Accelerating Recommendation System Training by Leveraging Popular Choices

Muhammad Adnan

Yassaman Ebrahimzadeh Maboud

Divya Mahajan*

Prashant Nair

48th International Conference on Very Large Databases

THE UNIVERSITY OF BRITISH COLUMBIA

Microsoft*

Recommendation Systems are Ubiquitous

amazon

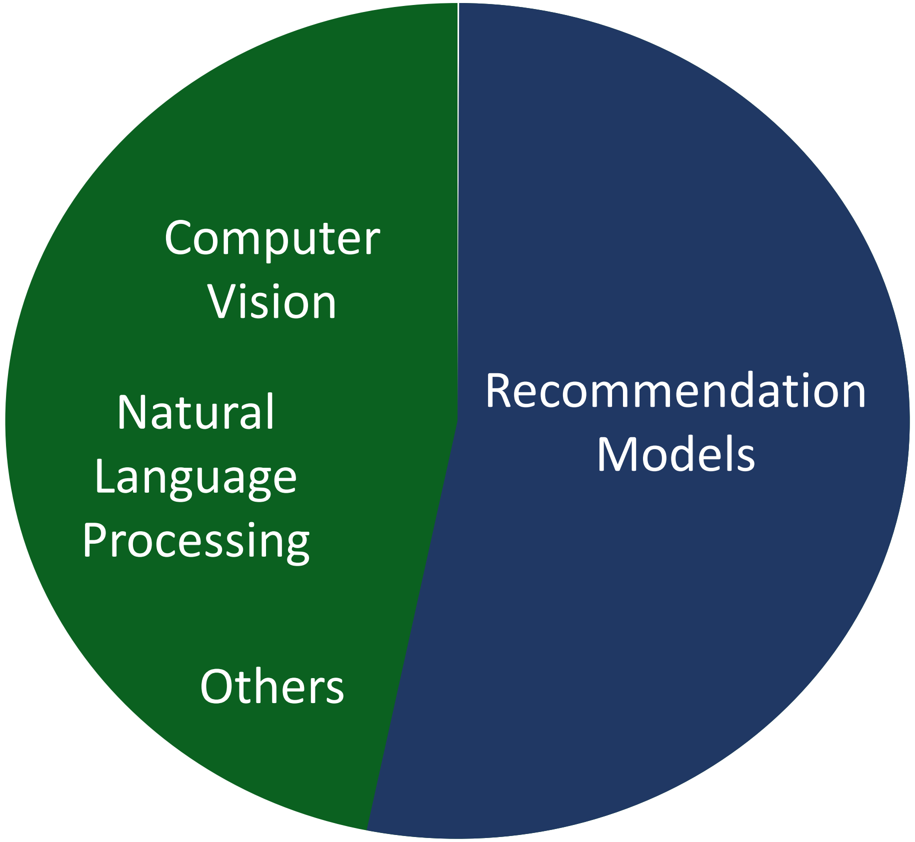
NETFLIX

Meta

YouTube

Bai du

Recommendation Systems in Industry

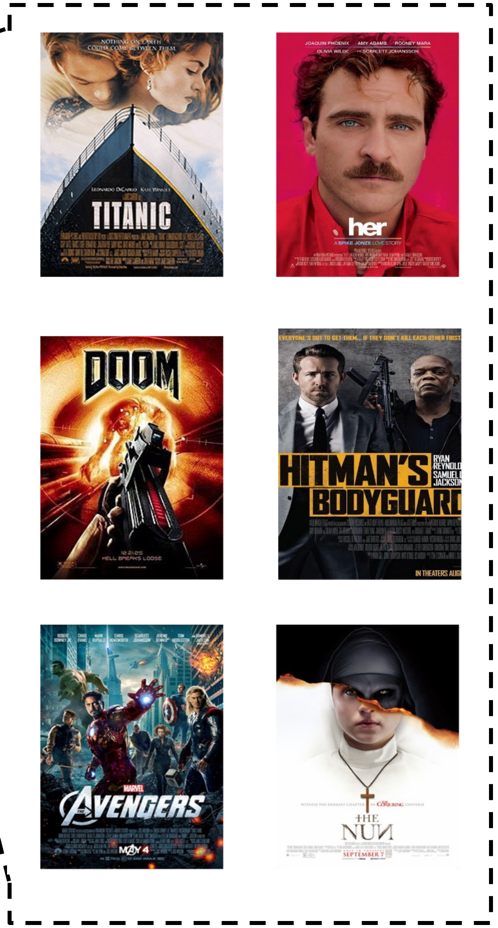


~50% of Training Demand

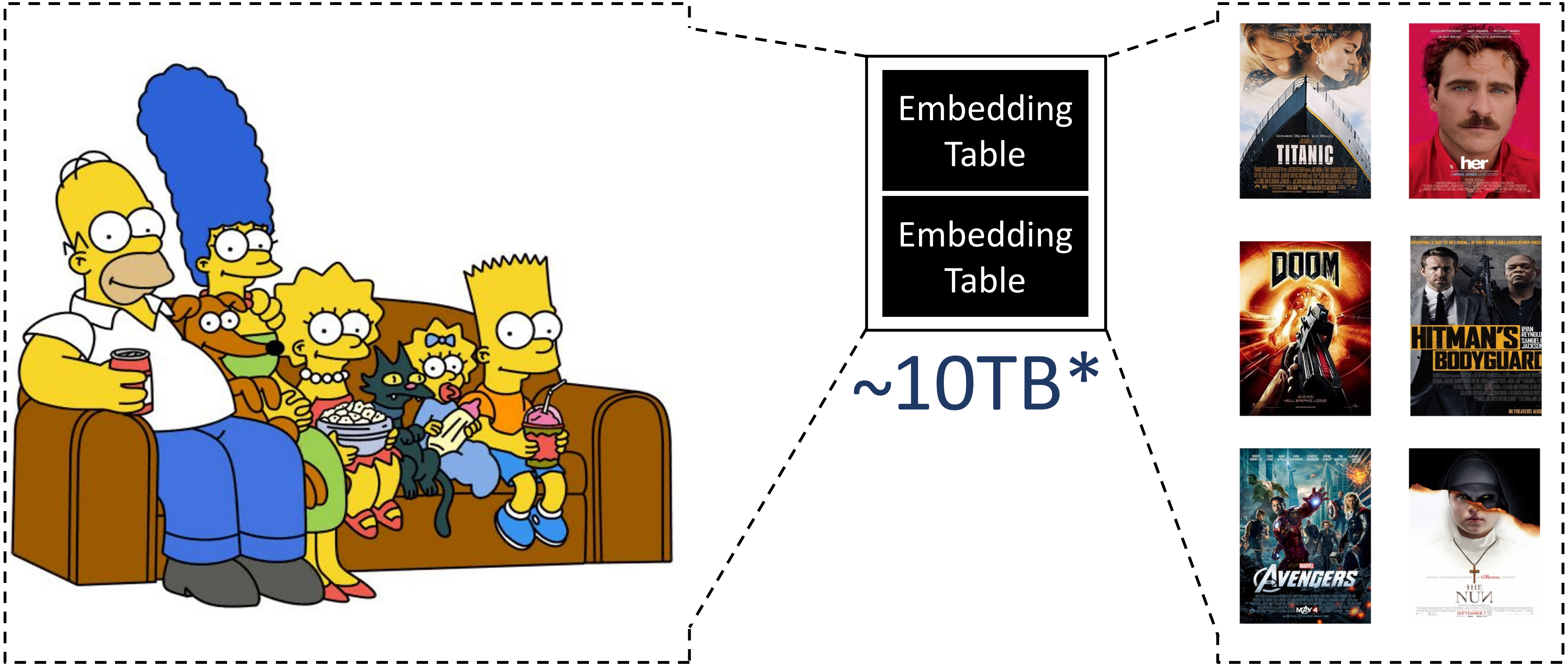
Naumov et al. arXiv'20.

Targeted Recommendation For Each of You

Targeted Recommendation For Each of You

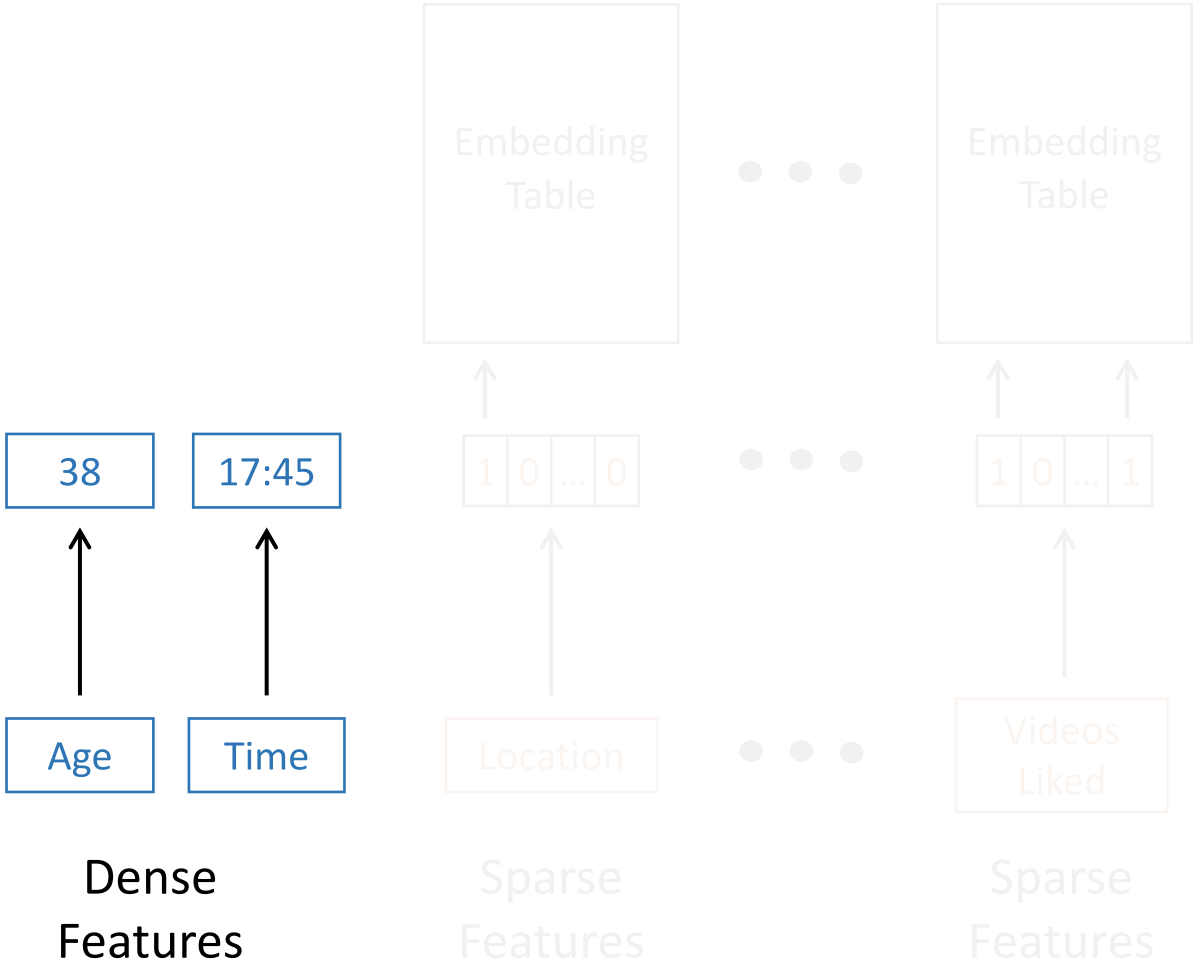


Users and Items Representation

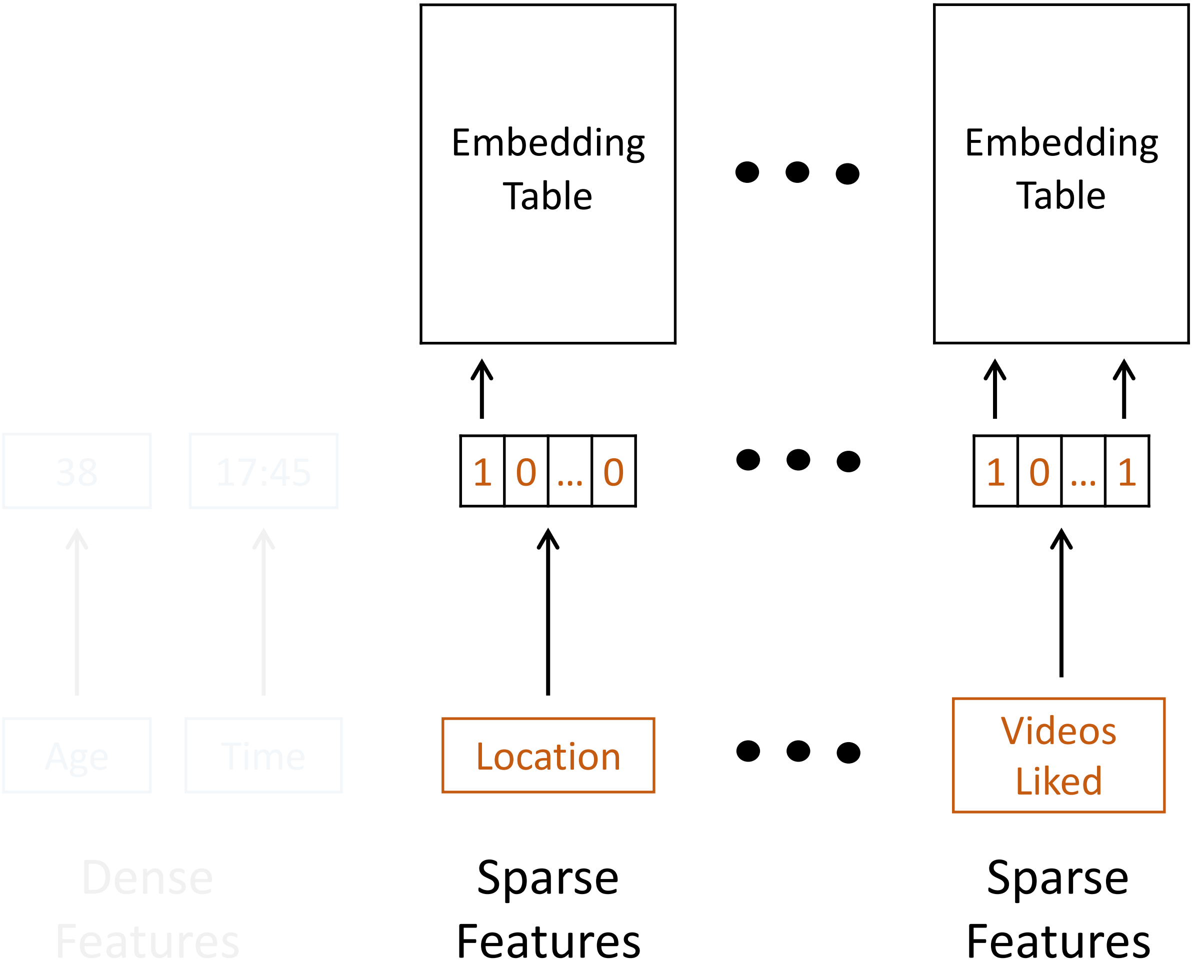


Zhao et al. MLSys'20.

