Orthogonal Adaptative Networks for Structured Word Embedding Compression

Amruddin Nekpai*

Department of Information Technologies, Friendship University of Russia, Moscow, Russia

*Corresponding author: Amruddin Nekpai, Amruddin.salaar1212@gmail.com

Abstract

This comparative study represents the Orthogonal Adaptive Neural Network (OA-NN), a novel dimensionality reduction method that synergizes orthogonal linear projections with adaptive nonlinear transformations to preserve both local and global structures in high-dimensional data. The dataset which we have used in this paper, Glove (Global Vectors for Word Representation), is an embedded dataset with 400000 embedded word to vectors. We curated a subset of 1000 word to vectors with the mentioned dataset from 400000 (word to vectors) to reduce the time duration and balance between comprehensiveness and computational efficiency. The proposed method is rigorously compared against three established techniques: principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP). we have used quantitative metrics (MSE, Trustworthiness, Continuity) and qualitative visual analysis for testing OA-NN method and and also for comparison,the OA-NN showed a (2-5) better continuity rate, and an equal continuity rate compared to UMAP while applying Glove 300D dataset.

Keywords

Data Science, Machine Learning, Dimensionality Reduction, Data Analysis, PCA, t-SNE, UMAP

1. Introduction

In this paper we have conducted a comparative study on traditional methods like PCA, t-SNE, and UMAP and introduced a novel method named Orthogonal Adaptive Neural Network (OA-NN). Traditional methods still have widely usage and are crucial techniques of data science, machine learning, data visualization, and data analysis for reducing dimensionality, visualization, and preprocessing the data. Traditional methods which I have named here for reducing high dimensionality like principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold estimation and projection, have been effective with both linear data and nonlinear data, they are very famous and still we use these methods. For example, if we want to reduce or visualize a high dimensional dataset containing linear data in it, for achieving high efficiency and MSE the best method to work with it , is PCA [1],which reduces the high dimensional data to a low dimensional data, later if needed we can reconstruct the reduced data back to original data but when it came to work with nonlinear data PCA is not effective capturing nonlinear relationships, but traditional methods like t-SNE and UMAP are nonlinear algorithms for dimensionality reduction and visualization while t-SNE effectively preserve local structured data [2] but struggles preserving global structured data, UMAP in other side effective for preserving for both global structured data [3] local structured data. Each of these traditional techniques have their strength and weaknesses and the usage of every method depends on type of datasets and applications.in this study we present a novel method named Orthogonal Adaptive Inverse Neural Network (OA-NN),this method combines orthogonal linear projections with adaptive nonlinear transformations to address challenges which traditional methods have, and enforces orthogonality in its projection matrices to stabilize training and ensure invertibility, while adaptively gating linear and nonlinear pathways to capture complex data relationships. effective with both linear and nonlinear dataset relationships and can preserve local and global structured data effectively.in here the dataset which we have tested and for every method is a Glove (Global Vectors for Word Representation) [4], 50D-300 D embedded text to vectors dataset. The dataset contains 400000 embedded word to vectors data and to apply the method on that kind of big dataset it takes a lot of time and computationally expensive because of that we curate a 2000 word to vectors embedded dataset with 100 semantically diverse indices to represent key lexical categories ,from 400000 text vectors embedded dataset to reduce computational load, complexity and reduce the time [4].the method is compared against PCA, t-SNE and UMAP using Mean Squared Error (MSE) for reconstruction fidelity, Trustworthiness for local neighborhood preservation, Continuity for global structure consistency.A direct comparison of the OA-NN method against established dimensionality reduction techniques (PCA, t-SNE, and UMAP) is presented in Table 1.

Table 1. Comparison of Dimensionality Reduction Methods.

