Vector embeddings in non-business information systems

Claude Charest
Athabasca University
COMP 601: Survey of Computing and Information Systems
Dr. Richard Huntrods
September 29, 2024

This report examines the expanding applications of vector embeddings beyond traditional business environments, focusing on their increasing relevance in various scientific and social domains. Vector embeddings, which represent discrete data in continuous vector spaces, have gained prominence alongside recent advancements in machine learning and artificial intelligence.

Our research is based on a review of academic literature and current industry practices. We analysed scholarly articles, conference papers, and reputable online resources to understand the evolution and applications of vector embeddings in non-business contexts. This approach allowed us to explore how the growing adoption of vector embeddings is prompting a reevaluation of conventional data methodologies across different fields.

The report traces the development of vector embeddings from their early concepts to current applications, investigating their use in areas such as natural language processing, computer vision, and bioinformatics. We explore various embedding techniques and their implementation, considering both their potential benefits and challenges.

Our findings indicate that vector embeddings are expanding the capabilities of machine learning in non-business domains, enabling more sophisticated analysis and decision-making in complex scientific and social contexts. The technology shows promise in enhancing analysis across diverse fields, from computational linguistics to molecular biology.

We anticipate further innovations in areas like personalised medicine, environmental conservation, and space exploration as embedding techniques continue to evolve. To address current limitations and ethical concerns, we recommend further research into developing more interpretable and bias-free embedding models.

This report highlights the transformative potential of vector embeddings in advancing our understanding and interaction with complex data across a range of scientific and societal challenges, driven by ongoing progress in machine learning and data representation techniques.

Introduction

Vector embeddings are a mathematical representation technique used in data processing and machine learning. This approach involves mapping data points to vectors in a multi-dimensional space, where each dimension corresponds to a feature or attribute of the data. By representing complex information as coordinates, vector embeddings allow for numerical operations on diverse types of data, such as text, images, or user behaviours. This representation can potentially reveal patterns and relationships within the data that might not be immediately apparent in its original form.

The core principle of vector embeddings is that similar data points should have similar representations in the embedding space, while dissimilar points should be further apart. This property facilitates various applications, from natural language processing and computer vision to recommendation systems and bioinformatics.

A key strength of vector embeddings lies in their ability to represent complex, high-dimensional data in a more compact and analyzable form (Bengio et al., 2014). This representation facilitates efficient similarity comparisons, enabling tasks like semantic search, where results are based on meaning rather than exact keyword matches (Zhang et al., 2024). Such capabilities make vector embeddings

particularly valuable for a wide range of applications, including data classification, anomaly detection, and advanced information retrieval systems.

As machine learning techniques continue to advance, the prevalence of vector data is increasing (*Vector Database Market Size, Share & Trends Report, 2030*, n.d.). This trend underscores the importance of exploring vector embeddings in depth. This subject was chosen in the hope to provide a comprehensive overview of vector embeddings, their applications, and their role in shaping modern information systems. By examining these topics, we seek to contribute to the ongoing dialogue about effective data representation and processing in the evolving landscape of machine learning and artificial intelligence.

Background

Vector embeddings, fundamental to modern information retrieval and natural language processing, have a rich history dating back to the 1970s. The concept originated with the Vector Space Model introduced by Gerard Salton and his team at Cornell University (Salton et al., 1975). This model laid the groundwork for representing words and documents in a way that captures semantic relationships.

A significant breakthrough came in 1990 with Latent Semantic Analysis (LSA) (Deerwester et al., 1990). LSA used singular value decomposition to create dense vector representations, revealing latent semantic structures in text data. This paved the way for more sophisticated embedding techniques. The field advanced significantly in 2003 when Yoshua Bengio and colleagues introduced neural probabilistic language models (Bengio et al., 2003). Their work demonstrated the potential of neural networks in creating meaningful word embeddings, setting the stage for future innovations.

