Microsoft

Embedding Based Recall: Practice, Progress and Perspectives

Zheng Liu, Jianxun Lian, Xing Xie

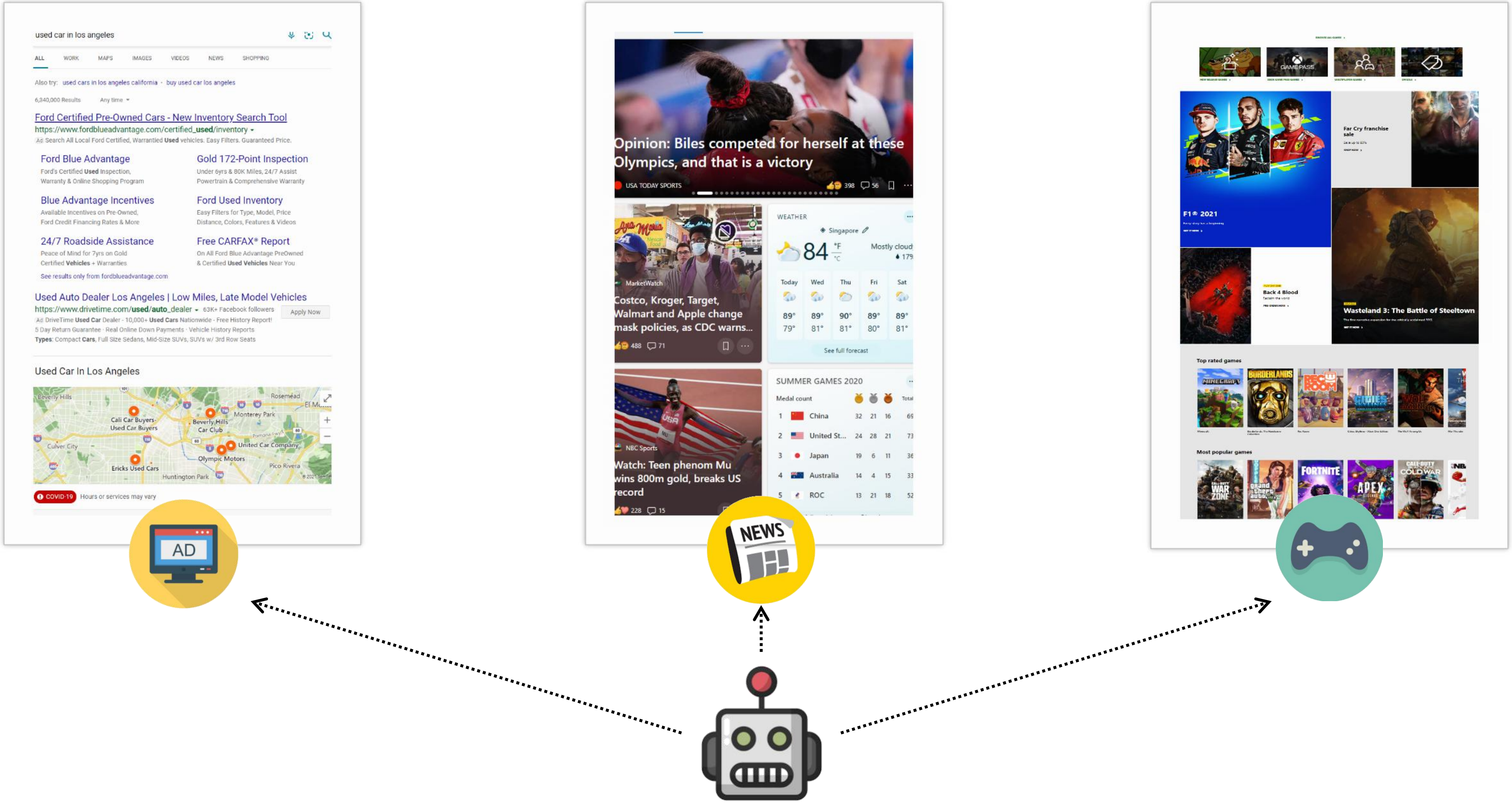
Social Computing Group, MSRA

Aug 15th, 2021

Reinforced Anchor Knowledge Graph Generation for News Recommendation Reasoning, Liu et. al. KDD 2021

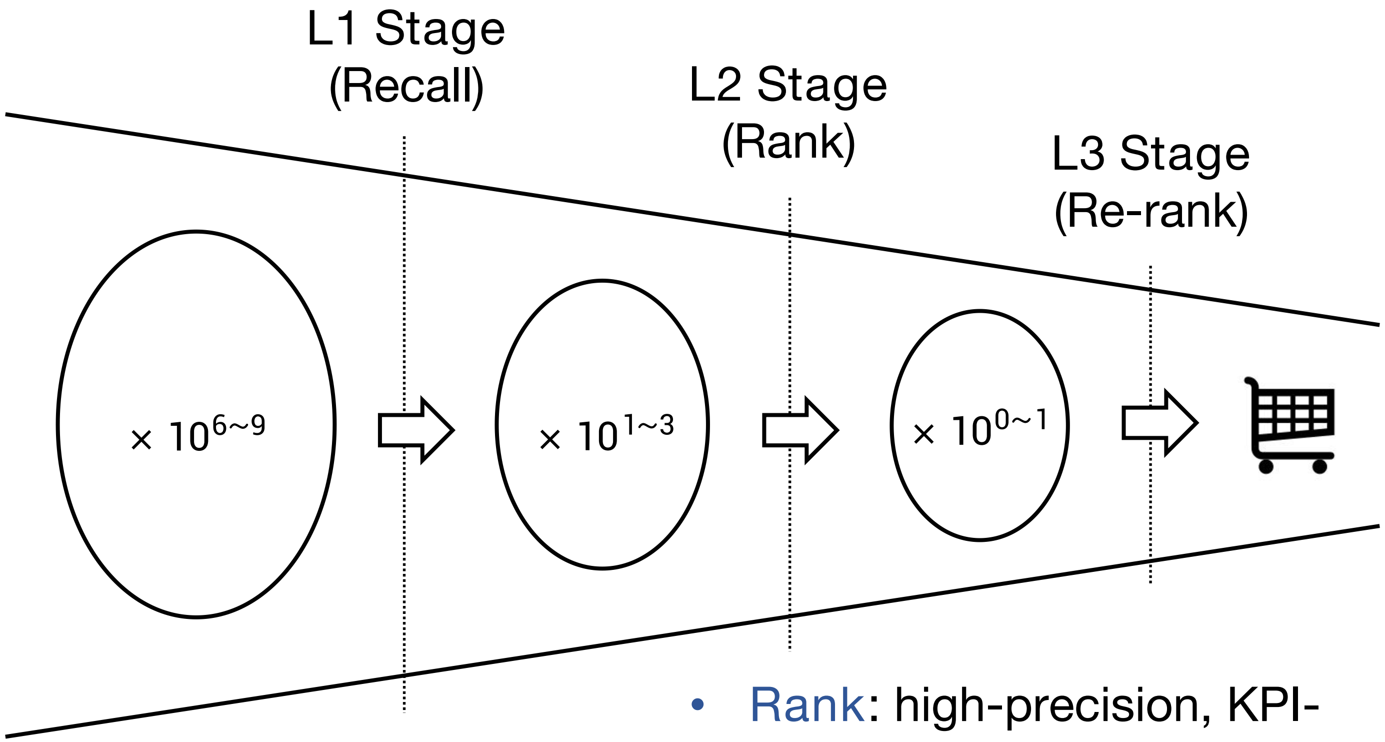
Outline

- Overview
  - Multi-Stage Pipeline
  - EBR: Pros and Cons
- Embedding learning algorithms
  - Negative Augmentation
  - Hard Negative Sampling
  - Diversified representation
  - Training as knowledge distillation
- Things beyond learning algorithms
  - Efficiency issues
  - Combo of sparse and dense



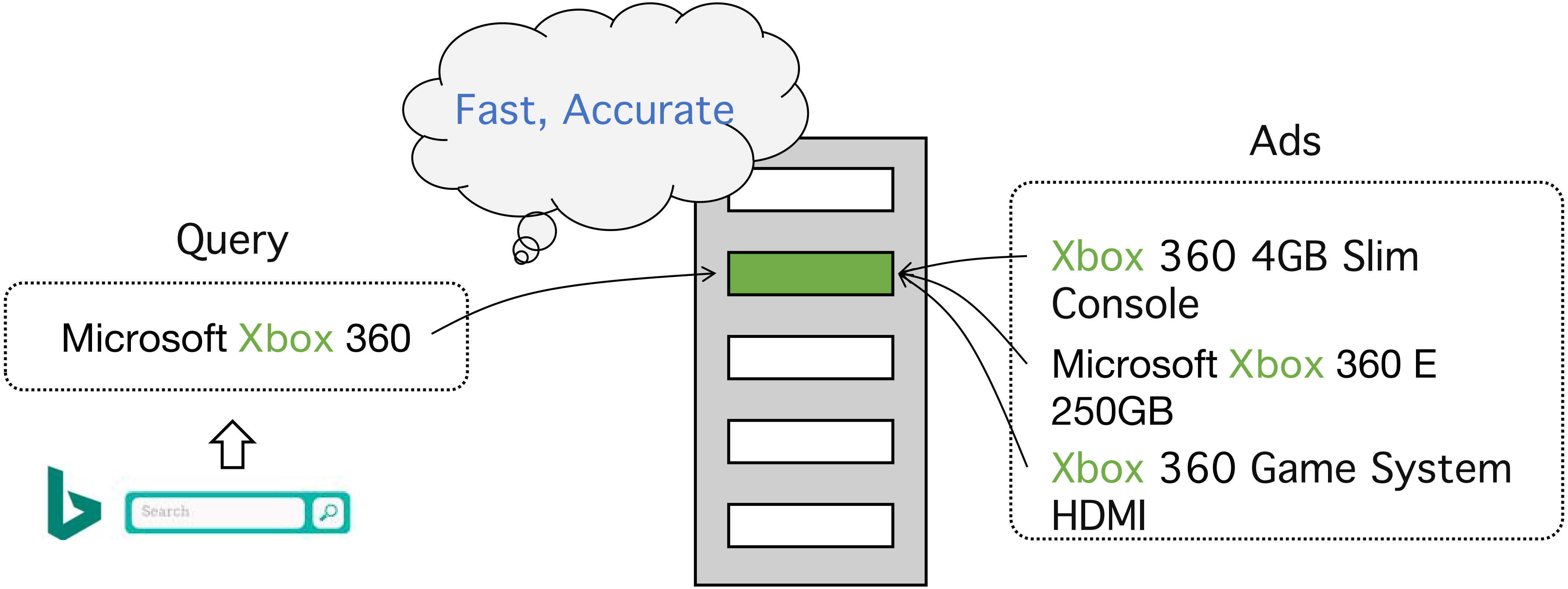
used car in los angeles
ALL
WORK
MAPS
IMAGES
VIDEOS
NEWS
SHOPPING
Ford Certified Pre-Owned Cars - New Inventory Search Tool
https://www.fordblueadvantage.com/certified_used/inventory
Ford Blue Advantage
Gold 172-Point Inspection
Blue Advantage Incentives
Ford Used Inventory
24/7 Roadside Assistance
Free CARFAX® Report
Used Auto Dealer Los Angeles | Low Miles, Late Model Vehicles
https://www.drivetime.com/used/auto_dealer
Used Car In Los Angeles
AD
Opinion: Biles competed for herself at these Olympics, and that is a victory
USA TODAY SPORTS
WEATHER
Singapore
84
Costco, Kroger, Target, Walmart and Apple change mask policies, as CDC warns...
Watch: Teen phenom Mu wins 800m gold, breaks US record
SUMMER GAMES 2020
Medal count
China
Japan
Australia
ROC
NEWS
F1® 2021
Far Cry franchise sale
Back 4 Blood
Wasteland 3: The Battle of Steeltown
Top rated games
Most popular games

