Evaluating EBR beyond historical search results

Yu Cao

21 Sep 2023
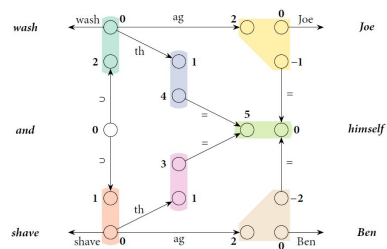
Yu Cao

Applied researcher @ Content Understanding

PhD in Linguistics (formal & computational)

pterosdiacos.github.io

Standard Evaluation

Ingredients

- A set of documents D
- A set of queries Q
- A relevance judgment for each (q, d) in $Q \times D$
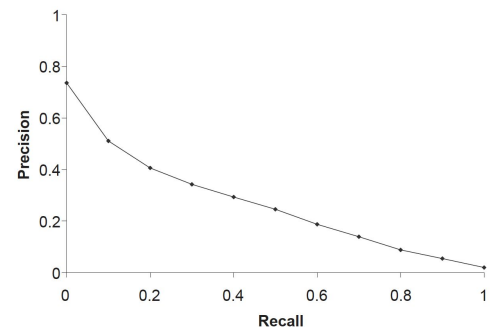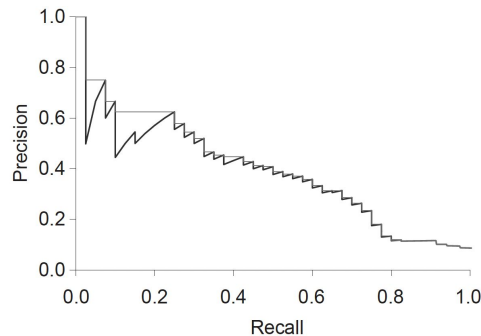
Ranked retrieval

$$q_i \mapsto d_{j1} < \cdots < d_{jn}$$

Standard Evaluation

Precision @ k   = (# of relevant items @ k) / k

Recall @ k   = (# of relevant items @ k) / (total # of relevant items)

(Interp.) precision–recall curve

Chris Manning et al. 2008. *Introduction to Information Retrieval*.

Standard Evaluation

Mean average precision

$$\frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk})$$

Search results from top to d_k

Relevance score for q_j and d_m

Normalized discounted cumulative gain @ k

$$\frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_k \sum_{m=1}^{k} \frac{2^{R(j,m)} - 1}{\log(1 + m)}$$

Normalization factor

Chris Manning et al. 2008. *Introduction to Information Retrieval.*

Standard Evaluation

Ingredients

- A set of documents D
- A set of queries Q
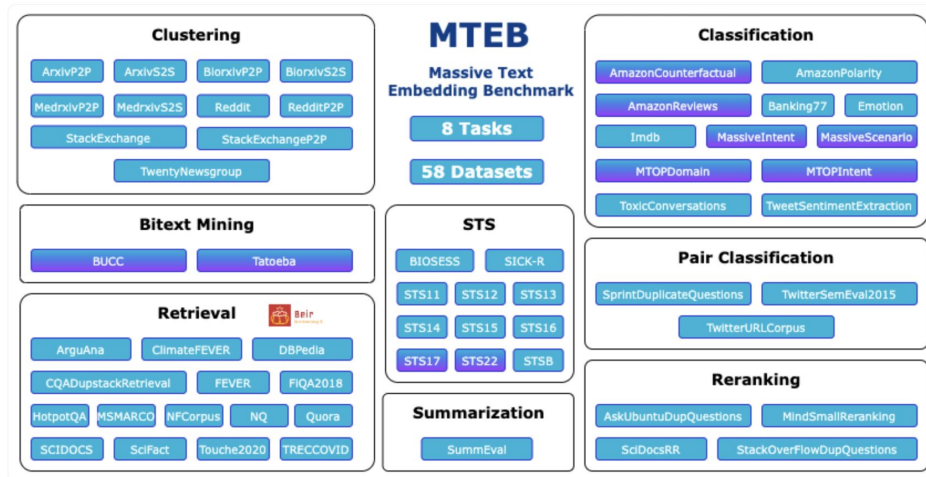- A relevance judgment for each (q, d) in Q×D

Where do I get these from?

- Human annotated datasets
- Generic benchmarks
  - MTEB, BEIR, …
- Historical user engagement data

Standard Evaluation

MTEB

a massive benchmark for measuring the performance of text embedding models on diverse embedding tasks.

Standard Evaluation

BEIR

a homogenous benchmark for diverse sentence or passage level IR tasks.

