UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Authors:
Leland McInnes John Healy James Melville

Paper available at: https://paperswithcode.com/paper/umap-uniform-manifold-approximation-and

Presentation by Alix Bernard

Content

Introduction

UMAP is a dimensional reduction technique,
such techniques seek to produce a low dimensional representation of high dimensional data while preserving relevant structure.

Dimension reduction is a fundamental technique for both visualization and pre-processing for machine learning. There exist two categories of algorithm for dimension reduction:

- Those seeking to preserve distance structure within the data
  (e.g.: PCA, MDS)

- Those seeking to preserve local distances over the global one
  (e.g.: t-SNE, LargeVis)

UMAP is part of the second category (i.e. preserving local distances) and is competitive with t-SNE.

UMAP is based on strong mathematical foundations notably topology wise and thus it can be scaled to significantly larger real data set sizes than are feasible for t-SNE.

Moreover, UMAP arguably preserves more global structure than t-SNE and has a superior run time performance especially for higher dimension data sets.

Theoretical Foundations

The foundations of UMAP are largely based on manifold theory and topological data analysis. Here only a shallow overview will be given – for more details please refer to the paper section 2 and subsequent references from it.

To get a topological representation of high dimensional data, local fuzzy simplicial sets are patched together – in case of low dimensional data an equivalent representation is obtained.

Building the fuzzy topological representation is broken into two steps:

- Approximate a manifold on which the data is assumed to lie
- Construct a fuzzy simplicial set representation of approximated manifold

In order to approximate the manifold, custom distances are defined for each element *x* of the data set *X* yielding a family of discrete metric spaces that need to be merged into a consistent global structure.

To do so the metric spaces are converted into fuzzy simplicial sets – simplicial sets will not be detailed here.

The classical notion of a fuzzy set is defined by a carrier set and a map called membership function such that the membership strength of an element is no longer a bivalent *true* or *false* property.

The underlying idea is to use such sets where all elements have membership strength of at least some value *a* comprised between *0* (excluded) and *1*.

Finally optimization of low dimensional representation has to be done but to make local connectedness requirement the distance to the nearest neighbor (NN) is used via a parameter defining the expected distance between NN.

Let *Y* be the low dimensional representation of the data *X.*
Some fuzzy set cross entropy C is used to compare two fuzzy sets.
To optimize the embedding of *Y* with respect to the cross entropy *C* stochastic gradient descent is used (in a similar way as t-SNE) but to do so some more approximation of necessary objects may be required.

This optimization minimize the error between the two topological representations

Computational View

From a computational point of view UMAP is a *k*-neighbors graph based algorithm – as is t-SNE – and can be described in two phases:

1. Construction of a particular weighted *k*-neighbors graph.
   Considering a specific *k* a graph, with vertices the data set *X,* is computed and weighted regarding the distance to the *k*-nearest neighbors (*k*NN) for each element *x*.

2. Computation of low dimensional layout of this graph.
   A force directed graph layout algorithm is used with a set of attractive forces on the edges and a set of repulsive ones applied among the vertices.
   The algorithm apply those forces in an iterative way and eventually reach convergence.

Implementation

Overview of the UMAP algorithm – for more details please refer to the paper section 4.1.

Algorithm 1 UMAP algorithm

function UMAP(X, n, d, min-dist, n-epochs)
  for all $x \in X$ **do**
    fs-set$[x] \leftarrow$ LocalFuzzySimplicialSet(X, x, n)
  top-rep $\leftarrow \bigcup_{x \in X}$ fs-set$[x]$ ▷ We recommend the probabilistic t-conorm
  $Y \leftarrow$ SpectralEmbedding(top-rep, d)
  $Y \leftarrow$ OptimizeEmbedding(top-rep, Y, min-dist, n-epochs)
  return Y

For a practical implementation of this algorithm an approximate *k*NN calculation is required – the authors recommend the Nearest-Neighbor-Descent algorithm c.f. [16] in paper.

An efficient optimization via stochastic gradient descent is also required – c.f. [45] & [33] in paper.