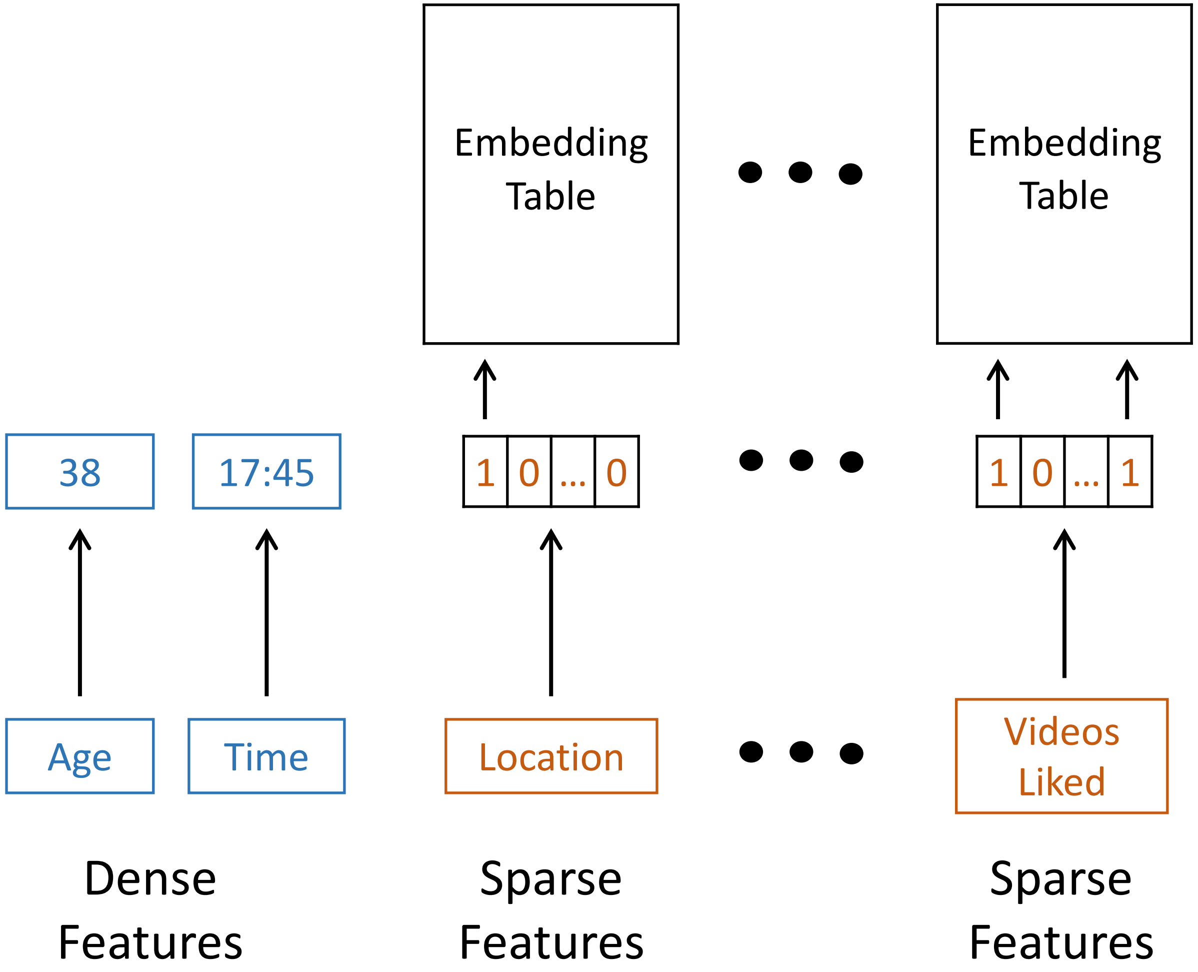
Deep Learning Recommendation Models



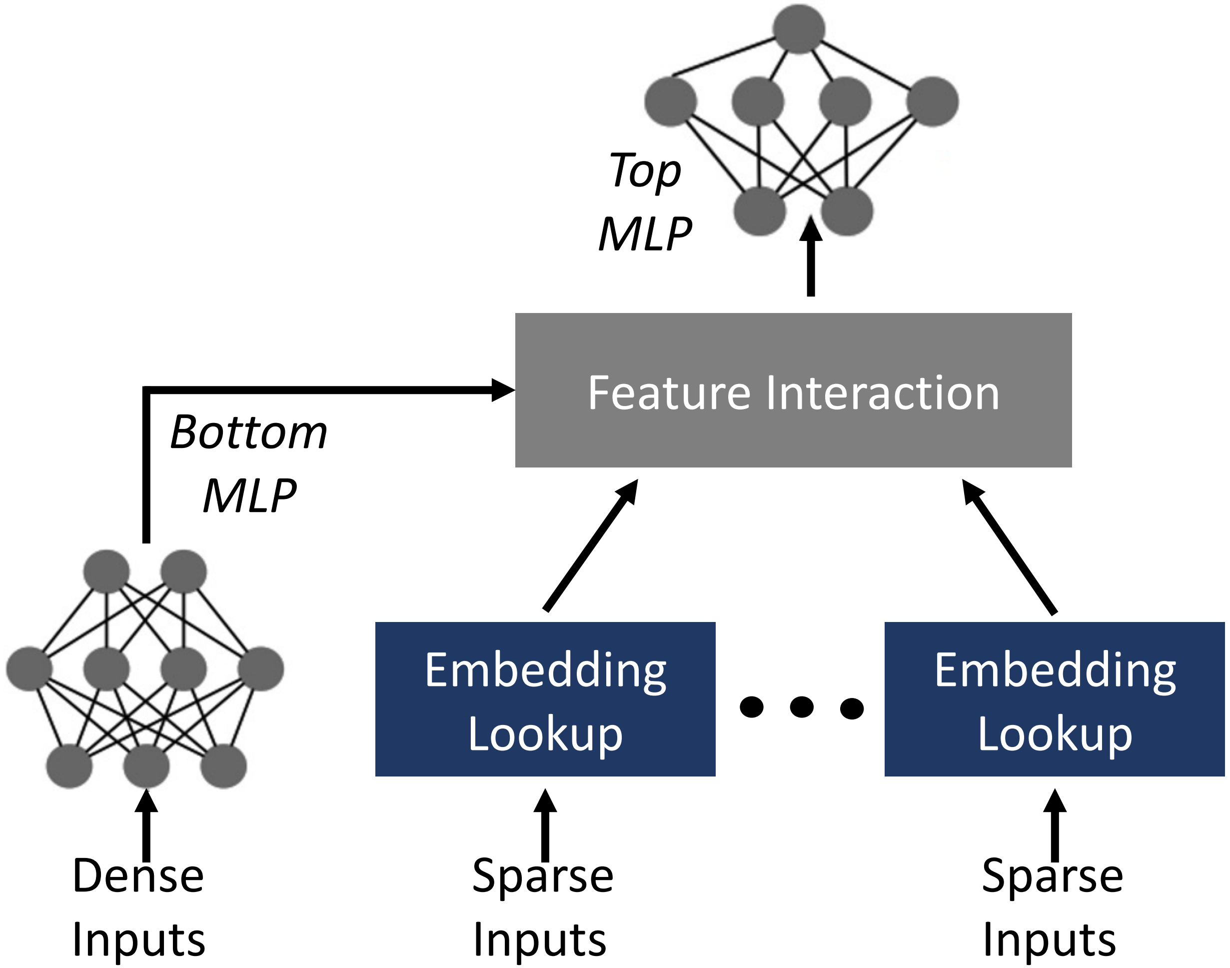
Deep Learning Recommendation Models



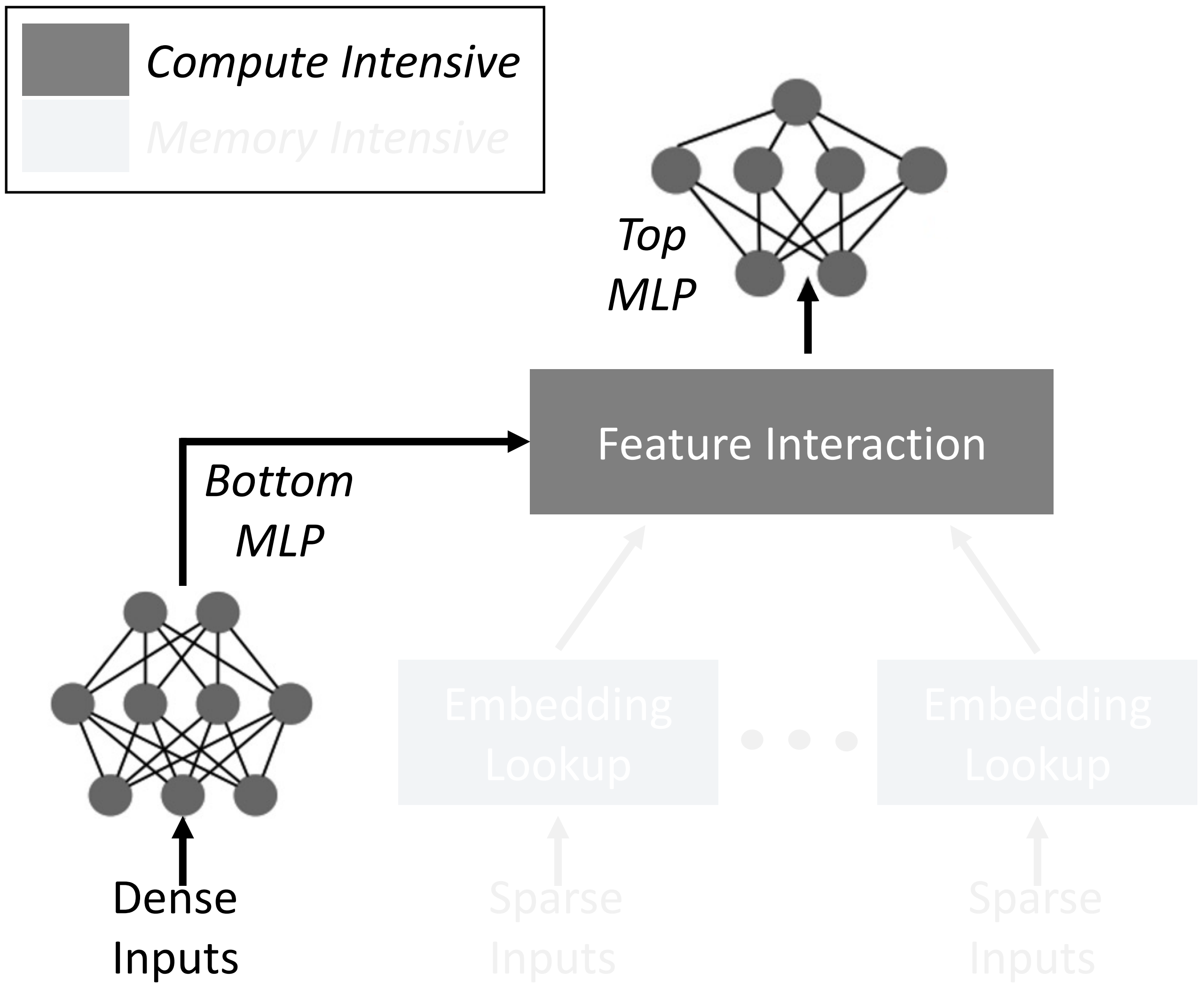
Deep Learning Recommendation Models



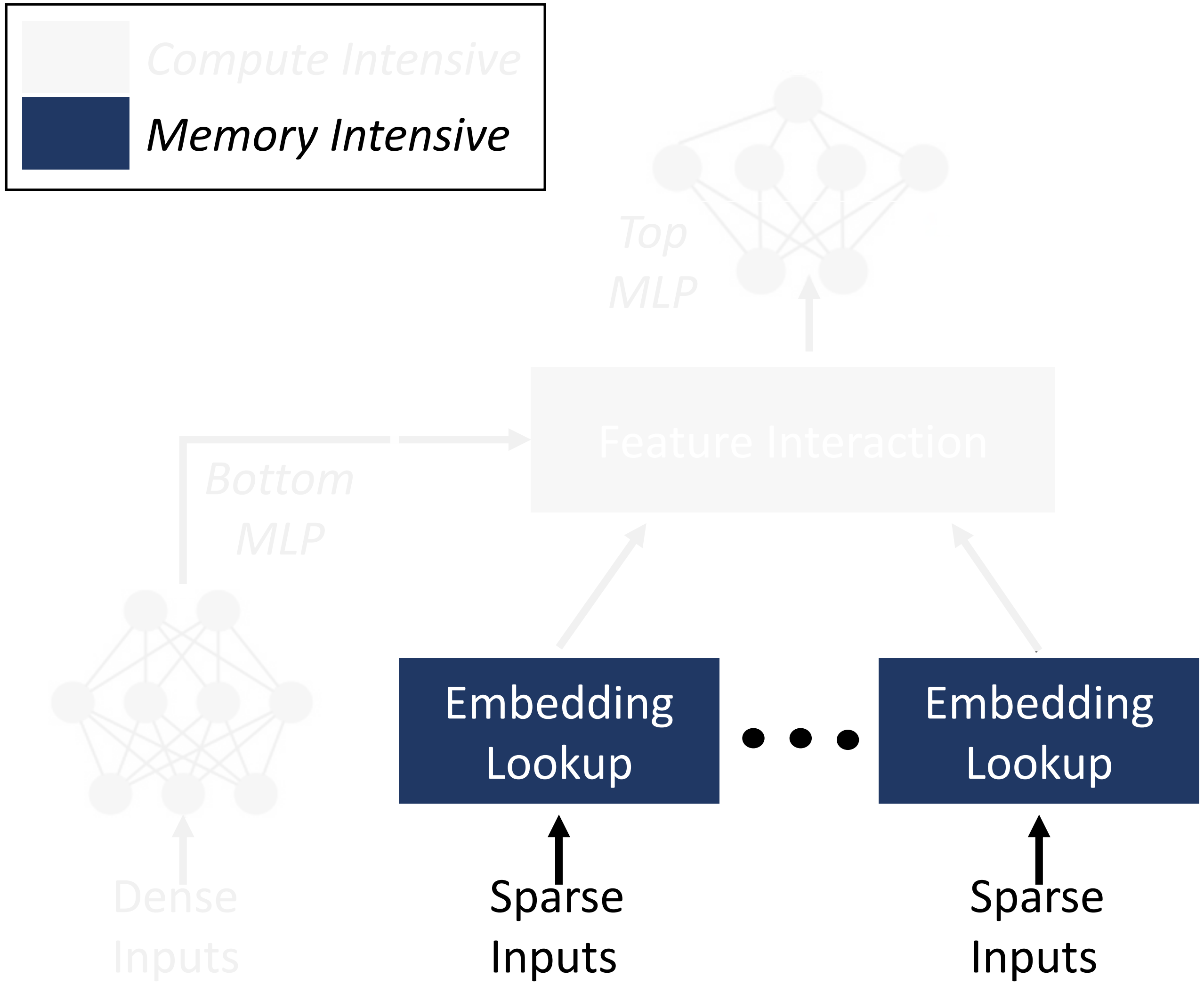
Recommendation Model: High-level Overview



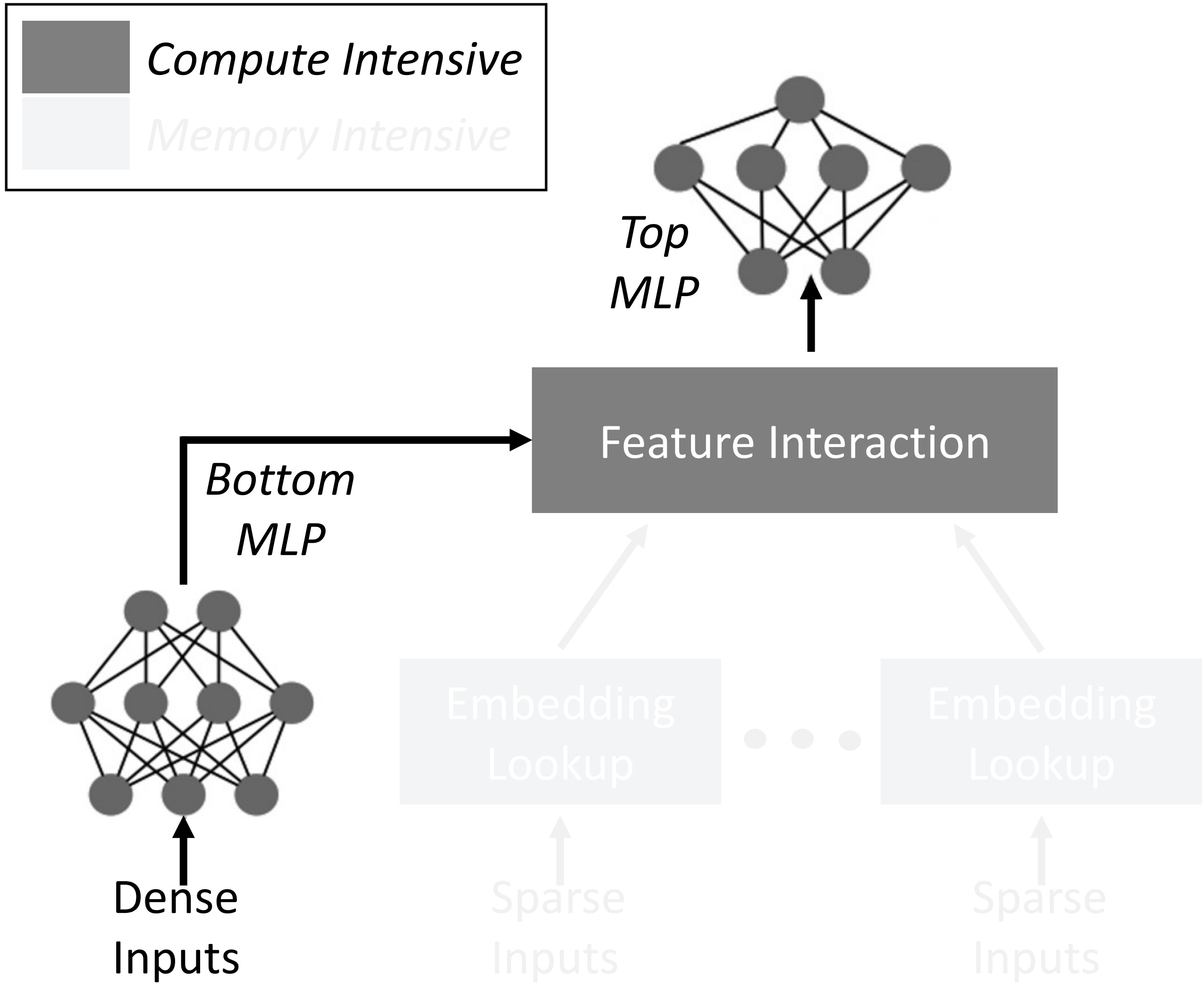
Training Hybrid Execution Mode



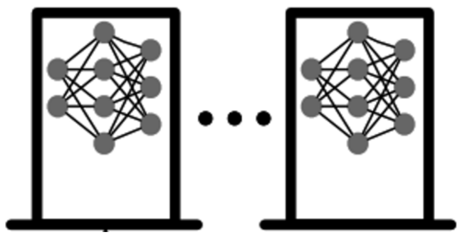
Training Hybrid Execution Mode



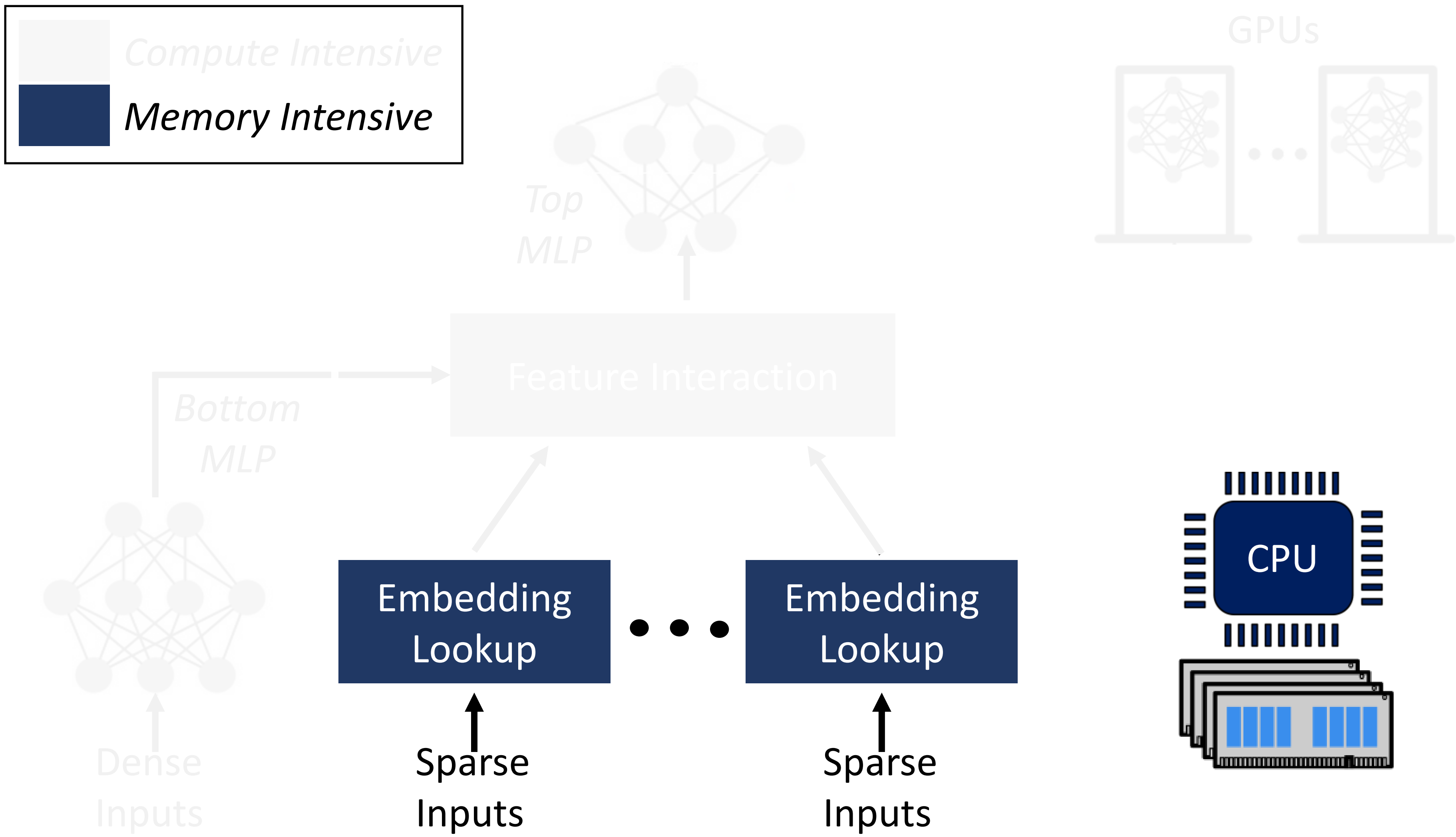
Training Hybrid Execution Mode



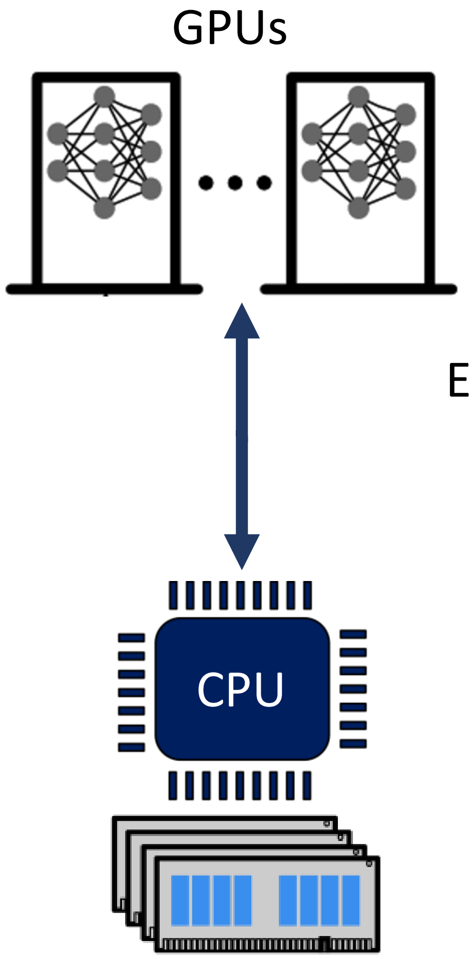
GPUs



Training Hybrid Execution Mode



Hybrid Execution Mode - Inefficiencies



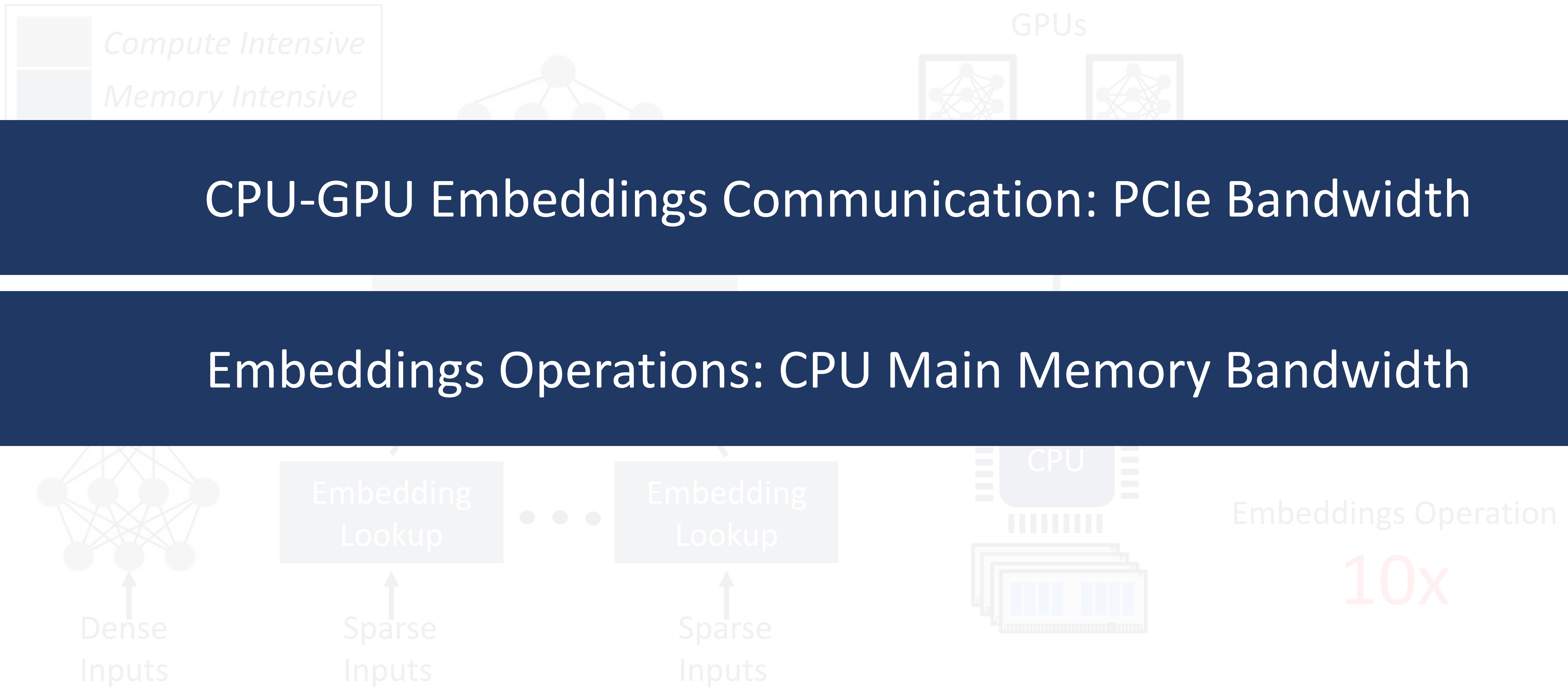
Embeddings Communication

22%

Embeddings Operation

10x

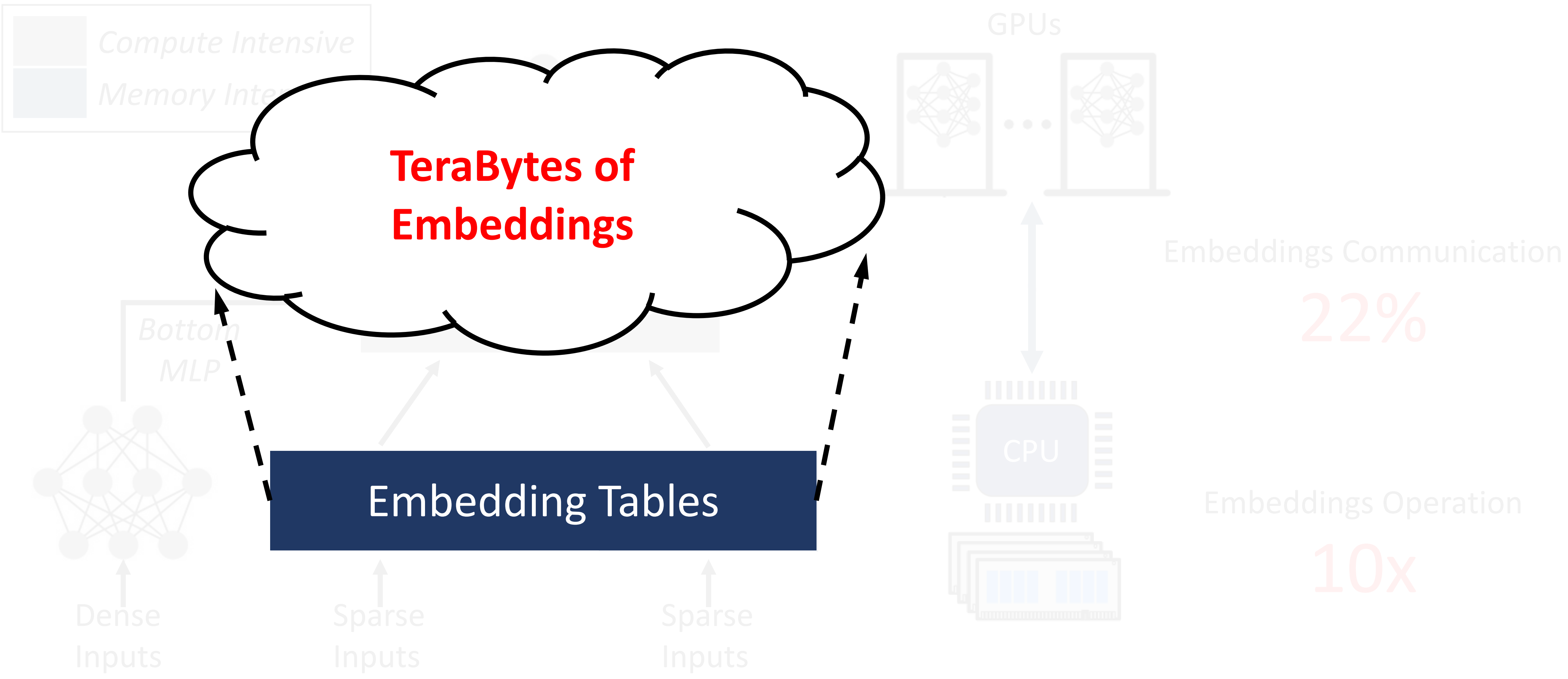
System Bottlenecks



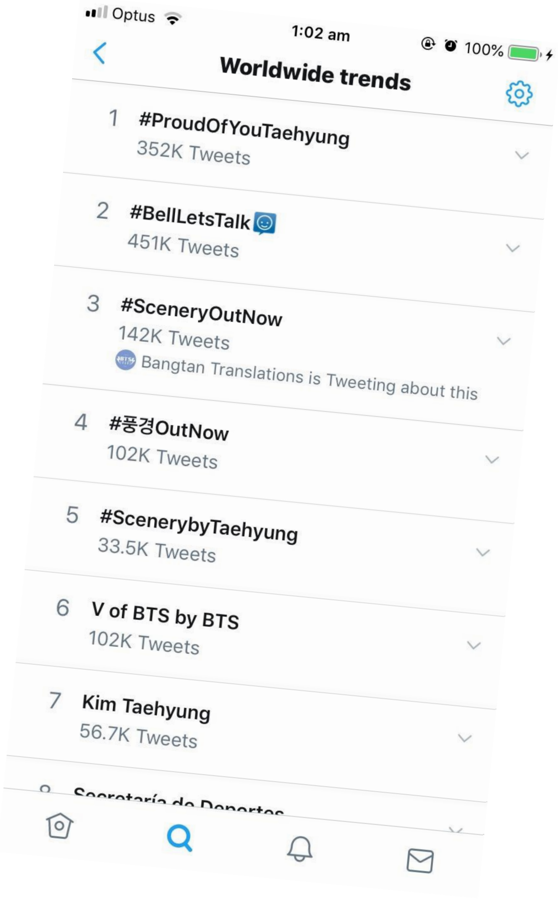
CPU-GPU Embeddings Communication: PCIe Bandwidth

