腾讯云开发者社区 | 腾讯云向量数据库 Tencent Cloud VectorDB | Andon

腾讯云向量数据库

RAG七天入门课 第三节

相似性检索的关键 — Embedding

腾讯云高级算法工程师

赵九州

BBTalk

第三节

相似性检索的关键

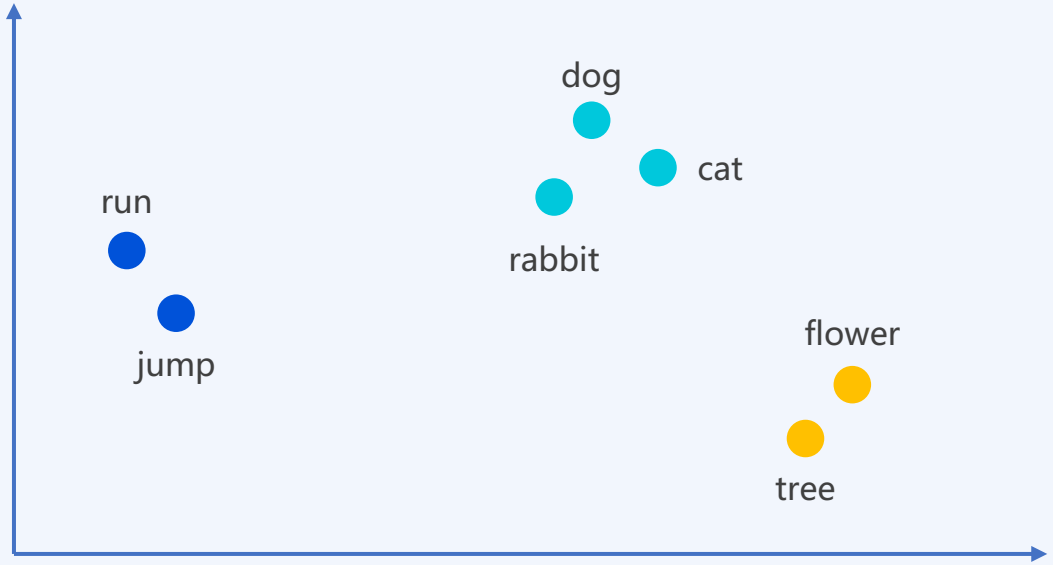
- Embedding

文本Embedding技术

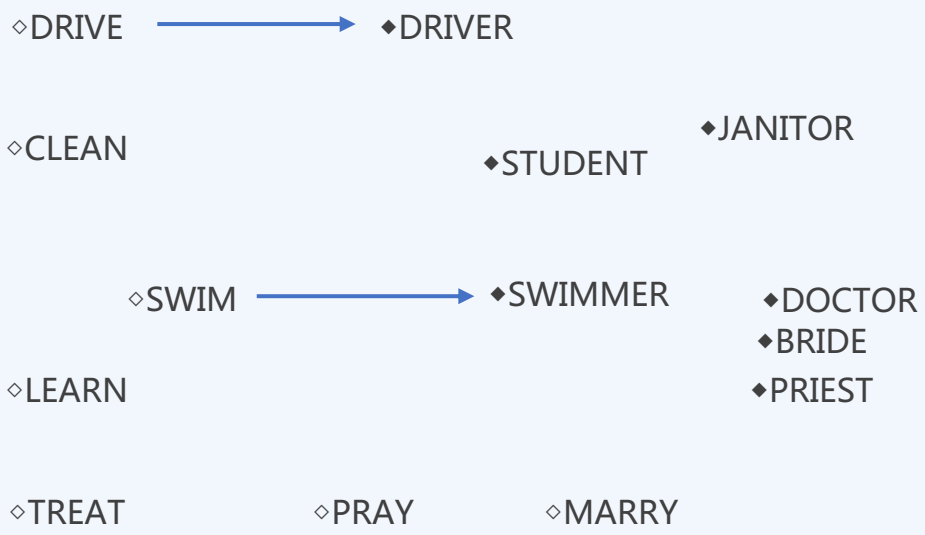
文本Embedding是将整个文本转化为实数向量的技术。

Embedding的优点是可以将离散的词语或句子转化为连续的向量，这样就可以使用数学方法来处理词语或句子，从而捕捉到文本的语义信息，文本和文本的关系信息。

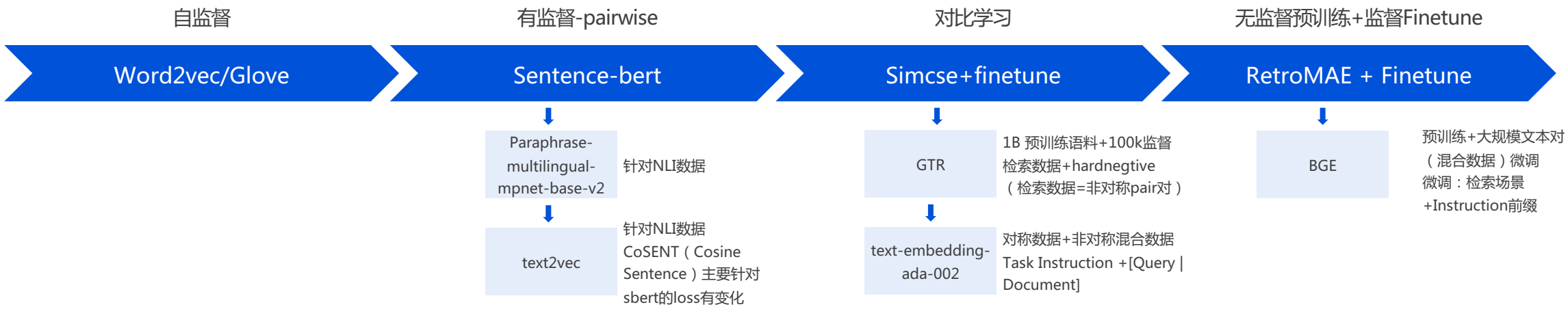
◉ 优质的Embedding通常会让语义相似的文本在空间中彼此接近



◉ 优质的Embedding相似的语义关系可以通过向量的算术运算来表示



文本Embedding模型的演进与选型



目前的向量模型从单纯的基于 NLI 数据集（对称数据集）发展到基于混合数据（对称+非对称）进行训练，即可以做 QQ 召回任务也能够做 QD 召回任务，通过添加 Instruction 的方式来区分这两类任务，只有在进行 QD 召回的时候，需要对用户 query 添加上 Instruction 前缀。

VDB 通用 Embedding 模型

模型选择

Massive Text Embedding Benchmark (MTEB) Leaderboard. To submit, refer to the MTEB GitHub repository 🤗 Refer to the MTEB paper for details on metrics, tasks and models.

Overall | Bitext Mining | Classification | Clustering | Pair Classification | Reranking | Retrieval | STS | Summarization