Method	Type	Adaptiveness to Data Linearity	Key Strength	Key Weakness / Limitation	Performance on Glove 300D (MSE & Continuity)
OA-NN	Hybrid	Both Linear & Non-Linear	Balances local/global structure preservation	Novel method, less widely validated	MSE: 0.001798-0.0059; Continuity: 2-5% better than t-SNE
PCA	Linear	Only Linear	Efficient for linear global structures	Fails with nonlinear relationships	Not specified (implied inferior)
t-SNE	Non-Linear	Non-Linear Only	Excellent for visualization of local clusters	Not adaptive; limited to visualization	Lower continuity vs. OA-NN
UMAP	Non-Linear	Non-Linear Only	Preserves global structure better than		

2. Main Section

2.1 Traditional Dimensionality Reduction Methods

2.1.1 Principal Component Analysis (PCA)

PCA method is a well-known and famous linear method for dimensionality reduction. the method was first represented by K. Pearson in 1901, and it was the great and a valuable find till now, using linear algebra for solving and make a confusing and big data much easier, maintaining the basic and important information without noises. PCA has widely usage in face recognition, dimension reduction, data analysis and neuroscience, visualization, machine learning and…etc. [5]. in here we are focusing on dimensionality reduction and how the PCA method works to reduce dimensionality, for this purpose I am considering a high dimensional dataset X = [X_1, X_2…. Xn], where every column is a single observation like X_1 and X_2. before applying PCA method we must clean and make the data more understandable for applying PCA method, because high dimensional data has different scales of information in it, and we must make the data suitable and comparable for analysis, we can call this a preprocessing or standardization. standardization starts with finding of average values of the independent features and the standard deviations of the features [6,7] after the standardization process, we get mean value close to 0 and standard deviation, value close to 1. the formulas for Mean value μ, standard deviation σ of the features and The standardization Z, given by

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (1)$$

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \mu\right)^2} \quad (2)$$

$$Z = \frac{X_i - \mu}{\sigma} \quad (3)$$

Where X_i – is the value of the feature with i observation. After that the process continues through covariance matrix, to find the relationship between every feature and variable in a dataset. The covariance matrix mathematically given by

$$Cov(x) = \frac{1}{m-1}\sum (x_i - \mu)(x_i - \mu)^{\mathrm{T}} \quad (4)$$

Where X_i – is the feature value vector for observation i; $\mu - isthe$ vector of mean value for i observation and m is the is the total number observation. After that we perform the eigen decomposition on covariance matrix to find the eigenvalues and eigenvectors. Eigenvectors represent the principal component in our data and the eigenvalue indicates total amount of variance (95%-99%), which is explained by each principal component [8]. The eigenvector with the must eigenvalue captures the big part of our information and will be selected the principal component. While selecting the number of components or a subset principal component, we reduce the dimensionality for required dimensionality, for example, to 2 or 3 principal components (dimension). if needed we can reconstruct the reduced data back to original data [7,8].

2.1.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

This method is One of the well-known and widely used non-linear method for exploring and visualizing complex and high dimensional data into two or three dimensions. The t-SNE method was first published in 2008 by Dutch researcher Lawrence van der Maaten and neural network wizard Geoffrey Hinton [2, 9]. they have been applied t-SNE on a real-world dataset up to 30 million examples [10]. While applying the t-SNE algorithm on a dataset (high dimensional), the algorithm starts to measuring using a Gaussian distribution to find how two points are alike or close to each other (data point xi with data point X_j and data point x_j with data point x_k and ..etc.) in the dataset, they call this measurement pairwise similarities, then these similarities are converted to conditional probabilities, $P_{j\,|\,I,}$,which forms later the joint probabilities (indicating that two point are neighbor and have close similarities). In low dimensionality we also have similar process (pairwise similarities → conditional probabilities $q_{i|j}$ → joint probabilities → KL divergence minimization). of creating a similar set of joint probabilities using students' t-distribution, it then minimizes KL divergence between the high and low dimensional joint probabilities, it ensures the closeness of low dimensional data with the original data. the method is effective maintaining local structures (neighboring points) in the data and not effective maintaining the global structures, requires precise tuning of hyperparameters [11]. It means the points which are close to each other in high dimensional will remain closer in low dimensional space. The conditional probabilities, mathematically defines by

$$P_{j|i} = \frac{\dfrac{\exp\left(-\left\|x_i - x_j\right\|^2\right)}{2\sigma_i^2}}{\sum_{k\neq 1}\exp f_0\left(\dfrac{\exp\left(-\left\|x_i - x_k\right\|^2\right)}{2\sigma_i^2}\right)} \quad (5)$$

Where σ_i is the variance of the Gaussian distribution that is centered on datapoint X_i. in low dimensional we have same points Y_i and Y_j and conditional probability $Q_{i|j}$.the similarities in map point Y_j to map point Y_i by

$$Q_{j|i} = \frac{\dfrac{\exp\left(-\left\|y_i - y_j\right\|^2\right)}{2\sigma_i^2}}{\sum_{k\neq 1}\exp f_0\left(\dfrac{\exp\left(-\left\|y_i - y_k\right\|^2\right)}{2\sigma_i^2}\right)} \quad (6)$$