Overview: Multi-Stage Pipeline

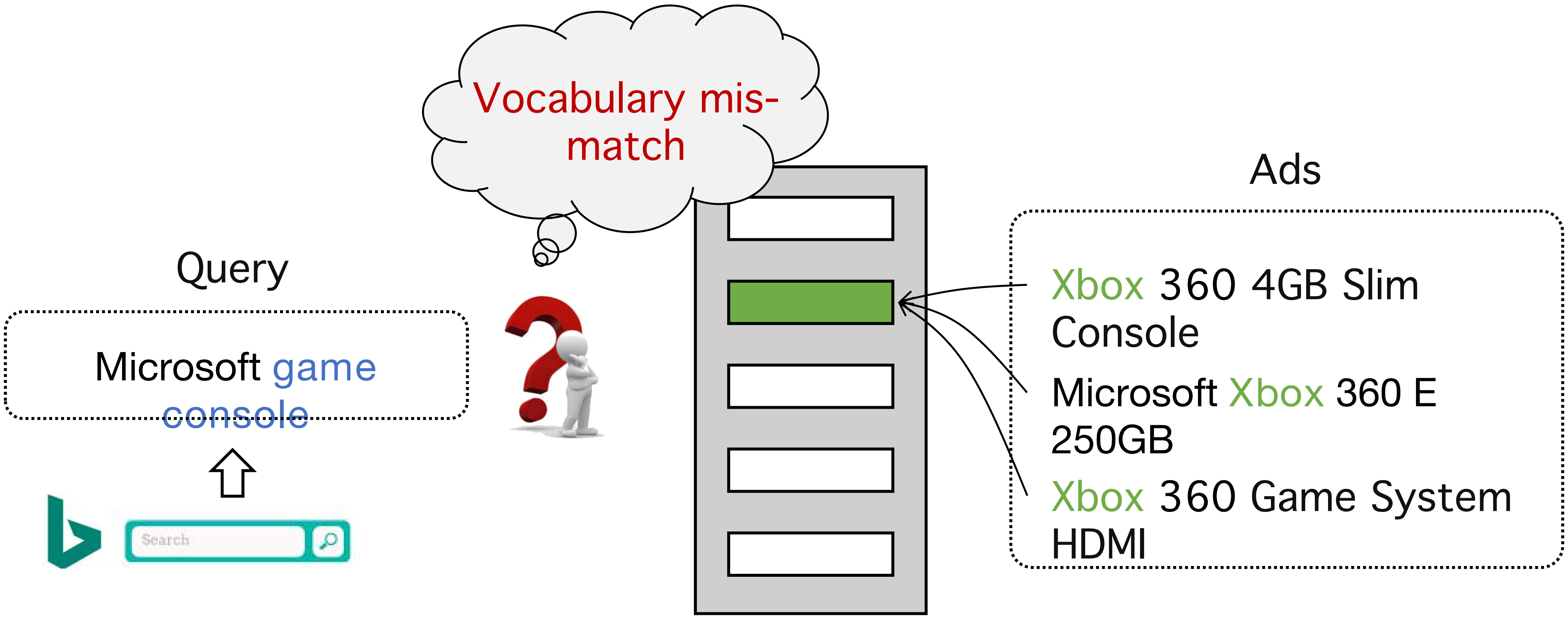


- Rank: high-precision, KPI-oriented
- Recall: fast, accurate, comprehensive

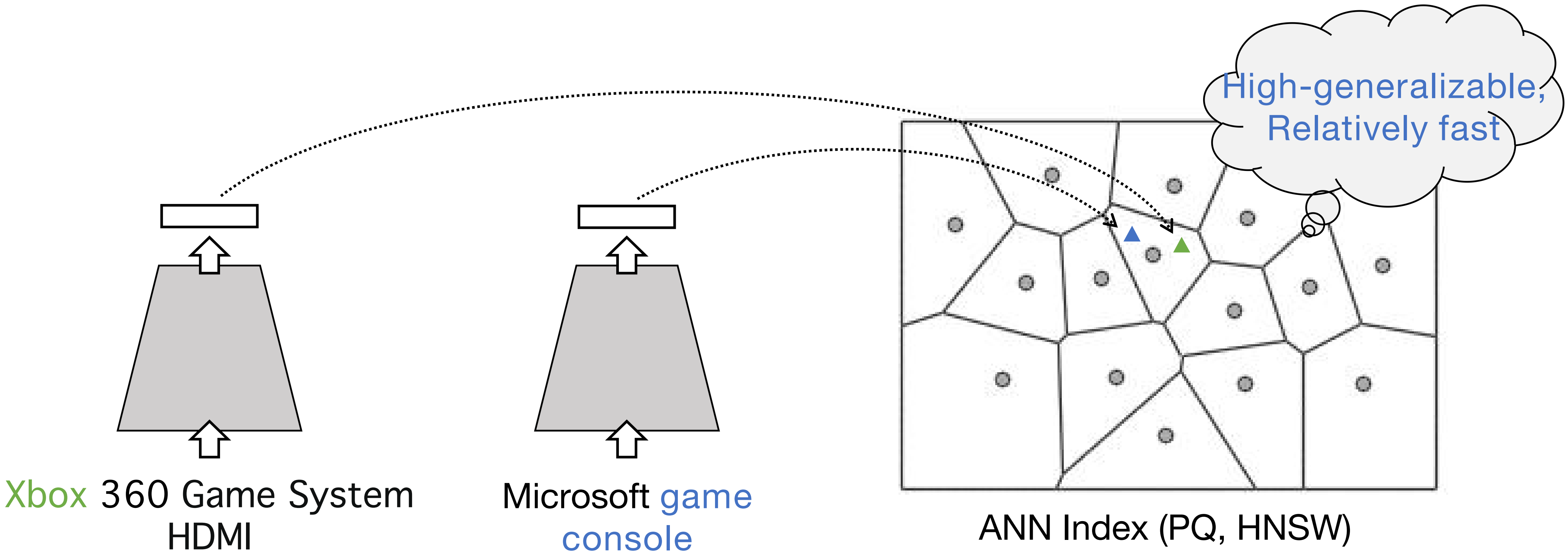
Overview: Multi-Stage Pipeline



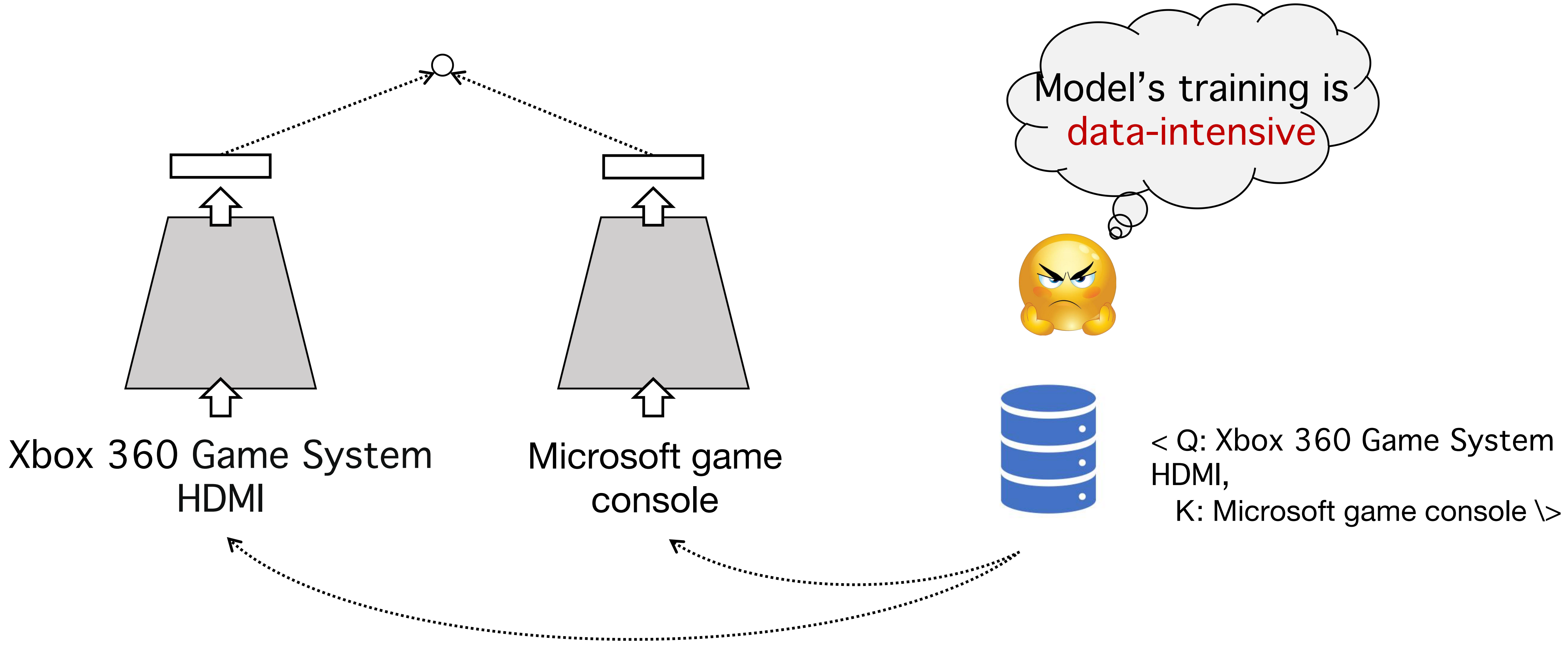
Overview: Multi-Stage Pipeline



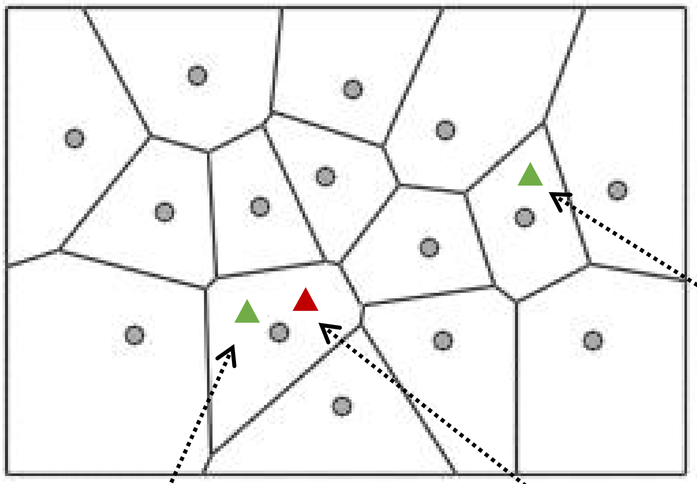
Overview: EBR, Pros and Cos



Overview: EBR, Pros and Cons



Overview: EBR, Pros and Cons



Xbox 360 Game
System HDMI

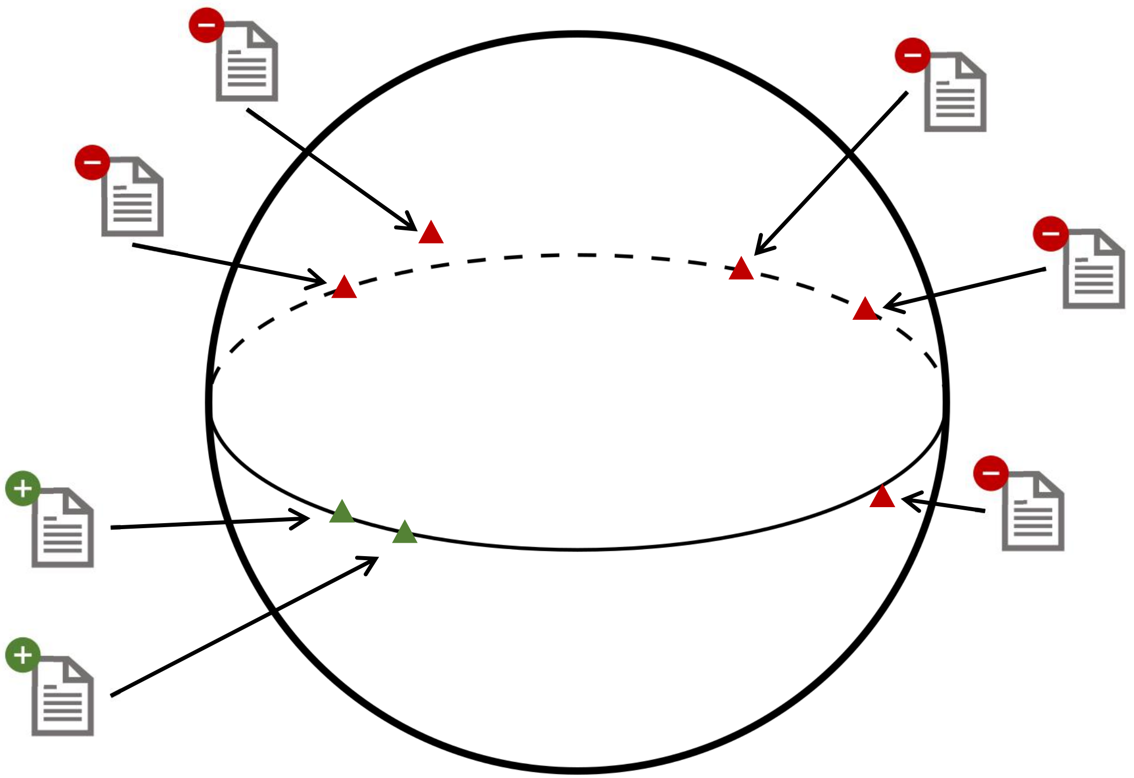
Nintendo switch
console

... the Xbox 360 console hard
drive ...
Hard drive will be at least 20GB
model
... supports HD graphics in 16 x 9
wide-
screen, with anti-aliasing

Overview: EBR, Pros and Cons

$$\underbrace{\ell_{\text{align}} \triangleq \mathop{\mathbb{E}}_{(x,x^+)\sim p_{\text{pos}}} \|f(x) - f(x^+)\|^2}_{\text{Alignment}}$$

