ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Forget early exaggeration in t-SNE: early hierarchization preserves global structure

John A. Lee[12], Edouard Couplet[1], Pierre Lambert[1], Ludovic Journaux[3], Dounia Mulders[1], Cyril de Bodt[1], and Michel Verleysen[1] *

1- UCLouvain - IREC/MIRO
Avenue Hippocrate 55, 1200 Brussels, Belgium

2- UCLouvain - ICTEAM/ELEN
Place du Levant 3, 1348 Louvain-la-Neuve, Belgium

3- Institut Agro Dijon - Laboratoire d'Informatique de Bourgogne
Boulevard Docteur Petitjean 26, 21079 Dijon, France

Abstract. As a local method of dimensionality reduction, t-SNE requires careful initialization in order to preserve the data global structure to the best extent. In regular t-SNE, the low-dimensional embedding is initialized either randomly or with PCA; next, gradient descent refines the embedding coordinates in two phases. In the first one, called *early exaggeration*, attractive forces between points are artificially strengthened to delay any detrimental effect of repulsive forces while points are still poorly organized. In this paper, a novel initialization of t-SNE is proposed. It works by hierarchizing the data points into a space-partitioning binary tree and successive runs of t-SNE with 4, 8, 16, ..., N points. Between two runs, the prototypical point in each tree branch is split into its two children prototypes, with some little random noise, and the embedding is rescaled to account for the increased population. Experimental results show the effectiveness of the method. The proposed method is compatible with any method of neighbor embedding (t-SNE, UMAP, etc.) provided early exaggeration can be disabled and initial coordinates can be fed into.

1 Neighbor embedding

Since 2008, the field of dimensionality reduction (DR) has gained much popularity thanks to Student t-distributed SNE (t-SNE) [1], a method of neighbor embedding that extends and modifies the original stochastic neighbor embedding (SNE) [2]. As compared to older methods like principal component analysis (PCA) [3] and multidimensional scaling (MDS) [4], for instance, t-SNE is no longer a global method based on variance or distances. Instead, it is a local method that that works with similarities and is almost immune to the curse of dimensionality and yields impressive results. Starting from 2015, accelerated versions of t-SNE have become available [5, 6], allowing to embed up to millions of points and then extending applicability to many domains. As a remarkable example, computational biology has widely adopted t-SNE and similar methods to embed and visualize cell data (e.g., single cell transcriptomics [7]). In that field, and in many others, t-SNE is used primarily for DR, although it gets used

*J.A.Lee is a Research Director with the Belgian F.R.S.-FNRS.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

more and more to identify clusters [8, 9], as it actually outperforms genuine methods of clustering in many respects, thanks to its inductive bias and strong repulsive forces. However, t-SNE is not perfect and, being a local method, it is known to poorly preserve the global structure of data [10].

This paper addresses this issue in a novel way, by chaining successive runs of t-SNE with growing number of points $(4, 8, 16, \dots N)$. Each run gets initialized with the previous one, by splitting each point into two and adding little random noise. Such divisions are made possible by first hierarchizing the data points into a space-partitioning binary tree (which is typically used in accelerated t-SNE anyway [5]). Each run of t-SNE inherits a trace of the global structure and, provided gradient descent works with a reasonable learning rate, this 'big picture' is gently modified, refined with some more local structure, and passed on to the next run. This approach bears some similarity with multiresolution pyramids in image processing [11], e.g., for non-rigid image registration [12].

The rest of this paper is organized as follows. Section 2 is a short reminder of t-SNE, also stating the requirements for the proposed initialization. Section 3 describes the hierarchical initialization of t-SNE. Experimental results are reported and discussed in Section 4. Section 5 concludes and sketches perspectives.

