Milvus

Build Up the Unstructured Data Service

Jun Gu
09.2020
Speaker bio

Jun Gu

Database engineer, SME

Voting member in Technical Advisory Council (TAC)
Partner, Chief Evangelist

Career history

Education
Zilliz: Who we are

• Open source software company based in Shanghai
• Mission: Reinvent data science
• Main contributor of Milvus project
Unlock the treasure of unstructured data

AI algorithms transform image, video, voice, natural language into vectors, and enables understanding and utilization of unstructured data at scale.
The flow-based AI applications

The most popular way
- Flexible
- Easy to compose, web-based UI
- Sample pipelines

The challenge
- Data fragmentation

The sample pipelines for video processing
The unstructured data service (UDS) for AI

Unstructured Data
image, video, voice, natural language

Model Inference Runtime
TensorRT, ONNX RT, TFRT

Milvus

Vectors
High dense + Sparse

Attributes
(will be in 0.11)

Multimodal
(will be in 0.14)

Scoring
(will be in 0.16)

Object Storage

Result Set
image, video, voice, natural language

Inference Layer

Data Service Layer

© 2020 Zilliz. All rights reserved.
Why Milvus: Vectors are different

Numbers

Arithmetic operation

+ − × ÷

Number comparison

a <=> b

1–10

1–5

6–10

1  2  3  4  5  6  7  8  9  10

Vectors

Similarity (eg. Euclidean distance)

\[ d(A, B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \]

Similarity comparison

TopK(A) = \arg\min_{B \in Y} (d(A, B))
Milvus: The big picture

- **X86**: supports SSE4.2, AVX2, AVX512
- **GPU**: Pascal microarchitecture or later, CUDA 10.0 or later
- **Arm**: requires aarch64
- **Kunpeng**: tested on Kunpen 920 with Centos 7.x
- **Loongson**: tested on Loongson with docker container
- **RSIC-V**: in early development
Milvus: Distributed deployment

1. TopK Query Request
2. Map Request to Sub-Requests
3. TopK Sub-Request
4. TopK Sub-Results
5. Merge TopK Results
6. TopK Query Response
Milvus: The ANN benchmark

Milvus: 0.8.0
OS: Ubuntu 18.04
ECS: AWS c5.4xlarge (16c, 32GB), Intel XeonPlatinum 8275CL
Data set: sift-128-euclidean (1 million vectors)
More info: https://milvus.io/docs/benchmarks_aws

Special thanks to ANN-Benchmarks (developed by Martin Aumueller, Erik Bernhardsson and Alec Faitfull)
Milvus: The journey

The idea → Milvus 0.1 → 1st seed user

Open Source → Joined LF AI

2018.10  2019.04  2019.06

The most active AI projects in Linux foundation

Milvus

- Vector
- Contributions: 263
- Commits: 14178
- Data Sources: 10984
- Repositories: 11

© 2020 Zilliz. All rights reserved.
Progress

Unstoppable momentum since its debut.

- Commits: 5.9K
- GitHub stars: 3.9K
- Contributors: 104
- Release: 16
- Users: 200+
- Patents filed: 19
Milvus Features & benefits

The world’s most advanced, our target 😊
Use case: Intelligent writing assistant

1. Corpus Data natural language
2. Data Cleansing
3. Feature engineering
4. Encoder TextCNN
5. Extract paragraph, summary
6. Encoder InferSent
7. Milvus
8. Object URI
9. Vector ID
10. Object Storage
11. Writing Intention
12. Result An auto-generated essay
Use case: News recommendation on mobile

1. **Daily batch News title**
2. **Feeding News title**
3. **Encoder SimBert**
4. **Milvus**
5. **Object URI**
6. **Object Storage**
7. **Reading Preference**
8. **Recommended News**
Use case: Image search for company trademark

- **55 million** images
- **Search elapsed time:**
  - **20 ms** on cloud GPU server
Use case: Pharmaceutical molecule analysis

- **800 million** molecules
- Search elapsed time: **500 ms** on single server

![Diagram showing molecular analysis process](image)

- Encoder
- RDKit
- Milvus
- Molecular Similarity
- Tanimoto similarity
- Substructure
- Superstructure
- Candidate List

Molecular Formula: CC(=O)Nc1ccc(S(=O)(=O)NCC(=O)N2CCS(=O)CC2)cc1

Molecular fingerprint: 1024 bits

00001100...10000000
Useful Links

- https://milvus.io
- https://github.com/milvus-io/milvus
- https://milvusio.slack.com
- https://twitter.com/milvusio
- https://zhuanlan.zhihu.com/ai-search

Performance benchmark:
https://milvus.io/docs/benchmarks_aws

Live demo:
https://milvus.io/scenarios

- Content-based image retrieval system (以图搜图)
- Q&A chatbot powered by NLP (智能客服机器人)
- Molecular analysis (化合物分析)
Thanks!