$$\underbrace{\ell_{\text{uniform}} \triangleq \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d.}{\sim} p_{\text{data}}} e^{-2\|f(x)-f(y)\|^2}}_{\text{Uniformity}}$$

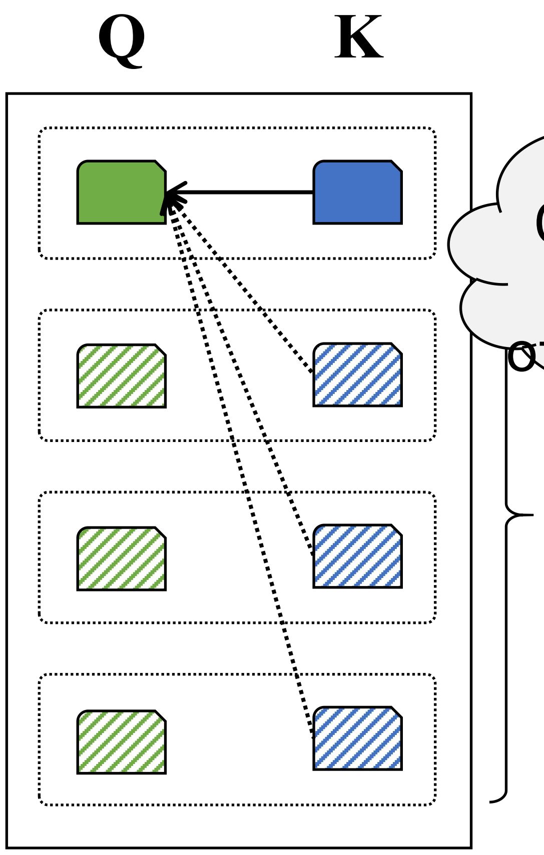


[1] SimCSE, Gao et.al.
[2] Understanding Contrastive Learning …
ICML 2020, Wang et. al.

Outline

- Overview
  - Multi-Stage Pipeline
  - EBR: Pros and Cons
- Embedding learning algorithms
  - Negative Augmentation
  - Hard Negative Sampling
  - Training as distillation
  - Diversified representation
- Things beyond learning algorithms
  - Efficiency issues
  - Combo of sparse and dense

Algos: Negative Augmentation



batc
h

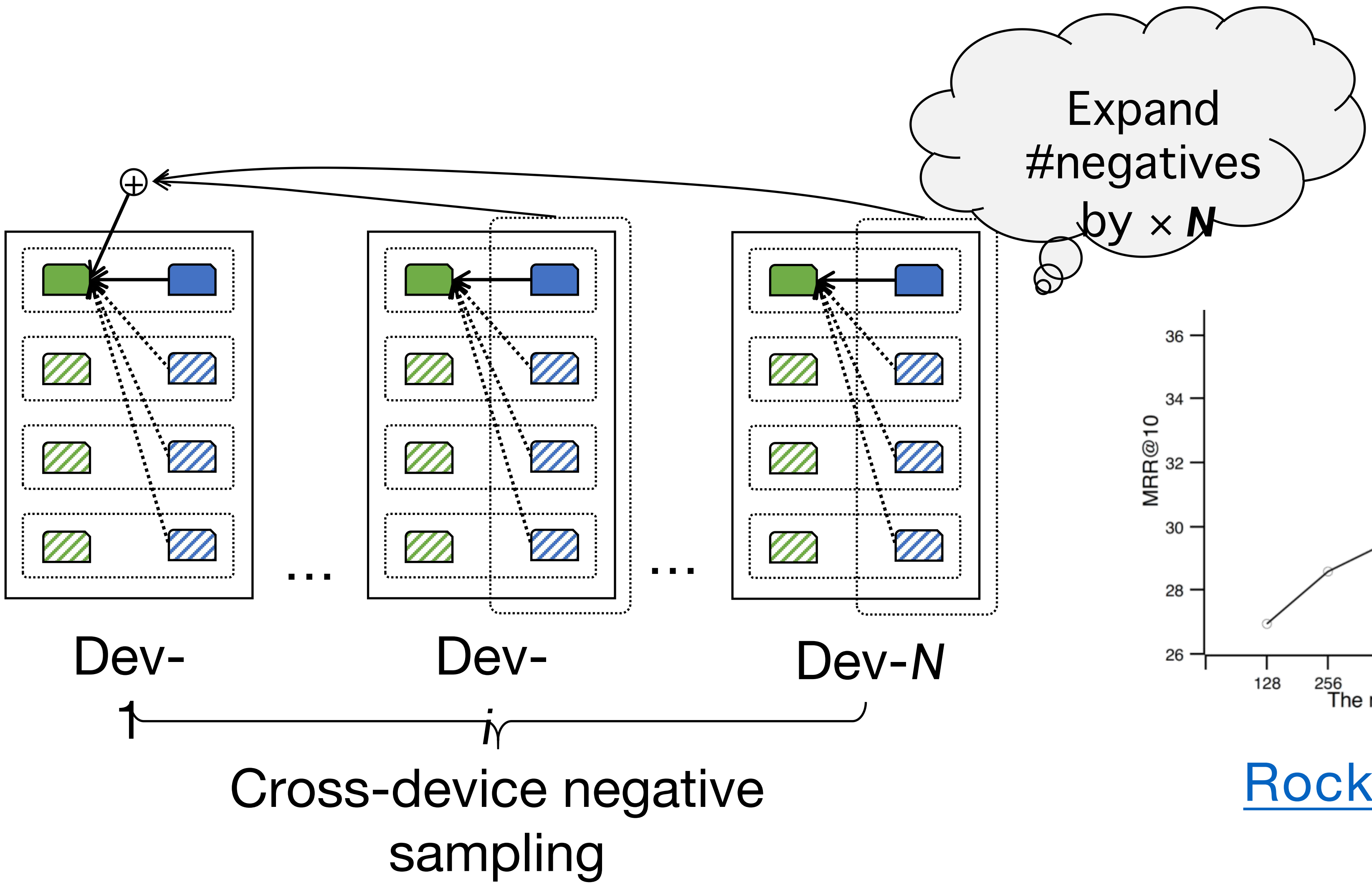
One's positive
used
others' negative

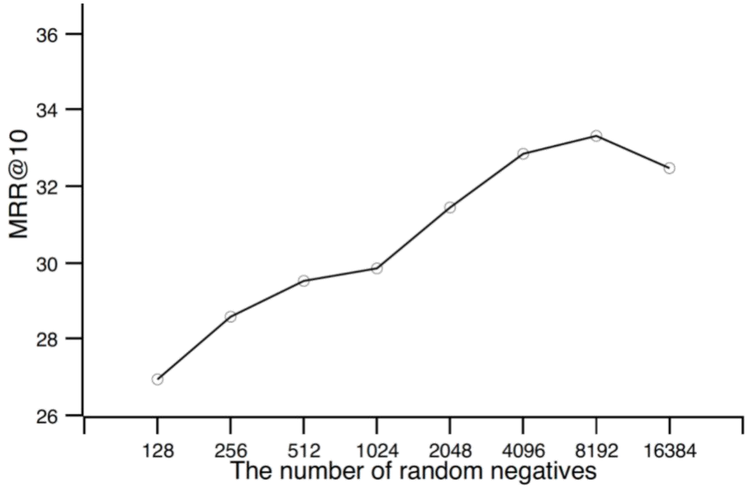
Larger batch size
->
higher accuracy

Type	#N	IB	Top-5	Top-20	Top-100
Gold	7	✓	51.1	69.1	80.8
Gold	31	✓	52.1	70.8	82.1
Gold	127	✓	55.8	73.0	83.1

[DPR, Karpukhin et. al.]

