



NEUKRON: Constant-Size Lossy Compression of Sparse Reorderable Matrices and Tensors



Taehyung Kwon*



Jihoon Ko*



Jinhong Jung



Kijung Shin

Sparse matrices from Web applications



Friendship in Social Media





Counting Clicks on Ads **By Search Engine**

	AD	AD	AD	AD
	5	1	0	0
2	0	1	0	2
3	0	1	0	0
4	0	1	0	3



Publication Records from Academic Databases



Real-world sparse matrices are large-scale

- Real-world sparse matrices often containing billions of rows or columns
 - \Rightarrow requires heavy memory or network I/O usage
 - \Rightarrow compressing these large sparse matrices is important!



Our goal: constant-size compression



- Given: a sparse and reorderable matrix $A \in \mathbb{R}^{N \times M}$ / a constant k = O(1)
- Find: a model Θ whose size is at most k
- To minimize: the approximation error $\|A \widetilde{A}_{\Theta}\|_{F}^{2}$

Overview of NEUKRON

 Recurrent Neural Network: having a constant number of parameters but also expressive power



• Reordering:

extract and exploit structural patterns for better compression

	1	2	3	4	
Q	0	0	1	0	
B	1	0	1	0	
OC	0	1	1	0	
0	1	1	1	1	

_		1	2	3	4
	Op	1	1	1	1
	θ	1	0	1	0
	Q	0	1	1	0
	O _A	0	0	1	0

Model of NEUKRON

• Encode the position in a sequence by recursively dividing the input matrix



Model of NEUKRON

- Feed the sequence to LSTM to compute seed matrices
- Approximate the entry by multiplying the outputs of the LSTM cells



Order optimization

- Many real-world sparse matrices are reorderable
 - \Rightarrow Exploit structural patterns for compression!



Order optimization

• Step 1. Find similar pairs of slices using Min-Hashing



• Step 2. Exchange slices with the neighboring slices when loss decreases



Overall training procedure



ts Conclusion

Experimental settings

10 real-world datasets: 6 sparse matrices and 4 sparse tensors (up to 233M non-zeros)









Email Communication

Twitch Watch History

Publication Record

And Others...

• 9 SOTA competitors

Experimental settings

- 10 real-world datasets: 6 sparse matrices and 4 sparse tensors
- 9 SOTA competitors
 - Factorization-based matrix compression
 - T-SVD, CMD, CUR
 - Co-clustering-based matrix compression
 - ACCAMS, bACCAMS
 - Kronecker product-based matrix compression
 - KronFit
 - Factorization-based tensor compression
 - CP, Tucker
 - Lossless tensor compression
 - CSF (Compressed Sparse Fiber)

NEUKRON is compact and accurate

- The outputs of NEUKRON are up to 5 orders of magnitude smaller
- The approximation error was up to 10.1X smaller in the outputs of NEUKRON



NEUKRON is scalable

• Compression by NEUKRON scaled linearly with the number of non-zeros



Ablation Study

- All components of NEUKRON are effective
 - the variants of NEUKRON with missing components (NEUKRON-H, -A, -F, -I) were outperformed by the original NEUKRON, equipped with all components



Conclusion

• We propose NEUKRON, a lossy compression algorithm for reorderable and sparse matrices and tensors



Code and datasets are available at https://github.com/kbrother/NeuKron





NEUKRON: Constant-Size Lossy Compression of Sparse Reorderable Matrices and Tensors



Taehyung Kwon*



Jihoon Ko*



Jinhong Jung



Kijung Shin