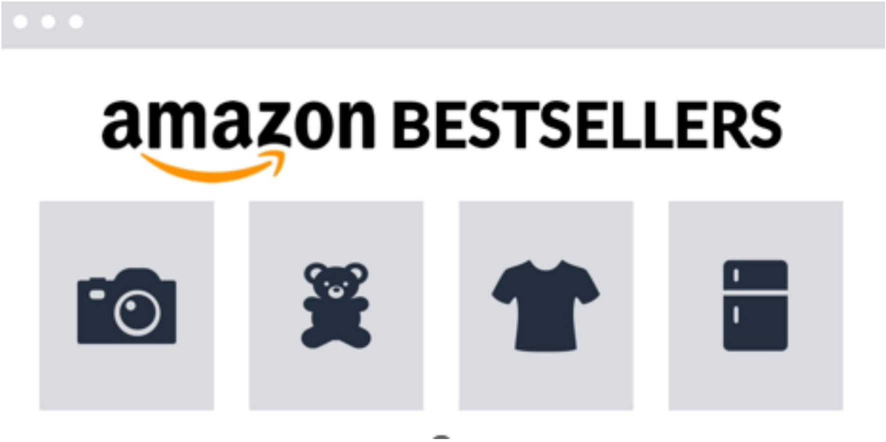
Embeddings Operations: CPU Main Memory Bandwidth

Are All Embeddings Equal?

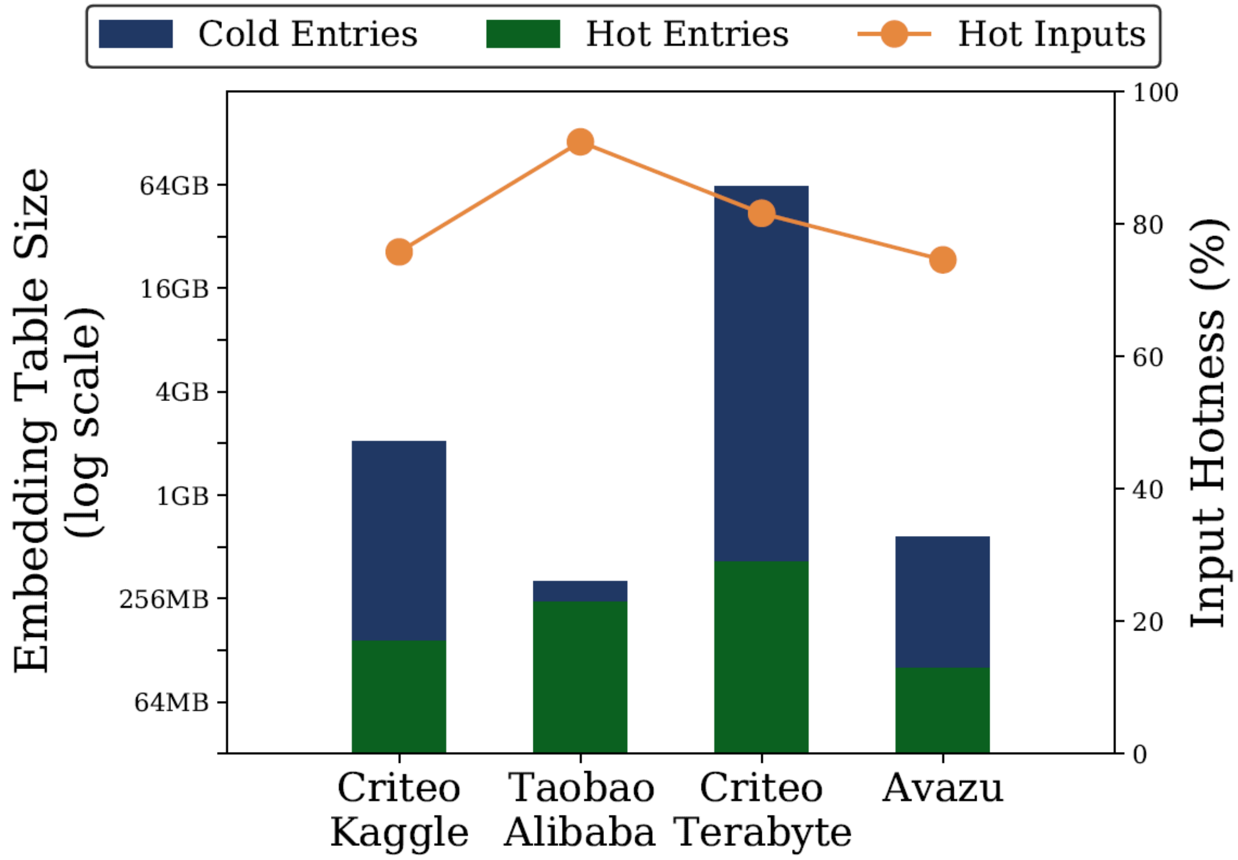


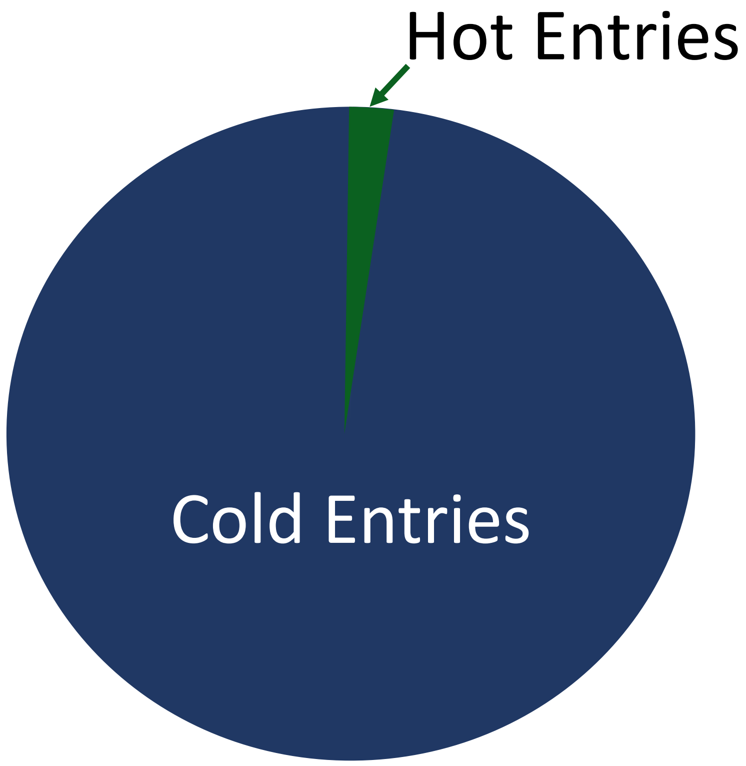
Are All Embeddings Equal?