2 Student t-distributed stochastic neighborhood

Neighbor embedding (NE), including methods like stochastic neighbor embedding (SNE) [2], t-distributed SNE (t-SNE) [1], uniform manifold approximation and projection (UMAP) [6], etc., work by preserving pairwise affinities (a.k.a. similarities). In SNE, these soft neighborhoods are softmax ratios, i.e., Gaussian functions that are normalized into discrete neighborhood probabilities in both the data and embedding spaces (HD & LD), whose mismatch is measured with Kullback-Leibler divergences. In t-SNE, the HD Gaussian affinities are symmetrized, while the LD affinities are Student t hyperbolic functions that are normalized jointly. If $\mathbf{X} = [\mathbf{x}_i]_{1\leq i\leq N}$ and $\mathbf{Y} = [\mathbf{y}_i]_{1\leq i\leq N}$ denote the HD data and their LD embedding, then the pairwise affinities are

$$p_{j|i} = \frac{\exp -\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}\,, \quad p_{i|i} = 0\,, \quad \text{and } p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}\,, \quad (1)$$

where bandwidth σ_i is such that the entropy $H_i = \log K_\star = -\sum_{j\neq i} p_{j|i}\log p_{j|i}$ is the same around all $\mathbf{x}_i$ and set by perplexity $K_\star$. In the embedding space, these symmetrized entropic affinities are matched by

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}\,, \quad q_{ii} = 0 \quad (2)$$

through minimizing the joint KL divergence $\mathrm{KL}(P||Q) = \sum_{i\neq j} p_{ij}\log(p_{ij}/q_{ij})$. Minimization is carried out with gradient descent (and momentum) in three stages: initialization, early exaggeration, and fine tuning. The gradient of the KL divergence is $\nabla_{\mathbf{y}_i}\mathrm{KL}(P||Q) = 4\sum_{j\neq i}(p_{ij} - q_{ij})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|)^{-1}(\mathbf{y}_i - \mathbf{y}_j)$,

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

where p_{ij} and q_{ij} are responsible for attractive and repulsive forces between $\mathbf{y}_i$ and $\mathbf{y}_j$, respectively. In legacy t-SNE, $\mathbf{Y}$ is initialized either randomly (Gaussian distribution with variance much lower than 1) or a rescaled PCA projection (also with low variance). Initialization with Laplacian eigenmaps (LE) instead of PCA has been proposed for UMAP [10]. Early exaggeration (EE) temporarily and artificially magnifies the attractive forces by replacing p_{ij} with αp_{ij}, where $\alpha = 4$ or even higher values [9]. The intuitive effect of EE is that it makes clusters tighter in $\mathbf{Y}$, thereby making inter-cluster gaps broader; EE has been related to spectral clustering and power iterations with the graph Laplacian matrix associated with symmetric transition probability matrix $\mathbf{P} = [p_{ij}]_{1 \leq i,j \leq N}$ [9]. Usually, t-SNE runs for 1000 iterations, with about one quarter or one fifth for EE and the rest for fine tuning with the actual values of p_{ij}. By magnifying p_{ij}, EE also increase the gradient magnitude, leading to swift and drastic motions of points in the embedding in the early iterations, and thus to possible scrambling and disorganization of the global data structure (inter-cluster arrangement). In the fine-tuning stage, the attractive forces go back to their nominal values and repulsive forces get comparatively stronger then, thereby accentuating cluster gaps or even separating spurious clusters due to noise or data sampling. The absence of global structure with random initialization or its likely scrambling by EE after initialization with PCA or LE makes t-SNE a strongly local method, generating debates and controversies among its users and research communities [10].

To prepare the ground for the proposed initialization, simple requirements should be fulfilled: the implementation of t-SNE (or any other similar method of NE for that matter) can be run with the call: $\mathbf{Y}^{(T)} = \texttt{tsne}(\mathbf{X}, \mathbf{Y}^{(0)}, K_\star, T)$, where t-SNE runs *without* EE, starting from specified initial coordinates $\mathbf{Y}^{(0)}$ for T iterations with perplexity $K_\star$.