If the points y_i and y_j correctly model the similarities with high dimensional data points x_i and x_j ,it will be $Q_{i|j} = P_{i|j}$. The SNE tries the minimize the differences between these probabilities. The mismatch between conditional probabilities is quantified using Kullback-Leibler (KL) divergence, the less our cost function C is the less we have mismatch between conditional probabilities. The cost function is defined by

$$C = \sum_i KL(P \,//\, Q) = \sum_i \sum_j P_j N_i \log \frac{P_j N_i}{Q_j N_i} \quad (7)$$

2.1.3 UMAP (Uniform Manifold Approximation and Projection)

UMAP is the alternative method of t-SNE, the method was developed by Leland McInnes and John Healy in 2018. this algorithm uses topological principles and preserve both global and local data structures, have a faster performance time and works effectively with local and global structured data compare to t-SNE [12,13].in this method like also like t-SNE the method start with calculating pairwise distances $d(X_i, X_j)$ between data point X_i and X_j to identify the local neighborhood points using Euclidean distance and cosine distance. Mathematically Euclidean distance is given by

$$d(X_i, X_j) = \sqrt{\sum_{m=1}^{n}\left(X_{i,m} - X_{j,m}\right)^2} \quad (8)$$

where $X_{i,m}$ and $X_{j,m}$ are the m-th features of points X_i and X_j Cosine distance given by

$$d(X_i, X_j) = 1 - \frac{\left\|X_i - X_j\right\|}{\left\|X_i\right\| \vee \left\|X_j\right\|} \quad (9)$$

where $// X_i //$ is the magnitude of X_i .after sorting out the distances with all other points X_j in order, we select the top K with the smallest distances, these are the K-nearest neighbors of X_i , we store and use these indices and distances later. the method proceeds with defining local Radius σ_i , fuzzy simplicial sets $P_{j|i}$, symmetric probabilities $P_{j|i}$ and $P_{i|j}$, graph construction, Initialize Low-Dimensional Embedding, Defining Low-Dimensional Probabilities Q_{ij}, Cost Function C, Gradient update $Y_i^{(t+1)}$ and Final Low-Dimensional Embedding (2D or 3D).

$$\sum_{j=1}^{k} \exp(-\frac{d\left(X_i, X_j\right)-\rho_i}{\sigma_i}) = \log_2(k) \quad (10)$$

Where is the distance with the nearest neigbor of X_i and σ_i is the bandwidth parameter for X_i .

$$P_{j|i} = \exp(-\frac{d\left(X_i, X_j\right)-\rho_i}{\sigma_i}) \quad (11)$$

Where $P_{j|i}$ is the probability that X_j is a neighbor of X_i .

$$P_{ij} = P_{j|i} + P_{i|j} - P_{j|i} \quad (12)$$

Where P_{ij} is Symmetric probability representing the connection strength between X_i and X_j , given by

$$Q_{ij} = \left(1 + a\left|Y_i - Y_j\right|^{2b}\right)^{-1} \quad (13)$$

Where Y_i, Y_j is the Low-dimensional embeddings of X_i , X_j ; $\left|Y_i - Y_j\right|$ represents Euclidean distance between Y_i and Y_j ; a and b Hyperparameters controlling the shape of the kernel. Later in this method we have the cost function C , which minimize the entropy between low Q_{ij} and high P_{ij} dimensional probabilities. C is given by

$$C = \sum_{i \neq j}\left[P_{ij}\left(\log\left(P_{ij}/Q_{ij}\right)\right) + \left(1-P_{ij}\right)\log\left(1-P_{ij}/1-Q_{ij}\right)\right] \quad (14)$$

$$Y_i^{(t+1)} = Y_i^{(t)} - \left(\eta \partial C/\partial Y_i\right) \quad (15)$$

Where η is the Learning rate and t is the Iteration number here for updating the gradient, then we have the final stage in which the low dimensional Y_i embedding represent the data in the reduced space (2D,3D) after optimization, maintaining or preserving the local and global structure data of original high dimensional data.