Dataset	Website	BEIR-Name	Public?	Type	Queries	Corpus	Rel D/Q	Down-load
MSMARCO	Homepage	msmarco	✅	train dev test	6,980	8.84M	1.1	Link
TREC-COVID	Homepage	trec-covid	✅	test	50	171K	493.5	Link
NFCorpus	Homepage	nfcorpus	✅	train dev test	323	3.6K	38.2	Link
BioASQ	Homepage	bioasq	❌	train test	500	14.91M	4.7	No
NQ	Homepage	nq	✅	train test	3,452	2.68M	1.2	Link
HotpotQA	Homepage	hotpotqa	✅	train dev test	7,405	5.23M	2.0	Link
FiQA-2018	Homepage	fiqa	✅	train dev test	648	57K	2.6	Link
Signal-1M(RT)	Homepage	signal1m	❌	test	97	2.86M	19.6	No
TREC-NEWS	Homepage	trec-news	❌	test	57	595K	19.6	No
Robust04	Homepage	robust04	❌	test	249	528K	69.9	No

Standard Evaluation

Historical user engagement

In a search session user engaged with item x but not item y

Engagement: *conversion, clicks, ...*

Standard Evaluation

Historical user engagement

Engagement → relevance

No/low engagement → no/low relevance ?

Scenario 1
- *User never reached page 2*
- *User jumped to page 3 skipping page 2*

Scenario 2
- *Null & low queries*

Standard Evaluation

Scenario 1
- *User never reached page 2*
- *User jumped to page 3 skipping page 2*

Scenario 2
- *Null & low queries*

…

Risks

- Reinforcing the current bias/failures
- No insights into model's performance on N&L

…

Beyond history

- Well-behaving of query embedding
- Well-behaving of item embedding
- Query & item embeddings cooperate
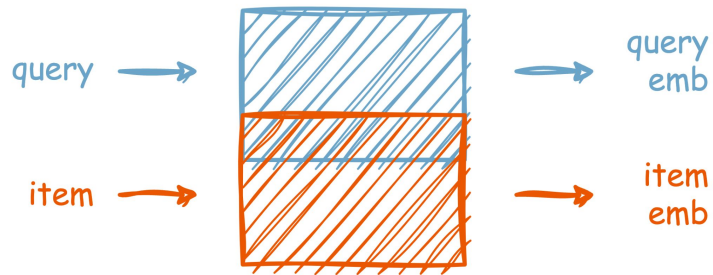
Before we proceed
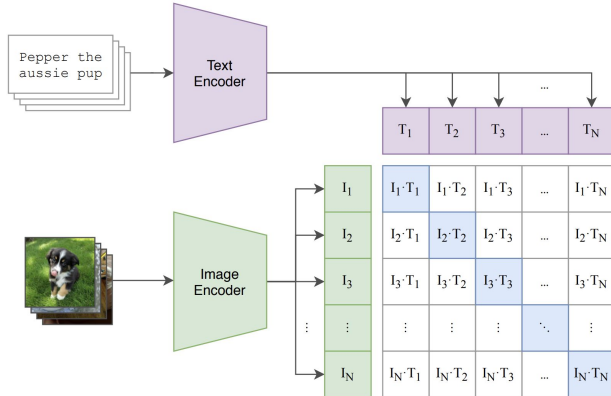
Is the similarity from one tower meaningful?

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." ICML. PMLR, 2021.

How eBay's New Search Feature Was Inspired By Window Shopping

Mar 31, 2023 | By: Senthilkumar Gopal, Shubhangi Tandon, Christopher Miller, Deepika Srinivasan, Rui Kong, Selcuk Kopru and Srinivas Bhagavathula

A new feature generates customer delight by using modern computer vision techniques to drive new search paradigms through visual discovery.

innovation.ebayinc.com/tech/engineering/how-ebays-new-search-feature-was-inspired-by-window-shopping/

Before we proceed
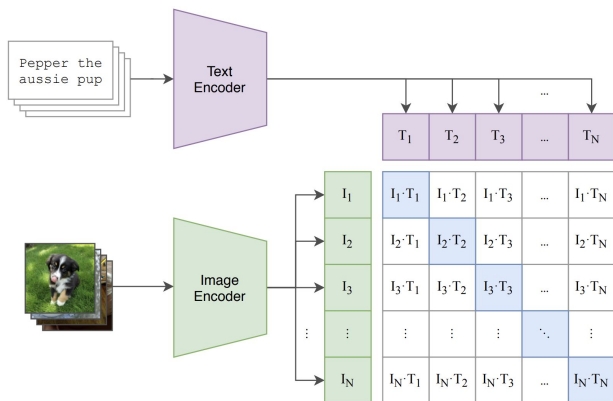
Is the similarity from one tower meaningful?

