CPU = Intel(R) Xeon(R) Gold 6326 CPU @ 2.90GHz

Note: smetric and recalls are 1-shot tested and subject to relatively high variance

sentence-transformers/all-MiniLM-l6-v2 params=23M, dim=384

L2 distance

int8 quant (PQ)

data = maths embeddings (13.9k - 61), films embeddings (20k - 1814)

search for k=10

Maths Embeddings
13.9k embeddings, 61 retrievals

Index	Parameters Notes	smetric@9	recall@10	recall@1	retrieval time per one query	retrieval time total	index building time	database size (MB)
FAISS Flat	exact search, baseline	45.00	100.0	100.0	632 µs ± 657 ns	250 ms ± 44.7 ms	37.7 ms ± 42.6 µs	20.36
Custom Flat	exact search	45.00	100.0	100.0	23 ms ± 179 µs	51.6 ms ± 721 µs	4.62 ms ± 6.59 µs	20.36
FAISS PQ	m=2	6.57	14.9	0.0	41.1 ms ± 213 µs	50.4 ms ± 536 µs	5.9 s ± 61.3 ms	0.027
Custom PQ	m=2	6.33	13.3	6.6	1.38 ms ± 148 µs	26.9 ms ± 180 µs	3.17 s ± 170 ms	0.027
FAISS PQ	m=4	11.95	21.3	16.4	82 ms ± 224 µs	90.3 ms ± 938 µs	8.75 s ± 31.9 ms	0.053
Custom PQ	m=4	11.41	21.5	26.2	2.11 ms ± 529 µs	28.4 ms ± 611 µs	3.5 s ± 162 ms	0.053
FAISS PQ	m=8	18.36	35.7	47.5	164 ms ± 380 µs	171 ms ± 652 µs	14.9 s ± 37.1 ms	0.106
Custom PQ	m=8	17.30	32.8	34.4	4.77 ms ± 701 µs	35.2 ms ± 141 µs	5.31 s ± 130 ms	0.106
FAISS PQ	m=16	21.70	43.1	54.1	328 ms ± 4.7 ms	341 ms ± 2.68 ms	12 s ± 831 ms	0.212
Custom PQ	m=16	22.16	42.3	55.7	3.58 ms ± 10.2 µs	47.9 ms ± 132 µs	8.46 s ± 279 ms	0.212
Custom PQ w/ MiniBatchKMeans	m=16	22.11	42.8	52.5	3.51 ms ± 13.2 µs	49.2 ms ± 119 µs	2.09 s ± 65.2 ms	0.212
FAISS PQ	m=32	31.75	61.6	59.0	169 µs ± 70.4 ns	21.3 ms ± 175 µs	24.9 s ± 28.4 ms	0.424
Custom PQ	m=32	32.43	61.3	70.5	6.44 ms ± 46.5 µs	120 ms ± 39.9 µs	14.7 s ± 99.4 ms	0.424
Custom PQ w/ MiniBatchKMeans	m=32	33.18	62.8	63.9	6.43 ms ± 4.25 µs	118 ms ± 52.1 µs	3.41 s ± 114 ms	0.424
FAISS IVF	nlist=100, nprobe=4	38.10	81.8	83.6	35.4 µs ± 53.2 ns	10.3 ms ± 38.6 µs	1.16 s ± 7.16 ms	
Custom IVF	nlist=100, nprobe=4	38.87	84.3	86.9	72.4 ms ± 108 µs	2.49 s ± 5.37 ms	1.5 s ± 53.1 ms	
Custom IVF w/ MiniBatchKMeans	nlist=100, nprobe=4	36.67	80.7	80.3	72.3 ms ± 271 µs	2.48 s ± 6.28 ms	283 ms ± 33.3 ms	
CustomOptimized IFV	nlist=100, nprobe=4	37.49	81.0	82.0	571 µs ± 1.15 µs	41.8 ms ± 33.5 µs	1.99 s ± 126 ms	
FAISS IVF	nlist=256, nprobe=8	41.49	91.1	91.8	76.6 µs ± 26.1 ns	10.4 ms ± 75.2 µs	1.16 s ± 6.82 ms	
Custom IVF	nlist=256, nprobe=8	39.92	86.9	91.8	72.8 ms ± 300 µs	2.45 s ± 3.16 ms	2.7 s ± 169 ms	
Custom IVF w/ MiniBatchKMeans	nlist=256, nprobe=8	38.79	84.3	90.2	74.5 ms ± 806 µs	2.5 s ± 8.95 ms	366 ms ± 79.5 ms	
CustomOptimized IFV	nlist=256, nprobe=8	39.67	86.2	88.5	826 µs ± 24.2 µs	22.8 ms ± 18.1	3.22 s ± 163 ms	
FAISS IVF	nlist=256, nprobe=32	44.41	98.0	100.0	115 µs ± 68.3 ns	10.4 ms ± 81.6 µs	1.97 s ± 10.1 ms	
Custom IVF	nlist=256, nprobe=32	44.31	98.0	98.4	81 ms ± 1.23 ms	2.72 s ± 13.3 ms	2.71 s ± 114 ms	
CustomOptimized IFV	nlist=256, nprobe=32	44.31	98.4	98.4	1.99 ms ± 23.3 µs	145 ms ± 122 µs	3.15 s ± 213 ms	
CustomOptimized IFV w/ MiniBatchKMeans	nlist=256, nprobe=32	44.64	99.0	98.4	1.38 ms ± 1.09 µs	193 ms ± 1.86 ms	509 ms ± 68.1 ms	
FAISS IVFPQ	nlist=256, nprobe=32, m=32	30.51	56.4	67.2	165 µs ± 38.2 ns	580 µs ± 83.7 µs	1.87 s ± 16.4 ms	