Algos: Negative Augmentation





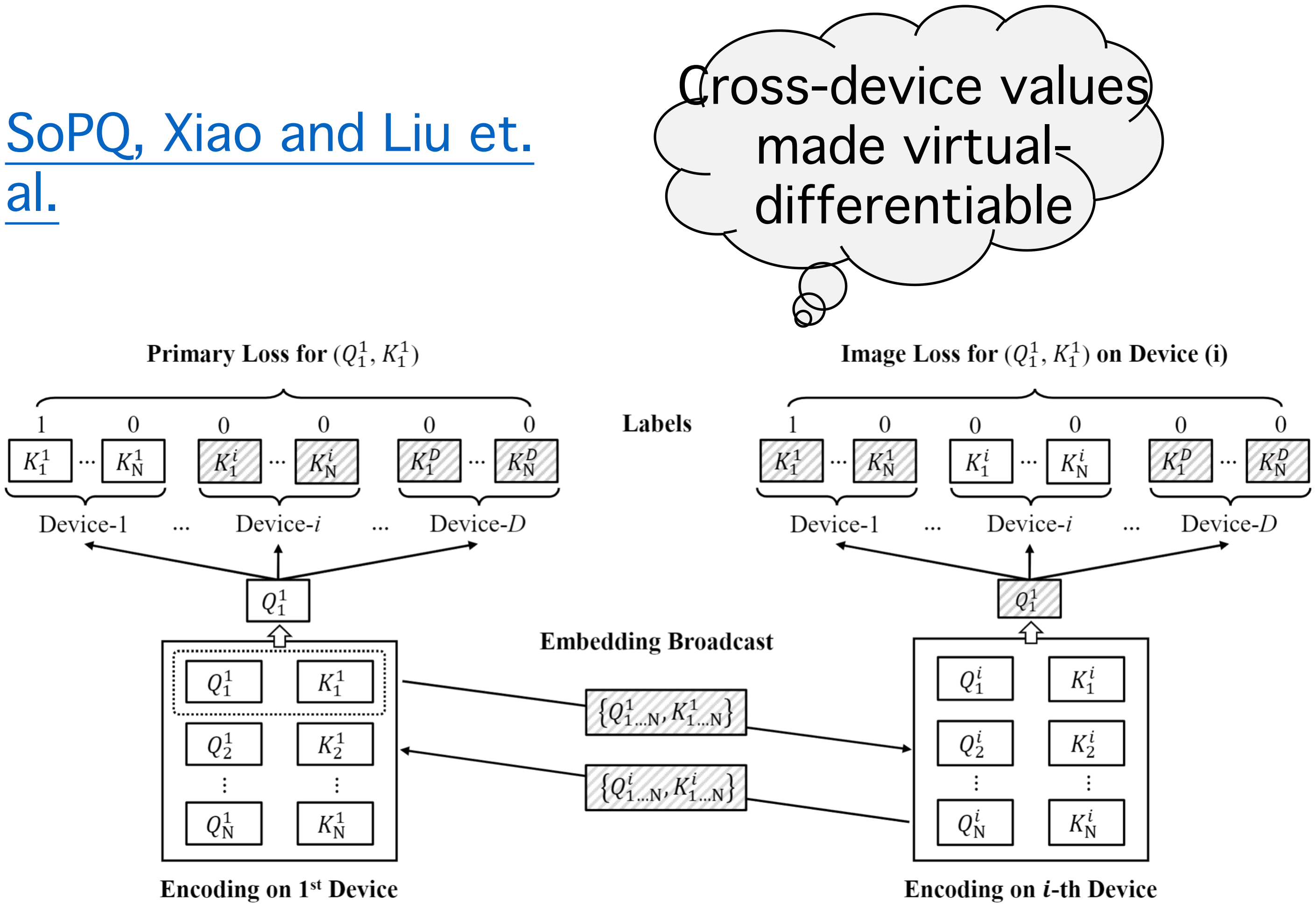
[RocketQA, Ding et. al.]

Algos: Negative Augmentation

Virtual Differentiable Cross-Device Sharing (V-DCS)

[SoPQ, Xiao and Liu et. al.]

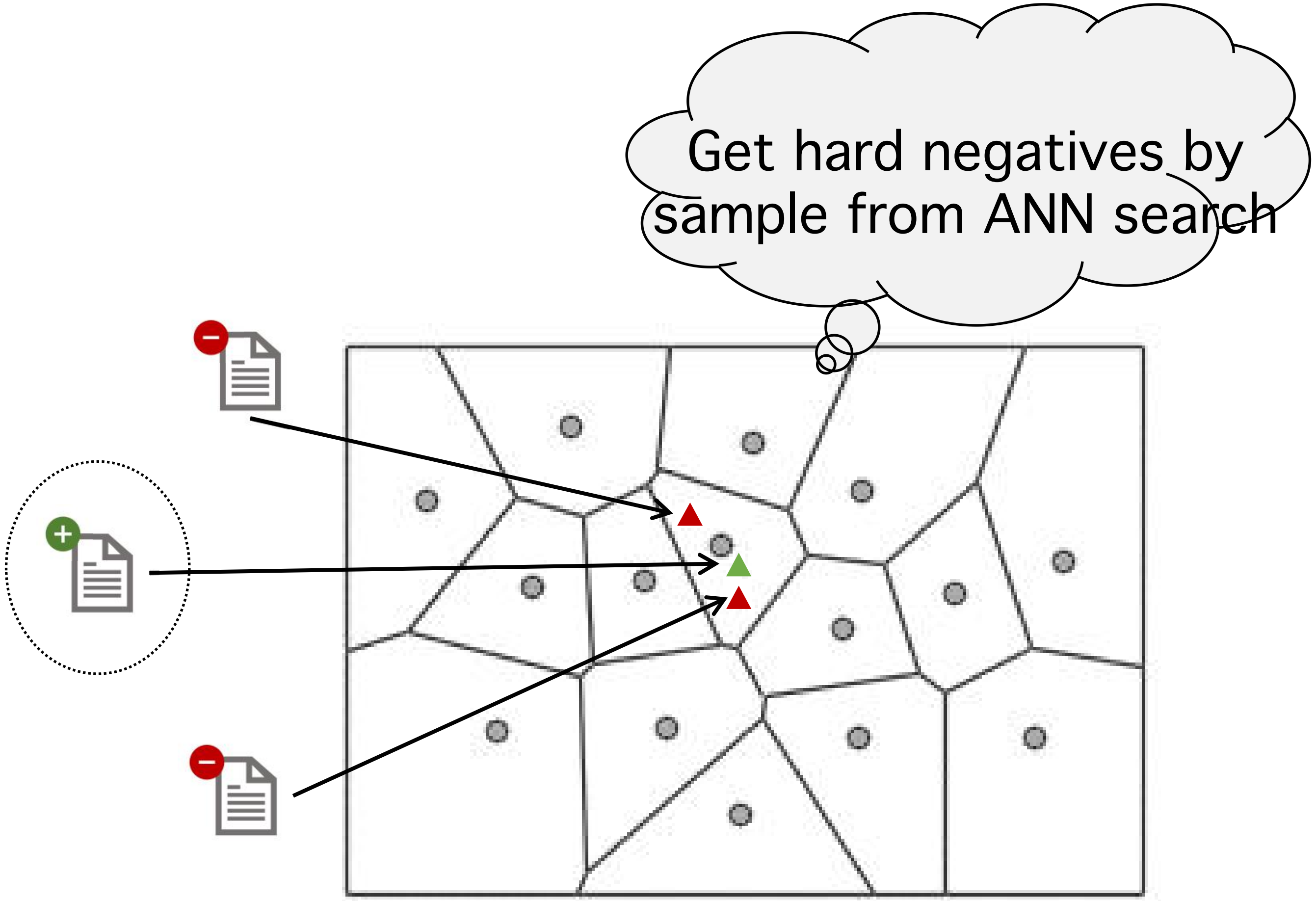
1. Generate embeddings for each batch, one batch/device
2. Broadcast embeddings to all devices
3. Compute the global NCE-loss symmetrically on all devices, based on the broadcasted embeddings
4. Back-propagate and reduce the gradients on all devices



Algos: Hard Negative Sampling

Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval (ANCE, Xiong et. al.)