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." ICML. PMLR, 2021.

Semantic Similarity Based Label Augmentation for Visual Classification

Yu Cao

02 Mar 2023 (modified: 02 May 2023)   Submitted to Tiny Papers @ ICLR 2023   Readers: 🌐 Everyone   Show Bibtex   Show Revisions

Keywords: Label Augmentation, CLIP, Semantic Similarity, Visual Classification

TL;DR: How semantic similarity can be used to augment labels for training visual classification models.

Abstract: Real-world applications may present visual categories for which examples are many but a definition is elusive. When data augmentation helps little and hand-crafted heuristics fail to warrant weak supervision, similarity remains a simple but effective guide for augmenting training labels. This paper showcases, for recognizing clutters in e-commerce photography, how similarity between visual representations pre-trained to capture semantics provides quality supervision. Our approach significantly improves on the label propagation baseline, with classification precision and recall above 0.8 in most cases.

Add   Public Comment

openreview.net/forum?id=bRI_3OFg4o

Beyond history

- Well-behaving of query embedding

  (dis-) similar queries should have (dis-)similar embeddings

similar
q_{00}, q_{01}
q_{10}, q_{11}
q_{n0}, q_{n1}

dissimilar
q_{00}, q_{01}
q_{10}, q_{11}
q_{n0}, q_{n1}

Beyond history

- Well-behaving of query embedding

  (dis-) similar queries should have (dis-)similar embeddings
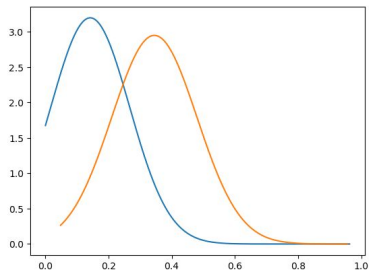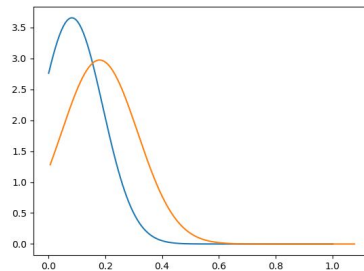
Model A acc. 0.80   Model B acc. 0.68

Beyond history

- Well-behaving of item embedding

  *Similar embeddings should imply **similar items***
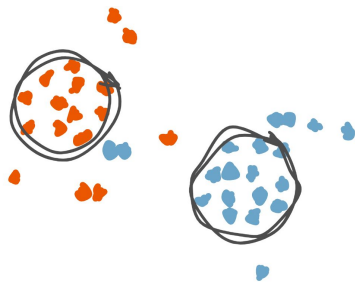
Beyond history

- Well-behaving of item embedding

  *Similar embeddings should imply **similar items***
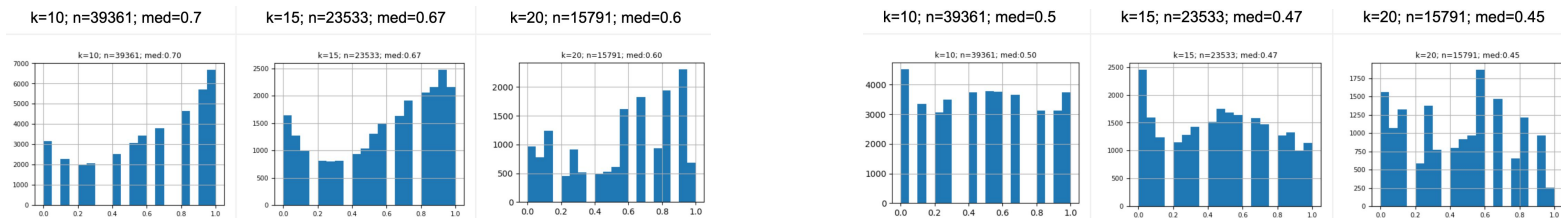
| Model A | K-neighbor product ID agreement | Model B |

Beyond history

- Query & item embeddings cooperate
  - A set of items I
  - A set of queries Q
  - Org knowledge about queries & items