CPU = Intel(R) Xeon(R) Gold 6326 CPU @ 2.90GHz:

Note: all numbers are 1-shot tested and subject to relatively high variance

Alibaba-NLP/gte-base-en-v1.5 params=137M, dim=768

L2 distance

int8 quant (PQ)

data = embeddings1M

search for k=10

embeddings1M
999 950 embeddings, 50 retrievals

Index	Parameters Notes	smetric@9	recall@10	recall@1	retrieval time per one query	retrieval time total	index building time	database size (GB)
FAISS Flat	exact search, baseline	45.00	100.0	100.0	249 ms ± 47.5 µs	4.66 s ± 168 ms	3.46 s ± 36.9 ms	2.861
Custom Flat	exact search	45.00	100.0	100.0	2.18 s ± 30 ms	25.8 s	1.21 s ± 9.76 ms	2.861
FAISS PQ	m=32	27.28	50.6	42.0	21.8 ms ± 2.66 ms	44.3 ms ± 18.8 ms	33.2 s	0.0298
Custom PQ w/ MiniBatchKMeans	m=32	27.76	54.4	44.0	185 ms ± 237 µs	14.4 s	1mn 29 s	0.0298
FAISS IVF	nlist=256, nprobe=32	43.74	96.4	96.0	31.7 ms ± 117 µs	140 ms ± 1.96 ms	33.7 s	2.861+
CustomOptimized IFV w/ MiniBatchKMeans	nlist=256, nprobe=32	44.36	98.2	96.0	594 ms ± 3.91 ms	54 s	17.6 s	2.861+
CustomOptimized IFV w/ MiniBatchKMeans, PCA	nlist=256, nprobe=32, pca_d=128	31.54	59.8	46.0	107 ms ± 4.69 ms	7 s	42 s	0.4768+
CustomOptimized IFV w/ MiniBatchKMeans, PCA	nlist=256, nprobe=32, pca_d=384	43.86	94.6	94.0	439 ms ± 144 µs	26 s	1mn 38 s	1.430+
FAISS PQ w/ OPQ	m=32	31.28	58.4	56.0	14.6 ms ± 130 µs	32.4 ms ± 345 µs	10mn	0.0298
FAISS HNSW	M=8	31.42	67.4	80.0	14.8 ms ± 138 µs	15.3 ms ± 117 µs	20 s	
FAISS HNSW	M=32	39.22	86.4	86.0	15 ms ± 114 µs	15.7 ms ± 205 µs	43 s	
FAISS HNSW	M=64	40.02	88.2	88.0	14.5 ms ± 1.3 ms	15.8 ms ± 106 µs	53 s	

CPU = Intel(R) Xeon(R) Gold 6326 CPU @ 2.90GHz

Note: all numbers are 1-shot tested and subject to relatively high variance

Alibaba-NLP/gte-base-en-v1.5 params=137M, dim=768

IP distance (dot product)

int8 quant (PQ)

data = embeddings1M

search for k=10

embeddings1M
999 950 embeddings, 50 retrievals

Index	Parameters Notes	smetric@9	recall@10	recall@1	retrieval time per one query	retrieval time total	index building time	database size (GB)
FAISS Flat	exact search, baseline	45.00	100.0	100.0	240 ms ± 1.23 ms	4.9 s ± 556 ms	3.4 s ± 15.5 ms	2.861
FAISS IVF	nlist=256, nprobe=32	43.66	96.4	96.0	29.2 ms ± 17.6 µs	134 ms ± 1.73 ms	33 s	
CustomOptimized IFV w/ MiniBatchKMeans	nlist=256, nprobe=32	40.72	90.6	88.0	94.9 ms ± 2.06 ms	10.7 s ± 172 ms	19 s	2.861
CustomOptimized IFV w/ MiniBatchKMeans, PCA	nlist=256, nprobe=32, pca_d=384	35.58	71.8	68.0	43.1 ms ± 1.26 ms	5.38 s ± 6.57 ms	1 mn 33 s	1.4304
FAISS HNSW	M=32	38.00	84.4	82.0	728 µs ± 266 µs	2.84 ms ± 1.67 ms	42 s	