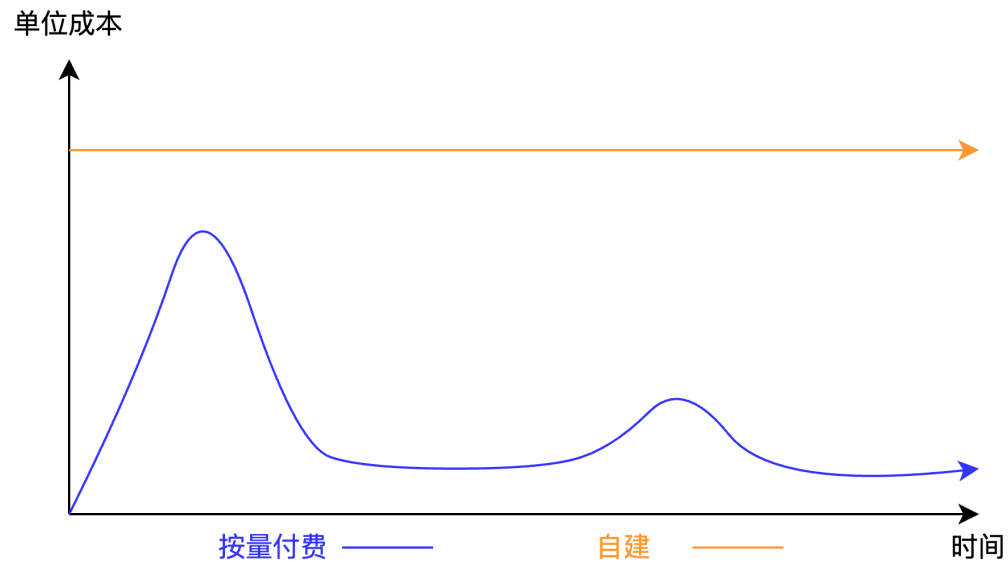
English | Chinese | French | Polish

Overall MTEB Chinese leaderboard (C-MTEB) 🔮🇨🇳

- **Metric:** Various, refer to task tabs
- **Languages:** Chinese
- **Credits:** FlagEmbedding

Rank	Model	Model Size (GB)	Embedding Dimensions	Max Tokens	Average (35 datasets)	Classification Average (9 datasets)	Clustering Average (4 datasets)	Pair Classification Average (2 datasets)
1	OpenSearch-text-hybrid		1792	512	68.71	71.74	53.75	88.1
2	stella-mrl-large-zh-v3.5-1792	1.3	1024	512	68.55	71.56	54.32	88.08
3	stella-large-zh-v3-1792d	1.3	1024	512	68.48	71.5	53.9	88.1
4	Baichuan-text-embedding		1024	512	68.34	72.84	56.88	82.32
5	stella-base-zh-v3-1792d	0.41	768	1024	67.96	71.12	53.3	87.93
6	Dmeta-embedding-zh	0.41	768	1024	67.51	70	50.96	88.92
7	xiaobu-embedding	1.3	1024	512	67.28	71.2	54.62	85.3
8	alime-embedding-large-zh	1.3	1024	512	67.17	71.35	54	84.34
9	acge-large-zh	0.65	1024	1024	67	73.39	55.89	81.38
10	gte-large-zh	0.65	1024	512	66.72	71.34	53.07	84.41