As seen in *Algorithm 1*, UMAP uses 4 hyper-parameters:

- *n*, the number of neighbors to consider for approximating the local metric
- *d*, the target embedding dimension
- *min-dist*, the desired separation between close points in the embedding space
- *n-epochs*, the number of training epochs to used when optimizing the low dimensional representation

n can be interpreted as the local scale to approximate the manifold as roughly flat, it also represents some degree of trade-off between fine grained and large scale manifold features.

min-dist determines how closely points can be packed together in low dimensional representation, it is an aesthetic parameter to increase if UMAP is used for visualization.

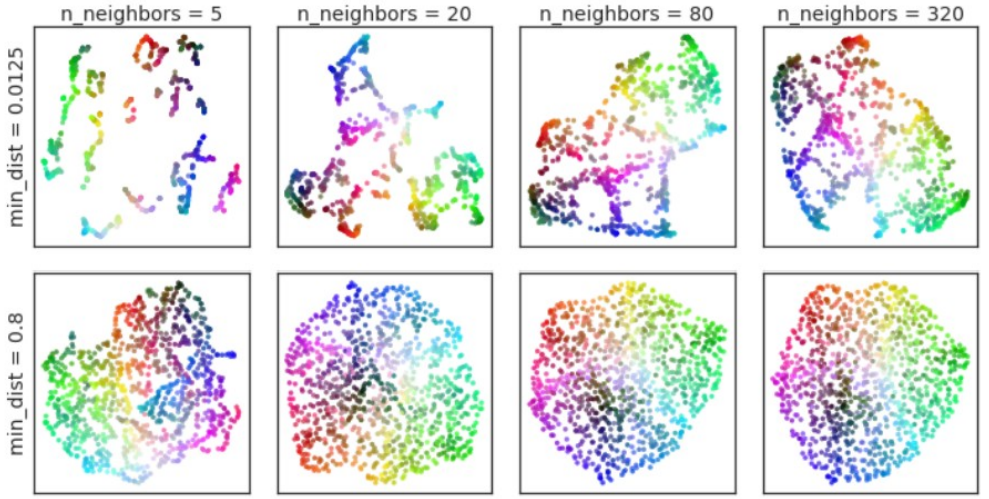


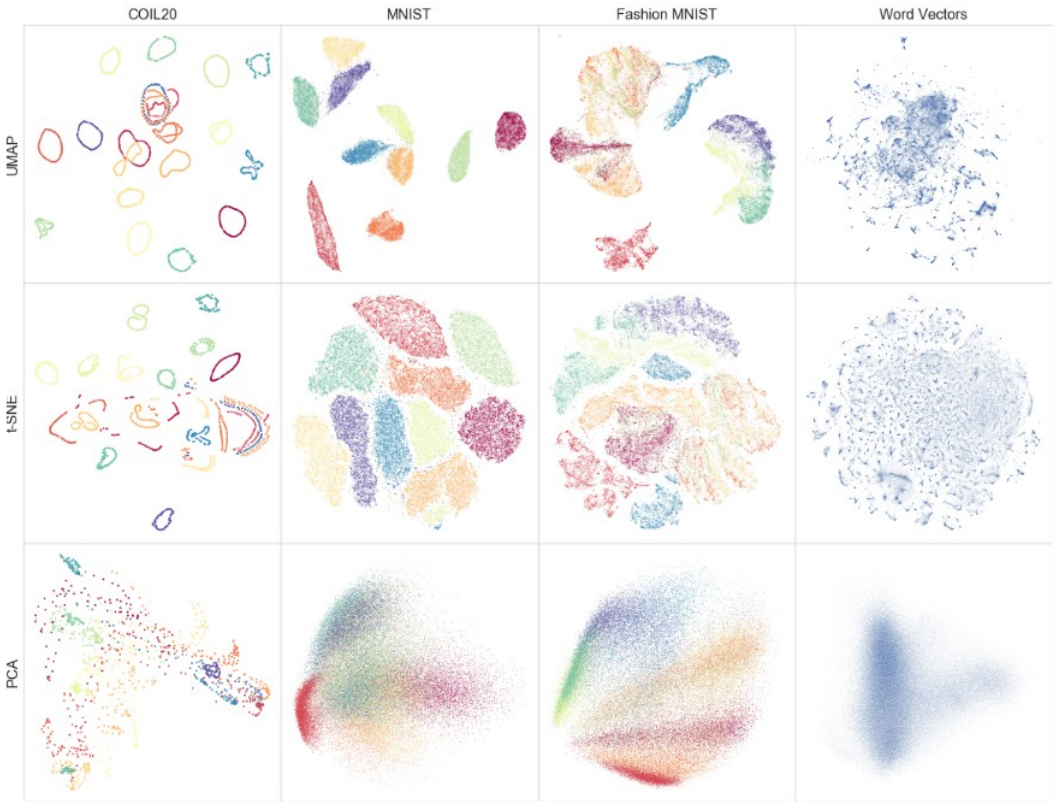
Figure 1[(1)] (cropped): Variation of UMAP hyper-parameters *n* and *min-dist* result in different embeddings. The data is uniform random samples from a 3-dimensional color-cube […]