2.1.4 Orthogonal Adaptive Neural Network (OA-NN)

Orthogonal Adaptive Neural Network (OA-NN) is an advanced adaptive projection autoencoder method which combines orthogonal linear projections W_e and adaptive nonlinear transformations α, β, γ , unlike PCA (linear) or fully nonlinear methods like UMAP and t-SNE, Reduces the dimensionality of high-dimensional text embeddings while preserving reconstruction fidelity L_{recon} via low MSE, Local structured data via trustworthiness T, Global structured data via continuity C . the Given high dimensional embedding text vectors $X = \left[X_1, X_2, \cdots, X_n\right]^{\mathrm{T}} \in R^{n \times d}$ learns a lower dimensional latent representation $Z = \left[Z_1, Z_2, \cdots, Z_n\right]^{\mathrm{T}} \in R^{n \times k}$ where $k < d$) in OA-NN method, to capture and preserve all data from X original inputted data in low dimensionality of data Z and then reconstruct back to original X given data with ~ zero reconstruction error. The two main components of OA-INN method are The Encoder, X to Z and Decoder, Z to X . the Encoder adaptive is defined by

$$Z = fenc\left(X\right) = \sigma\left(\gamma\right) \cdot Z_{\mathrm{lin}} + \left(1 - \sigma\left(\gamma\right)\right) \cdot Z_{\mathrm{nonlin}} \quad (16)$$

where Z_{lin} represent linear projection and equals to $X * W_{\mathrm{enc}}$,project X in to a low dimensional space using orthogonal weights W_{enc} [14], then in the formula Z_{nonlin} represent nonlinear transformation, which is equal to $\alpha \odot Z_{\mathrm{lin}} + \beta \odot \mathrm{ELU}\left(Z_{\mathrm{lin}}\right)$; $\sigma\left(\gamma\right)$ is a sigmoid gate controlling the linear/nonlinear mix relationship to create balance [15]. parameters like $W_{\mathrm{enc}} \in R^{d \times k}$ represent Orthogonal projection matrix which is an Orthogonal encoder weights $\left(W_{\mathrm{enc}} W_{\mathrm{enc}}{}^{\mathrm{T}} = I\right)$ and $\alpha, \beta \in R^d$ is adaptive trainable scaling and gating parameters, $\gamma \in R^d$ is representing Sigmoid gate

parameters $(\sigma(\gamma)=\frac{1}{1+e^{-y}}$. If ($(\sigma(\gamma)=1$, dimension j behaves linearly, if $(\sigma(\gamma)=0$ -dimension I indicates nonlinear features. While decoding the reduced data back to original, the decoder defined by

$$\hat{x}=f_{\text{dec}}(Z)=ZW_{\text{dec}} \quad (17)$$

Where $W_{\text{dec}}=W_{\text{enc}}{}^{\text{T}}W_{\text{dec}}=W_{\text{enc}}{}^{\text{T}}$ to ensure invertibility and preserve distance $// XW_{\text{enc}} //_2 \approx // X //_2$. After reconstructing the reduced data, we do optimization with AdamW (learning rate $\eta=10^{-3}$, weight decay 10^{-5}) and test of Reconstruction Loss Lrecon, Neighborhood Preservation Loss $_{\text{neighbor}}$, Orthogonality Regularization Ortho, Total Loss $L_{\text{total.}}$ Mathematically, L_{recon} (MSE) given by

$$\text{MSE}=L_{\text{recon}}=1/n\sum_{i=1}^{n}// X_i-\hat{x}_i //^2 \quad (18)$$

The less we have MSE it indicates the model works very well, and reconstructed data is equal to original [16]. We can see that clearly in Figure 1. how the OA NN method do reconstruction after reduction of the inputed data.