A revolutionary moment arrived in 2013 with Word2Vec, introduced by Tomas Mikolov and his team at Google (Mikolov et al., 2013). Word2Vec utilised shallow neural networks to create dense vector representations of words based on their context in large text corpora. This method produced remarkably effective embeddings that captured both semantic and syntactic relationships between words. The application of embeddings expanded beyond text data. In 2014, Perozzi et al. introduced DeepWalk, applying embedding techniques to graph-structured data (Perozzi et al., 2014). This development led to further innovations like node2vec, broadening the applicability of embedding methods (Grover & Leskovec, 2016). The versatility of embedding techniques across different data types was further demonstrated in 2016 when Asgari and Mofrad introduced ProtVec for biological sequences (Asgari & Mofrad, 2015).

This report will explore current state-of-the-art embedding techniques, their applications across various domains, and the challenges and opportunities they present for future research and development.

Discussion

Applications of vector embeddings

Vector embeddings have found widespread applications across various non-business domains, revolutionising how we process and analyse complex data. In natural language processing (NLP), word embeddings from attention transformers have become a fundamental tool (Devlin et al., 2019). These embeddings capture semantic relationships between words, enabling machines to understand context and nuance in human language (Mikolov et al., 2013). This has led to significant improvements in machine translation, sentiment analysis, and text classification tasks (Pennington et al., 2014).

In the field of bioinformatics, vector embeddings have been instrumental in advancing our understanding of biological sequences (Asgari & Mofrad, 2015). Techniques like BioVectors and ProtVec represent DNA, RNA, and protein sequences as dense vectors, capturing functional and structural similarities. This has accelerated research in protein function prediction, drug discovery, and genetic analysis. For instance, embeddings can help identify proteins with similar functions or predict how genetic mutations might affect protein behaviour (Rives et al., 2020).

Computer vision has also benefited greatly from vector embeddings. Image embeddings, often generated by convolutional neural networks, represent visual content in a high-dimensional space. This enables sophisticated image recognition, object detection, and even image generation tasks (Radford et al., 2021). A notable advancement in this field is CLIP (Contrastive Language-Image Pre-training), which creates a common embedding space for both images and text. CLIP's ability to align visual and textual representations has revolutionised various computer vision tasks, enabling zero-shot learning and more flexible image-text interactions (Goh et al., 2021). This approach has significantly enhanced capabilities in image classification, semantic image search, and cross-modal understanding, demonstrating the power of using a common embedding space in computer vision applications (Shen et al., 2021).

Recommendation systems across various domains have been enhanced by vector embeddings. Whether recommending scientific articles to researchers (Cohan et al., 2020), music tracks to listeners, or products to consumers, embeddings help capture complex preferences and similarities that go beyond simple categorical matching.

In the realm of scientific research, embeddings have been used to represent complex phenomena in physics, chemistry, and materials science. For example, materials scientists use embeddings to represent atomic structures, accelerating the discovery of new materials with desired properties (Xie & Grossman, 2018).

These diverse applications demonstrate the versatility and power of vector embeddings in non-business contexts, enabling more sophisticated analysis and decision-making across a wide range of scientific and technological domains.

Embedding techniques

Creating effective vector embeddings involves various techniques, often tailored to the specific data type and application. The key to success lies in choosing appropriate training objectives that capture the essence of the data and its relationships (Bengio et al., 2014).

Transfer learning models like BERT use a masked language modelling objective during pre-training. The model learns to predict masked words in a sentence, which helps it understand context and relationships between words (Devlin et al., 2019). This pre-training objective creates rich, contextual embeddings that can be fine-tuned for specific tasks, adapting the learned representations to various downstream applications (Ethayarajh, 2019).

Deep learning models, particularly autoencoders, create embeddings with a reconstruction objective. The model learns to compress input data into a lower-dimensional representation and then reconstruct it. This training objective is effective for creating embeddings of complex, high-dimensional data, as it forces the model to capture the most salient features for accurate reconstruction (Hinton & Salakhutdinov, 2006).