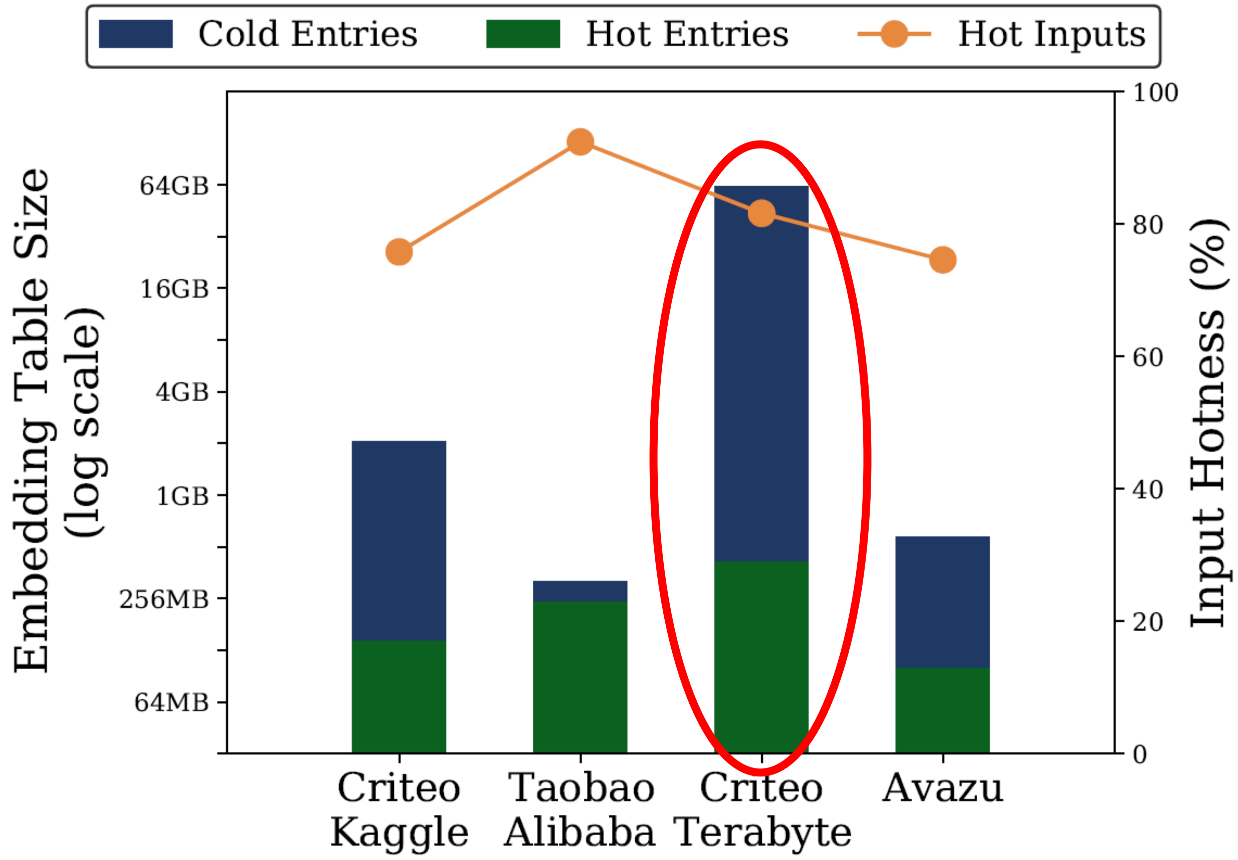


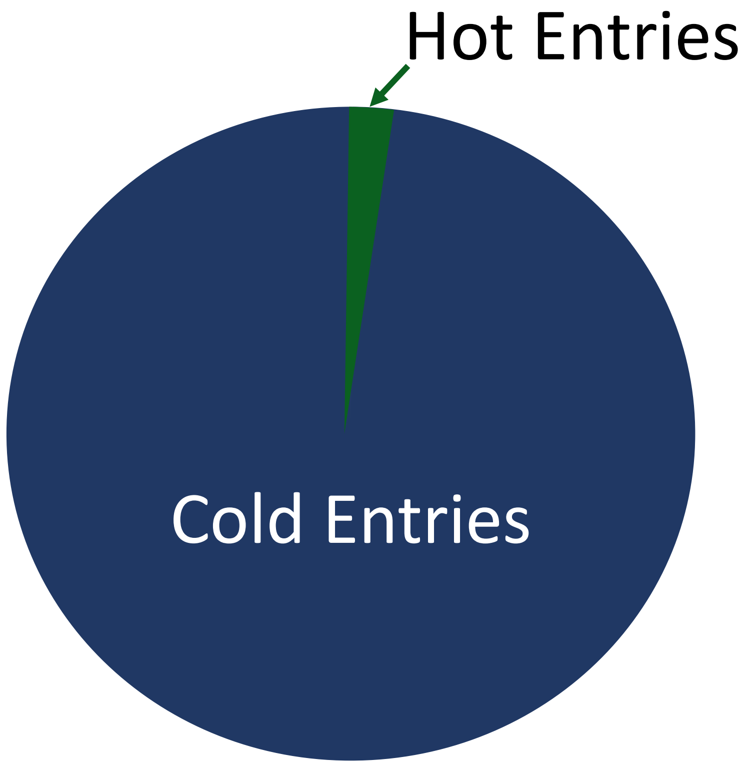
High Popularity in Training Data





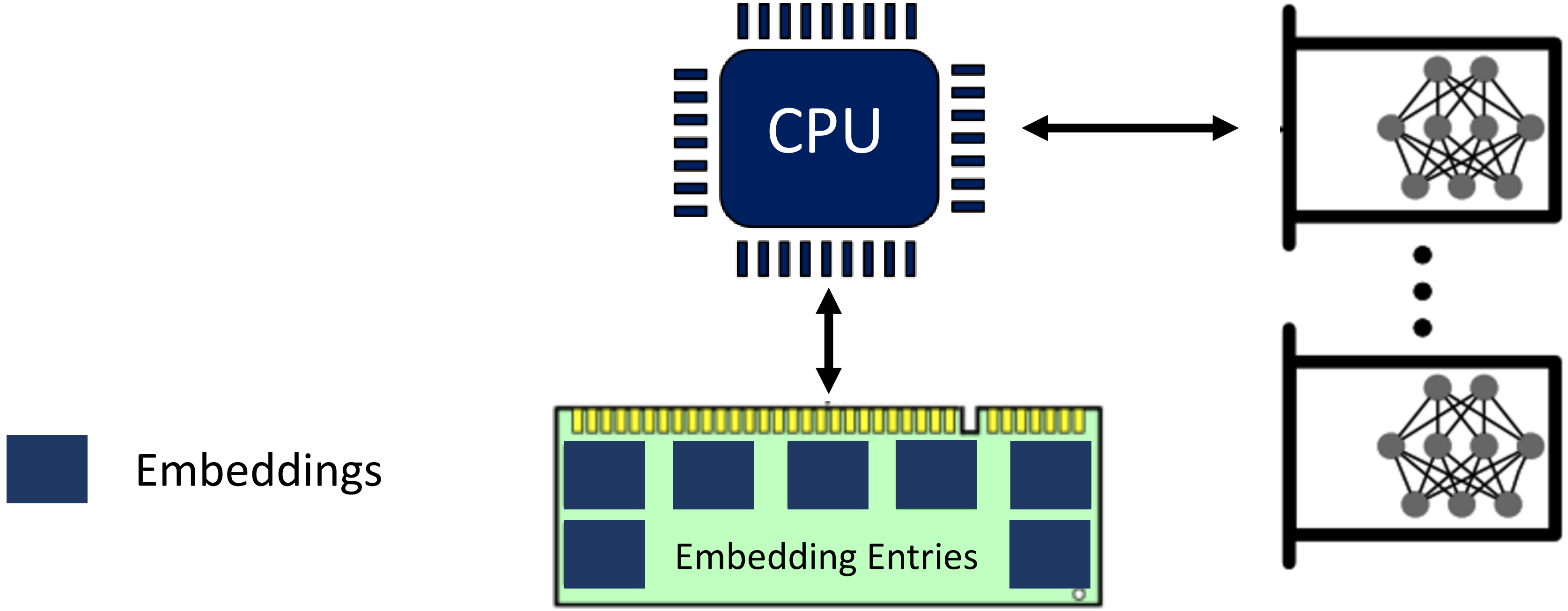
High Popularity in Training Data



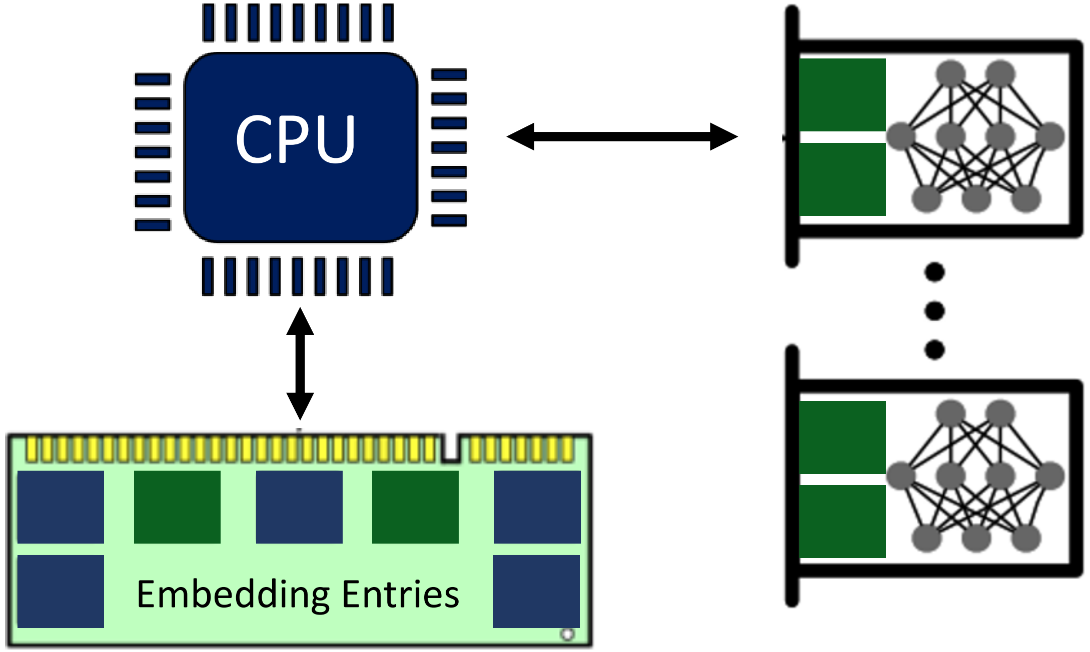


Criteo Terabyte → Hot Entries → ~512 MB (0.7%) → 82% Hot Inputs

Embedding Layout across Memories



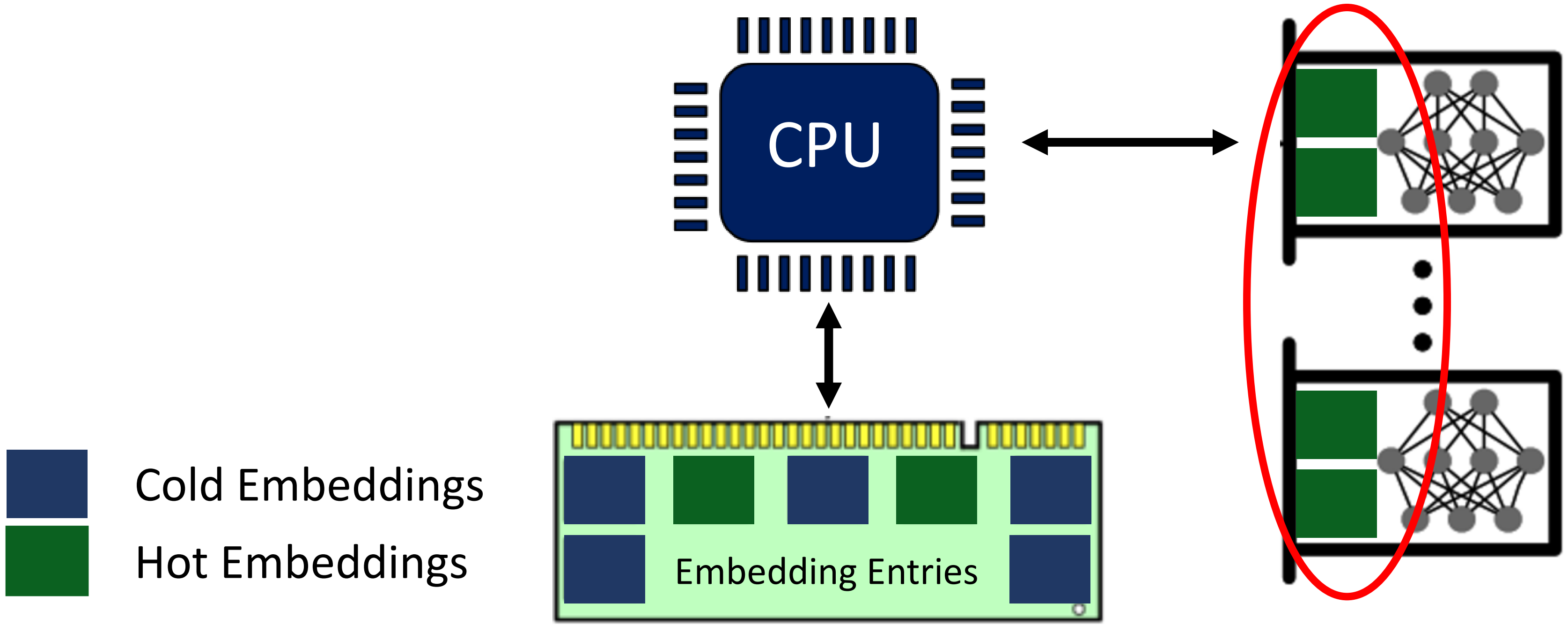
Embedding Layout across Memories



Cold Embeddings

Hot Embeddings

Embedding Layout across Memories

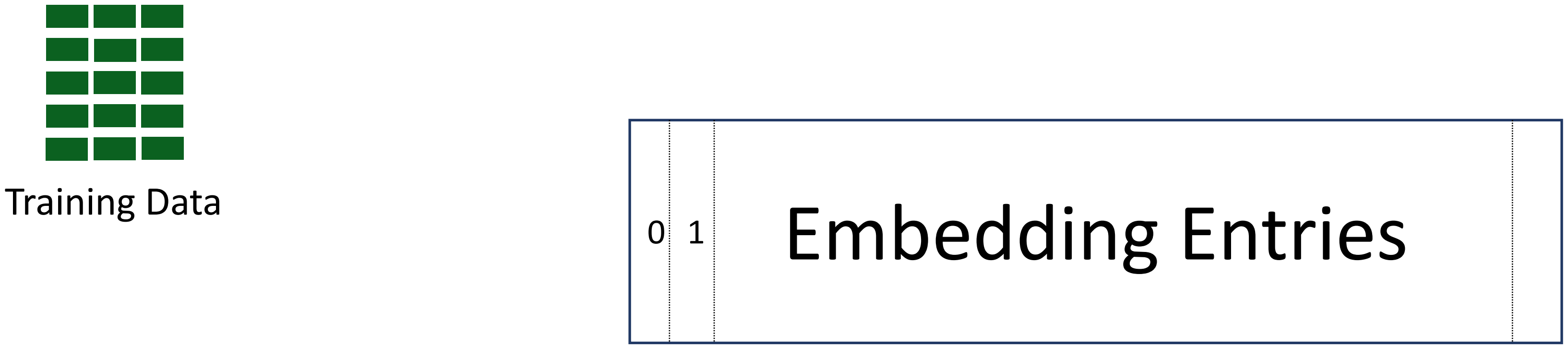


Challenges – Embedding Layout

Converting popularity into a quantifiable metric

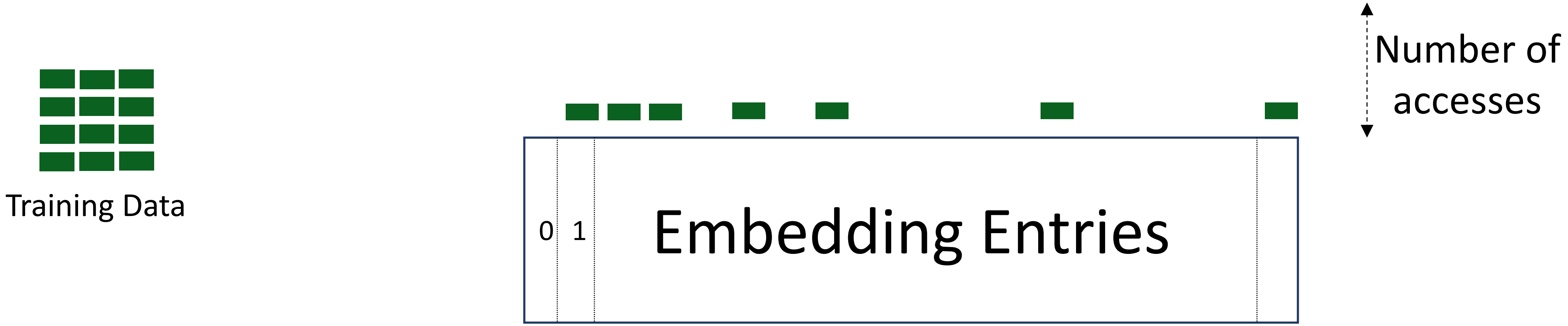
Challenges – Embedding Layout

Converting popularity into a quantifiable metric



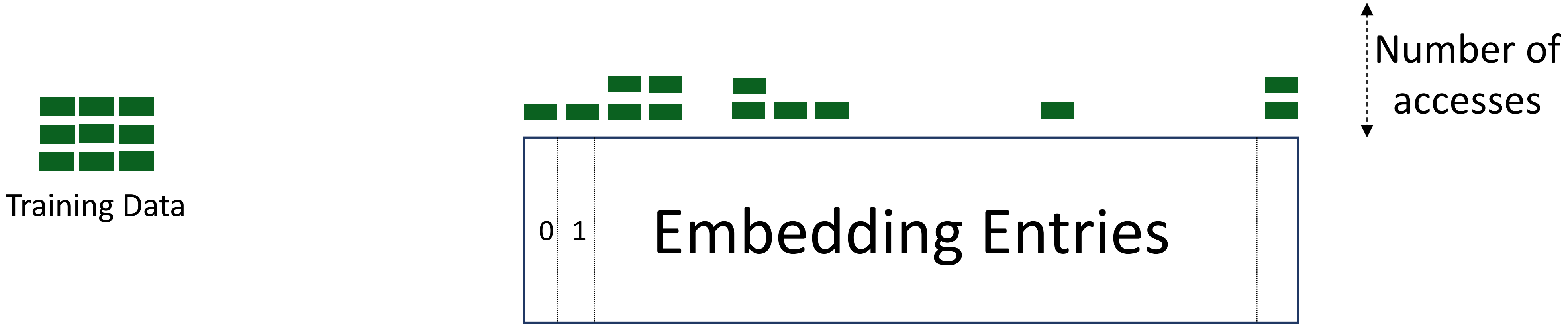
Challenges – Embedding Layout

Converting popularity into a quantifiable metric



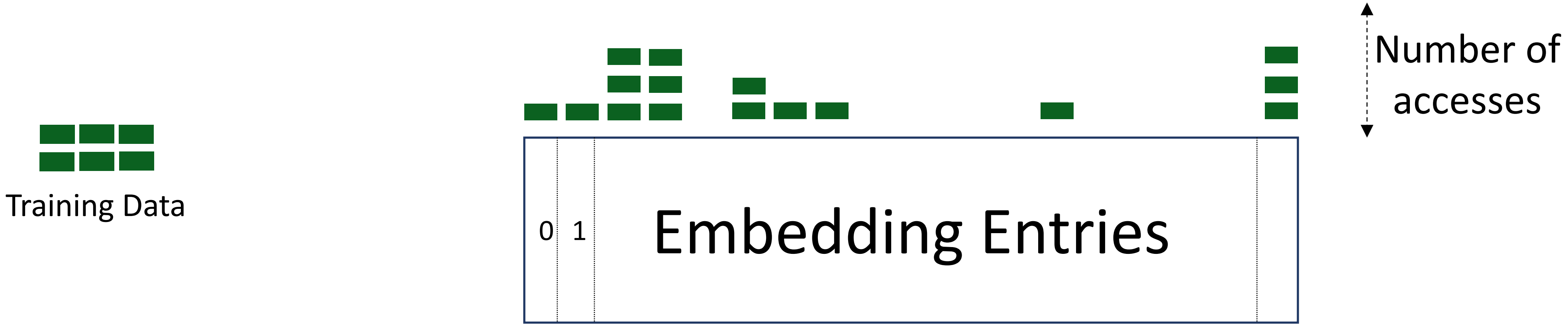
Challenges – Embedding Layout

Converting popularity into a quantifiable metric



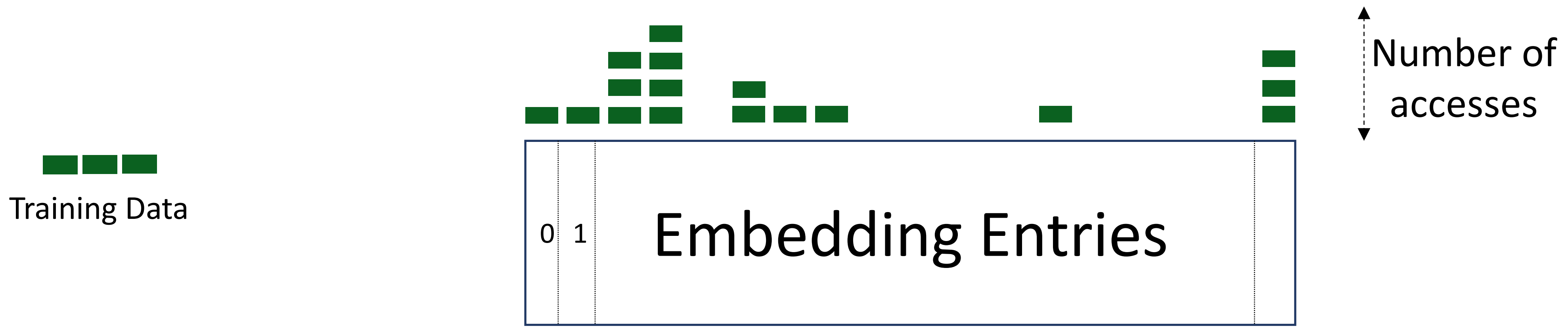
Challenges – Embedding Layout

Converting popularity into a quantifiable metric



Challenges – Embedding Layout

Converting popularity into a quantifiable metric



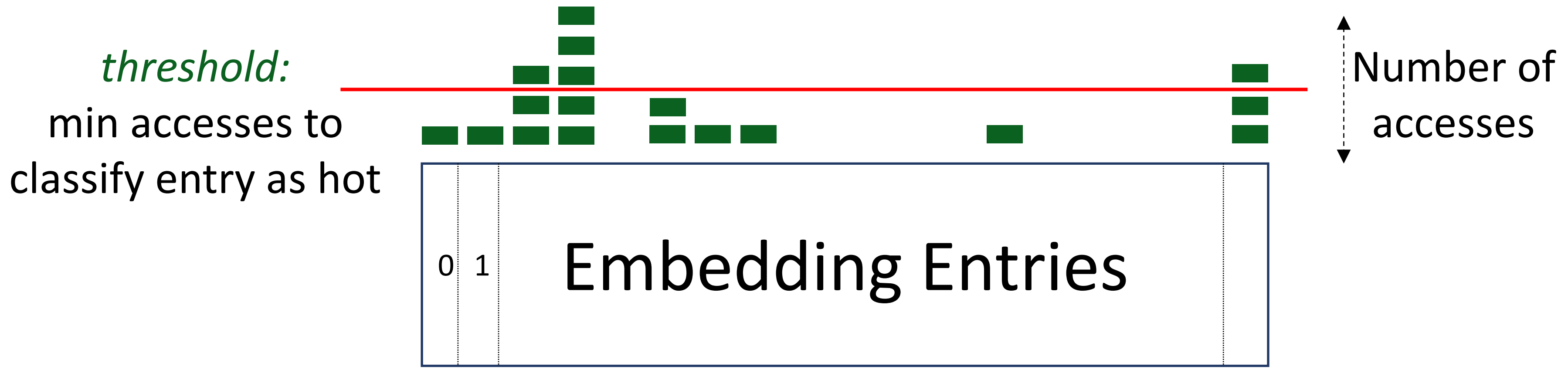
Challenges – Embedding Layout

Converting popularity into a quantifiable metric

Training Data

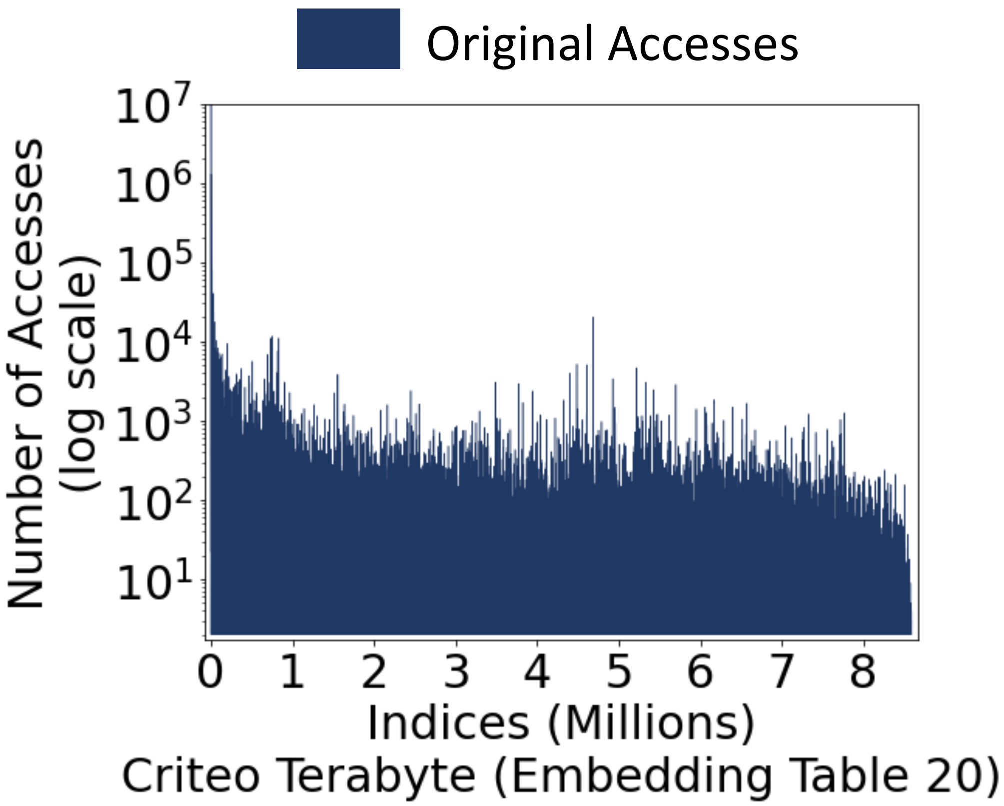
Challenges – Embedding Layout

Converting popularity into a quantifiable metric



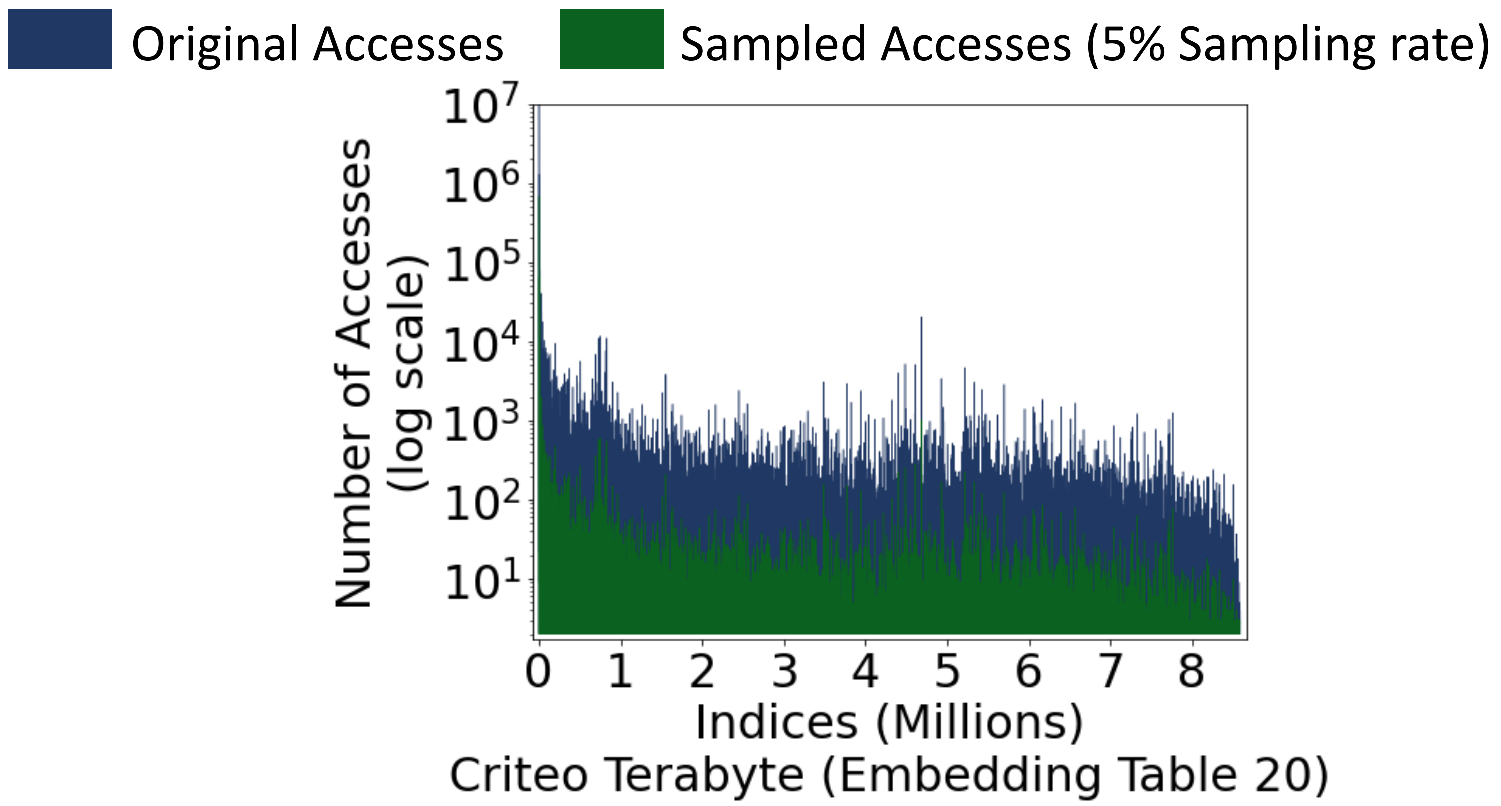
Challenges – Embedding Layout

Hot embeddings identification without parsing entire training data



Challenges – Embedding Layout

Hot embeddings identification without parsing entire training data

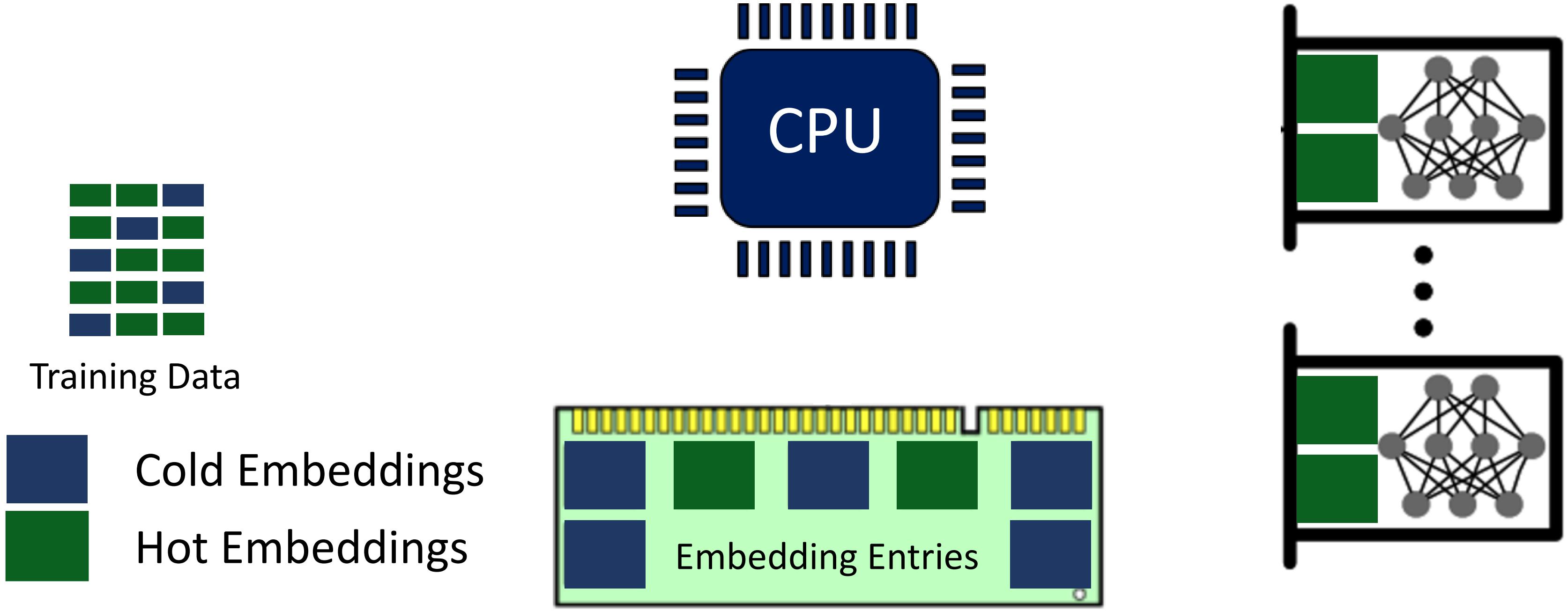