3 Early hierarchization to initialize successive t-SNE runs

Starting from $\mathbf{X}$, a space-partitioning binary tree can be built. In HD space, vantage-point trees, relying on distances, are typically preferred over (non-binary) kd-trees, working with coordinates. Each node of the tree splits a set of N points into two non-overlapping regions with $\lfloor N/2 \rfloor$ and $\lceil N/2 \rceil$ points, respectively. Each branch then gets further divided into thinner ones, up to reaching the tree leafs where just one or two points are held (if the data size N is not a power of 2). In each node, the subset $\mathbf{X}_{(k,l)}$ of $\mathbf{X}$, with $k < \lfloor \log_2 N \rfloor$ and $1 \leq l \leq 2^k$, can be summarized with one prototypical point $\mathbf{x}_i \in \mathbf{X}_{(kl)}$. An easy choice for the prototype is the point in $\mathbf{X}_{(kl)}$ that lies the closest to the average of all points in $\mathbf{X}_{(kl)}$. Let $\bar{\mathbf{x}}_{(kl)}$ denote the prototype of node $\mathbf{X}_{(kl)}$. Gathering all prototype on each level k, we get data subsets $\bar{\mathbf{X}}_{(k)} = [\bar{\mathbf{x}}_{(kl)}]_{1 \leq l \leq 2^k}$ of size $N_{(k)} = \min(2^k, N)$. On the leaf level $k = \lfloor \log_2 N \rfloor$, the full data set $\mathbf{X}$ is eventually considered.

Our proposal runs t-SNE without EE successively on $\{\bar{\mathbf{X}}_{(k)}, \mathbf{X}\}_{2 \leq k < \lfloor \log_2 N \rfloor}$ with perplexities $K_{(k)} = \min(2^{k-1}, K_\star)$ where $K_\star$ is user-defined. Notice that the first run starts with four points ($k = 2$), which are initialized randomly

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

with small variance, and a perplexity of 2. A call to t-SNE writes down into $\mathbf{Y}_{(k)}^{(T)} = \texttt{tsne}(\bar{\mathbf{X}}_{(k)}, \mathbf{Y}_{(k)}^{(0)}, K_{(k)}, T)$ for $T = 1000$ iterations of gradient descent. The learning rate is fixed to $N_{(k)} \sum_{k \neq l} q_{kl}(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}$, which is a very crude scalar estimate of the inverted Hessian matrix. In between two successive runs k and $k+1$, point mitosis is carried out by assuming that the children of $\bar{\mathbf{x}}_{(k,l)}$ in $\bar{\mathbf{X}}_{(k)}$ are $\bar{\mathbf{x}}_{(k+1,2l)}$ and $\bar{\mathbf{x}}_{(k+1,2l+1)}$ in $\bar{\mathbf{X}}_{(k+1)}$. If $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ denote i.i.d. vectors of Gaussian noise with small variance, then $\mathbf{y}_{(k+1,2l)}^{(0)} = 2^{1/2}\mathbf{y}_{(k,l)}^{(T)} + \boldsymbol{\epsilon}_1$ and $\mathbf{y}_{(k+1,2l+1)}^{(0)} = 2^{1/2}\mathbf{y}_{(k,l)}^{(T)} + \boldsymbol{\epsilon}_2$, where $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \sim \mathcal{N}(\mathbf{0}, (\nu/6)^2)$. As the embedding can be freely centered, $\bar{R} = \sum_{l=1}^{N_{(k)}} \|\mathbf{y}_{(k,l)}^{(T)}\|^2 / N_{(k)}$ is the average squared radius and $\delta^2 = \bar{R}/N_{(k)}$ is then a crude approximation to the expected squared distance between one point's closest neighbor. The next T iterations then adjust those offsprings' locations in the embedding space, from slightly random to an arrangement that reduces the KL divergence.