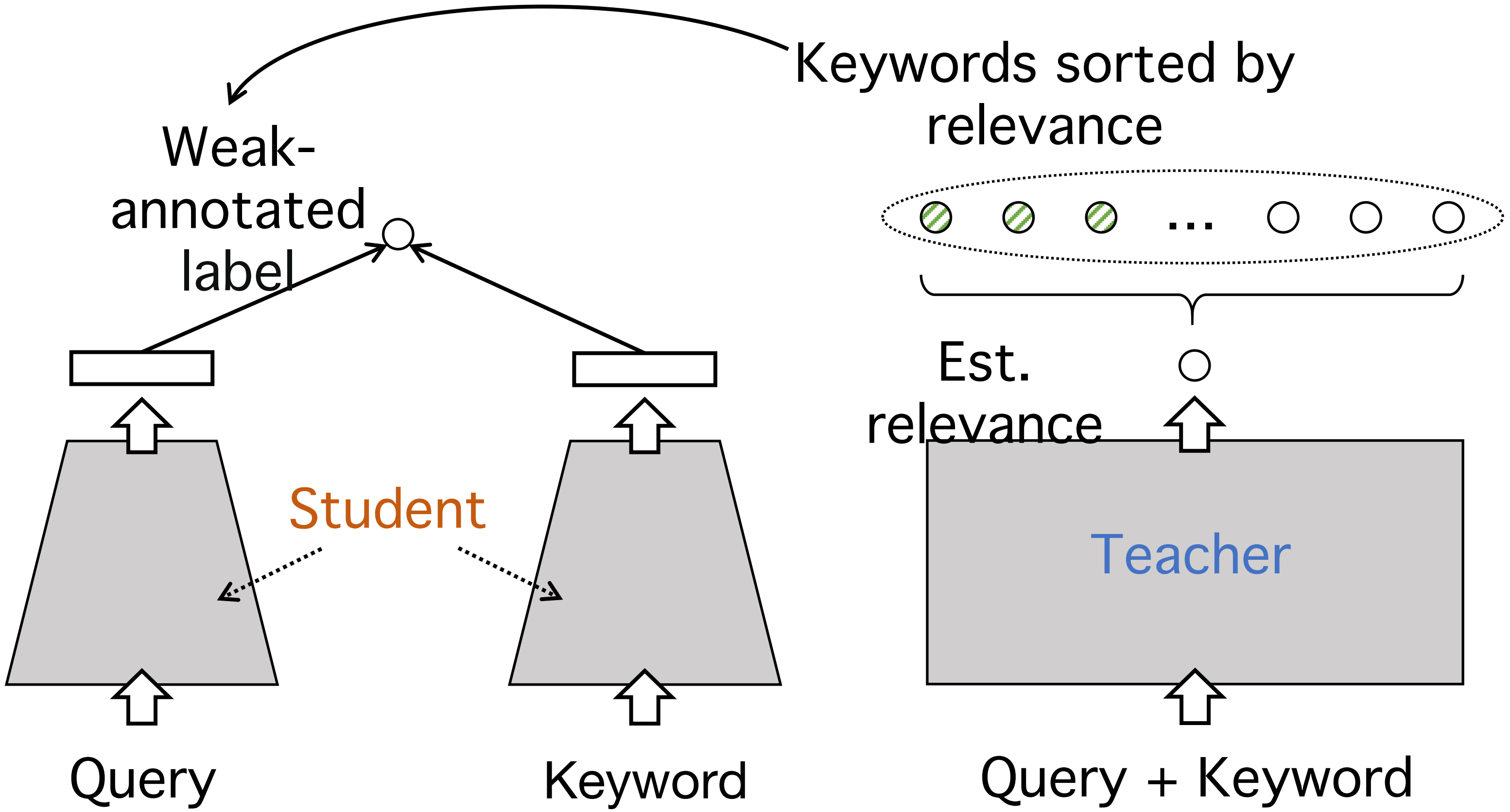
1. Learn embedding model with in-batch negative
2. Build ANN index and get hard negatives
3. Update embedding model with hard negative
4. Repeat 2. and 3. until converge



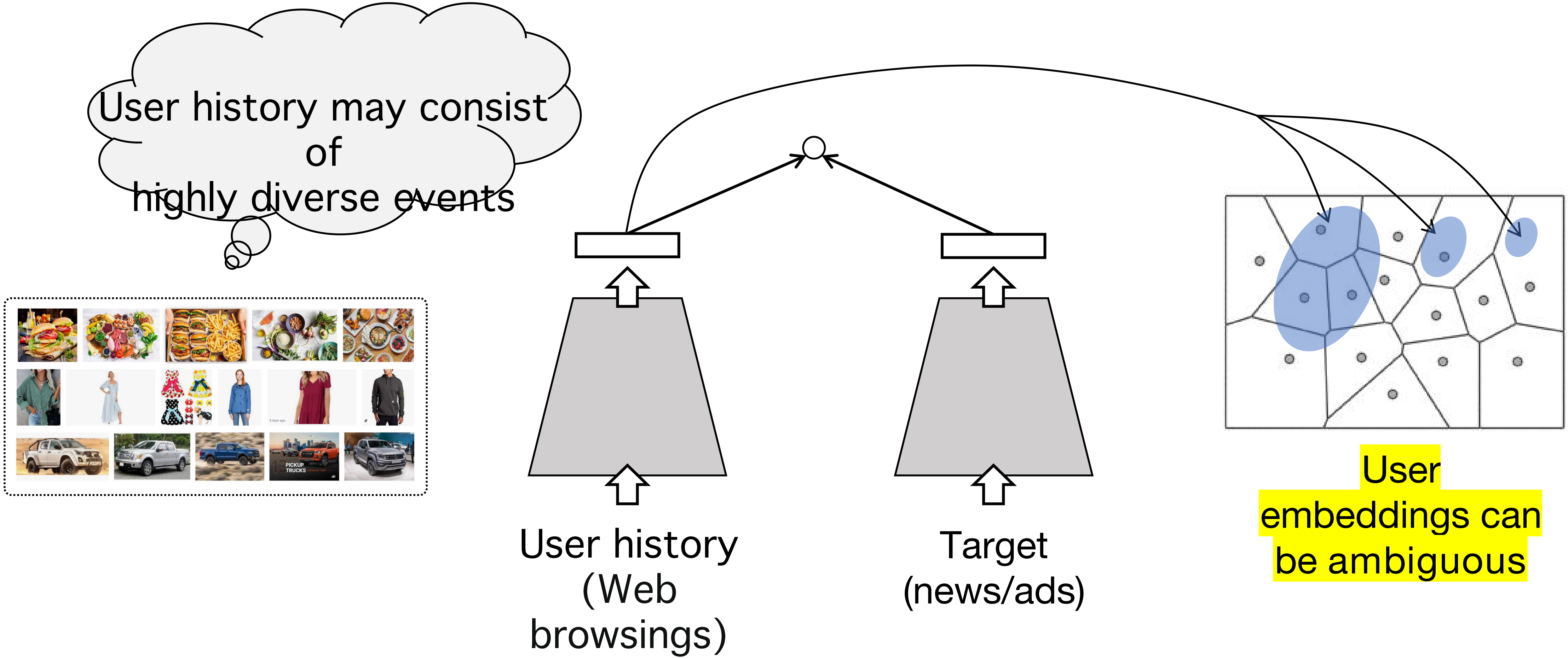
Algos: Training as distillation

Training as distillation
- Training teacher model with labeled data
- Annotate unlabeled data with teacher
- Train student with labeled and weak-annotated data

[RocketQA, Ding et. al.](#)
[Weak Annotation, Li et. al.](#)



Algos: diversified representation

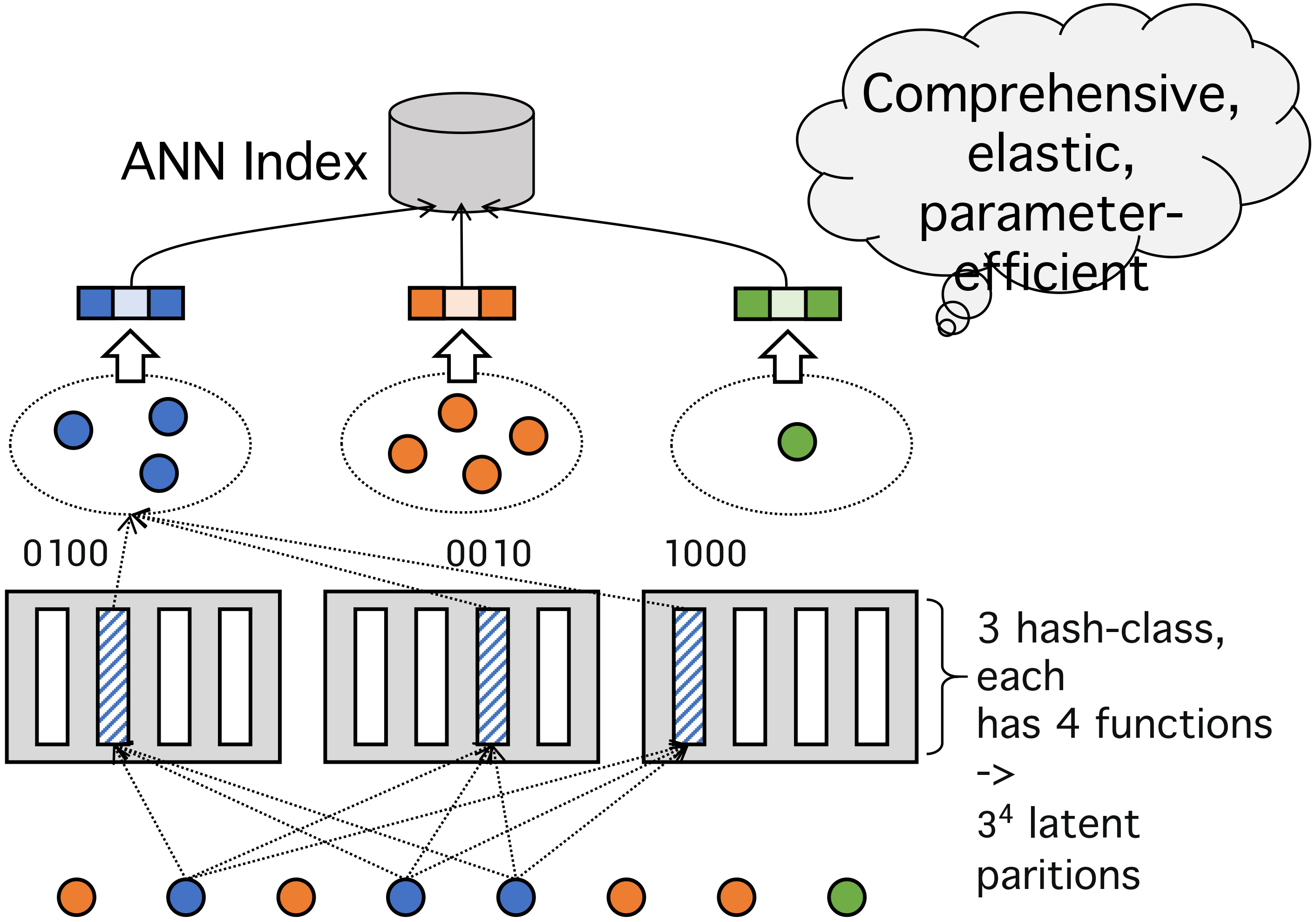


Algos: diversified representation

Elastic Multi-embedding Retrieval (Bloom-filter style interest extractor)

- Generate item embeddings
- Compute item embeddings' membership via learned hash
- Group items based on binary codes
- Aggregate items with the same binary codes for user embeddings

[Octopus, Liu et. al.]()