Contrastive learning objectives, as seen in approaches like SimCLR (Chen et al., 2020) for visual representations and CLIP (Radford et al., 2021) for multimodal embeddings, have emerged as powerful tools. These methods train embeddings to be similar for semantically related inputs and dissimilar for unrelated ones, proving particularly effective in self-supervised learning scenarios.

By carefully selecting and optimising these training objectives, we can create embeddings that effectively capture relevant features and relationships in the data, making them useful for a wide range of downstream tasks (Bengio et al., 2014).

Special requirements

Vector embeddings in non-business domains present unique challenges and requirements, necessitating special considerations in their development and operation. The development of effective vector embeddings in specialised domains demands extensive subject-matter expertise. For instance, in bioinformatics, a deep understanding of protein structures and functions is crucial for creating meaningful protein embeddings (Asgari & Mofrad, 2015). The necessity for domain-specific knowledge underscores the importance of interdisciplinary collaboration between data scientists and domain experts.

Many scientific and specialised domains face issues of data sparsity and high dimensionality (Altman & Krzywinski, 2018). This necessitates the development of techniques that can learn from limited datasets or leverage transfer learning. Additionally, vector data has different information system requirements compared to traditional relational data, necessitating specialised database systems and query processing techniques (Zhang et al., 2024). To address these challenges, vector databases must offer specific capabilities such as scalability to handle high-dimensional data, efficient indexing structures for fast similarity search, support for sparse vector representations, and the ability to handle continuous updates and insertions (Han et al., 2023).

In fields such as healthcare or legal applications, understanding the reasoning behind predictions is crucial (Tjoa & Guan, 2021). This requirement drives the need for interpretable embedding techniques

and explainable AI methods. Developers must balance the trade-off between model performance and interpretability, often necessitating novel approaches to maintain both.

Ethical considerations are paramount, especially when dealing with sensitive data in healthcare or criminal justice . Ensuring embeddings don't encode or amplify biases in training data is critical. This requires ongoing monitoring, bias detection techniques, and potentially the development of fairness-aware embedding methods (Mehrabi et al., 2022).

The application of vector embeddings in specialised domains often comes with unique requirements. Depending on the field and use case, these requirements might involve rapid processing speeds, high accuracy, or both. Meeting these demands requires careful consideration of the specific needs of each application and the implementation of tailored strategies. When these challenges are successfully addressed, vector embeddings have the potential to significantly enhance capabilities in various non-business domains.

Comparison with business information systems

Vector embeddings in business and non-business information systems share fundamental principles but differ significantly in their applications, scale, and challenges. Both domains use embeddings to represent complex data efficiently, enabling similarity calculations and various machine learning tasks.

In business contexts, embeddings often support customer segmentation, product recommendations, and fraud detection, dealing with large-scale structured data from customer interactions and transactions (Grbovic & Cheng, 2018). These systems prioritise real-time performance and scalability to handle millions of users or products simultaneously.

Non-business applications, such as scientific research, typically involve more specialised and diverse data types, like genetic sequences or climate models. Here, the focus is on capturing intricate domain-specific relationships and supporting exploratory analysis rather than immediate commercial outcomes (Asgari & Mofrad, 2015).

The privacy and ethical considerations vary considerably between these domains. Business systems must rigorously protect customer data and avoid discriminatory biases, while non-business applications, especially in healthcare or social sciences, may have additional requirements related to research ethics or public policy implications (Mehrabi et al., 2022).

Interpretability is crucial in both areas but for different reasons. In business, it aids decision-making and ensures regulatory compliance, while in scientific applications, it's essential for advancing understanding and validating research findings (Doshi-Velez & Kim, 2017).

The development cycle also differs significantly. Business embeddings might update frequently based on rapidly changing market conditions, while scientific embeddings might evolve more slowly, aligning with the deliberate pace of research and peer review processes.

