

← Atrás

Compartir:



## Cuantización vectorial

**Cuantificación vectorial:** Una técnica para la compresión de datos y la búsqueda eficiente de similitudes en el aprendizaje automático.




La cuantificación vectorial (VQ) es un método utilizado en el aprendizaje automático para la compresión de datos y la búsqueda eficiente de similitudes. Implica la conversión de datos de alta dimensión en representaciones de menor dimensión, lo que puede reducir significativamente la sobrecarga computacional y mejorar la velocidad de procesamiento. VQ se ha aplicado en varias formas, como la cuantización ternaria, la cuantización de bits bajos y la cuantización binaria, cada una con sus ventajas y desafíos únicos.

El objetivo principal de VQ es minimizar el error de cuantificación, que es la diferencia entre los datos originales y su representación comprimida. Investigaciones recientes han demostrado que los errores de cuantificación en la norma (magnitud) de los vectores de datos tienen un mayor impacto en el rendimiento de la búsqueda de similitudes que los errores en la dirección. Este conocimiento ha llevado al desarrollo de la cuantificación explícita de normas (NEQ), un paradigma que mejora las técnicas de VQ existentes para la búsqueda máxima de productos internos (MIPS). NEQ cuantifica explícitamente las normas de los elementos de datos para reducir los errores en la norma, lo cual es crucial para MIPS. En el caso de los vectores de dirección, NEQ puede reutilizar las técnicas VQ existentes sin modificaciones.

Artículos recientes sobre cuantización vectorial han explorado varios aspectos de la técnica. Por ejemplo, el artículo "Ternary Quantization: A Survey" de Dan Liu y Xue Liu proporciona una visión general de los métodos de cuantificación ternaria y su evolución. Otro artículo, "Word2Bits - Quantized Word Vectors" de Maximilian Lam, demuestra que los vectores de palabras cuantificados de alta calidad se pueden aprender usando solo 1-2 bits por parámetro, lo que resulta en un ahorro significativo de memoria y almacenamiento.

Soluciones ▾

### Industrias

-  Agricultura →
-  Procesamiento de audio →
-  Autónomo y Robótica →

### Casos de estudio

#### Empresas



Biomédico

Chatea con Rayos X.  
Adiós, SQL



Multimedia

Reduzco el tiempo de

#### IA generativa



Multimedia

Consultas 100 veces más  
rápidas



GenAI

Base de datos sin conexión

#### Startups





-50% menos de c  
GPU y 3 veces má



5 veces más rápid


 **Biomédica y Cuidado de la Salud** →

 **IA generativa y RAG** →

 **Multimedia** →


 **Seguridad y protección** →

Reduzca el tiempo de  
preparación de datos  
hasta en un 80 %


 **Biomédico**  
+18% más de precisión  
RAG


**FORTUNE 500** **Tecnología Médica**  
Búsqueda rápida de IA en  
40M+ documentos

Base de datos sin servidor  
para el asistente de código

 **GenAI**  
RAG para asistente de IA  
multimodal

5 veces más rápido  
5 veces menos recu

 **Uberwa**  
Preparación de da  
5 veces más rápida

 **King miao**  
+19,5% en precisio  
modelo

Compañía ▾

## Compañía

 **Acerca de** →

Conozca nuestra empresa, sus miembros y  
nuestra visión

 **Contáctenos** →

Nuestro equipo responda a todas sus  
preguntas


 **Carreras** →


Construye cosas geniales que importen.  
Desde cualquier lugar


Docs


Recursos ▾


## Recursos


 **Blog** →  
Artículos de opinión y artículos de  
tecnología


 **Tutoriales** →  
Más información sobre cómo usar la  
pila de ActiveLoop


 **Noticia** →  
Seguimiento de los principales hitos  
de la empresa

 **Documento académico de  
Deep Lake** →  
Lea el artículo académico publicado  
en CIDR 2023


 **LangChain (Cadena de  
idiomas)** →  
Procedimientos de LangChain con  
Deep Lake Vector DB


 **Glosario** →  
Explicación de los términos  
principales de 1000 ML


 **Notas** →  
¿Ves qué hay de nuevo?

 **Libro blanco de Deep Lake** →  
Vea cómo su empresa puede  
beneficiarse de Deep Lake

## Cursos gratuitos de GenAI

 **LangChain y DBs vectoriales  
producción**  
Lleve las aplicaciones de IA a la  
producción

 **Entrena y afina los LLM**  
LLMs desde cero con todos los  
métodos

 **Cree aplicaciones RAG con  
LlamaIndex y LangChain**  
Estrategias avanzadas de recup  
de datos multimodales

 **Despliegue aplicaciones RAG**  
Despliegue aplicaciones RAG en  
la nube



código en el libro de códigos de VQ.

### **¿Cuáles son algunas de las aplicaciones de la cuantización vectorial?**

La cuantificación vectorial tiene varias aplicaciones prácticas, como el procesamiento de textos, la clasificación de imágenes y la estimación de medias distribuidas. En el procesamiento de textos, los vectores de palabras cuantificados se pueden utilizar para representar palabras en tareas de procesamiento del lenguaje natural. En la clasificación de imágenes, VQ se puede aplicar al modelo de bolsa de características. En la estimación de medias distribuidas, se pueden utilizar cuantificadores eficientes en varios problemas de optimización.

