Performance Analysis of Embedding Cache Models for Optimizing Data Movement in Recommendation Systems

문현우이, 구건재 고려대학교

ty5623@korea.ac.kr, gunjaekoo@korea.ac.kr



Embedding Vector Embedding Vector Table N Embedding Tables	Memory capacity shortage issues are severe in GPU *K. Balasubramanian, A. Alshabanah, J. D. Choe, and M. Annavaram. cDLRM: Look ahead caching for scalable training of recommendation models. Pages 263–272. Association for Computing Machinery, Inc, 9 2021		20000 40000 60000 80000 100000 120000 140000 Embedding Indices	Hierarc	Cold Vector Cold Vector Cold Vector hical embedding Ching model	
	Motivation		Experimer	t Setup		
SSD CPU Mem	ory Sorted by access count		Experiment Setup	Datas	set Setup	
V0 Original CPU Cache	GPU Memory	CPU	11th Gen Intel i5-11400 (12) @ 4.400GHz	Dataset	Criteo Al Labs Kaggle A	
V1 Index Index 1 ID 0	CV0 Original GPU Cache Index Index	Memory	Samsung 32GB DDP/	Indices Range	100000	
V2 ← 2 ID 1	$\begin{array}{c c c c c c c c c c c c c c c c c c c $			Sparse Features	26	
25 ID 2 -	$\begin{array}{c c c c c c c c c c c c c c c c c c c $	GPU	NVIDIA T1000 8GB	Sparse Feature Size	64	
V26 26 ID 3	CV3 Metadata Table Vectors	Storage	SK Hynix P31 1TB PCle 3.0 4-lane (32GT/s)	GPU Cache Size	1.5%	
···· 50 ID 4 V49 Metadata Table	Vectors	OS	Ubuntu 22.04	Recommendation system setup		
V50	Vector 50 mapped to GPU vector 2	NVIDIA-FS version	2.15	Ca	che Size	
CPU Memory Cache and Stora	GPU Memory Cache	GDS release version	1.5.0.59	= Maximum ve	ctor # in cache / Total	
The CPU memory capacity is not unlimited!What is the optimal memory capacity?Analyze the impact of SSD I/O time on model performanceAnalyze the impact of GPU cache size on model performance			Hardware setup		embedding vectors #	

GPU cache hit rate and memory usage

CPU cache hit rate and performance comparison



Conclusion

- Analyze the performance of memory hierarchy in embedding cache models
- GPU cache performance improvement saturates at a certain memory capacity.
- The hit rate of CPU memory cache is maintained even with low





memory capacity.

 Confirming that the miss penalty of CPU memory cache hierarchy is a major bottleneck for overall model performance.

SSD I/O time is a major bottleneck of overall performance degradation

*이 성과는 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2021R1C101272)



Labs Kaggle AD