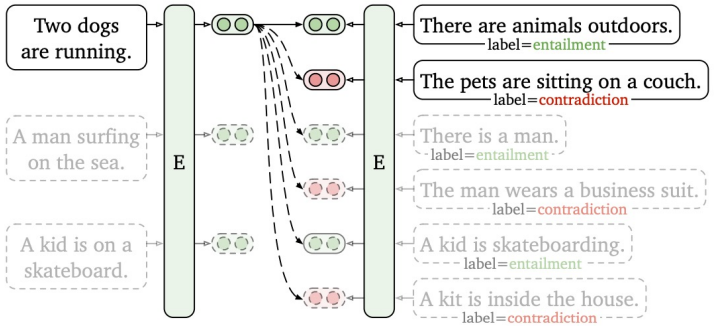
GPU资源



VDB 垂类 Embedding 模型

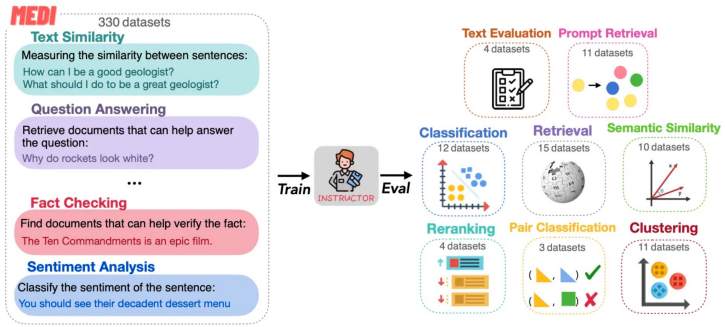
用户提供垂类文档数据，VDB对模型进行微调，助力垂类应用效果更进一步

优化1：对比学习拉近同义文本的距离，推远不同文本的距离

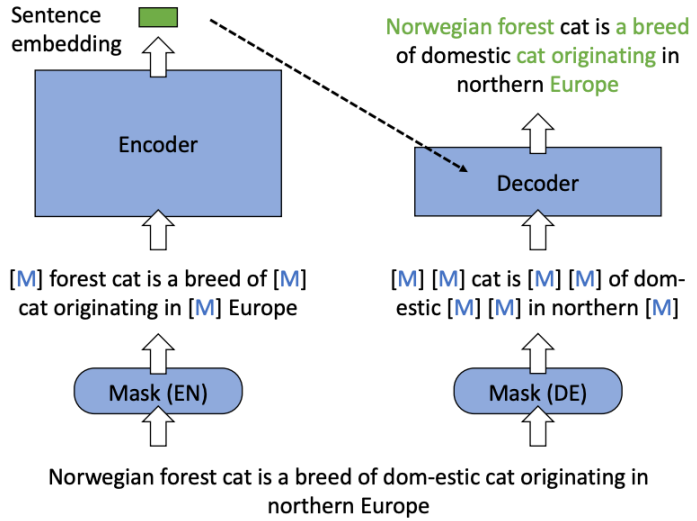


$$L_q = -log\frac{exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{k} exp(q \cdot k_i / \tau)}$$

优化2：短文本匹配和长文本匹配使用不同prompt，提升非对称类文本效果



优化3：预训练阶段提升基座模型面向检索的能力，对比学习阶段提高负样本数



效果对比

模型	维度	文档检索(MRR@10)	FAQ检索(MRR@10)	相似句(Spearman)
openai	1536	83.4%	90.65%	75.56%
bge-large-zh(开源SOTA)	1024	83.08%	91.09%	77.18
ancse-v1	768	27.1%	76.82%	80.69%
ancse-v2	1024	84.05%	93.25%	82.11%
ancse-v2-白化	256	83.71%	93.25%	82.07%
ancse-v3	1024	85.23%	94.58%	82.81%

存储、检索向量数据

为何需要一个专用的向量数据库

1. 查询方式与传统数据库存在区别

2. 简单易用，无需关心细节

3. 为相似性检索设计，天生性能优势

腾讯云向量数据库的优势

“首家”