Outline

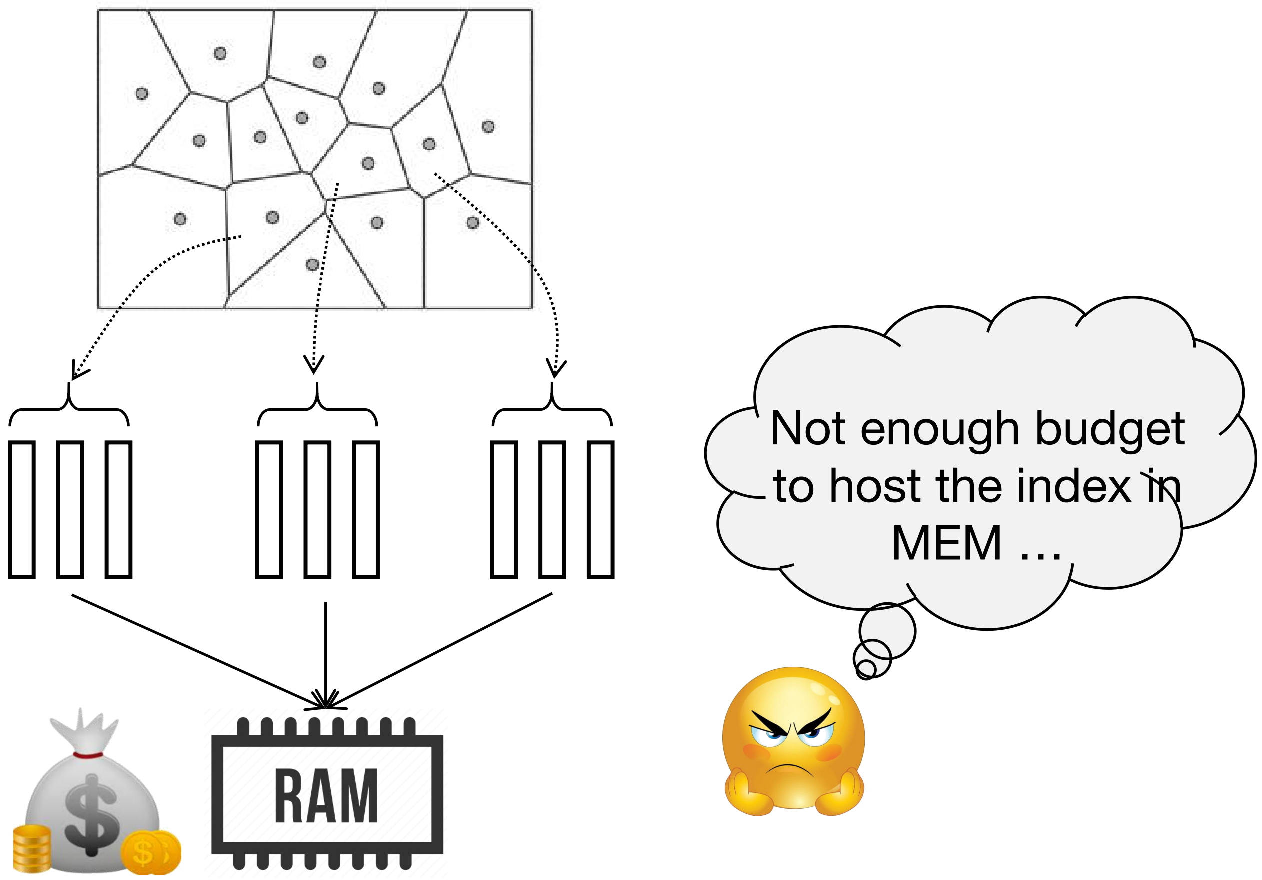
- Overview
  - Multi-Stage Pipeline
  - EBR: Pros and Cons
- Embedding learning algorithms
  - Negative Augmentation
  - Hard Negative Sampling
  - Diversified representation
  - Training as knowledge distillation
- Things beyond learning algorithms
  - Efficiency issues
  - Combo of sparse and dense

Things beyond: Efficiency

Desired properties about ANN (HSWN, PQ, ANNOY …)

- Accurate (high recall)
- Fast (low latency)
- Light (low mem cost)

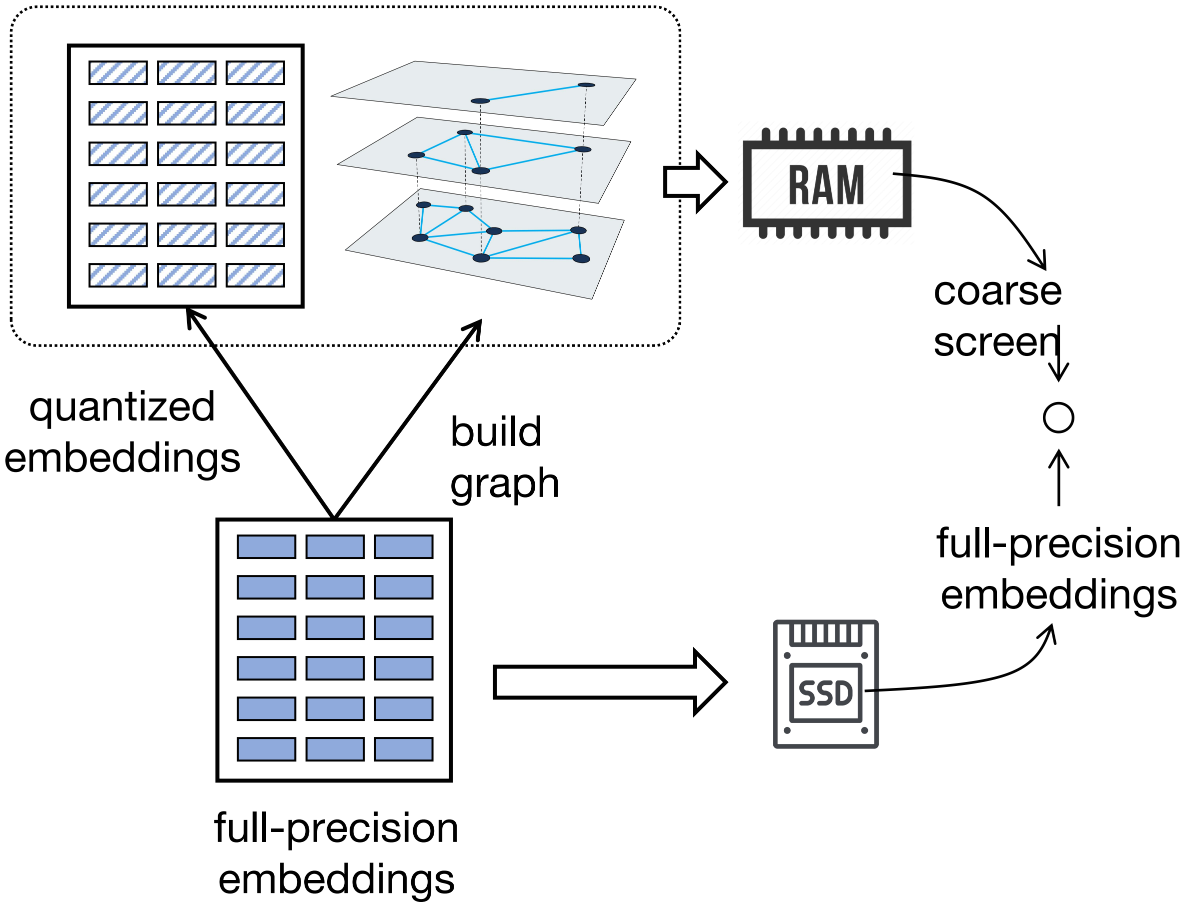
[FAISS, Facebook AI Research]



Things beyond: Efficiency

[DiskANN, Subramanya et. al.]

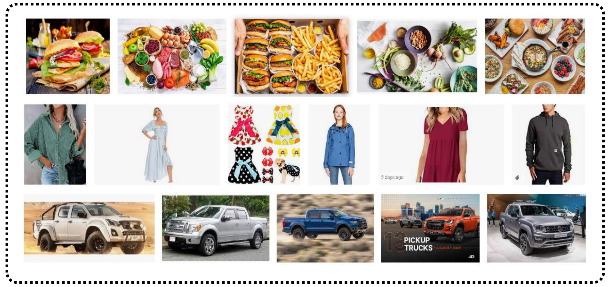
- Bi-granularity structure: Vamana Graph + PQ, quantized embeddings and graph in MEM, full-precision embeddings in SSD
- Using full-precision embeddings to build graph
- Using quantized embeddings for coarse-screened search
- Using full-precision embeddings for refinement