Conceptually, the above proposal resembles some previous approaches to the problem of global structure preservation with methods of NE. The problem arises from the considerable gap between initialization, where the global structure gets captured by PCA or LE, for instance, and subsequent local processing with a perplexity $K_\star \ll N$. The issue gets even worse in accelerated methods of NE, where perplexities remain within the range from 5 to 100 while N can grow to millions. Moreover tails of p_{ij} are typically truncated to make P sparse, making fast NE even more local. Previous approaches include multi-scale NE [13, 14], where several runs are carried out similarly, although all runs involves all points in $\mathbf{X}$ and decreasing perplexities $(\lfloor N/2 \rfloor, \ldots, 8, 4, 2)$. An accelerated version of multi-scale NE [14] with space-partitioning trees introduces the idea of succesive subsampling of $\mathbf{X}$ in the gradient computation only. The time complexities of iterations are $\mathcal{O}(N^2 \log_2 N)$ and $\mathcal{O}(N \log_2^2 N)$, respectively. Notice, however, that those approaches start from large scale and *add* smaller ones progressively, whereas the proposed method *switches* from one scale to the next, with the risk of progressively forgetting the upper scales. The gain is a time complexity of $\mathcal{O}(\sum_k N_{(k)} \log_2 N_{(k)}) \approx \mathcal{O}(N \log_2 N)$ for Barnes-Hut t-SNE [5] or even lower for more recent accelerated variants [6]. The proposed method can also be related to B. Fritzke's work in the mid 90s, with growing-size neural networks [15]. From a biological standpoint, inspiration stems from cell division in an embryo, whereby large beings can get spatially organized from local interactions only.

4 Experiments, results, and discussion

In order to assess the proposed early hierarchization (EH) as an initialization for t-SNE, several data sets are embedded in 2D, after PCA down to 16 dimensions. Regular t-SNE is run for 1000 iterations with EE = 4; EH t-SNE is run for 500 iterations on each level, without EE. In addition to embeddings, the curves $R_{\mathrm{NX}}(K) = ((\frac{N-1}{KN} \sum_i |\nu_i^K \cap n_i^K|) - K)/(N - 1 + K)$ [13] are reported, where $1 \leq K \leq N$ is a neighborhood size and ν_i^K and n_i^K are the K-ary neighborhoods of $\mathbf{x}_i$ and $\mathbf{y}_i$, respectively. These curves allow inspecting both the local and global

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

structures. The minimum value is 0 (not better than a random embedding on average) and the maximum is 1 (perfect rendering of all K-ary neighborhoods from HD to 2D). The area under the curves (AUCs) compounds local and global.

Figure 1 shows embeddings and quality curves for eight (subsampled) data sets: (row 1) noisy circle as in [10] and MNIST, (row 2) COIL-20 and COIL-100, (row 3) phoneme and google, (row4) Frey faces and Mouse RNA. The subsamples contain $N = 1000 \sim 3000$ points. Data dimensions range between 10 (8 noisy dimensions for the circle) and 16384 (COIL). Local neighborhoods are almost equally well preserved with the three initializations, while the curves differ for the global structure where early hierarchization manages to dominate there in most cases. Quantitatively, the AUCs are better for early hierarchization in 8/10 data sets; PCA and random initializations come 2nd and 3rd, respectively.