通过信通院的标准化性能和规模测试

支持千亿级向量规模和最高500W QPS

自研

内核源自集团自研OLAMA引擎，

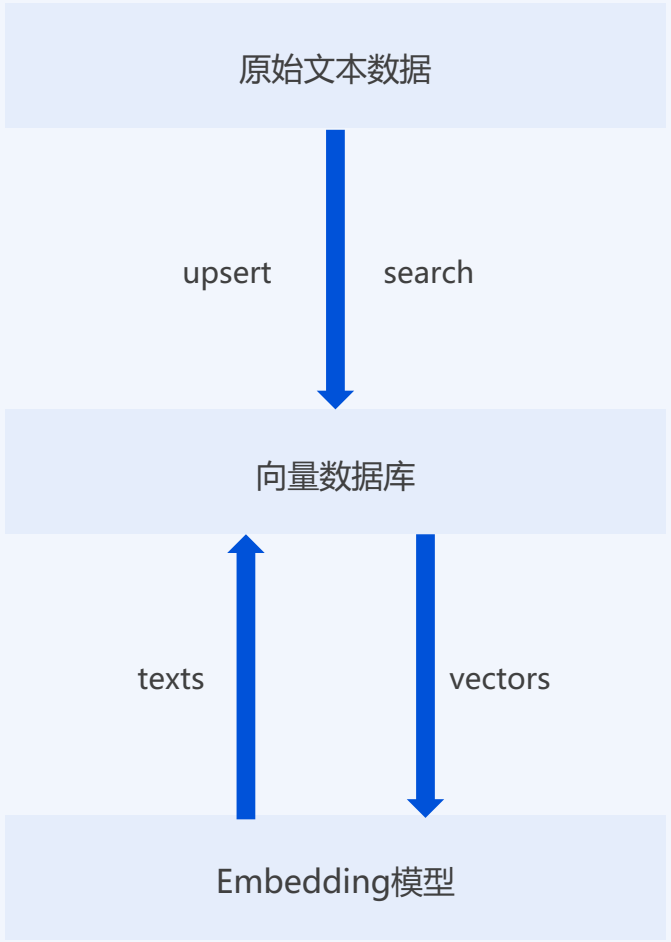
内部已有**40+**业务接入

性价比

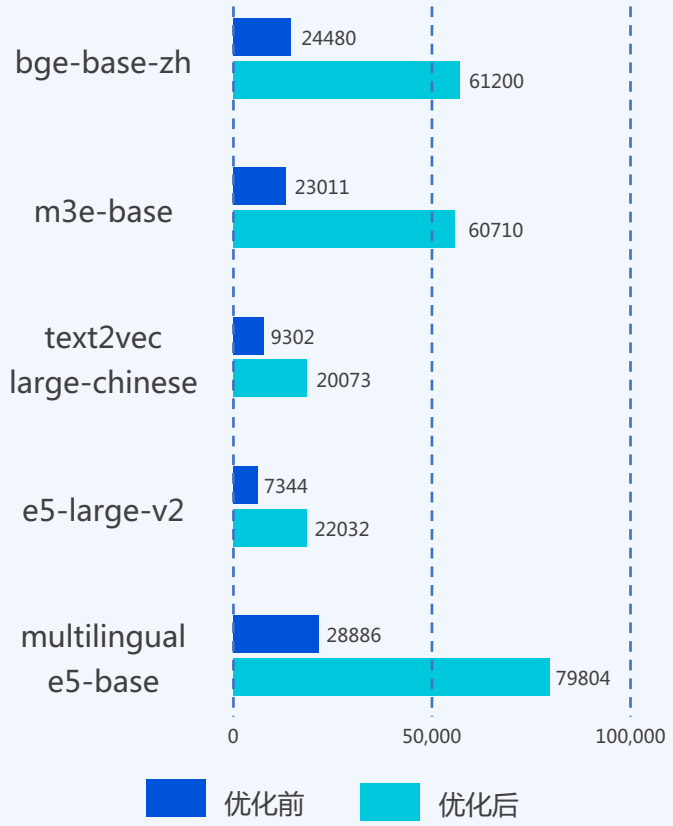
性能领先业内平均水平**1.5**倍

同时客户成本降低**20%**

VDB 优势

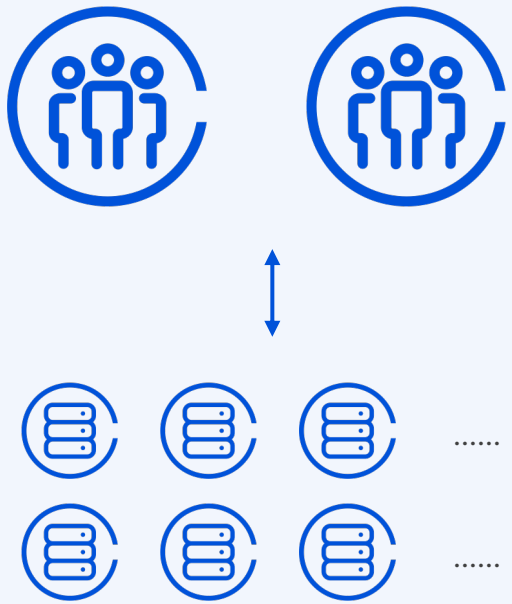


流程简化



*单张T4卡，每秒处理Token速率
*初步优化结果，不代表最终结果

模型简化



◉ 默认提供 1w Token/s，远超其他Embedding服务
◉ 在线、离线导入方式分离，满足不同场景下的Embedding需求

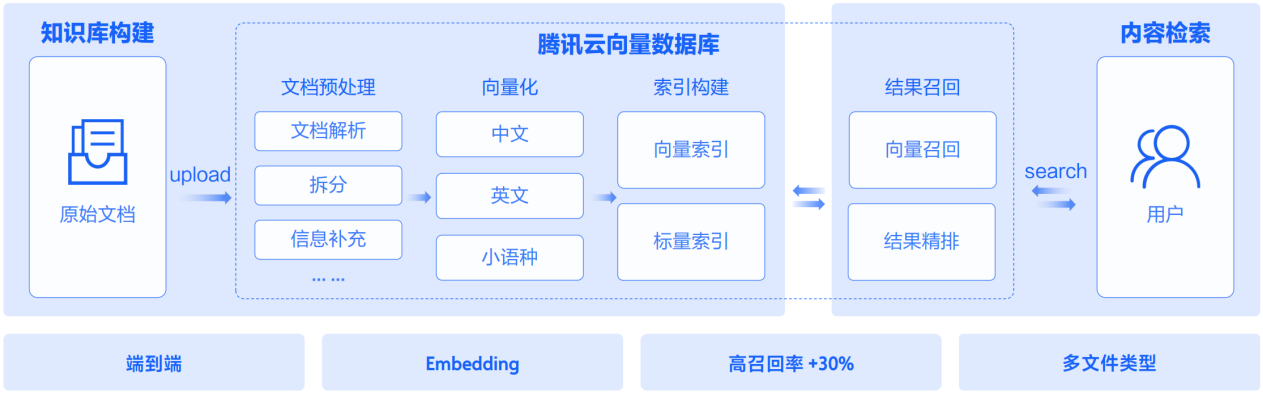