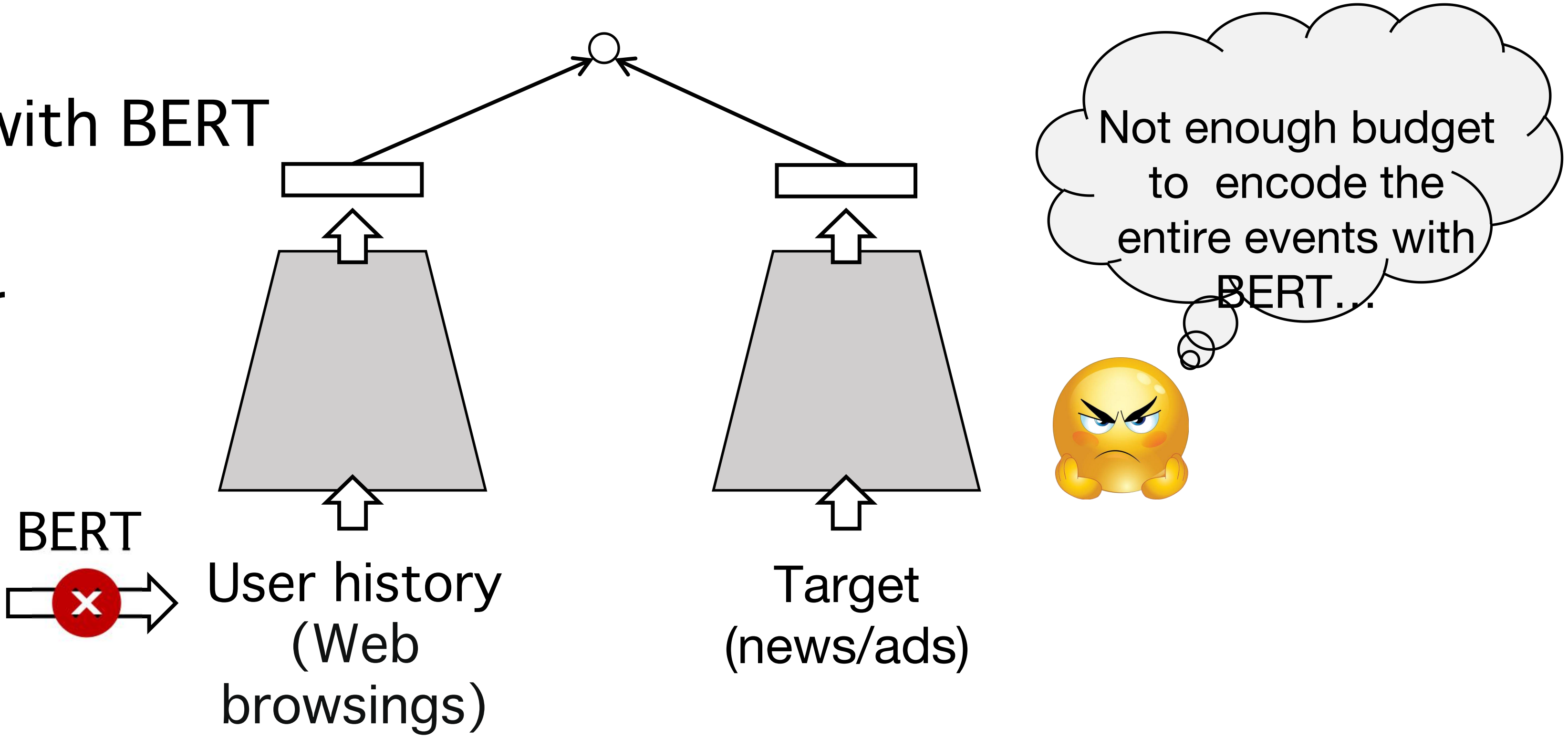


Things beyond: Efficiency

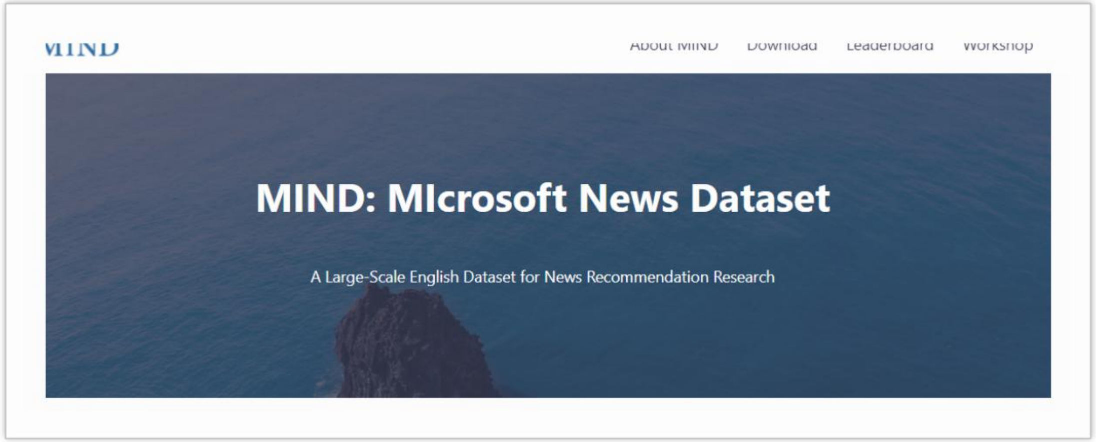
Applying PLMs for user modeling

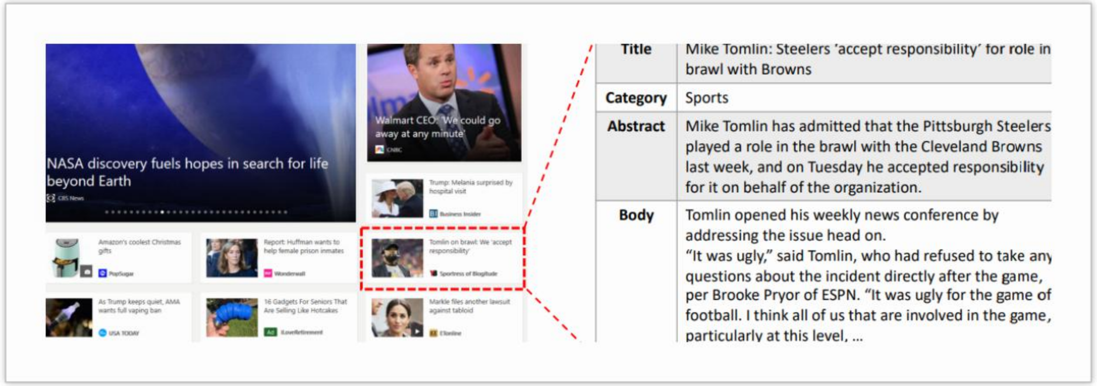
- Encode all events with BERT
- Aggregate events embeddings for user embeddings





Things beyond: Efficiency



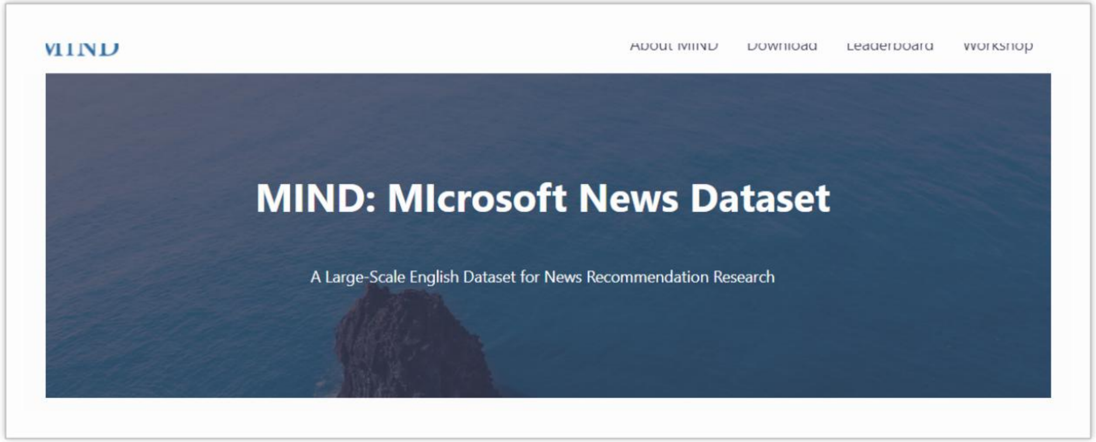


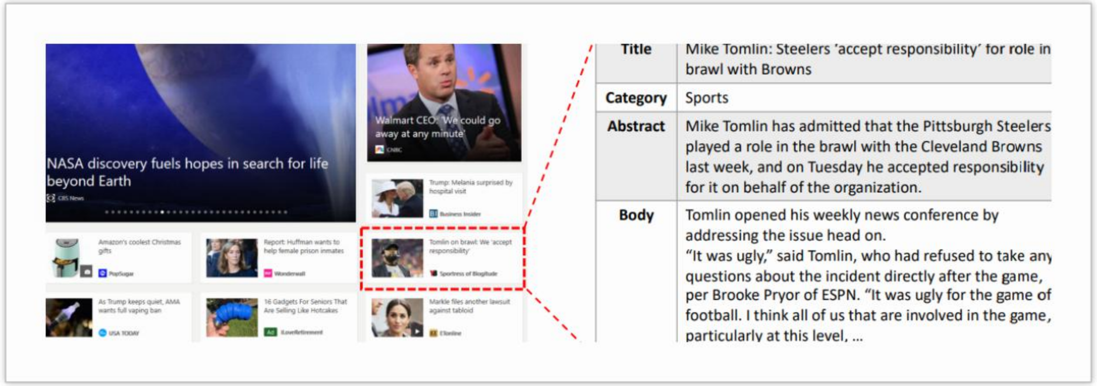
More features, model ensemble, complicated user models but insufficient BERT

Rank	Team	AUC	MRR	nDCG@5	nDCG@10
1	chenghuige	0.7131	0.3608	0.3960	0.4521
2	**oahciy**	**0.7096**	**0.3540**	**0.3883**	**0.4454**
3	Ravox	0.7048	0.3505	0.3845	0.4416
3	Qinne	0.7032	0.3496	0.3830	0.43976
3	gcc_microsoft	0.6979	0.3479	0.3806	0.4373

MIND Competition, July – Sept, 2020

Things beyond: Efficiency





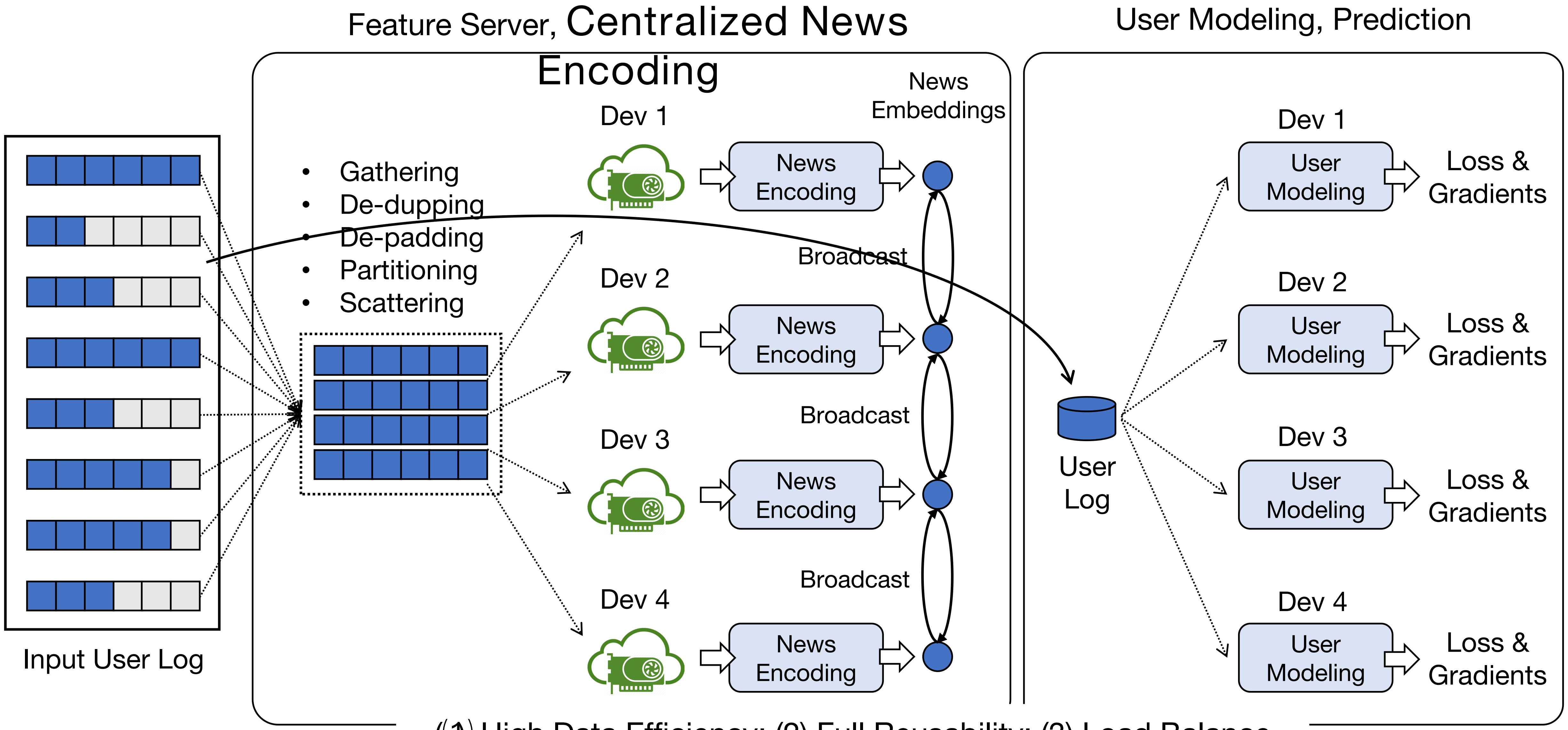
Rank	Team	AUC	MRR	nDCG@5	nDCG@10
1	chenghuige	0.7131	0.3608	0.3960	0.4521
2	**oahciy**	**0.7096**	**0.3540**	**0.3883**	**0.4454**
3	Ravox	0.7048	0.3505	0.3845	0.4416
3	Qinne	0.7032	0.3496	0.3830	0.43976
3	gcc_microsoft	0.6979	0.3479	0.3806	0.4373