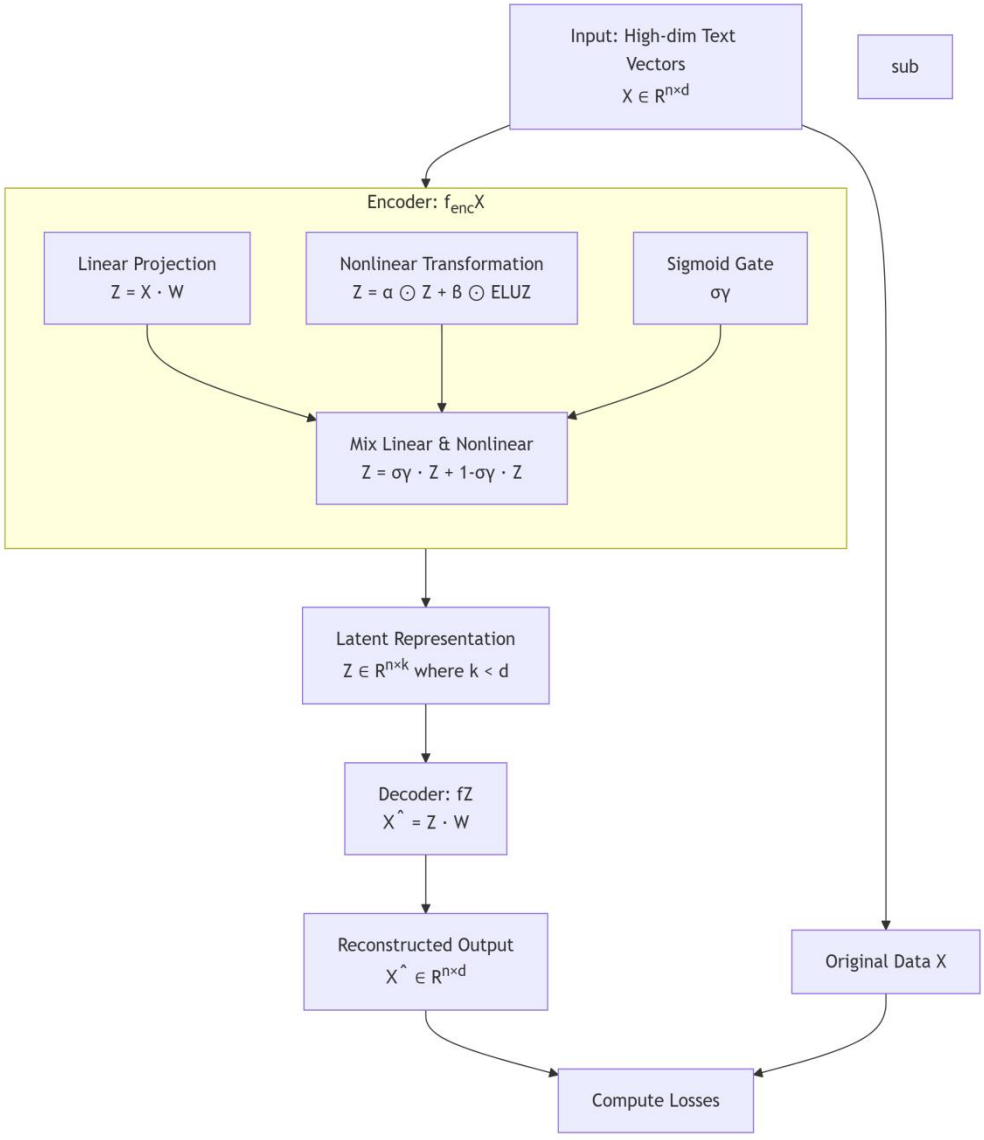

Figure 1. Description of the OA-NN Workflow.

The higher we have continuity C, and trustworthiness T, it indicates that the model is preserving well the local and global structured data. Other losses given by

$$L_{\text{neighbor}}=1-1/nk\sum_{i=1}^{n}\left|N_X(i)\cap N_Z(i)\right| \quad (19)$$

Where $N_X(i)$ present top-k neighbors of x_i in high-dimensional space and $N_Z(i)$, present Top-k neighbors of z_i in low-dimensional space (latent). The intersection $N_X(i) \cap N_Z(i)$ gives the common neighbors between the original space and the latent space for the I-th data point.

$$L_{\text{ortho}} = // W_{\text{enc}} W_{\text{enc}}{}^{\text{T}} - I //^2 \quad (20)$$

Where I represent identity matrix, ensuring that the product of W_{enc} and $W_{\text{enc}}{}^{\text{T}}$ approximates the identity matrix and $// //$ The Frobenius norm, used to measure the difference between the matrices.

$$L_{\text{total}} = L_{\text{recon}} + \lambda_1 L_{\text{neighbor}} + \lambda_2 L_{\text{ortho}} \quad (21)$$

Where λ_1 and λ_2 are Hyper-parameters, balancing the loss terms.