### **¿Cómo mejora la cuantificación vectorial el rendimiento del aprendizaje automático?**

La cuantificación vectorial mejora el rendimiento del aprendizaje automático al reducir la dimensionalidad de los datos de entrada, lo que a su vez reduce la sobrecarga computacional y mejora la velocidad de procesamiento. Al minimizar los errores de cuantificación y adaptarse a las necesidades específicas de diversas aplicaciones, VQ puede mejorar significativamente el rendimiento de los modelos de aprendizaje automático y permitir su implementación en dispositivos con recursos limitados.

### **¿Cuáles son algunos de los avances recientes en la investigación de la cuantificación de vectores?**

Los avances recientes en la investigación de la cuantificación vectorial incluyen el desarrollo de la cuantificación explícita de normas (NEQ), un paradigma que mejora las técnicas de VQ existentes para la búsqueda máxima de productos internos (MIPS). NEQ cuantifica explícitamente las normas de los elementos de datos para reducir los errores en la norma, lo cual es crucial para MIPS. En el caso de los vectores de dirección, NEQ puede reutilizar las técnicas VQ existentes sin modificaciones. Otros avances incluyen la exploración de métodos de cuantificación ternaria y el desarrollo de vectores de palabras cuantificados de alta calidad utilizando solo 1-2 bits por parámetro.

### **¿Puede proporcionar un estudio de caso de una empresa que utilice la cuantificación vectorial?**

Un estudio de caso de una empresa que muestra el uso de la cuantificación vectorial es Word2Vec de Google, que emplea técnicas de cuantificación para crear incrustaciones de palabras compactas y eficientes. Estas incrustaciones se utilizan en diversas tareas de procesamiento del lenguaje natural, como el análisis de sentimientos, la traducción automática y la recuperación de información.

## **Cuantización vectorial Lecturas adicionales**

1. Cuantización ternaria: una encuesta <http://arxiv.org/abs/2303.01505v1> Dan Liu, Xue Liu
2. Word2Bits - Vectores de palabras cuantificados <http://arxiv.org/abs/1803.05651v3> Maximilian Lam
3. Una limitación fundamental en la dimensión máxima del parámetro para una estimación precisa con datos cuantificados <http://arxiv.org/abs/1605.07679v1> Jiangfan Zhang, Rick S. Blum, Lance Kaplan, Xuanxuan Lu
4.  $\mathbb{U}_h$  invariante Cuantificación de órbitas conjuntas y haces vectoriales sobre ellas <http://arxiv.org/abs/math/0006217v1> J. Donin
5. Árboles de proyección aleatoria para la cuantización vectorial <http://arxiv.org/abs/0805.1390v1> Sanjoy Dasgupta, Yoav Freund
6. Cuantificación explícita de normas: mejora de la cuantificación vectorial para la máxima búsqueda de productos internos <http://arxiv.org/abs/1911.04654v2> Xinyan Dai, Xiao Yan, Kelvin K. W. Ng, Jie Liu, James Cheng
7. Cuantización vectorial minimizando la divergencia de Kullback-Leibler <http://arxiv.org/abs/1501.07681v1> Lan Yang, Jingbin Wang, Yujin Tu, Prarthana Mahapatra, Nelson Cardoso
8. Diseño de cuantificadores vectoriales optimizados para canales para mediciones de detección comprimida <http://arxiv.org/abs/1404.7648v1> Amirpasha Shirazinia, Saikat Chatterjee, Mikael Skoglund
9. Ajuste Tautológico del Mapa de Cuantización de Kostant-Souriau con Estructuras Geométricas Diferenciales <http://arxiv.org/abs/2003.11480v1> Tom McClain
10. RATQ: Un cuantificador universal de longitud fija para la optimización estocástica <http://arxiv.org/abs/1908.08200v3> Prathamesh Mayekar, Himanshu Tyagi

## Explore más términos y conceptos de Machine Learning

### → Indexación vec...

La indexación vectorial es una técnica utilizada para buscar y recuperar información de grande...

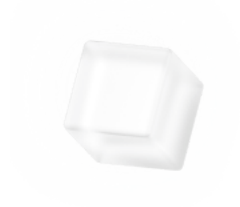
### → Modelo de esp...

El Modelo de Espacio Vectorial (VSM) es una poderosa técnica utilizada en el...

Subscribe to our  
newsletter for more  
articles like this

Your Email

Subscribe



Deep Lake. Database for AI.

#### Solutions

[Agriculture](#)

[Audio Processing](#)

[Autonomous Vehicles &  
Robotics](#)

[Biomedical & Healthcare](#)

[Multimedia](#)

[Safety & Security](#)

#### Company

[About](#)

[Contact Us](#)

[Careers](#)

[Privacy Policy](#)

[Do Not Sell](#)

[Terms &  
Conditions](#)

#### Resources

[Blog](#)

[Documentation](#)

[Deep Lake Whitepaper](#)

[Deep Lake Academic  
Paper](#)



Featured by