Challenges – Embedding Layout

All inputs in a mini-batch need to access the *hot embedding entries*

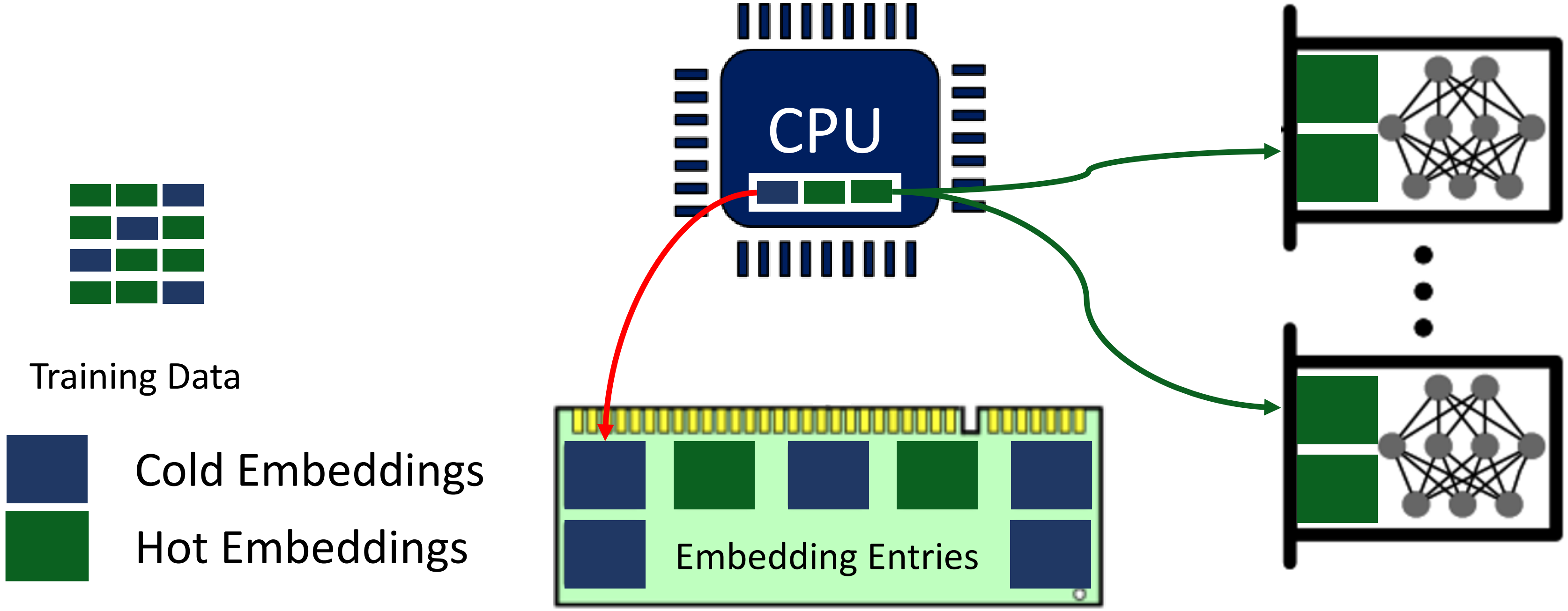
Challenges – Embedding Layout

All inputs in a mini-batch need to access the *hot embedding entries*



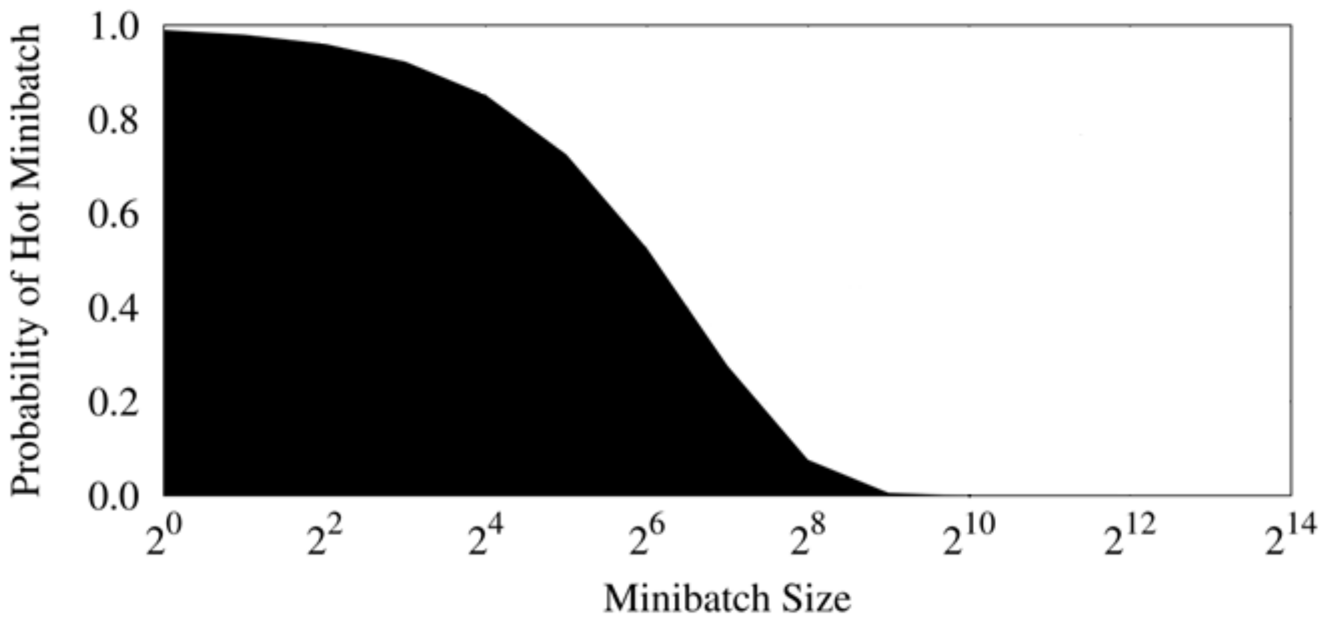
Challenges – Embedding Layout

All inputs in a mini-batch need to access the *hot embedding entries*



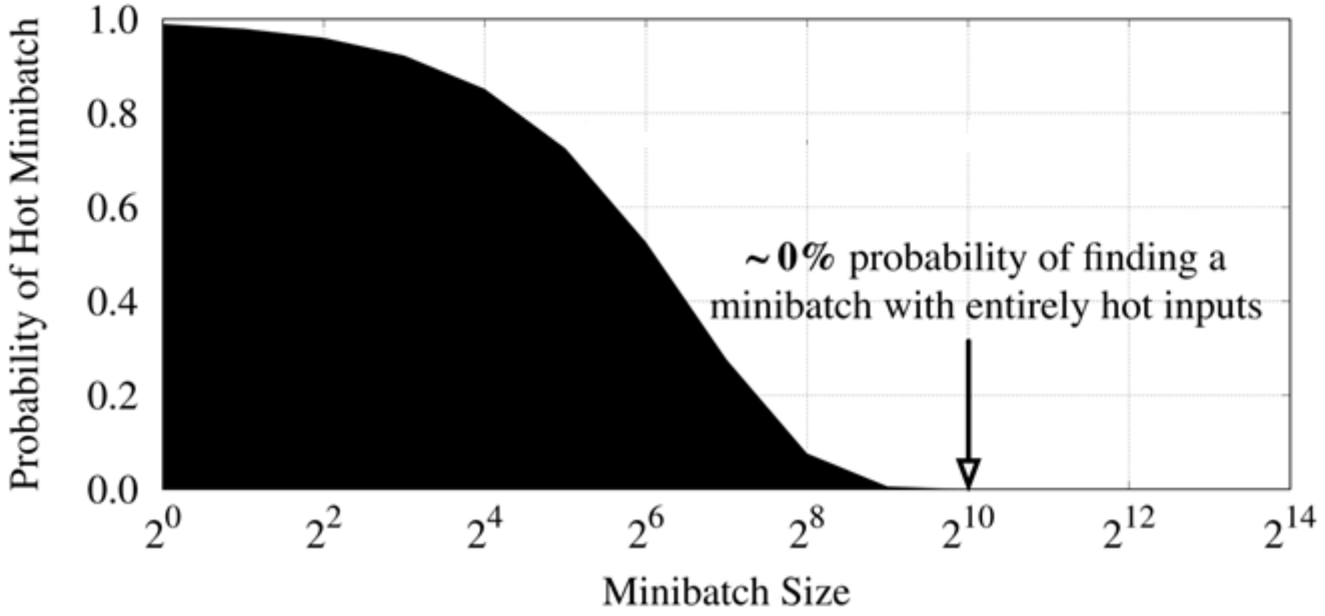
Challenges – Embedding Layout

All inputs in a mini-batch need to access the *hot embedding entries*



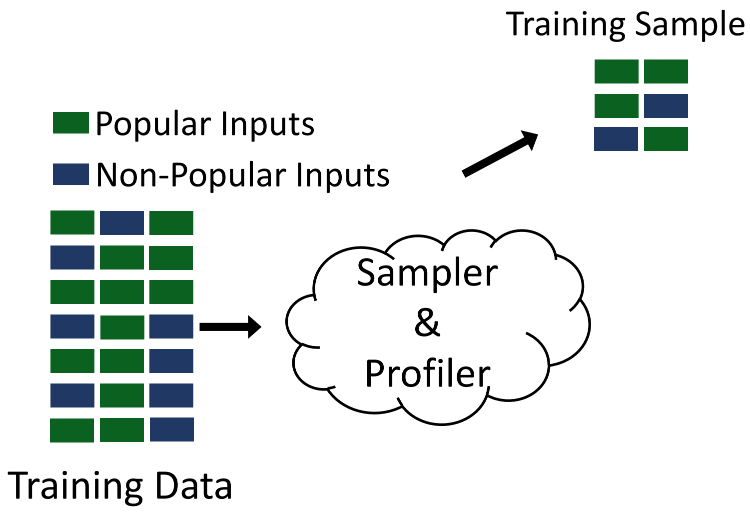
Challenges – Embedding Layout

All inputs in a mini-batch need to access the *hot embedding entries*

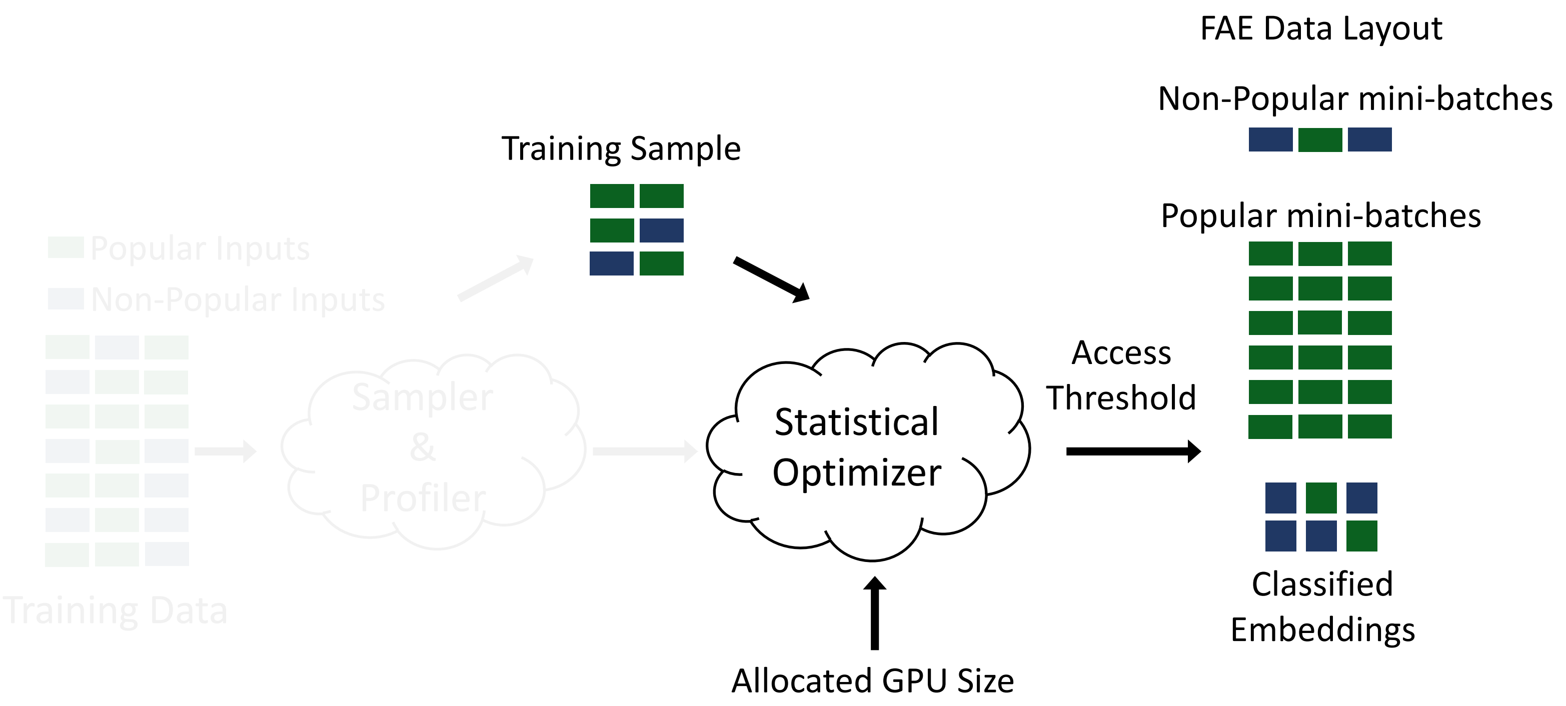


Inputs → *popular and non-popular categories* → mini-batches

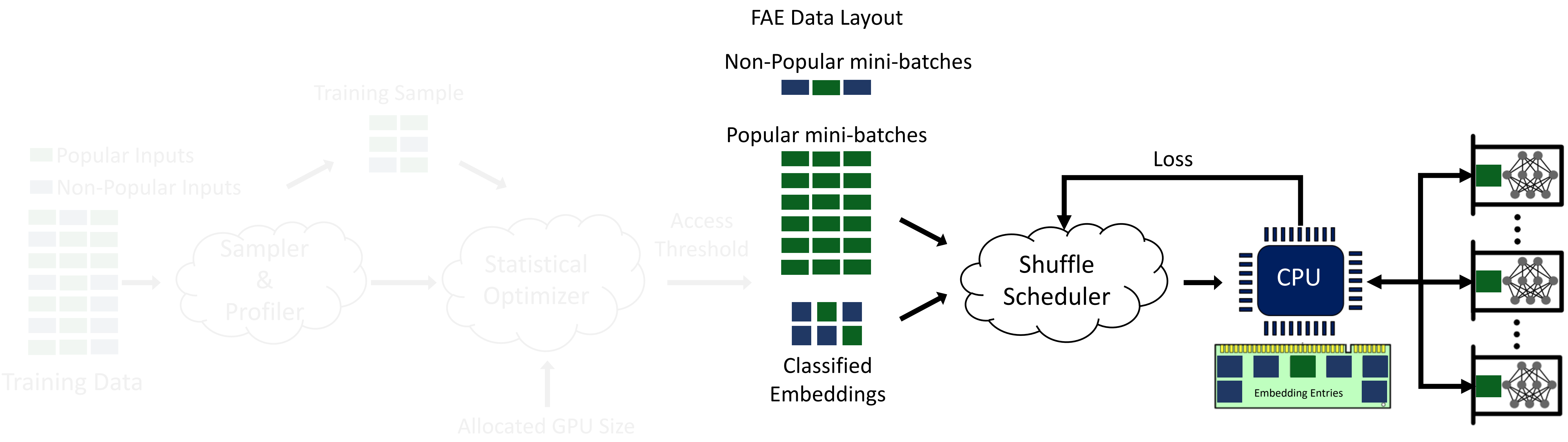
Frequently Accessed Embeddings (FAE)



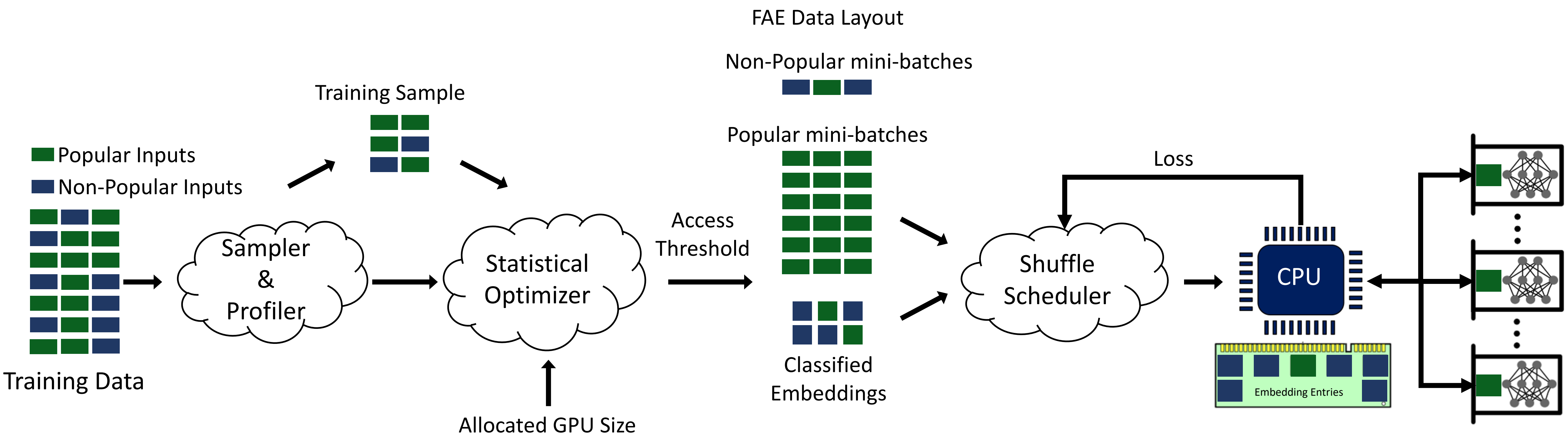
Frequently Accessed Embeddings (FAE)



Frequently Accessed Embeddings (FAE)



Frequently Accessed Embeddings (FAE)



FAE Framework-Mitigating System Bottlenecks

CPU-GPU Embeddings Communication: PCIe Bandwidth

No CPU-GPU Communication for Popular Mini-batches

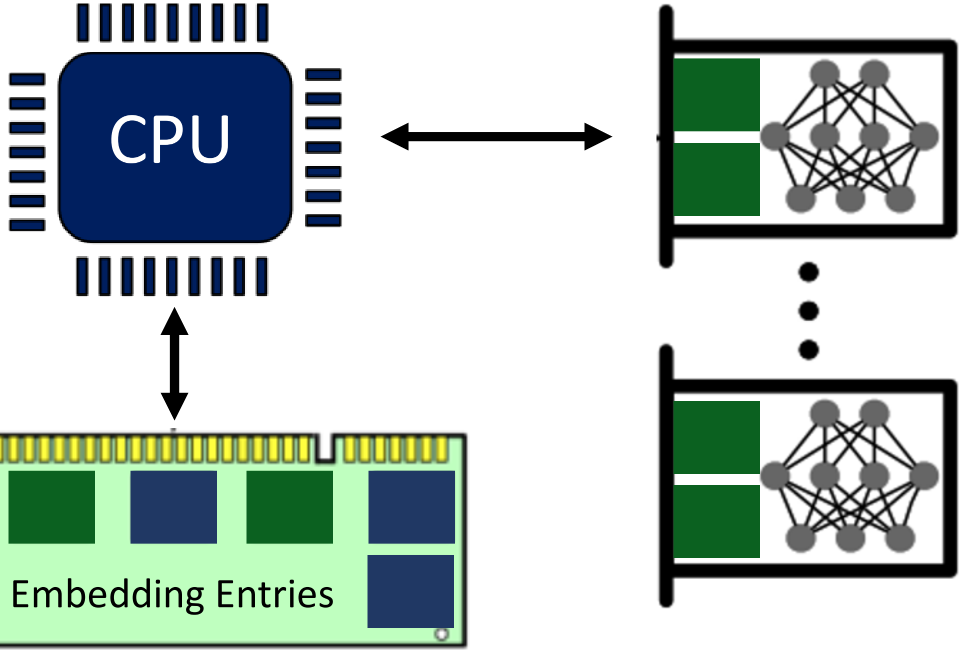
Embeddings Operations: CPU Main Memory Bandwidth

Popular Mini-batch → High Bandwidth GPU Memory

Evaluation: System Setup

CPU: Intel Xeon Silver 4116 (2.1GHz)

DRAM:768 GB DDR4 (2.7GB/s)



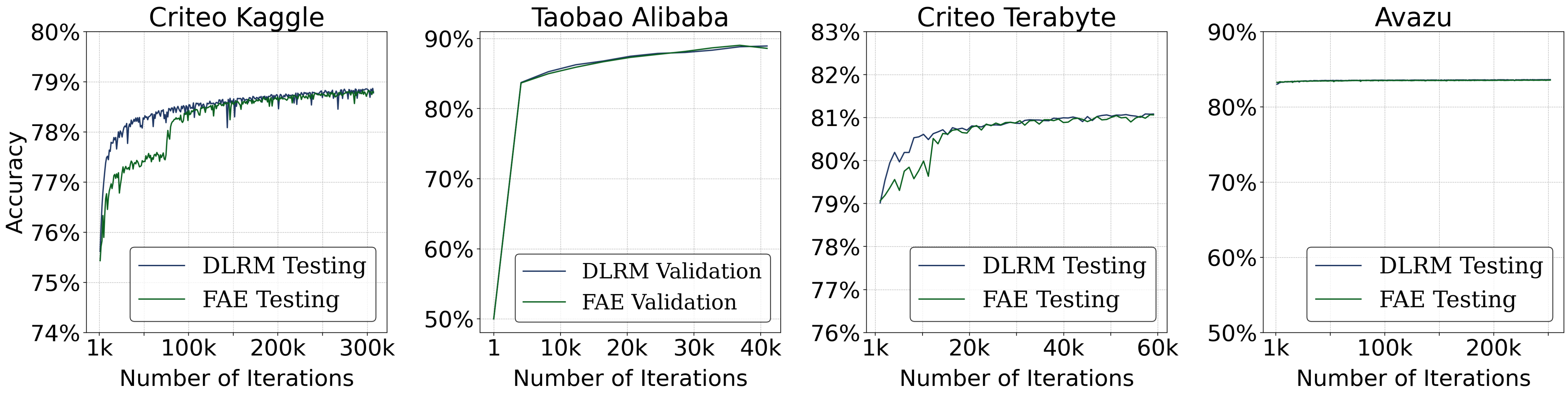
GPU: Nvidia Tesla V100 16GB *HBM* device memory

Baselines and Benchmarks

Baseline	XDL[1]	Open Source DLRM[2]		
Datasets	Criteo Terabyte	Criteo Kaggle	Taobao Alibaba	Avazu

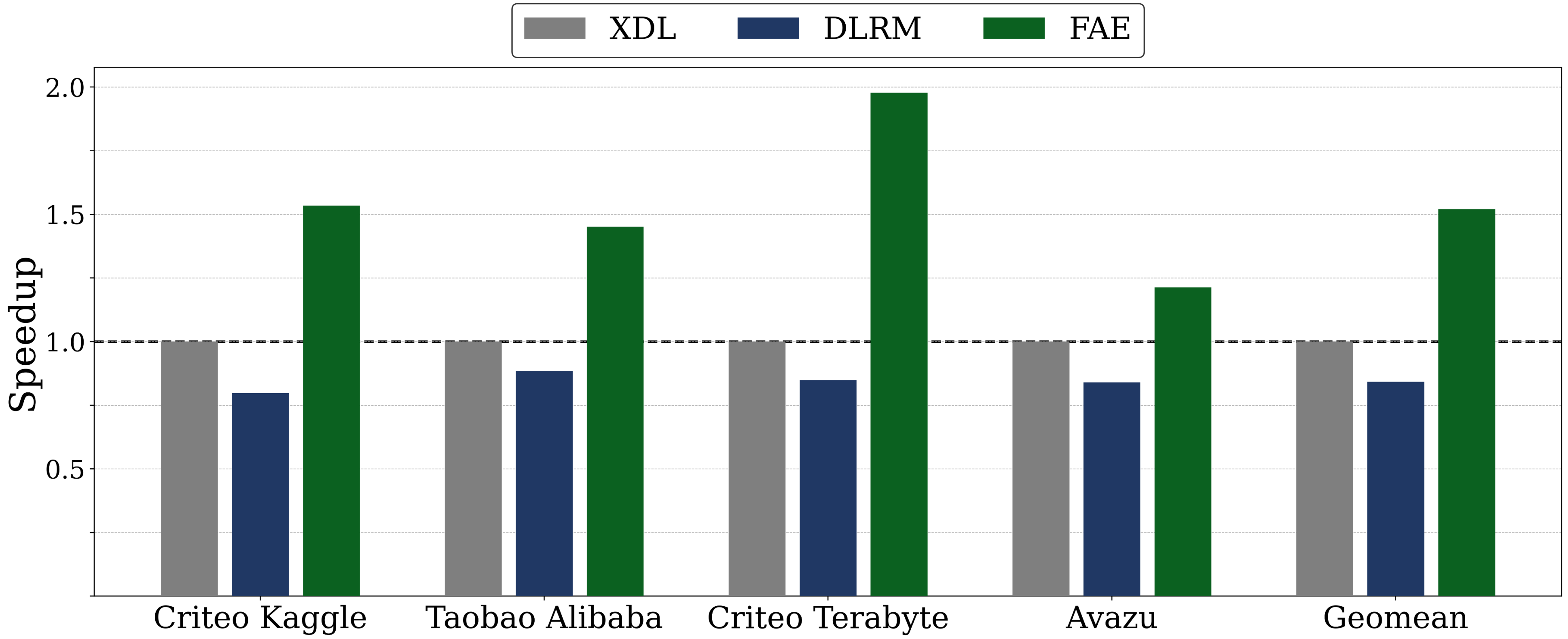
1 - Jiang et al. DLP-KDD'19
2 - Naumov et al. arXiv'19