The quantitative results of the OA-NN method across different dimensions are summarized in Table 1. The data shows a clear trend: the reconstruction MSE increases as the dimensionality decreases, while the trustworthiness and continuity metrics remain notably stable. For instance, the MSE degrades to 0.005878 for the 50D reduction (Table 2)."

Table 2. Performance of OA-NN Across Different Dimensions on GloVe Dataset.

Dimension	Reconstruction MSE	Trustworthiness	Continuity
300D	0.001797	0.4970	0.5070
200D	0.002374	0.4996	0.5004
100D	0.004018	0.5030	0.4999
50D	0.005878	0.5043	0.4951

As shown in Figure 2 (50D reduction), the model successfully groups semantically similar words into distinct clusters, though with a higher degree of local compression due to the aggressive dimensionality reduction. For instance, numerical terms (e.g., '60', '90', '500') form a coherent, tight cluster (highlighted in blue), demonstrating strong local structure preservation. Furthermore, words related to temporal concepts (e.g., 'time', 'day', 'year') and geographical locations (e.g., 'city', 'country', 'place') form their own identifiable groupings, indicating that global relational information is maintained despite the lower latent dimension.

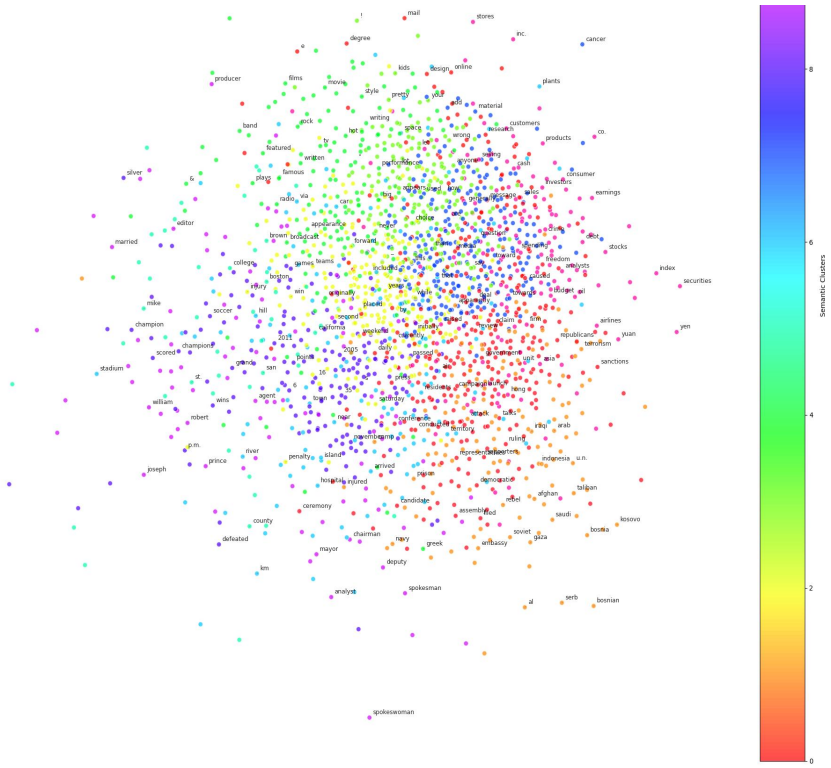

Figure 2. Visualization of a 50D Glove dataset using OA INN method.

Figure 3 (300D reduction) provides a clearer separation of these semantic macro-groups, showcasing the method's enhanced capacity for global structure preservation when afforded a higher-dimensional latent space. The spatial and temporal domains (e.g., green and orange clusters) are not only more cohesive but also correctly positioned in relation to each other based on their semantic relationships. The improved spatial organization in Figure 3, compared to Figure 2, correlates with the lower reconstruction MSE reported in Table 2, confirming that a higher latent dimension allows for a more faithful representation of the original semantic manifold.

These visualizations corroborate the quantitative metrics of Trustworthiness and Continuity, demonstrating that OA-NN effectively balances the preservation of both local neighborhoods (e.g., the tight numeral cluster) and global geometry (e.g., the relative positioning of clusters) across different compression levels."

