommendation systems. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 165–175.
- [20] Raphael Shu and Hideki Nakayama. 2017. Compressing word embeddings via deep compositional code learning. *arXiv preprint arXiv:1711.01068* (2017).
- [21] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. 2019. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1161–1170.
- [22] Dan Tito Svenstrup, Jonas Hansen, and Ole Winther. 2017. Hash embeddings for efficient word representations. *Advances in neural information processing systems* 30 (2017), 4928–4936.
- [23] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*. 1–7.
- [24] Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. 2009. Feature hashing for large scale multitask learning. In *Proceedings of the 26th annual international conference on machine learning*. 1113–1120.
- [25] Fumito Yamaguchi and Hiroaki Nishi. 2013. Hardware-based hash functions for network applications. In *2013 19th IEEE International Conference on Networks (ICON)*. IEEE, 1–6.
- [26] Xinyang Yi, Constantine Caramanis, and Eric Price. 2015. Binary embedding: Fundamental limits and fast algorithm. In *International Conference on Machine Learning*. PMLR, 2162–2170.
- [27] Caojin Zhang, Yicun Liu, Yuanpu Xie, Sofia Ira Ktena, Alykhan Tejani, Akshay Gupta, Pranay Kumar Myana, Deepak Dilipkumar, Suvadip Paul, Ikuhiro Ihara, et al. 2020. Model Size Reduction Using Frequency Based Double Hashing for Recommender Systems. In *Fourteenth ACM Conference on Recommender Systems*. 521–526.
- [28] Xiangyu Zhao, Haochen Liu, Hui Liu, Jiliang Tang, Weiwei Guo, Jun Shi, Sida Wang, Huiji Gao, and Bo Long. 2020. Memory-efficient Embedding for Recommendations. *arXiv preprint arXiv:2006.14827* (2020).
- [29] Xiangyu Zhao, Chong Wang, Ming Chen, Xudong Zheng, Xiaobing Liu, and Jiliang Tang. 2020. AutoEmb: Automated Embedding Dimensionality Search in Streaming Recommendations. *arXiv preprint arXiv:2002.11252* (2020).
- [30] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1059–1068.