Despite these differences, there's increasing cross-pollination of techniques between business and non-business domains, driving innovation in both areas. This exchange of ideas and methods continues to push the boundaries of what's possible with vector embeddings across various fields.

Conclusion

Vector embeddings have emerged as a powerful tool in non-business information systems, revolutionising data analysis across scientific and social domains. From natural language processing to bioinformatics and computer vision, these techniques enable more nuanced understanding of complex data. While challenges remain, particularly in interpretability and ethical considerations, the potential for further innovation is vast. As embedding techniques continue to evolve, we anticipate transformative applications in fields like personalised medicine, environmental conservation, and space exploration. Future research should focus on developing more interpretable and bias-free embedding models, ensuring that this technology can be leveraged responsibly to address a wide range of scientific and societal challenges.

References

Altman, N., & Krzywinski, M. (2018). The curse(s) of dimensionality. *Nature Methods*, *15*(6), 399–400. https://doi.org/10.1038/s41592-018-0019-x

Asgari, E., & Mofrad, M. R. K. (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLOS ONE*, *10*(11), e0141287. https://doi.org/10.1371/journal.pone.0141287

Bengio, Y., Courville, A., & Vincent, P. (2014). *Representation Learning: A Review and New Perspectives* (arXiv:1206.5538). arXiv. http://arxiv.org/abs/1206.5538

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). *A Neural Probabilistic Language Model*.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A Simple Framework for Contrastive Learning of Visual Representations*.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). *SPECTER: Document-level Representation Learning using Citation-informed Transformers* (arXiv:2004.07180). arXiv. http://arxiv.org/abs/2004.07180

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning*

(arXiv:1702.08608). arXiv. http://arxiv.org/abs/1702.08608

Ethayarajh, K. (2019). *How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings* (arXiv:1909.00512). arXiv. http://arxiv.org/abs/1909.00512

Goh, G., † N. C., † C. V., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal Neurons in Artificial Neural Networks. *Distill*, *6*(3), e30. https://doi.org/10.23915/distill.00030

Grbovic, M., & Cheng, H. (2018). Real-time Personalization using Embeddings for Search Ranking at Airbnb. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 311–320. https://doi.org/10.1145/3219819.3219885

Grover, A., & Leskovec, J. (2016). *node2vec: Scalable Feature Learning for Networks* (arXiv:1607.00653). arXiv. http://arxiv.org/abs/1607.00653

Han, Y., Liu, C., & Wang, P. (2023). *A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge* (arXiv:2310.11703). arXiv. http://arxiv.org/abs/2310.11703

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). *A Survey on Bias and Fairness in Machine Learning* (arXiv:1908.09635). arXiv. http://arxiv.org/abs/1908.09635

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online Learning of Social Representations. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 701–710. https://doi.org/10.1145/2623330.2623732

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual*

Models From Natural Language Supervision (arXiv:2103.00020). arXiv. http://arxiv.org/abs/2103.00020

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2020). *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences* (p. 622803). bioRxiv. https://doi.org/10.1101/622803

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., & Keutzer, K. (2021). *How Much Can CLIP Benefit Vision-and-Language Tasks?* (arXiv:2107.06383). arXiv. http://arxiv.org/abs/2107.06383

Tjoa, E., & Guan, C. (2021). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(11), 4793–4813. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2020.3027314

Vector Database Market Size, Share & Trends Report, 2030. (n.d.). Retrieved 29 September 2024, from https://www.grandviewresearch.com/industry-analysis/vector-database-market-report

Xie, T., & Grossman, J. C. (2018). Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, *120*(14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301

Zhang, Y., Liu, S., & Wang, J. (2024). Are There Fundamental Limitations in Supporting Vector Data Management in Relational Databases? A Case Study of PostgreSQL. *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 3640–3653. https://doi.org/10.1109/ICDE60146.2024.00280