MIND Competition, July – Sept, 2020

BERT Only	0.717 4	0.366 0	0.402 2	0.458 4

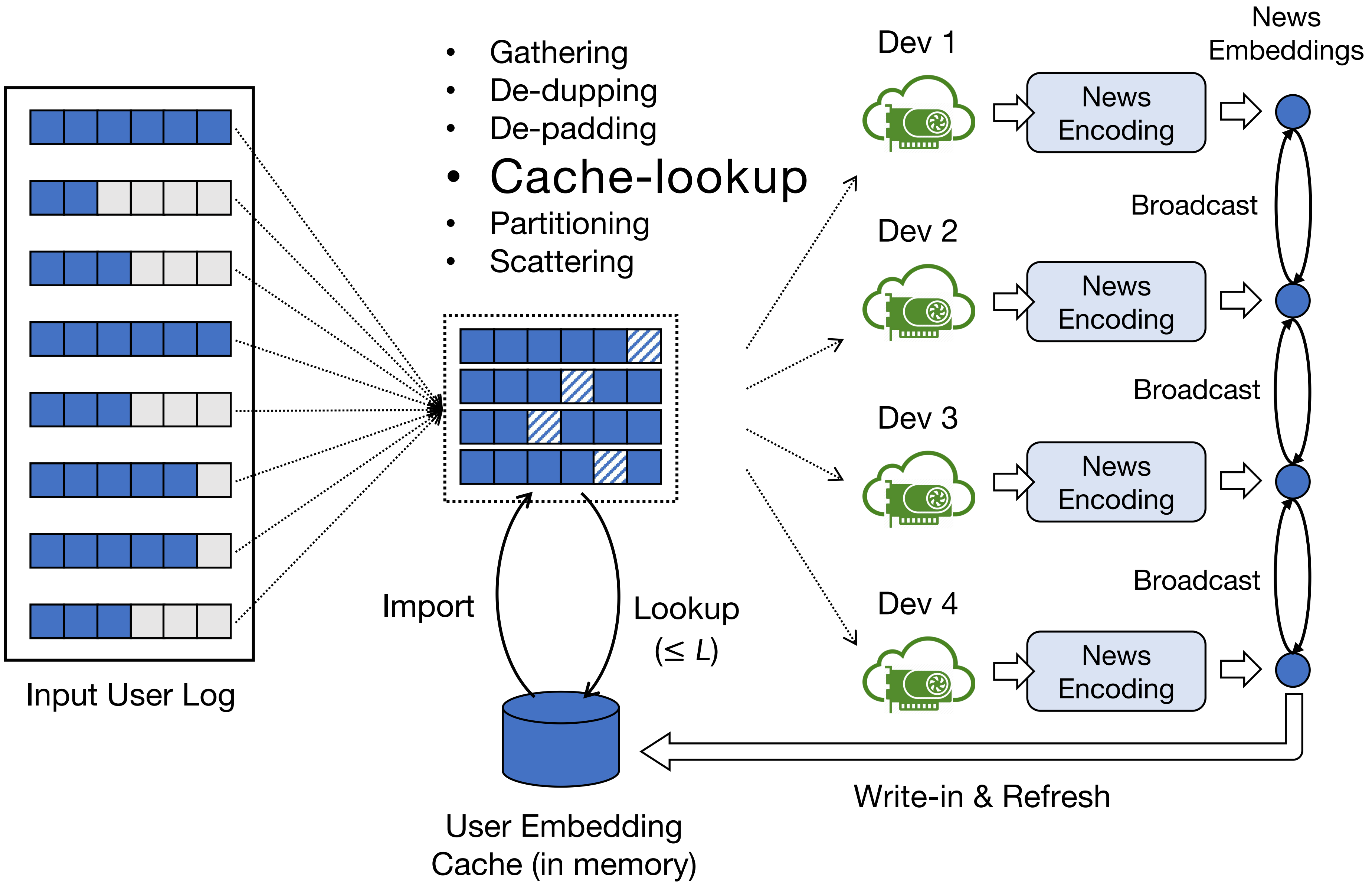
~4 hours, with 4*Nvidia V100 GPUs

SpeedyFeed: Paper, Xiao & Liu et. al., GitHub Repo



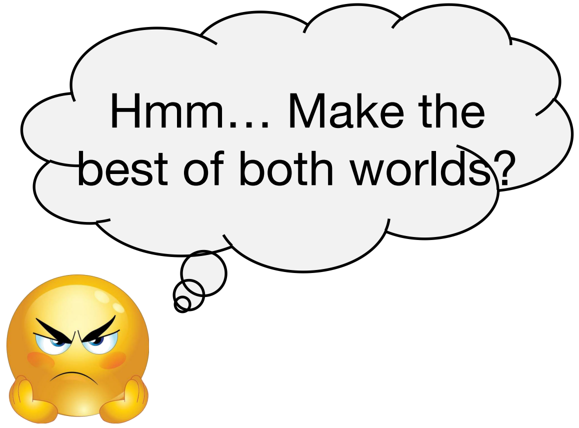
(1) High Data Efficiency; (2) Full Reusability; (3) Load Balance

A virtual super-big GPU device!


Gathering
De-dupping
De-padding
Cache-lookup
Partitioning
Scattering
Input User Log
Import
Lookup
(≤ L)
User Embedding
Cache (in memory)
Dev 1
Dev 2
Dev 3
Dev 4
News
Encoding
News
Encoding
News
Encoding
News
Encoding
News
Embeddings
Broadcast
Broadcast
Broadcast
Write-in & Refresh

Things beyond: Sparse, Dense Combo

	Pros	Cons
Sparse	• Efficient • Highly robust (in many cases)	• Inaccurate term selection • Mismatch of lexical feature
Dense	• Data driven • Highly generalizable	• Expensive (data, host) • Embedding ambiguity



Things beyond: Sparse, Dense Combo

Combo of Sparse and Dense Retrieval

- Use context-aware models (e.g., BERT) for term selection
- Use generation models (e.g., T5) for document expansion
- Do retrieval based on overlapped terms but estimate relevance based on embedding similarity

DeepCT, Dai et. al.

COIL, Gao et. al.

Uni-COIL, Lin et. al.

Sparse Representations			**MRR@10**	**Notes**
	Term Weighting	Expansion		
(1a)	BM25	None	0.184	copied from (Nogueira and Lin, 2019)
(1b)	BM25	doc2query–T5	0.277	copied from (Nogueira and Lin, 2019)
(2a)	DeepCT	None	0.243	copied from (Dai and Callan, 2019)
(2b)	DeepCT	doc2query–T5	?	no publicly reported figure
(2c)	DeepImpact	None	?	no publicly reported figure
(2d)	DeepImpact	doc2query–T5	0.326	copied from (Mallia et al., 2021)
(2e)	COIL-tok ($d = 32$)	None	0.341	copied from (Gao et al., 2021a)
(2f)	COIL-tok ($d = 32$)	doc2query–T5	0.361	our experiment
(2g)	uniCOIL	None	0.315	our experiment
(2h)	uniCOIL	doc2query–T5	0.352	our experiment
Dense Representations			**MRR@10**	**Notes**
(3a)	ColBERT		0.360	copied from (Khattab and Zaharia, 2020)
(3b)	ANCE		0.330	copied from (Xiong et al., 2021)
(3c)	DistillBERT		0.323	copied from (Hofstätter et al., 2020)
(3d)	RocketQA		0.370	copied from (Qu et al., 2021)
(3e)	TAS-B		0.347	copied from (Hofstätter et al., 2021)
(3f)	TCT-ColBERTv2		0.359	copied from (Lin et al., 2021)
Dense–Sparse Hybrids			**MRR@10**	**Notes**
(4a)	CLEAR		0.338	copied from (Gao et al., 2021b)
(4b)	COIL-full		0.355	copied from (Gao et al., 2021a)
(4c)	TCT-ColBERTv2 + BM25 (1a)		0.369	copied from (Lin et al., 2021)
(4d)	TCT-ColBERTv2 + doc2query–T5 (1b)		0.375	copied from (Lin et al., 2021)
(4e)	TCT-ColBERTv2 + DeepImpact (2d)		0.378	our experiment
(4f)	TCT-ColBERTv2 + uniCOIL (2h)		0.378	our experiment
(4g)	TCT-ColBERTv2 + COIL (2f)		0.382	our experiment

Things beyond: Sparse, Dense Combo

MSMARCO Document Ranking

- Negative expansion
- Enhancement of negative sampling
- Utilization of sparse feature
- Multiple rankers of different

Largest single leap forward in recent month. Should still be a lot of room for enhancement

MS MARCO Document Ranking Leaderboard

date		description	team	paper	code	type	MRR@100 (Dev)	MRR@100 (Eval)
2021/07/14	🏆	UniRetriever	Microsoft-Research-Asia and STCA-BingAdsSelection			full ranking	0.500	0.440
2021/06/24	🏆	Group-HNS-Retrieval+Multi-Granularity-Rerank	BUPT-University-MS-Recommenders			full ranking	0.496	0.436
2021/05/24	🏆	ANCE MaxP + LongP / SEED-Encoder+LongP (ensemble)	Soonhwan Kwon,Minyoung Lee, Samsung SDS AI Research			full ranking	0.487	0.427
2021/04/25	🏆	PROP_step400K base + doc2query top1000(ensemble v0.2)	Yingyan Li, Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xueqi Cheng - ICT, CAS	[paper]		full ranking	0.479	0.423
2021/04/28		Knowledge Retrieval	HuaweiPoissonLab, RUCIR			full ranking	0.482	0.423
2021/05/10		Knowledge Retrieval	HuaweiPoissonLab, RUCIR			full ranking	0.484	0.423
2021/05/26		Thinking Reranker (single)	Tongyuan - KCAI-Lab			full ranking	0.485	0.422
2021/07/08		COIL + RoBERTa	Luyu Gao, Carnegie Mellon University			full ranking	0.478	0.422
2021/04/27		ANCE BS+GL	Jiajia Ding*, Chunyu Li* - PingAn			full ranking	0.489	0.421
2021/04/18	🏆	ANCE + LongP (ensemble)	Soonhwan Kwon,Minyoung Lee, Samsung SDS AI Research			full ranking	0.481	0.420

Summary

- Overall speaking, EBR (dense) is the future, but SBR (sparse) will continue to thrive.
- Improve EBR algorithmically: negative sampling, weak-supervision, joint training of language model and user model, etc.
- Efficiency: training, serving, unified solution