  Ranked retrieval ($k \ll |I|$)

  $$q_i \mapsto d_{j1} < \cdots < d_{jk}$$

Beyond history

- **Org knowledge** about queries & items
  - Ground-truth category IDs of items
  - Predicted category IDs of queries
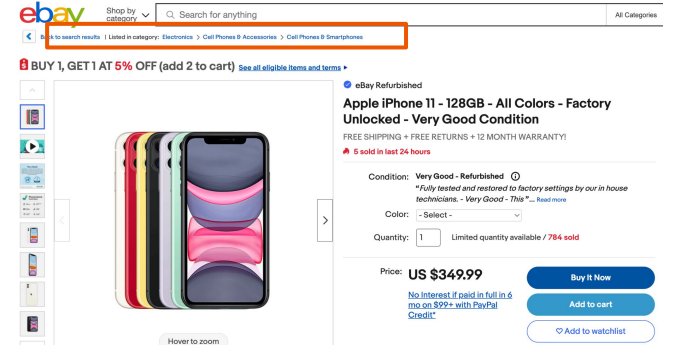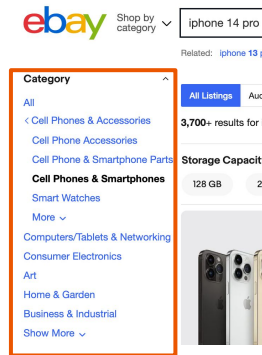
Beyond history

- **Org knowledge** about queries & items
  - Category IDs of queries & items
  - Predicted query-item relevance

Category agreement rate @ k

$$\frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{\sum_d \text{agree}(q_i, d)}{k}$$

Mean predicted relevance @ k

$$\frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{\sum_d f_{\text{rel}}(q_i, d)}{k}$$

Beyond history

- Example
  - query = "benchmade 417 fact"
  - Model A's recall

Rockys Messer Shop
https://www.rockys-messer-shop.de › ...

Benchmade 417 FACT, Axis

Benchmade 417 FACT, Axis ; Produktnummer: 610953151003 ; Rockys Messershop seit 2003 ; KOSTENLOSER VERSAND Deutschland ; Top Bewertung fast 100 %.

ACMA - Distributor For Fine Knives
https://acma.de › ... › Benchmade

Benchmade 417 FACT, Axis

Benchmade 417 FACT, Axis · Artikel-Nr.: BM417 · EAN/ UPC: 610953151003.

	title	cat_name	pred_rel
0	Benchmade Fact S30V Folding Knife Folder Plain...	[Factory Manufactured]	0.993635
1	417 AXIS Lock S30V Blade NylonFiber Handle Tac...	[Factory Manufactured]	0.017153
2	Benchmade Folding Pocket Knife Osborne Design ...	[Factory Manufactured]	0.771771
3	US SCOTT #417 MINT-SUPERB-OG-NH GRADED 98 W/ P...	[1901-40: Unused]	0.014371
4	Comp Cams 12-417-8 Xtreme Marine 218/224 Hydra...	[Camshafts]	0.003222
5	17 Issue of 417 Magazine top doctors, restuarn...	[Books]	0.012784
6	Powermaster Performance Alternator For 1972 Ch...	[Electronics]	0.002620
7	417 Casual Jacket Blue M 2200301029092	[Coats, Jackets & Vests]	0.211486
8	hoofdremcilinder Mini Mini Countryman R60 4175...	[Other Auto Tools & Supplies]	0.003259
9	AMAZING SPIDER-MAN #417 (9.0) SECRETS!! 1996	[Comics & Graphic Novels]	0.012171
10	69-78 Chevrolet 350 4-Bolt Camshaft M2S14	[Camshafts]	0.015529
11	B2851- 1993 Classic Four Sport Card #s 1-250 -...	[Trading Card Singles]	0.004843
12	Clevite Camshaft 717 fits 1965-69 Chevrolet 39...	[Camshafts]	0.003427
13	Amazing Spider-Man #417 FN 1996 Stock Image	[Comics & Graphic Novels]	0.007939
14	DSS Piston Set 3-2457-4125; 4.125" Bore 4cc Do...	[Pistons & Rings]	0.003108
15	SPA170X4 SPA or A Type Belt 170mm OD Vee Pulle...	[Other Business & Industrial]	0.007551
16	BATMAN #417 VF/NM - 1988 DC Comics - 1st app K...	[Comics & Graphic Novels]	0.006584
17	Diamond Racing Piston Kit 11828-8; 4.125" Bore...	[Pistons & Rings]	0.004296
18	QTY:1 170M4166 4168 4169 4170 4171 4174 4176 f...	[Other Metalworking Inspection & Measurement]	0.009681
19	1975 SSPC 416-629 EX-MT/NM Pick From List All ...	[Trading Card Singles]	0.007654

Beyond history

- Example
  - query = "benchmade 417 fact"
  - Model B's recall

⭐ Rockys Messer Shop
https://www.rockys-messer-shop.de › ... ⋮

Benchmade 417 FACT, Axis

Benchmade 417 FACT, Axis ; Produktnummer: 610953151003 ; Rockys Messershop
seit 2003 ; KOSTENLOSER VERSAND Deutschland ; Top Bewertung fast 100 %.