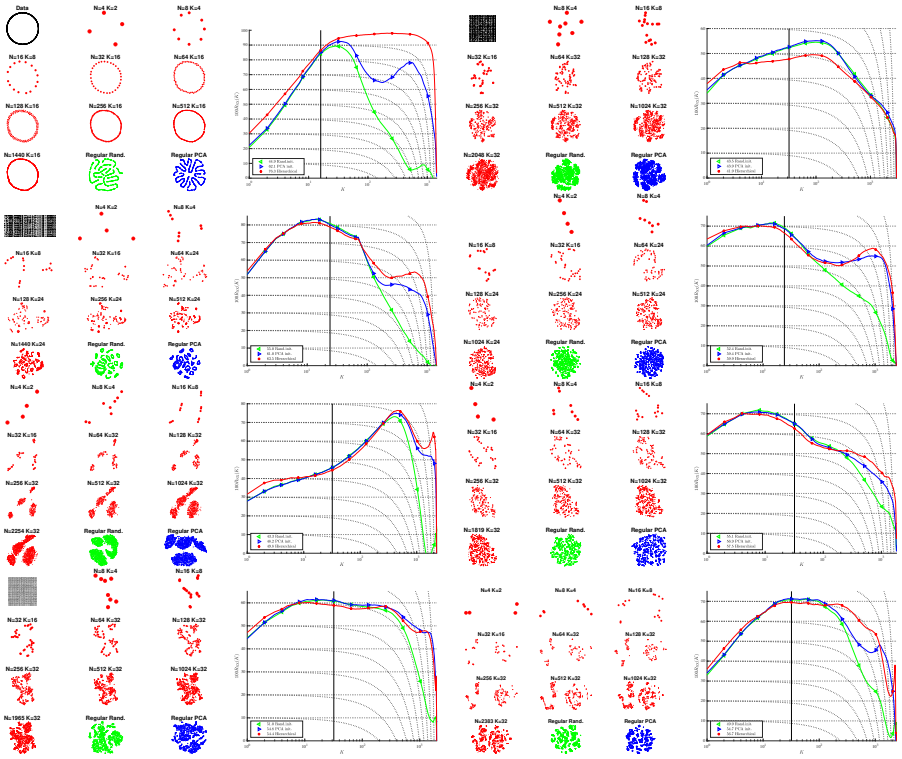

Fig. 1: Embeddings & quality curves: (row 1) noisy circle & MNIST, (row 2) COIL-20 & 100, (row 3) phoneme & google, (row4) Frey faces & Mouse RNA.

5 Conclusions and perspectives

Due to their local nature, t-SNE and most methods of NE critically depend on their initialization and a first few iterations to prevent scrambling the global

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

structure of data. Relying on PCA, Laplacian eigenmaps, or early exaggeration hardly bridges the broad gap between global and local. The proposed approach, coined *early hierachization*, relies shorter bridge spans and multiple intermediate pillars between global and local. This is achieved with space-partitioning binary trees, allowing a hierarchy of subsamples of sizes $4, 8, 16, \ldots, N$; the transition between two subsamples by point mitosis into its two children. Experimental results show that early hierarchization slightly outperforms the legacy random and PCA initializations of t-SNE. Early hierarchization also makes t-SNE simpler and more self-contained (re-use of trees to search for KNNs, no call to PCA or any other global method of DR like PCA), without changing its computational complexity. Future work will extend to more and bigger data sets.

References

[1] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[2] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

[3] I.T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[4] Ingwer Borg and Patrick J. F. Groenen. *Modern Multidimensional Scaling Theory and Applications*. Springer, New York, 2005.

[5] Laurens Van Der Maaten. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.*, 15(1):3221–3245, jan 2014.

[6] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.

[7] Dmitry Kobak and Philipp Berens. The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, 10(1):5416, 2019.

[8] Sanjeev Arora, Wei Hu, and Pravesh K. Kothari. An analysis of the t-SNE algorithm for data visualization. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1455–1462. PMLR, 06–09 Jul 2018.

[9] George C. Linderman and Stefan Steinerberger. Clustering with t-sne, provably. *SIAM Journal on Mathematics of Data Science*, 1(2):313–332, 2019.

[10] Dmitry Kobak and George C. Linderman. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology*, 39(2):156–157, 2021.

[11] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden. 1984, Pyramid methods in image processing. *RCA Engineer*, 29(6):33–41, 1984.

[12] M. Corvi and G. Nicchiotti. Multiresolution image registration. In *Proceedings., International Conference on Image Processing*, volume 3, pages 224–227 vol.3, 1995.

[13] John A. Lee, Diego H. Peluffo-Ordóñez, and Michel Verleysen. Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. *Neurocomputing*, 169:246–261, 2015.

[14] Cyril de Bodt, Dounia Mulders, Michel Verleysen, and John Aldo Lee. Fast multiscale neighbor embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 33(4):1546–1560, 2022.

[15] Bernd Fritzke. A growing neural gas network learns topologies. *Neural Information Processing Systems*, 7, 03 1995.