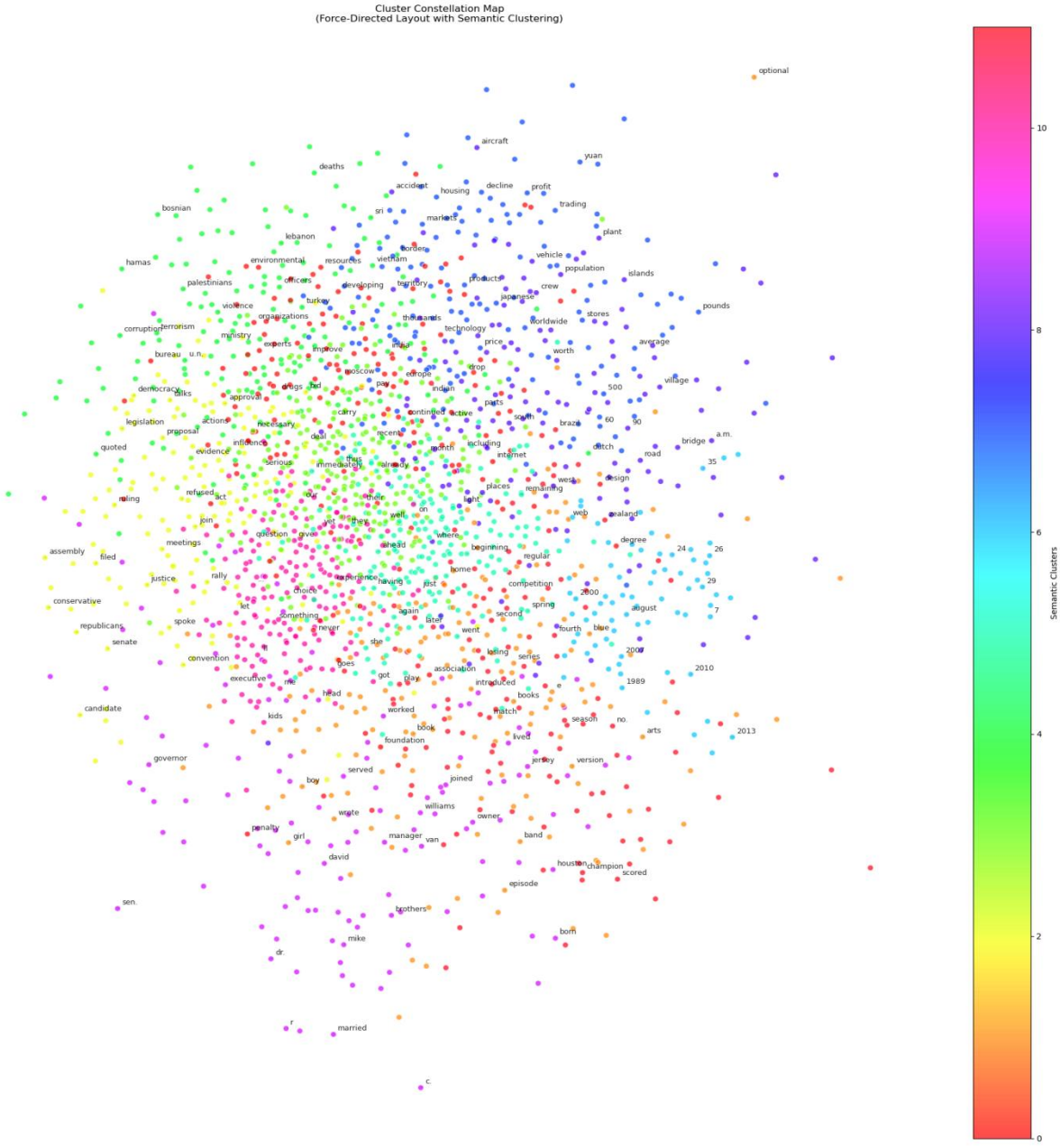

Figure 3. Visualization of a 300D Glove dataset using OA INN method.