🔺 ACMA - Distributor For Fine Knives
https://acma.de › ... › Benchmade ⋮

Benchmade 417 FACT, Axis

Benchmade 417 FACT, Axis · Artikel-Nr.: BM417 · EAN/ UPC: 610953151003.

	title	cat_name	pred_rel
0	BENCHMADE for practice	[Factory Manufactured]	0.139019
1	Benchmade Fact S30V Folding Knife Folder Plain...	[Factory Manufactured]	0.993635
2	Benchmade Mini Osborne Black G10 Black S30V Ha...	[Factory Manufactured]	0.664691
3	Benchmade Folding Pocket Knife Osborne Design ...	[Factory Manufactured]	0.771771
4	Benchmade B945BK-1 Mini Osborne Reverse Tanto ...	[Other Hunting Knives & Tools]	0.430670
5	Benchmade 945BK-1 Mini Osborne Knife	[Factory Manufactured]	0.032023
6	Benchmade 430BK REDOUBT, CPM-D2 Stahl	[Camping Knives]	0.021519
7	Benchmade 550 - Black Handle with Black Blade	[Factory Manufactured]	0.032269
8	Benchmade Skeletonized Fixed Blade 440C Plain ...	[Fixed Blade]	0.299674
9	Benchmade 595BK Mini Boost AXIS-Assisted Foldi...	[Factory Manufactured]	0.072558
10	Benchmade 940 S30V OSBORNE Green Handle Pock...	[Factory Manufactured]	0.065963
11	Benchmade 317-1 Weekender CPM-S30V 2.97" 1.97"...	[Factory Manufactured]	0.088668
12	Benchmade 940 Plain Satin Blade S30V Osborne G...	[Factory Manufactured]	0.053036
13	Benchmade 15002 Saddle Mountain Skinner 4.2" S...	[Sheaths]	0.013257
14	Discontinued Benchmade Mini Boost 595BK	[Folding Blade]	0.032185
15	Benchmade 940BK-2003 Osborne Axis Limited Edit...	[Factory Manufactured]	0.077107
16	Benchmade 570SBK-1 AXIS Folder 3.72" CPM-S30V ...	[Folding Blade]	0.045416
17	Benchmade 7 - Safety, Rescue Tool - Black	[Pliers & Hook Removers]	0.403247
18	BENCHMADE 2007 CONSUMER CATALOG	[Custom & Handmade]	0.089569
19	Spring Set for the Axis Lock on Benchmade 810 ...	[Factory Manufactured]	0.014799

Beyond history

- Org knowledge about queries & items
  - Category IDs of queries & items
  - Predicted query-item relevance
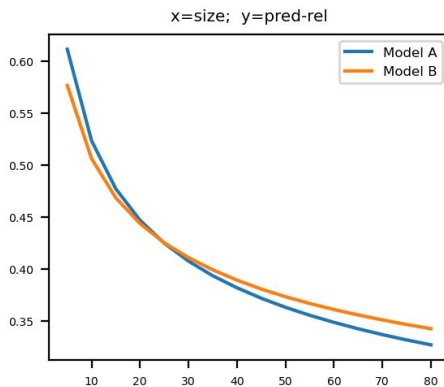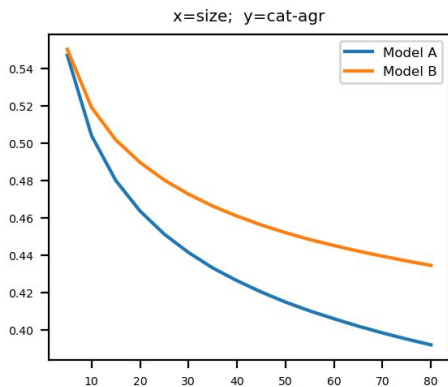
Beyond history

- Well-behaving of query embedding
  - query similarity ⇌ emb similarity

- Well-behaving of item embedding
  - item similarity ⇌ emb similarity

- Query & item embeddings cooperate
  - Category agreement
  - Predicted relevance

Questions

Thanks:)