Accuracy Comparison



FAE always achieves baseline accuracy

Performance Comparison



FAE → 1.5x in comparison to XDL

FAE → 1.8x in comparison to DLRM baseline with 4-GPUs

Conclusion

- FAE meets the baseline accuracy across all models and datasets

- Accelerates training
  - 1.5x compared to XDL
  - 1.8x compared to DLRM

Questions?

adnan@ece.ubc.ca

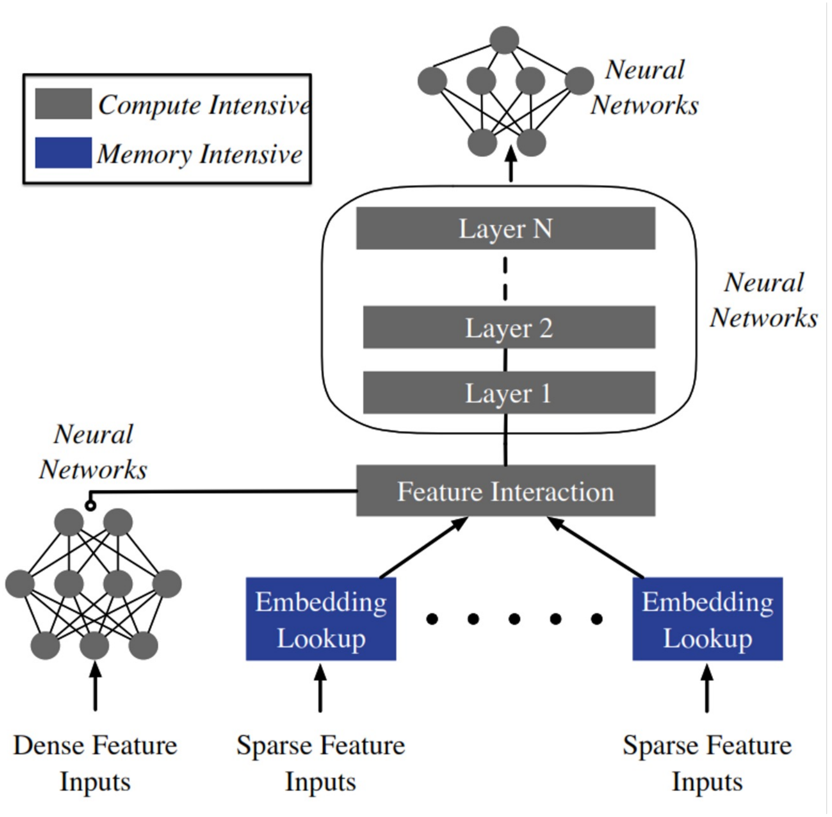
http://people.ece.ubc.ca/adnan/

madnan_92

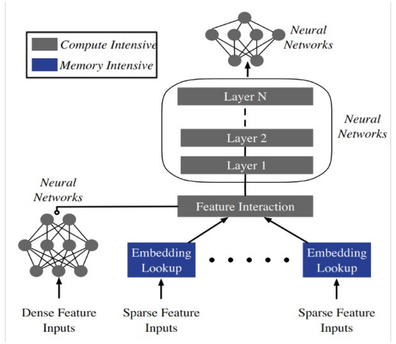
https://github.com/STAR-Laboratory/Accelerating-RecSys-Training

Backup Slides

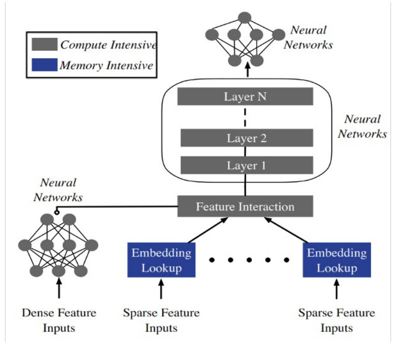
Hybrid Execution Training Flow

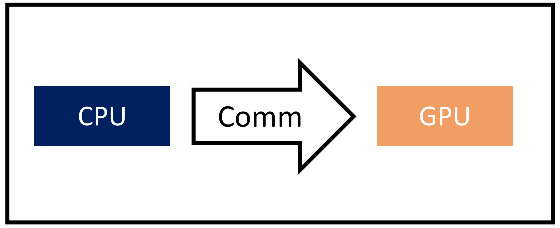


Hybrid Execution Training Flow

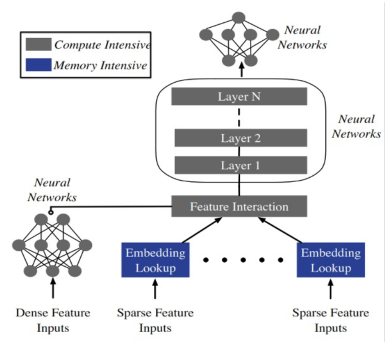


Hybrid Execution Training Flow





Hybrid Execution Training Flow

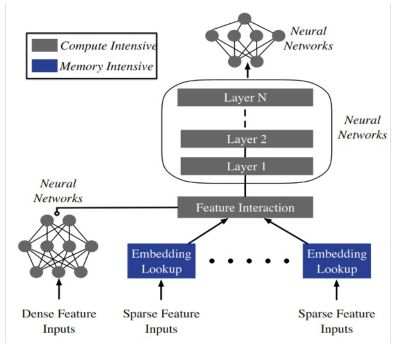


CPU → Comm → GPU

Read Training Mini-batch → Embedding Lookup → Transfer Embeddings → Neural Network Forward → Feature Interaction Forward → Neural Network Forward

Forward Path

Hybrid Execution Training Flow



CPU → Comm → GPU

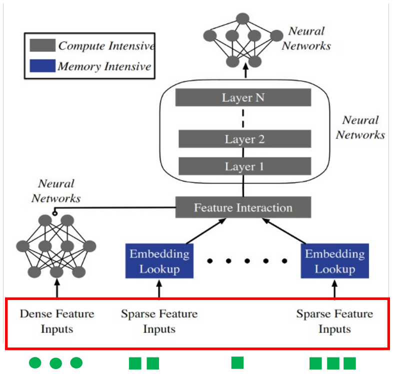
Read Training Mini-batch | Embedding Lookup | Transfer Embeddings | Neural Network Forward | Feature Interaction Forward | Neural Network Forward

Forward Path

Embedding Optimizer / Neural Network Optimizer | Transfer Emb Gradients | Embeddings Backward | Neural Network Backward | Feature Interaction Backward | Neural Network Backward

Backward Path

Hybrid Execution Training Flow



CPU

Comm

GPU

Read Training Mini-batch

Embedding Lookup

Transfer Embeddings

Neural Network Forward

Feature Interaction Forward

Neural Network Forward

Forward Path

Embedding Optimizer

Neural Network Optimizer

Transfer Emb Gradients

Embeddings Backward

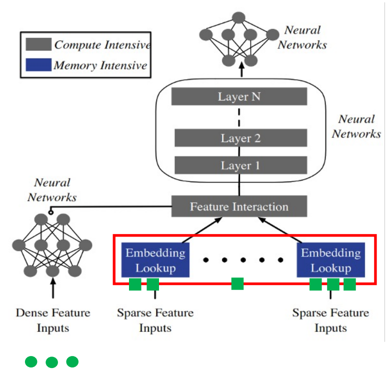
Neural Network Backward

Feature Interaction Backward

Neural Network Backward

Backward Path

Hybrid Execution Training Flow



CPU → Comm → GPU

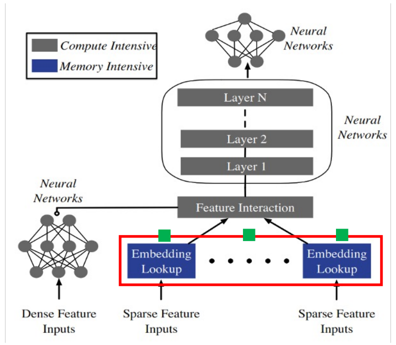
Read Training Mini-batch → Embedding Lookup → Transfer Embeddings → Neural Network Forward → Feature Interaction Forward → Neural Network Forward

Forward Path

Embedding Optimizer / Neural Network Optimizer ← Transfer Emb Gradients ← Embeddings Backward ← Neural Network Backward ← Feature Interaction Backward ← Neural Network Backward

Backward Path

Hybrid Execution Training Flow



CPU

Comm

GPU

Read Training Mini-batch

Embedding Lookup

Transfer Embeddings

Neural Network Forward

Feature Interaction Forward

Neural Network Forward

Forward Path

Embedding Optimizer

Neural Network Optimizer

Transfer Emb Gradients

Embeddings Backward

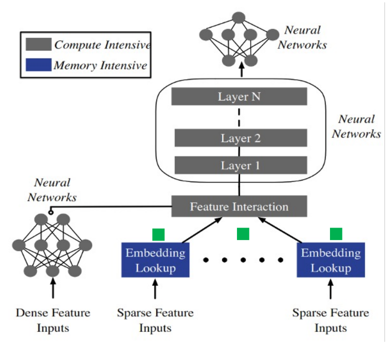
Neural Network Backward

Feature Interaction Backward

Neural Network Backward

Backward Path

Hybrid Execution Training Flow



CPU

Comm

GPU

Read Training Mini-batch

Embedding Lookup

Transfer Embeddings

Neural Network Forward

Feature Interaction Forward

Neural Network Forward

Forward Path

Embedding Optimizer

Neural Network Optimizer

Transfer Emb Gradients

Embeddings Backward

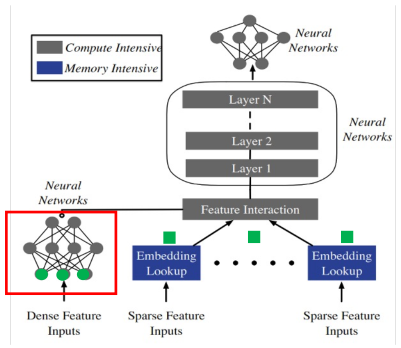
Neural Network Backward

Feature Interaction Backward

Neural Network Backward

Backward Path

Hybrid Execution Training Flow



CPU → Comm → GPU

Read Training Mini-batch

Embedding Lookup

Transfer Embeddings

Neural Network Forward

Feature Interaction Forward

Neural Network Forward

Forward Path

Embedding Optimizer

Neural Network Optimizer

Transfer Emb Gradients

Embeddings Backward

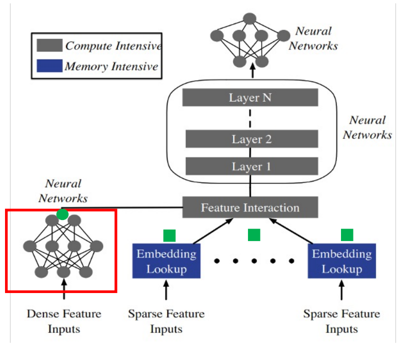
Neural Network Backward

Feature Interaction Backward

Neural Network Backward

Backward Path

Hybrid Execution Training Flow



CPU | Comm | GPU

Read Training Mini-batch | Embedding Lookup | Transfer Embeddings | Neural Network Forward | Feature Interaction Forward | Neural Network Forward

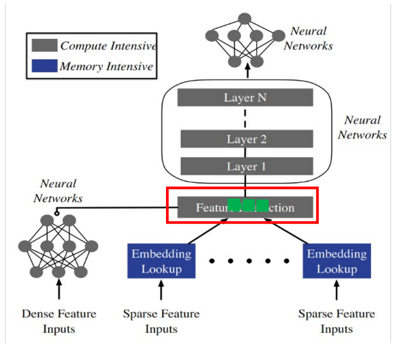
Forward Path

Embedding Optimizer | Transfer Emb Gradients | Embeddings Backward | Neural Network Backward | Feature Interaction Backward | Neural Network Backward

Neural Network Optimizer

Backward Path

Hybrid Execution Training Flow



CPU → Comm → GPU

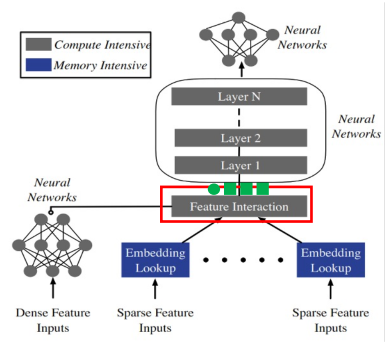
Read Training Mini-batch → Embedding Lookup → Transfer Embeddings → Neural Network Forward → Feature Interaction Forward → Neural Network Forward

Forward Path

Embedding Optimizer / Neural Network Optimizer ← Transfer Emb Gradients ← Embeddings Backward ← Neural Network Backward ← Feature Interaction Backward ← Neural Network Backward

Backward Path

Hybrid Execution Training Flow



CPU → Comm → GPU

Read Training Mini-batch → Embedding Lookup → Transfer Embeddings → Neural Network Forward → Feature Interaction Forward → Neural Network Forward

Forward Path

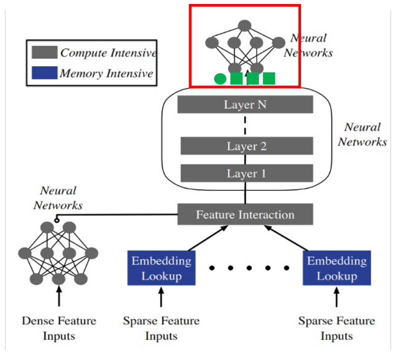
Embedding Optimizer

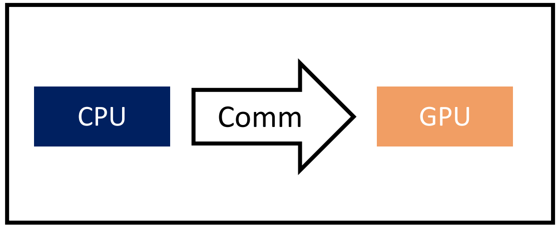
Neural Network Optimizer

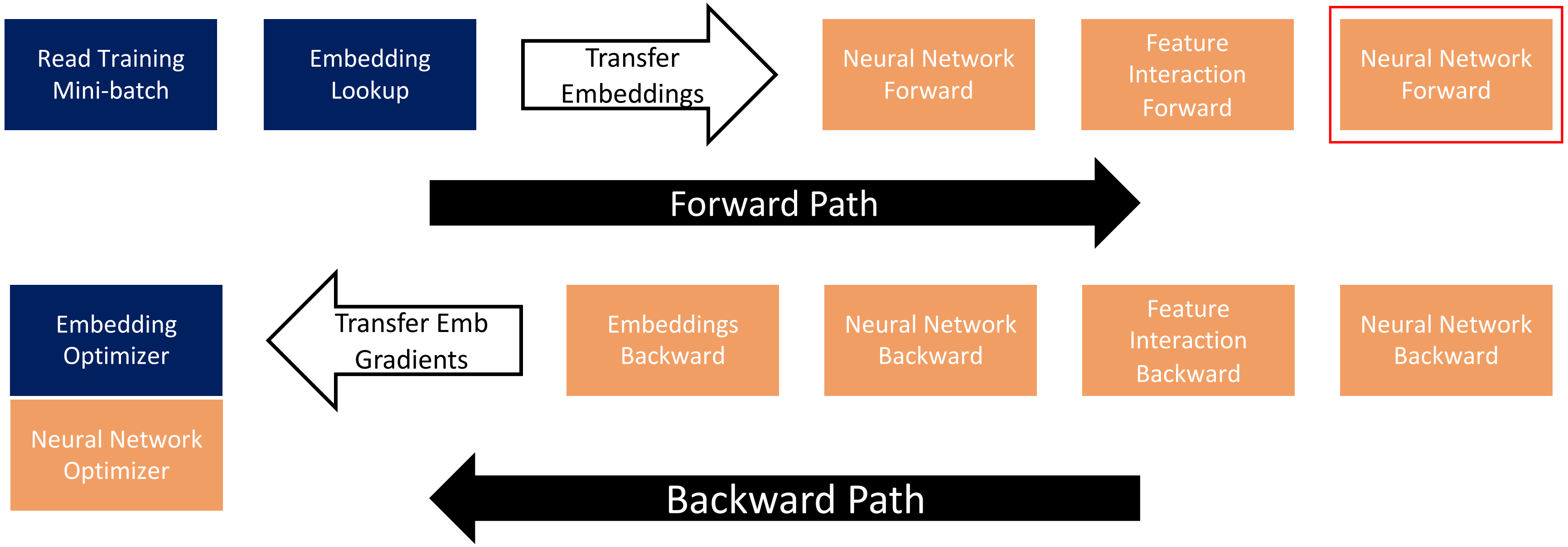
Transfer Emb Gradients ← Embeddings Backward ← Neural Network Backward ← Feature Interaction Backward ← Neural Network Backward

Backward Path

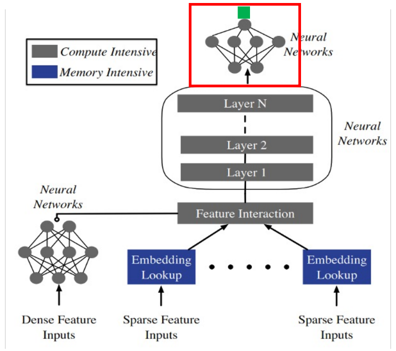
Hybrid Execution Training Flow







Hybrid Execution Training Flow



CPU → Comm → GPU

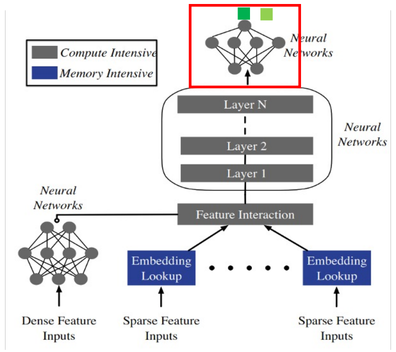
Read Training Mini-batch → Embedding Lookup → Transfer Embeddings → Neural Network Forward → Feature Interaction Forward → Neural Network Forward

Forward Path

Embedding Optimizer / Neural Network Optimizer ← Transfer Emb Gradients ← Embeddings Backward ← Neural Network Backward ← Feature Interaction Backward ← Neural Network Backward

Backward Path

Hybrid Execution Training Flow



CPU → Comm → GPU

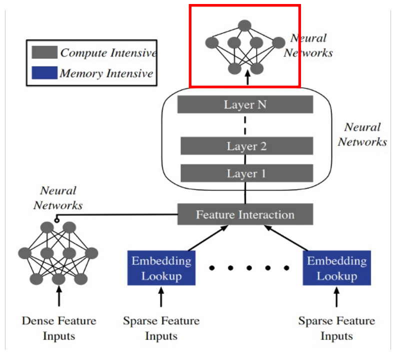
Read Training Mini-batch | Embedding Lookup | Transfer Embeddings → | Neural Network Forward | Feature Interaction Forward | Neural Network Forward

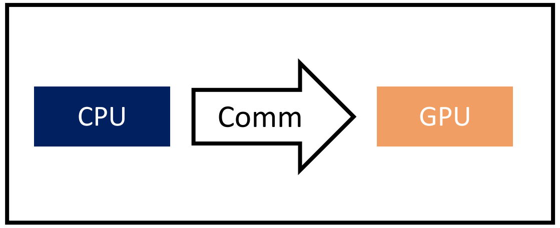
Forward Path →

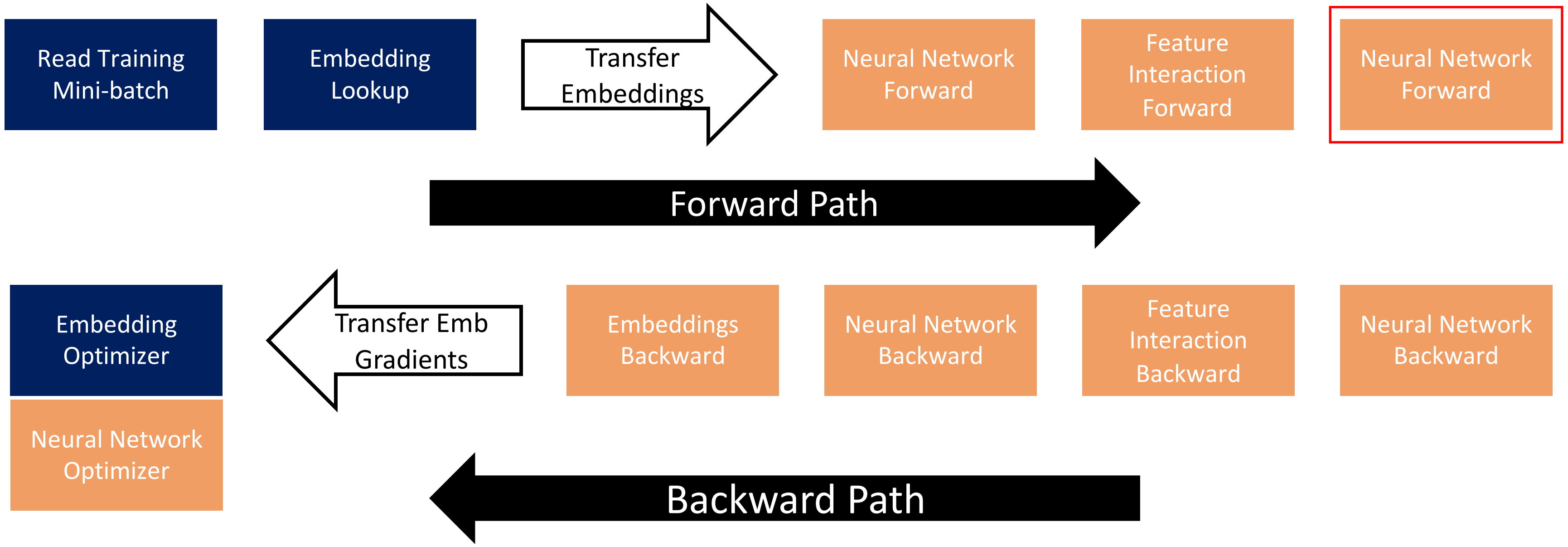
Embedding Optimizer / Neural Network Optimizer | ← Transfer Emb Gradients | Embeddings Backward | Neural Network Backward | Feature Interaction Backward | Neural Network Backward