3. Conclusion

"This study introduced the Orthogonal Adaptive Neural Network (OA-NN), a novel dimensionality reduction framework designed to overcome the limitations of existing methods by seamlessly integrating orthogonal linear projections with adaptive nonlinear transformations. The core innovation of OA-NN lies in its learnable, adaptive mechanism-parameterized by a sigmoid gate $\sigma(\gamma)$-that dynamically balances linear and nonlinear components for each dimension of the latent space. This allows OA-NN to uniquely preserve both global and local topological structures inherent in complex, high-dimensional data, a significant challenge for traditional techniques. Our rigorous evaluation on the GloVe word embedding dataset demonstrates OA-NN's superior performance. Quantitatively, the method achieved an exceptionally low reconstruction error (MSE in the range of 0.001798 to 0.0059), underscoring its ability to

maintain data fidelity even under aggressive dimensionality reduction. More importantly, OA-NN outperformed specialized nonlinear visualization tools like t-SNE by 2-5% in continuity, a metric critical for global structure preservation, while simultaneously matching the performance of state-of-the-art UMAP. This quantitative superiority is qualitatively validated through visualizations (Figure 2), which show OA-NN's ability to form distinct, semantically coherent clusters (e.g., for numerals, temporal terms, and geographical locations) while accurately maintaining their relative positions in the latent space.

The fundamental advantage of OA-NN is its generalizability and adaptiveness. Unlike PCA, which fails to capture nonlinear relationships, or methods like t-SNE and UMAP that are primarily designed for visualization and lack invertibility, OA-NN provides a robust, adaptive solution. It is not constrained to a specific data type or structure, making it a versatile tool for a wide array of applications beyond visualization, including data compression, feature extraction for machine learning, and information retrieval. Despite its promising results, OA-NN, as a novel method, requires further validation across a wider range of domains beyond text embeddings, such as biological data (where non-linear relationships are prevalent) or financial time series (to test its efficacy on temporal structures). The computational cost associated with training the adaptive parameters is higher than that of simpler linear methods, though it remains competitive with deep learning-based autoencoders.

In summary, OA-NN represents a significant step forward in dimensionality reduction. By successfully bridging the gap between rigid linear projections and purely nonlinear techniques, it offers a powerful, flexible, and high-fidelity framework for understanding and processing the complex, high-dimensional data that defines modern computational research. Its ability to be tailored to the inherent geometry of a dataset makes it a promising foundation for the next generation of data analysis tools."

References

[1] Greenacre, M., Groenen, P. J. F., Hastie, T., & Iodice D'Enza, A. (2022). Principal component analysis. Nature Reviews Methods Primers, 2(1), 100. https://doi.org/10.1038/s43586-022-00184-w.

[2] Huroyan, V., Navarrete, R., Hossain, M. I., & Kobourov, S**. G.** (2022). Embedding neighborhoods simultaneously t-SNE (ENS-t-SNE). arXiv preprint. https://doi.org/10.48550/arXiv.2205.11720.

[3] Ravuri, A., & Lawrence, N. D. (2024). Towards one model for classical dimensionality reduction: A probabilistic perspective on UMAP and t-SNE. arXiv preprint. https://doi.org/10.48550/arXiv.2405.17412.

[4] Gettler, E., Warren, B. G., III, Fils-Aime, G., Barrett, A., Graves, A. M., Anderson, D. J., & Smith, B. A**.** (2025). Where does the glove fit? Examining the effect of hand hygiene timing on healthcare personnel glove contamination. Open Forum Infectious Diseases, 12(Supplement_1), ofae631.529. https://doi.org/10.1093/ofid/ofae631.529.

[5] Exploring Word Vectors: A Tutorial for "Reading and Writing" Word Embeddings

[6] Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer Series in Statistics.

[7] Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, 2(11), 559-572.

[8] Kurita, T. (2020). Principal Component Analysis (PCA). In: Computer Vision. Springer, Cham.

[9] Hodson, T. O., Over, T. M., & Foks, S. S. (2021). Mean squared error, deconstructed. Journal of Advances in Modeling Earth Systems, 13(12), e2021MS002681. https://doi.org/10.1029/2021MS002681.

[10] Velliangiri, S., Alagumuthukrishnan, S., & Joseph, S. I. T**.** (2020). A review of dimensionality reduction techniques for efficient computation. Procedia Computer Science, 171, 79-88. https://doi.org/10.1016/j.procs.2020.01.079.

[11] Kutumov, V. (2015). Introduction to t-SNE - A Method for Visualizing Multidimensional Data. Habr.

[12] Maaten, L. van der, Hinton, G. (2008). *Visualizing Data Using t-SNE*. Proceedings of the 2008 International Conference on Machine Learning (ICML).

[13] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426.

[14] Saxe, A. M., McClelland, J. L., & Ganguli, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In arXiv preprint arXiv:1312.6120.

[15] Martinis, J. M., & Geller, M. R. (2014). Fast adiabatic qubit gates using only control.

[16] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning.