DATA / LARGE LANGUAGE MODELS

Vector Databases: Where Geometry Meets Machine Learning

A professor of semantic data processing channels Harry Potter to explain the role of advanced vector databases for better AI/ML performance.

Nov 21st, 2023 9:55am by Ivan Yamshchikov

Image via Unsplash.

In the past five years, the landscape of voice assistants and chatbots has undergone a remarkable transformation. Early on, people who interacted with chatbots likely noticed great fluency but an unexpected forgetfulness, akin to conversing with the iconic Pixar character Dory. In the last year, a pivotal shift occurred with the introduction of Large Language Models (LLMs). Modern top-

vector databases.

This article aims to help enterprise decision-makers refine their strategy for the adoption of AI and machine learning technologies, by helping them understand the role of advanced vector databases for better performance and scalability.

Vector Embeddings

So what are vector embeddings? Embeddings are words represented as vectors. It's a sequence of numbers that encodes information. Let's revisit Harry Potter to illustrate.

In the wizarding world of Harry Potter, there is the enchanted Weasley Clock. It tracks the whereabouts of each Weasley family member through the use of individual clock hands. Each hand is connected with one family member, while the direction encodes various states, such as "home," "in transit", or "mortal peril." At first, one might think wizards have a poor understanding of geometry; otherwise, why would distant locations be right next to each other on the dial? However, if you think about it, you could see how it makes sense: As Mr. Weasley travels home from work at the ministry, the clock-hand steadily moves through the different sectors ("ministry — in transit — home") instead of jumping back and forth across the dial.

A data scientist would call these movements of a hand on a dial "vector representations" for different concepts: Mr. Weasley is at the ministry, Mr. Weasley is in transit, Mr. Weasley is home. The direction of a clock hand corresponds with how the state that it encodes is related to other states. This is how vector representations capture semantic relationships. Think of vectors as "clock hands" with varying lengths; and the dial is not a two-dimensional surface, but a multidimensional space that is called the space of representations.

In math, the notion of proximity, or closeness, is vital for understanding geometry. We are all used to our old and comfortable three-dimensional space; and even in such an appallingly low number of dimensions, we try our best to use the notion of proximity to make our lives a bit simpler. For example, if you have a wardrobe, you might use one section for your clothes while leaving the other for your partner's clothes. You can further split your section into more segments, storing underwear in one place and T-shirts in the other. This stands to reason: once you

What is important is that you can use geometry to encode any property of your clothes. For example, if you are a bit too fond of Marie Kondo you could sort your dresses in such a way that similar colors hang closer to each other, so you can easily navigate through the color spectrum of your options. The distance between two dresses in your wardrobe encodes how similar the colors of these dresses are.

TRENDING STORIES

1. How to Set up and Run a Local LLM with Ollama and Llama 2
2. Stop Treating Your LLM Like a Database
3. How To Create Software Diagrams With ChatGPT and Claude
4. Supercharge Your RAG App With Agentic Hybrid Search
5. The Building Blocks of LLMs: Vectors, Tokens and Embeddings

Let's imagine that instead of sorting your wardrobe, you need to store something that is a notch bigger and a bit more complicated. Say, the whole knowledge that all of humanity managed to gather for the last millennia. Our three-dimensional wardrobe could not handle the information. The good news is that once you grasp the concept of distance, you do not have to stay in the three-dimensional space. You can have as many dimensions as you like!

Back to the Weasley Clock: Its dial was only encoding the whereabouts of the family. Let's say we want to manufacture a new version of the Clock that tracks the whereabouts as well as the mood of every family member. Instead of a two-dimensional dial, we could construct a sphere: now every hand has a projection on a traditional dial that encodes the location, while the vertical direction of a hand tells me whether the corresponding family member is happy, sad, angry or hesitant. There can be dozens, hundreds or even thousands of dimensions. What remains constant is the basic idea that similar things are closer to each other, while different things are further apart. This basic principle is used in a vector abase.

databases and language models go together like peanut butter and jelly. First, the idea to encode semantic similarity through the distance of embeddings in the representation space is one of the fundamental ideas in natural language processing. Thus, a vector database is a “native” way to encode information for your language model. You can encode any text into a vector in your representation space and this opens up a whole new world of opportunities.

Vector databases store embeddings of words or phrases, enabling LLMs to swiftly fetch contextually relevant information. When LLMs encounter a term, they can retrieve similar embeddings from the database, maintaining context and coherence.

For example, LLMs struggle with long passages, but vector databases allow them to access prior information. Retrieving embeddings of earlier sections ensures continuity and relevance in longer text generation. For some applications, you need your model to understand certain names. For example, you want your home assistant to know the preferences of your family members or you want a model that assists your legal team to recognize your company and its subsidiaries in legal documents.

Vector databases house embeddings for named entities, enhancing the ability of LLMs to recognize and utilize nouns in text accurately. Organizations can craft custom vector databases for specific domains. By training embeddings on domain-specific texts, LLMs can generate contextually relevant content tailored to the corresponding industry.

Vector databases can scale to accommodate vast amounts of embeddings, enabling LLMs to manage extensive datasets efficiently. Scalability is vital for chatbots, content generation, and question-answering systems. Finally, LLMs can operate in multiple languages; and so can vector databases. Storing embeddings for various languages facilitates seamless transitions between languages while maintaining a cross-lingual context.

Human Evaluation

Implementing an LLM and a vector database in production is a huge step forward, but it’s important to conduct safety checks to ensure that the solutions are safe, responsible and deliver actual business value.

databases, making them more applicable and useful across a range of business scenarios:

1. Contextual Relevance: AI might categorize terms based purely on the similarities encoded in your vector database. Human evaluators can ensure that the categorizations also make sense in the broader context in which the terms are usually used, adding a layer of real-world relevance.
2. Ethical and Cultural Nuances: Certain terms or phrases may be culturally sensitive or could have ethical implications. Human evaluation ensures that vector databases are sensitive to these issues, filtering out or re-categorizing potentially problematic content.
3. Industry-Specific Jargon: In specialized fields like law, healthcare or engineering, certain terms have specific meanings that general AI models may not grasp. Human experts in these domains can ensure that such terms are accurately represented in the vector database.
4. Ambiguity Resolution: Language is often ambiguous, and words can have multiple meanings based on the context. Human evaluation can help distinguish between these different meanings, ensuring that the vector database handles ambiguity more effectively.

Given that language, culture and context is always evolving, and new slang or terminology can quickly become relevant, periodic human evaluation ensures that the vector database stays current and relevant.

By efficiently storing and retrieving vector representations of words and phrases, these databases enhance an LLM's ability to maintain context, offer relevant responses, and manage extensive data. As organizations strive for higher efficiency and personalized services, investing in LLMs that utilize advanced vector databases is not just smart — it's strategic. Understanding this technology is crucial for business leaders aiming to maximize the benefits of AI in their enterprise. Now is the time to act, aligning AI adoption strategies with solutions that harness the full potential of vector databases.

TNS

Dr. Ivan Yamshchikov is a professor of semantic data processing and cognitive computing at the Center for AI and Robotics, Technical University of Applied Sciences Würzburg-Schweinfurt. He is the head of ecosystem strategy at Toloka AI. His research interests include...