(1): the number of the figure correspond to the number assigned in the paper as for the next figures in this presentation.

Practical Efficacy

Qualitative comparison on multiple data sets shows comparable quality of embedding to t-SNE for UMAP while reducing to 2 or 3 dimensions. UMAP is also arguably capturing more of the global and topological structure of the data sets than t-SNE.

Figure 2 (cropped): comparison of several dimension reductions algorithms [on multiple data sets]. [...]



Since UMAP uses stochastic processes, embeddings are different from run to run, therefore measuring how stable embeddings are is relevant.

To do so a Procrustes distance is used and the lower the distance between the two sets *X* and *Y* the more stable is the algorithm.

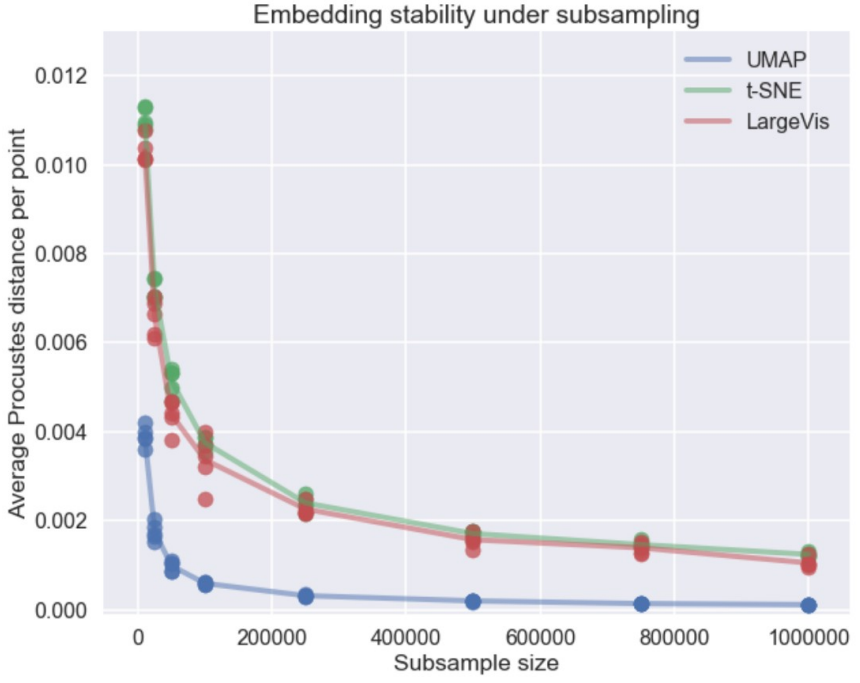


Figure 4: comparison of average Procrustes distance per point […] over a variety of sizes of sub-samples from the full Flow Cytometry data set.

Computational performance comparison have been done for different algorithms on multiple data sets – for details on the data sets refer to the paper section 5.

The *Table 1* shows that UMAP is superior to any of those other algorithms except for Pen Digits.

Table 1: Run-time of several dimension reduction algorithms on various data sets. […] Fastest run-time for each data set has been bolded.

