An embedding is a data representation that uses low-dimensional numerical vectors to capture the semantic meaning and relationships of non-numerical data like words, images, or graphs, making them understandable for machine learning models. These dense vectors are created through a process of mapping complex, real-world data into a vector space, where the distance between embeddings reflects the similarity of the original data points. For instance, in natural language processing (NLP), word embeddings allow for mathematical operations that represent relationships, like king - man + woman ≈ queen, to uncover deeper meaning and enable tasks such as sentiment analysis and machine translation.

What is the process?

- Data Transformation: Non-numerical data (text, images, graphs) is converted into numerical vectors, where each dimension represents a specific feature of the data.
- Vector Space: These vectors are placed into an n-dimensional space, creating a dense numerical representation.
- Semantic Relationships: The embedding process captures the nuances and context of the original data. For example, words with similar meanings will have embeddings that are closer together in the vector space.
- Machine Learning Use: These vectors can then be used by machine learning models for tasks like classification, search, recommendation systems, and more.

Key Characteristics and Benefits

- Information Density: Embeddings provide an information-dense representation, meaning they pack a lot of meaning into fewer dimensions than traditional methods.
- Semantic Similarity: The distance between two embeddings in the vector space correlates with the semantic similarity of the original inputs.
- Machine-Readable Format: Embeddings make complex data understandable to computers and algorithms, which only process numerical inputs.
- Capture of Relationships: Embeddings can capture intricate relationships within the data, such as co-occurrence patterns in text or structural relationships in graphs.

Examples of Embeddings

- Word Embeddings: Represent words as vectors, enabling tasks like sentiment analysis and machine translation.
- Image Embeddings: Map images into a vector space to group similar images and enable visual content retrieval.
- Graph Embeddings: Convert nodes and edges in graphs into numerical vectors to understand relationships and predict links.