共享GPU集群

参考文档链接：https://cloud.tencent.com/document/product/1709/98014

腾讯云向量数据库：消除大模型幻觉，加速大模型在企业落地

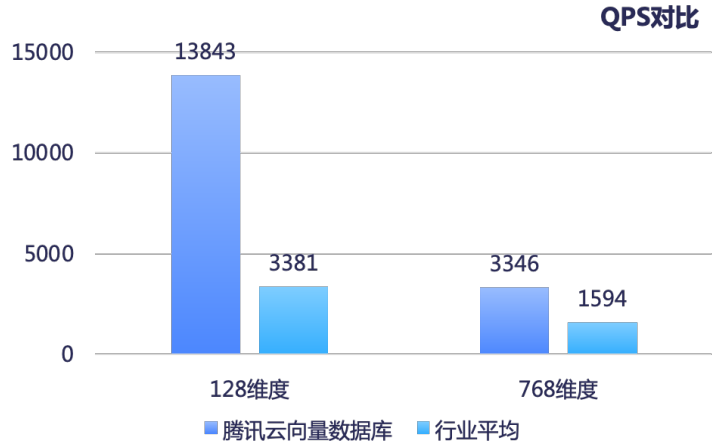
端到端AI套件，AGI时代的知识库解决方案

提供**一站式**知识检索方案，实现业界内**最高召回率、大幅降低开发门槛**，帮助企业快速搭建RAG应用，解决大模型幻觉问题



源自集团多年积累，产品能力行业领先

源自腾讯自研向量检索引擎OLAMA，集团内部**40+**业务线上使用，日均处理**1600亿次**检索请求



首家

『首家』通过中国信通院向量数据库标准测试

千亿

单索引支持最高**千亿级**超大数据规模

500万

单实例最高可达**500万 QPS**