	UMAP	FIt-SNE	t-SNE	LargeVis	Eigenmaps	Isomap
Pen Digits (1797x64)	9s	48s	17s	20s	**2s**	**2s**
COIL20 (1440x16384)	**12s**	75s	22s	82s	47s	58s
COIL100 (7200x49152)	**85s**	2681s	810s	3197s	3268s	3210s
scRNA (21086x1000)	**28s**	131s	258s	377s	470s	923s
Shuttle (58000x9)	**94s**	108s	714s	615s	133s	–
MNIST (70000x784)	**87s**	292s	1450s	1298s	40709s	–
F-MNIST (70000x784)	**65s**	278s	934s	1173s	6356s	–
Flow (100000x17)	**102s**	164s	1135s	1127s	30654s	–
Google News (200000x300)	**361s**	652s	16906s	5392s	–	–

UMAP shows better performance than t-SNE and LargeVis algorithms even when scaling with embedding dimension (Figure 5.b), ambient dimension (c.f. Figure 6 in paper), and number of samples (Figure 7).

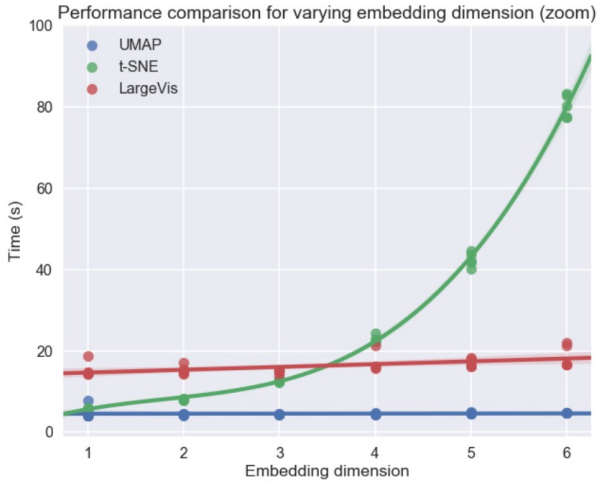
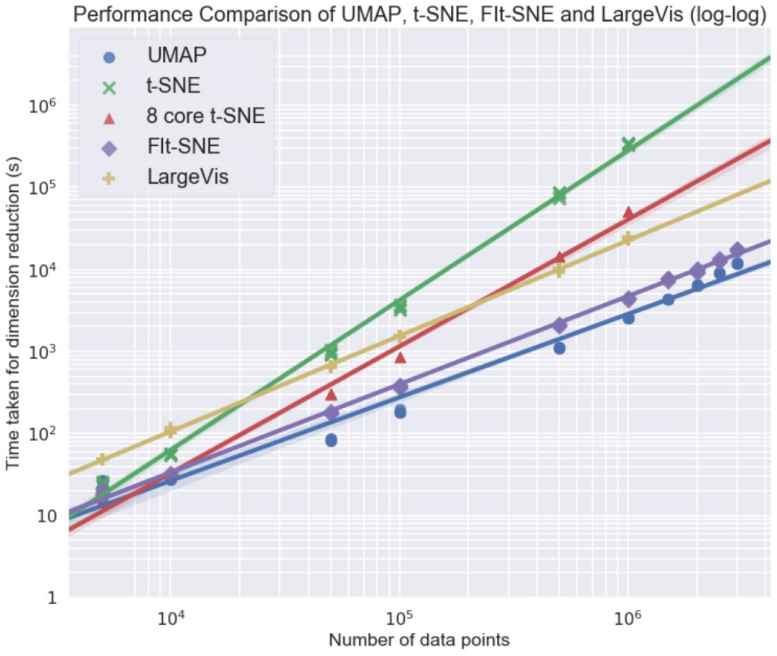


Figure 5.b: detail of scaling for embedding dimension of six or less. [...]

Figure 7: run-time performance scaling of t-SNE and UMAP on various sized sub-samples of the full Google News data set. [...]



Weaknesses

Despite being a very effective algorithm for visualization and dimension reduction UMAP make trade-off as well.

- UMAP lacks strong interpretability, its dimension embedding space has no meaning contrary to the PCA algorithm whose dimensions are the direction of greatest co-variance.
- It assumes manifold structures exist in the data, care must be taken for small sample sizes of noisy data and data with only large scale manifolds.
- It assumes that local distance is more important than long range one and therefore do not necessarily represent accurately global structure.
- Many approximations are made for computational efficiency, those approximations may have an impact on the result especially for small (< 500) data set sizes.

Conclusion

UMAP is a general purpose dimension reduction technique and the algorithm implementing it is faster than t-SNE and has better scaling.

It allows high quality embeddings of larger data sets than previously attainable, moreover its effectiveness in various scientific fields shows the strength of this algorithm.