← Backward Path

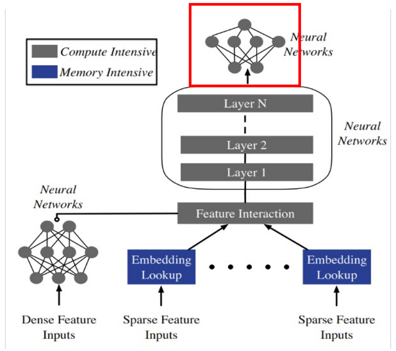
Hybrid Execution Training Flow







Hybrid Execution Training Flow



CPU → Comm → GPU

Read Training Mini-batch | Embedding Lookup | Transfer Embeddings | Neural Network Forward | Feature Interaction Forward | Neural Network Forward

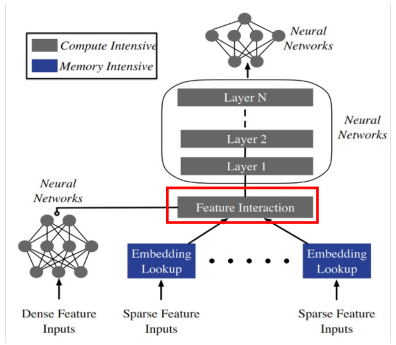
Forward Path

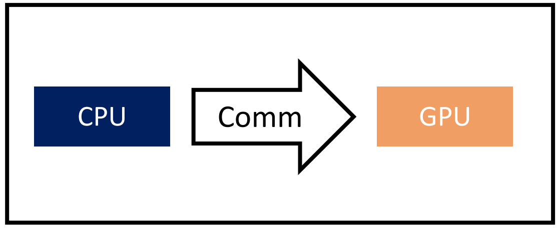
Embedding Optimizer | Transfer Emb Gradients | Embeddings Backward | Neural Network Backward | Feature Interaction Backward | Neural Network Backward

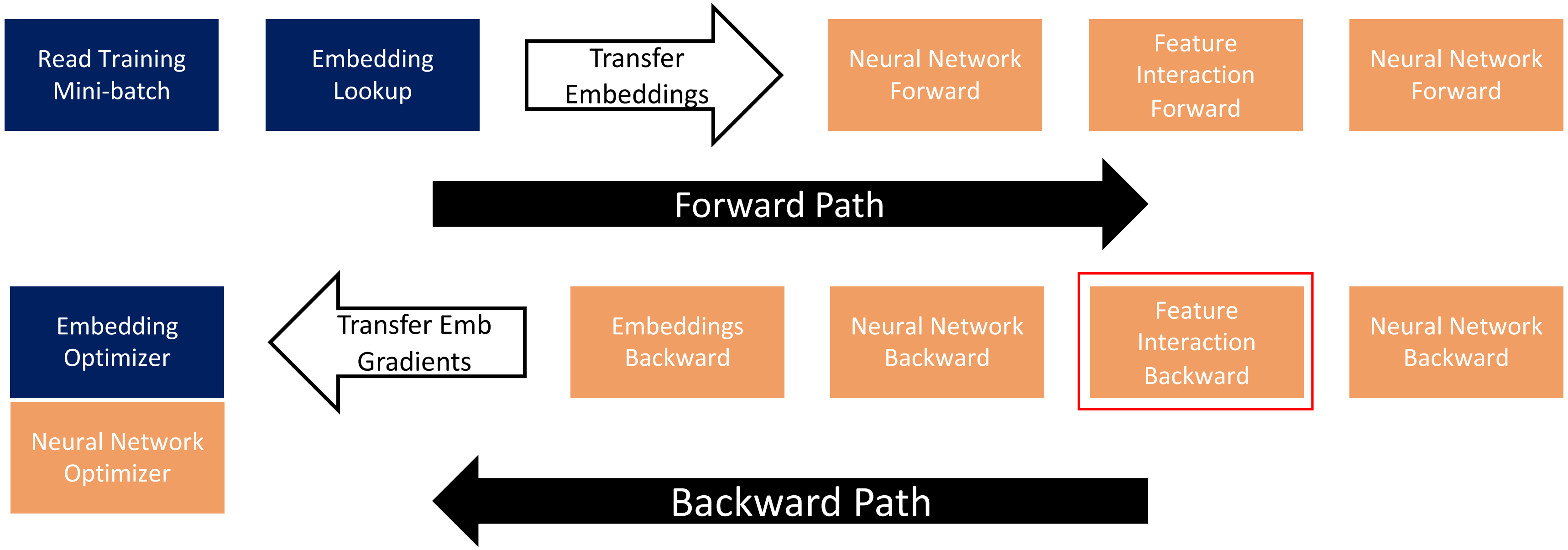
Neural Network Optimizer

Backward Path

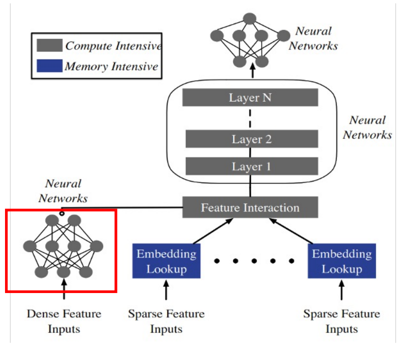
Hybrid Execution Training Flow







Hybrid Execution Training Flow



CPU → Comm → GPU

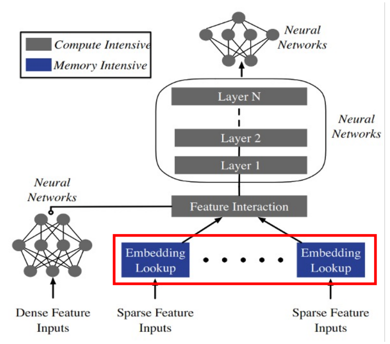
Read Training Mini-batch | Embedding Lookup | Transfer Embeddings → | Neural Network Forward | Feature Interaction Forward | Neural Network Forward

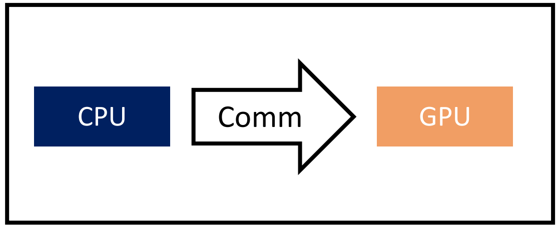
Forward Path →

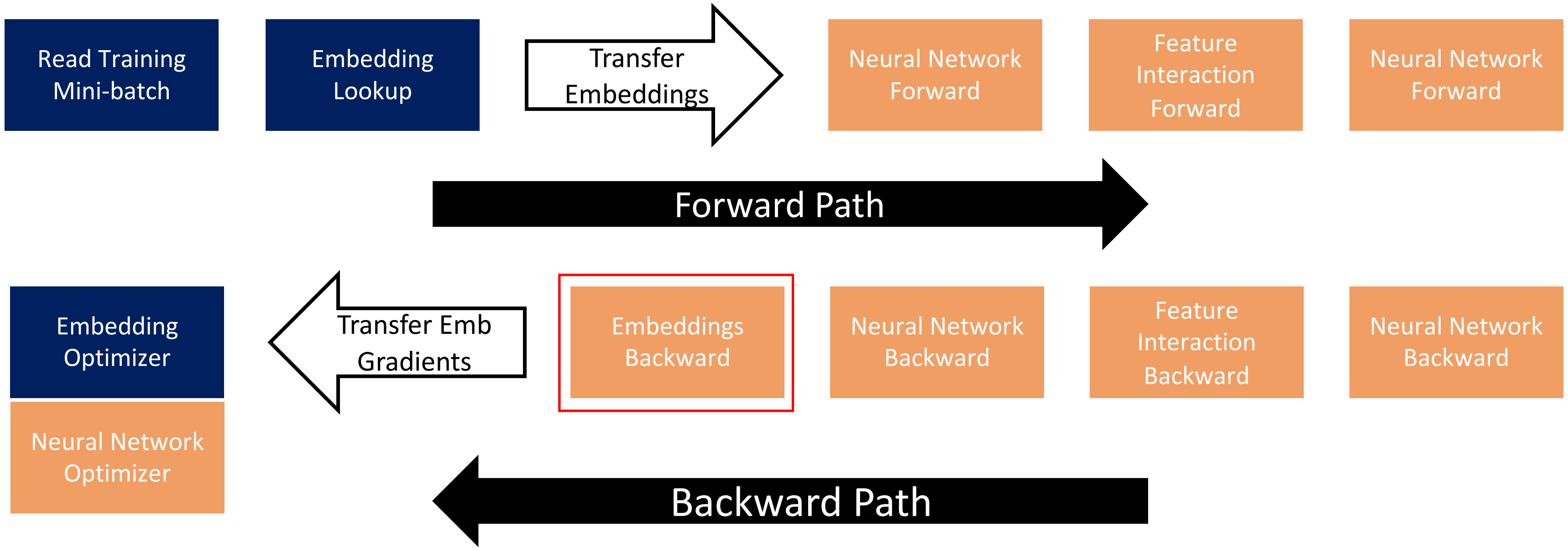
Embedding Optimizer / Neural Network Optimizer | ← Transfer Emb Gradients | Embeddings Backward | Neural Network Backward | Feature Interaction Backward | Neural Network Backward

← Backward Path

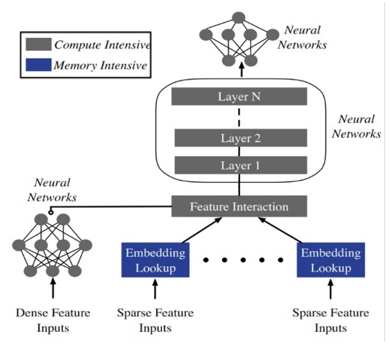
Hybrid Execution Training Flow







Hybrid Execution Training Flow



CPU

Comm

GPU

Read Training Mini-batch

Embedding Lookup

Transfer Embeddings

Neural Network Forward

Feature Interaction Forward

Neural Network Forward

Forward Path

Embedding Optimizer

Neural Network Optimizer

Transfer Emb Gradients

Embeddings Backward

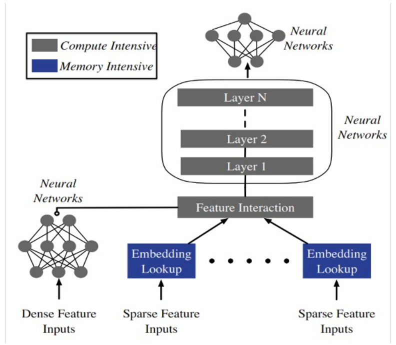
Neural Network Backward

Feature Interaction Backward

Neural Network Backward

Backward Path

Hybrid Execution Training Flow



CPU → Comm → GPU

Read Training Mini-batch

Embedding Lookup

Transfer Embeddings

Neural Network Forward

Feature Interaction Forward

Neural Network Forward

Forward Path

Embedding Optimizer

Neural Network Optimizer

Transfer Emb Gradients

Embeddings Backward

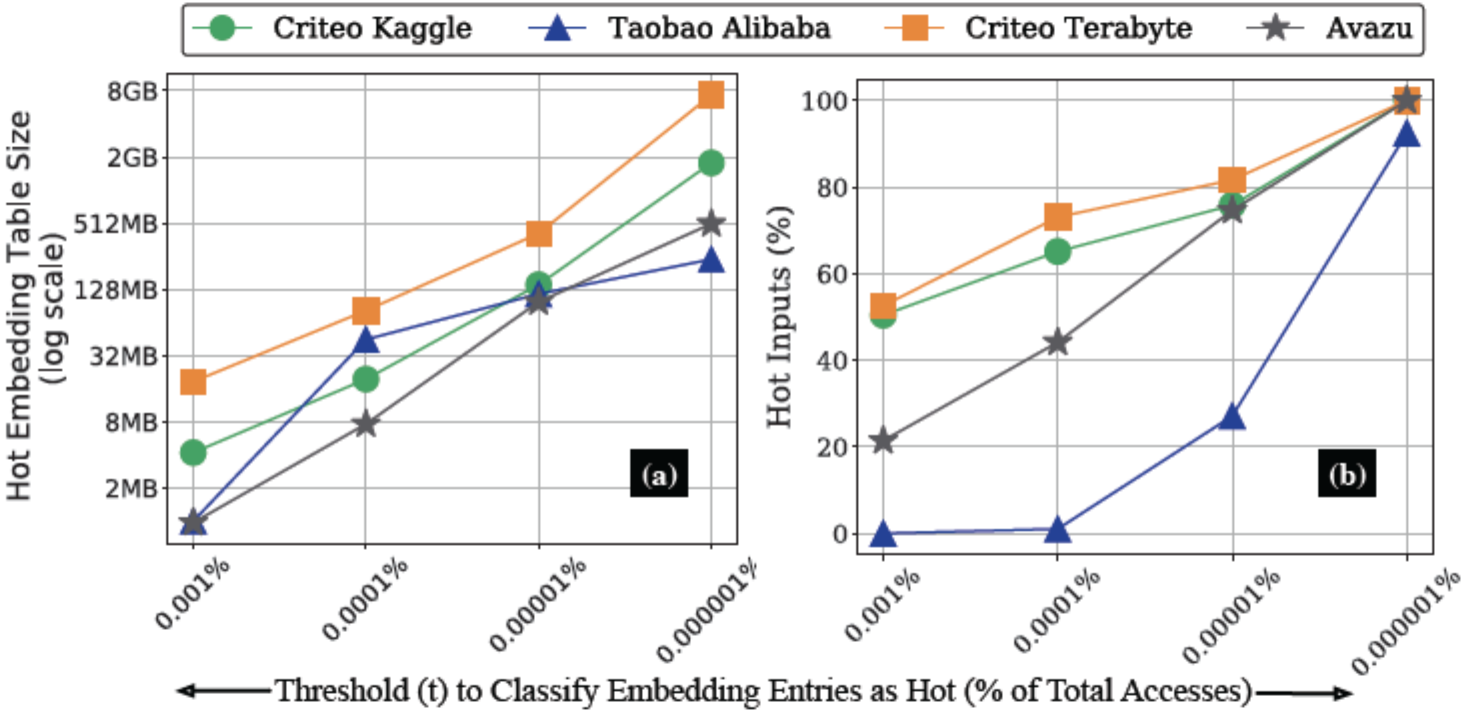
Neural Network Backward

Feature Interaction Backward

Neural Network Backward

Backward Path

Access Threshold vs Hot Inputs & Hot Embs



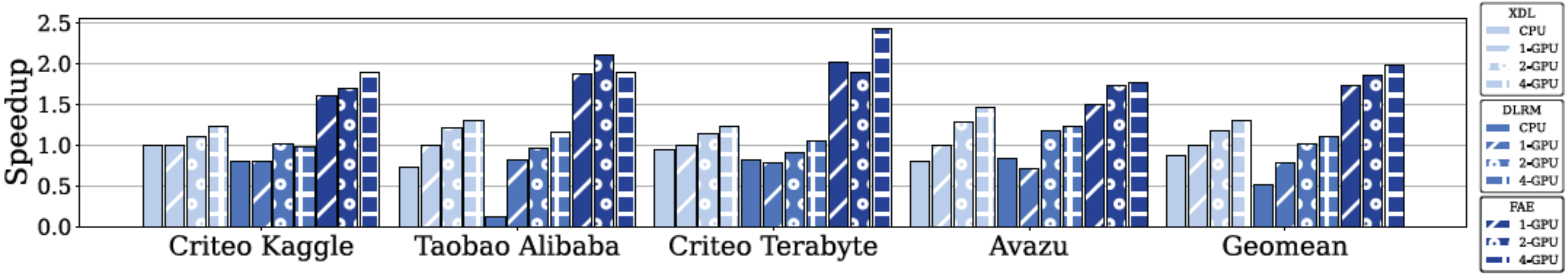
Recommendation Models

Workload	Dataset	Training Input		Model Features		Embedding Tables			Neural Network Configuration		
		Samples	Size	Dense	Sparse	Rows	Row Dim	Size	Bottom MLP	Top MLP	DNN
RMC1 (TBSM [4])	Taobao (Alibaba) [28]	10 M	1 GB	1	3	5.1M	16	0.3 GB	1-16 & 22-15-15	30-60-1	Attn. Layer
RMC2 (DLRM [2])	Criteo Kaggle [27]	45 M	2.5 GB	13	26	33.8M	16	2 GB	13-512-256-64-16	512-256-1	-
RMC3 (DLRM [2])	Criteo Terabyte [29]	80 M	45 GB	13	26	266M	64	63 GB	13-512-256-64	512-512-256-1	-
RMC4 (DLRM [2])	Avazu [30]	32.3 M	2.4 GB	1	21	9.3M	16	0.55 GB	1-512-256-64-16	512-256-1	-

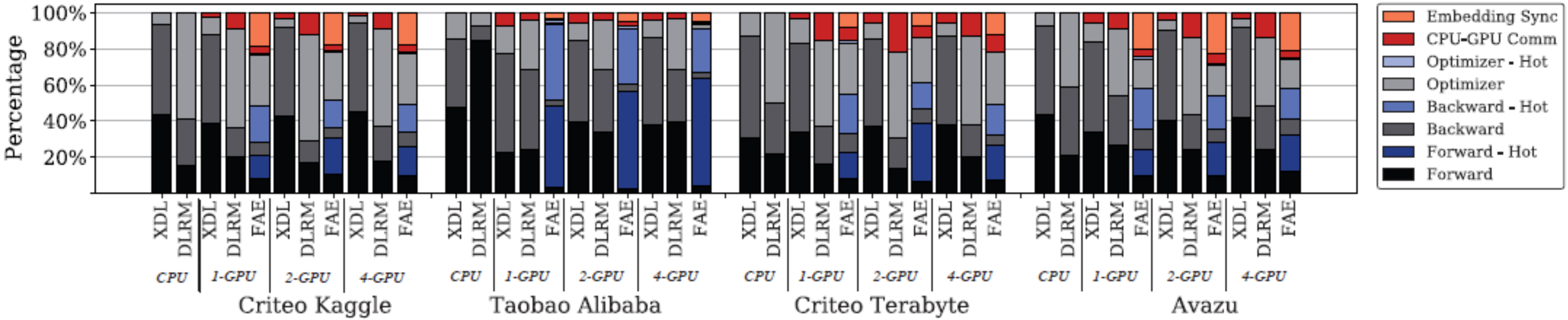
Accuracy Metric

Dataset	XDL			FAE		
	Accuracy (%)	AUC	Logloss	Accuracy (%)	AUC	Logloss
Criteo Kaggle	78.86	0.802	0.452	78.86	0.802	0.452
Taobao Alibaba	89.21	-	0.269	89.03	-	0.271
Criteo Terabyte	81.07	0.802	0.424	81.06	0.802	0.424
Avazu	83.61	0.758	0.390	83.60	0.758	0.391

Speedup Comparison

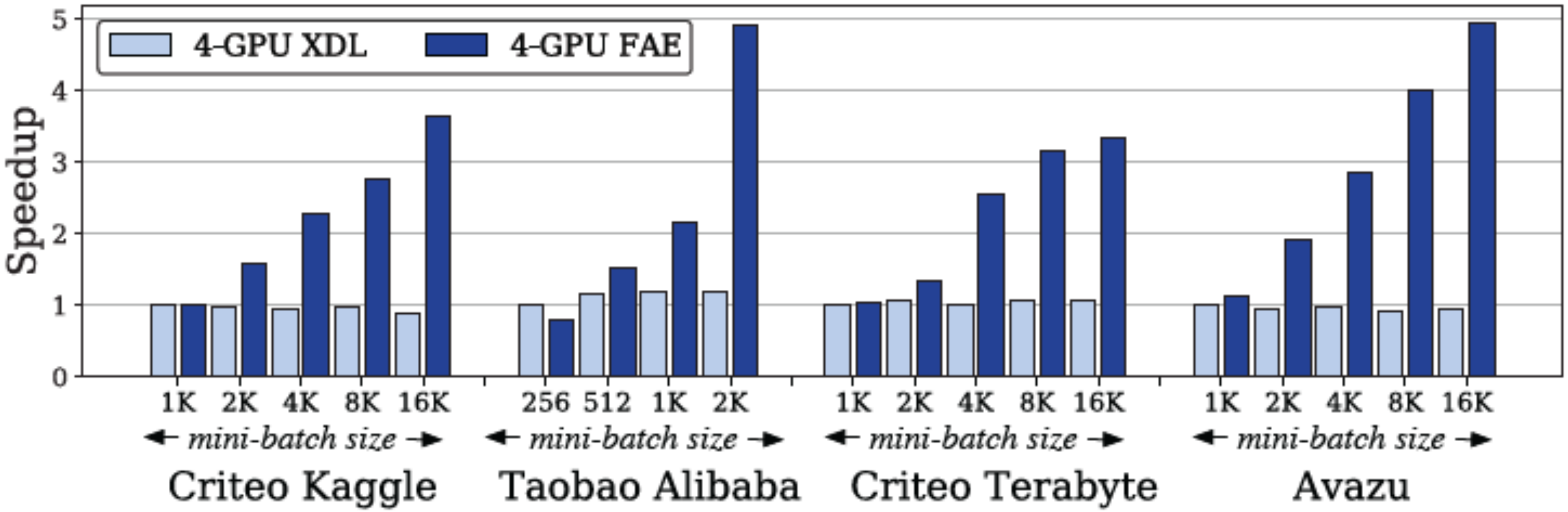


Latency Breakdown



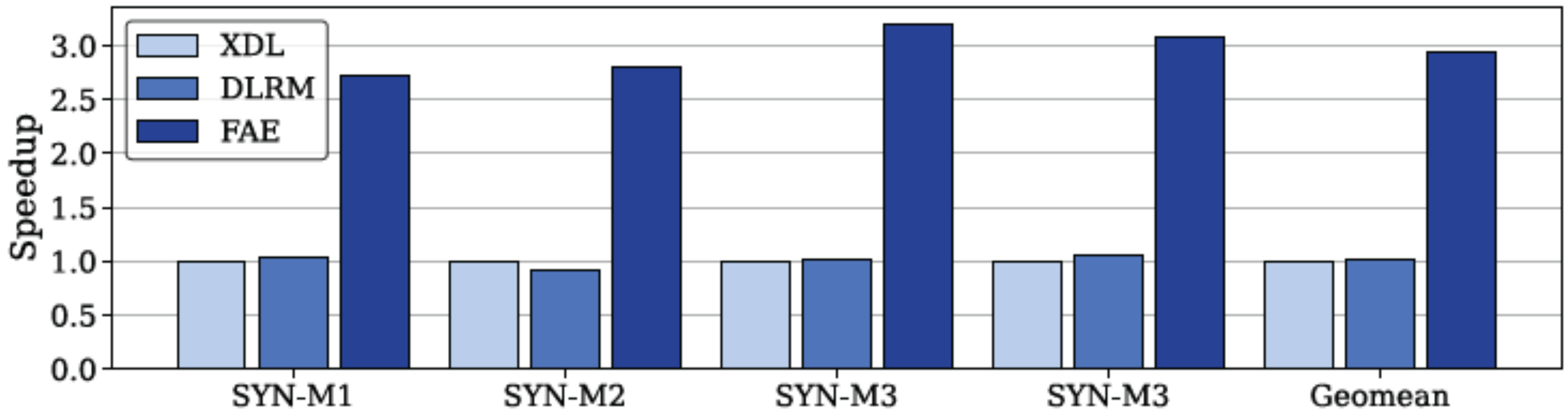
FAE reduces average total data transfer by 50%
and incurs a 13% overhead on the end-to-end training time

Scalability with Mini-batch Size



Synthetic Models

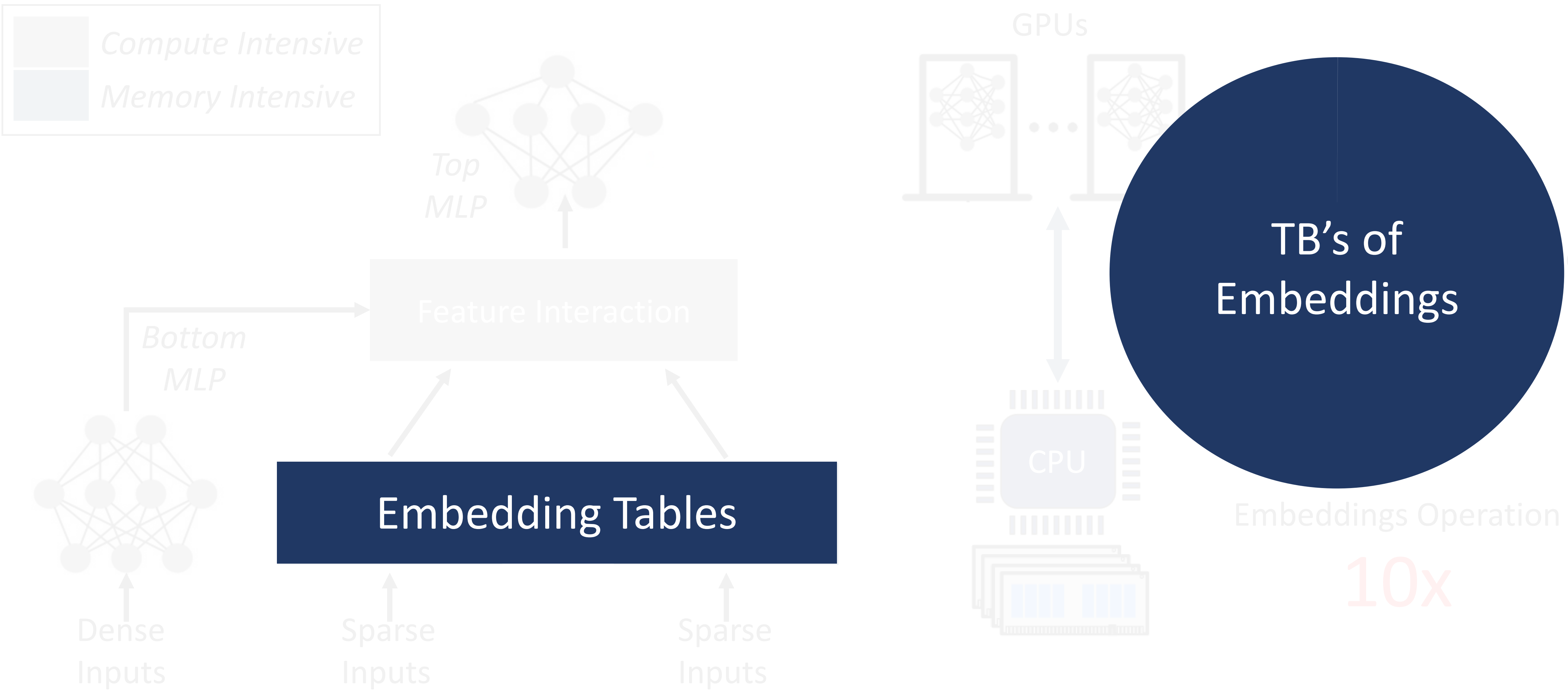
Dataset	Bottom MLP	Top MLP
SYN-M1	13-64	512-1
SYN-M2	13-512-64	512-256-1
SYN-M3	13-1024-512-64	512-1024-256-1
SYN-M4	13-1024-512-256-64	512-1024-512-256-1



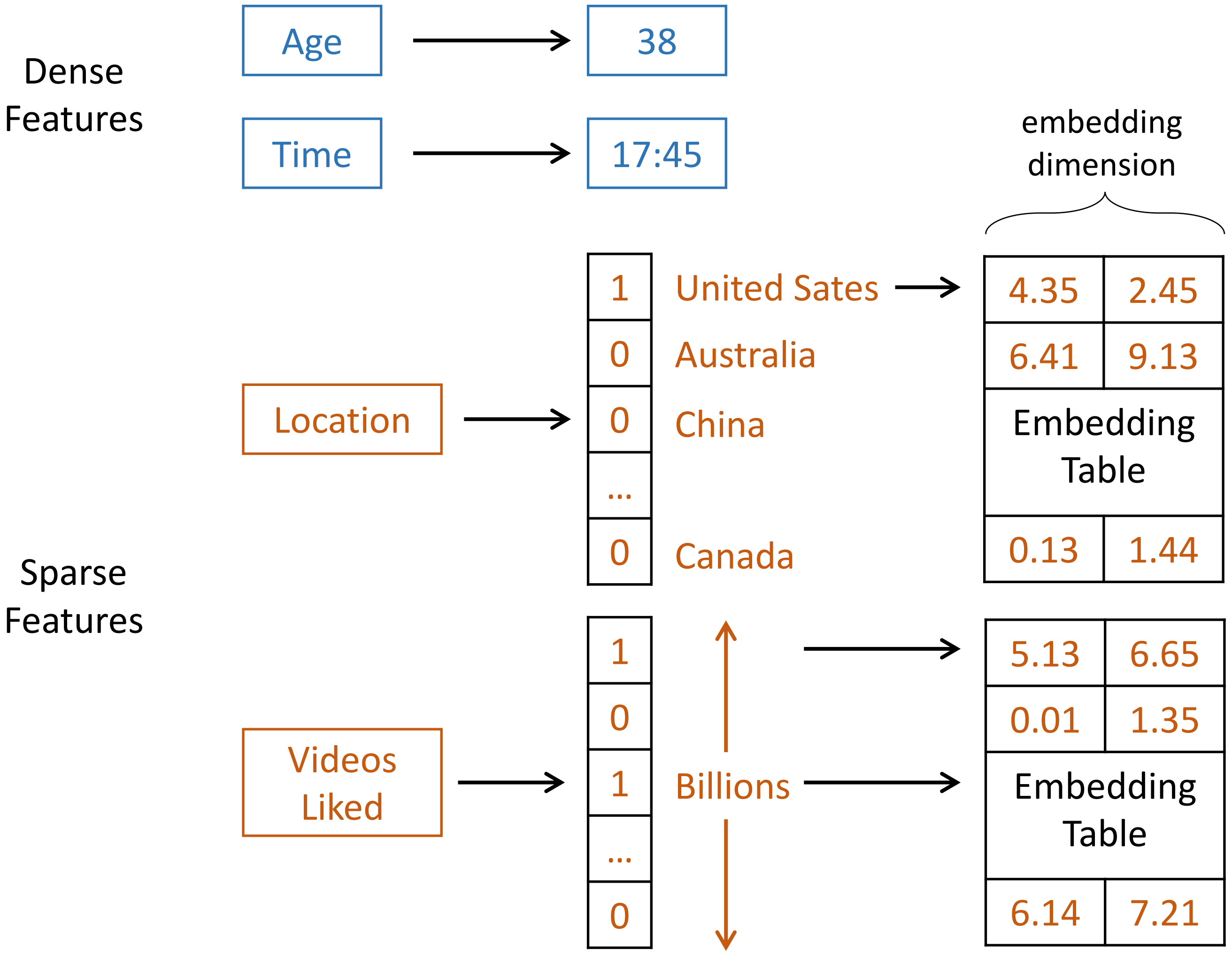
Challenges – Embedding Layout

- Training consecutively on popular and non-popular mini-batches can have impact on training accuracy.

Are All Embeddings Equal?



Deep Learning Recommendation Models



Deep Learning Recommendation Models

Dense Features

Age → 38

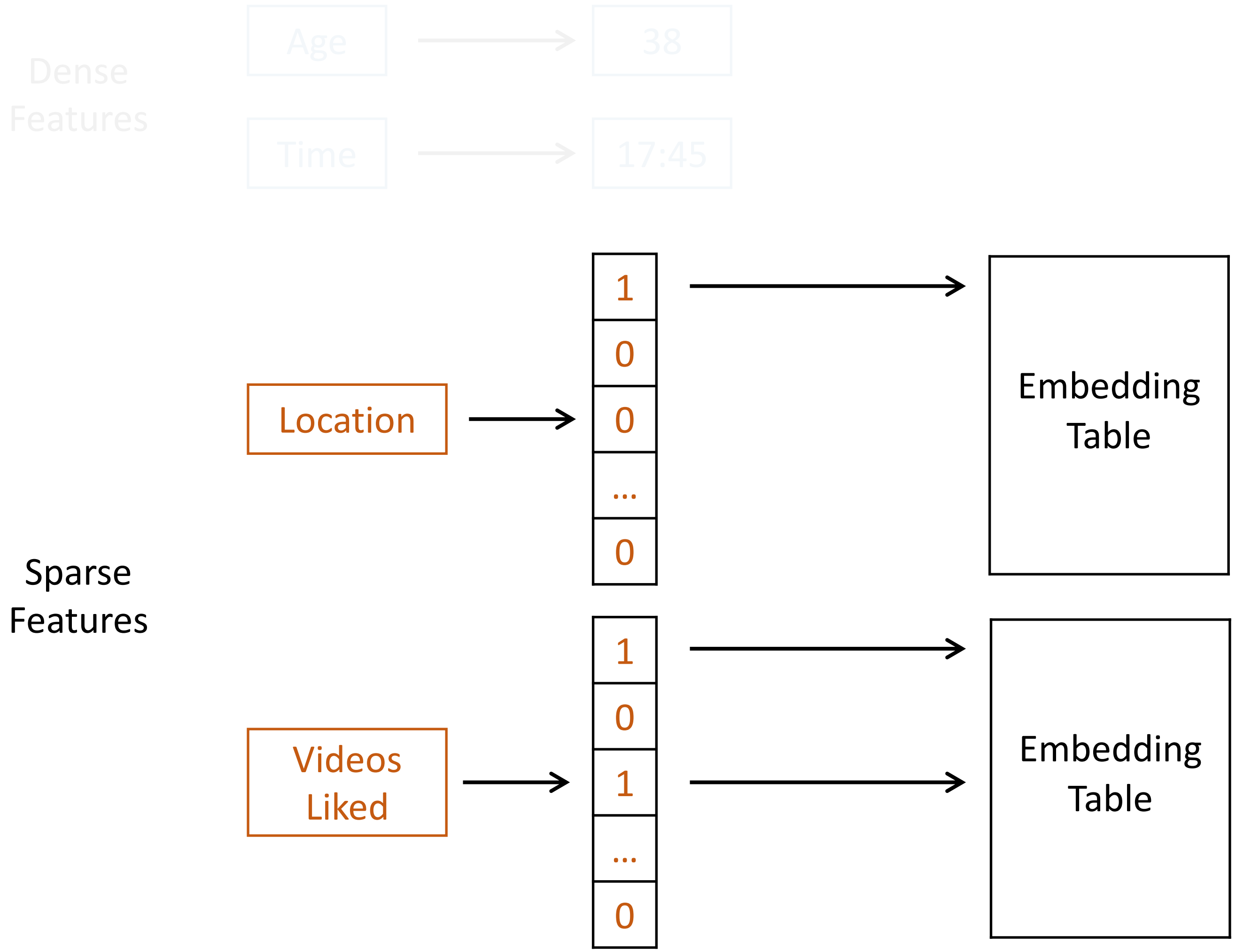
Time → 17:45

Sparse Features

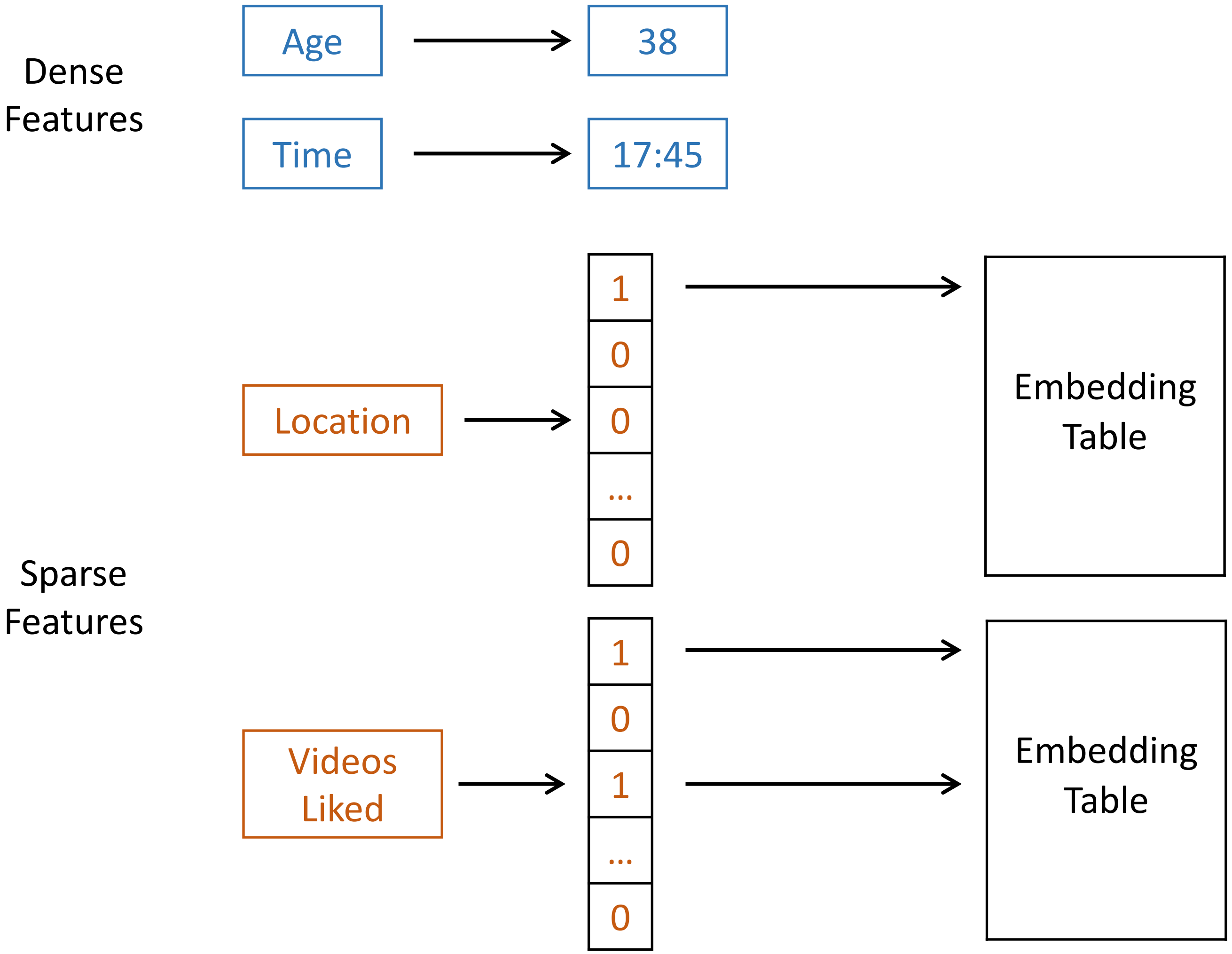
Location → 1, 0, 0, …, 0 → Embedding Table

Videos Liked → 1, 0, 1, …, 0 → Embedding Table

Deep Learning Recommendation Models

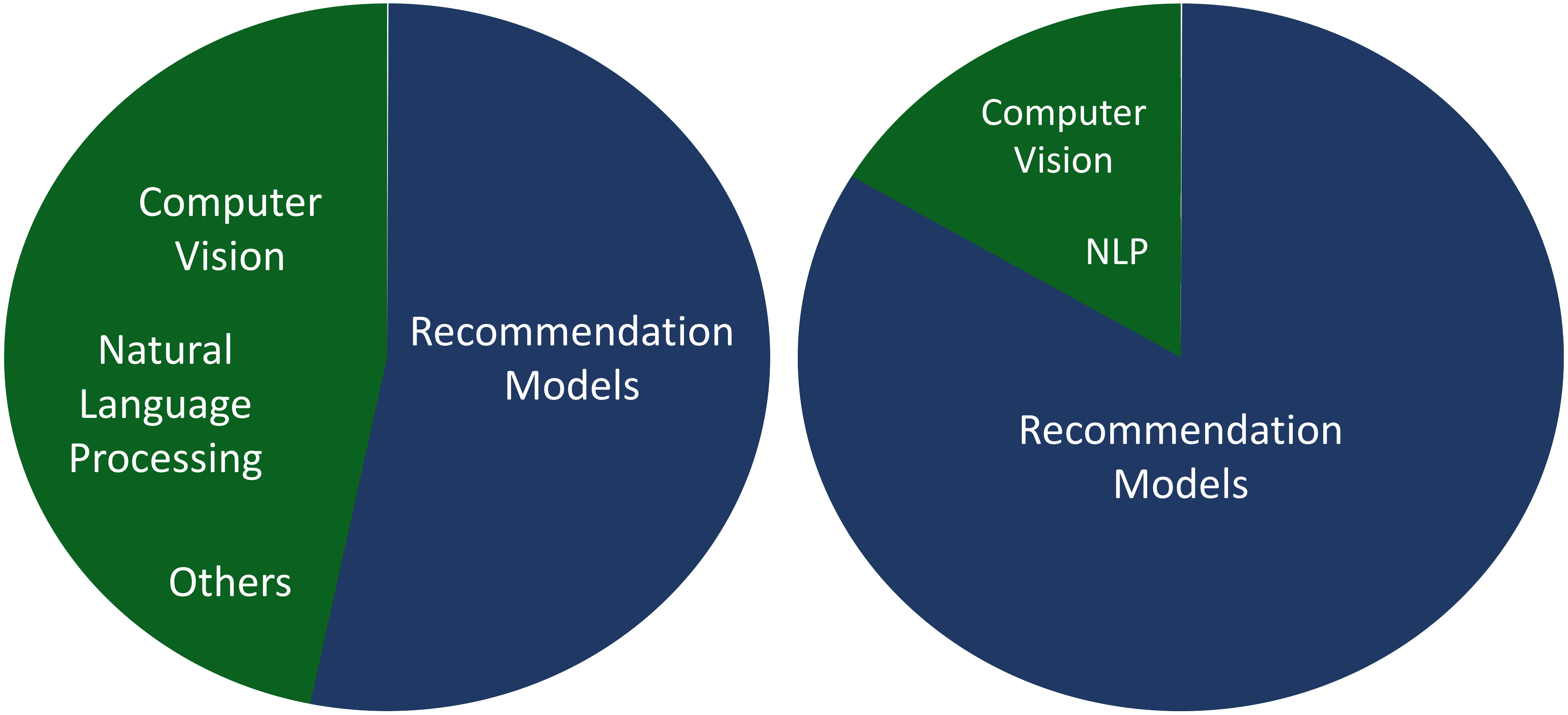


Deep Learning Recommendation Models



Recommendation Systems in Industry

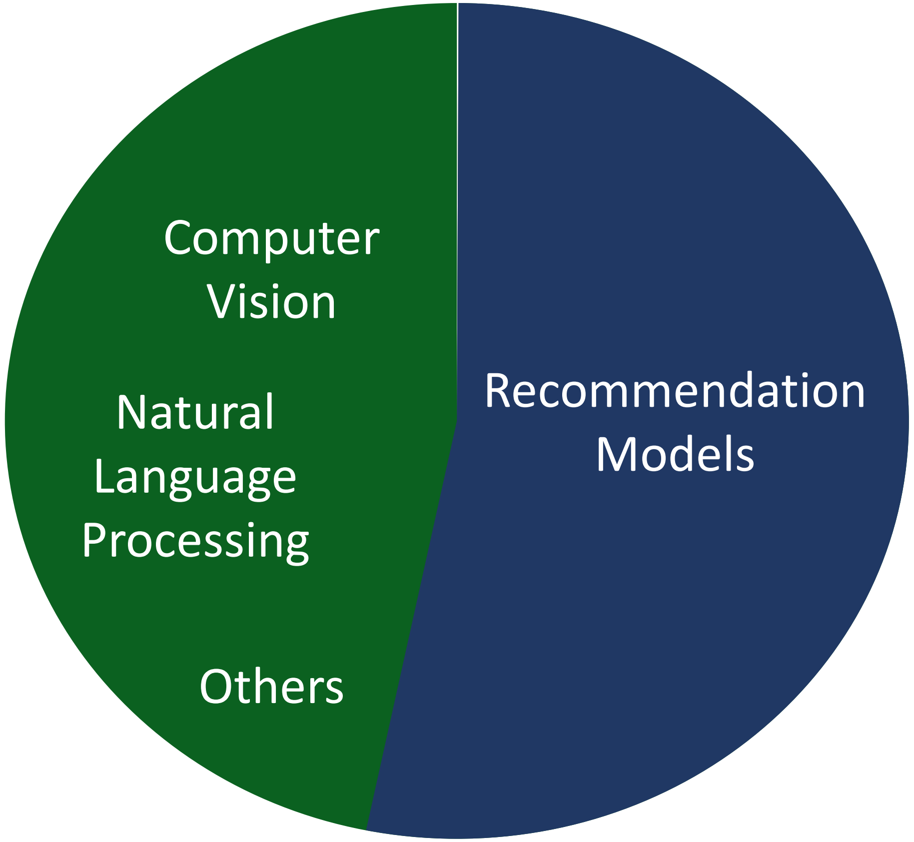
Meta



~50% of Training Demand ~80% of Inference Demand

Naumov et al. arXiv'20.

Recommendation Systems in Industry



~50% of Training Demand

~80% of Inference Demand

Naumov et al. arXiv’20.