Report

Karolina Maruszak, Aleksandra Warchoł, Piotr Wodecki

May 2021

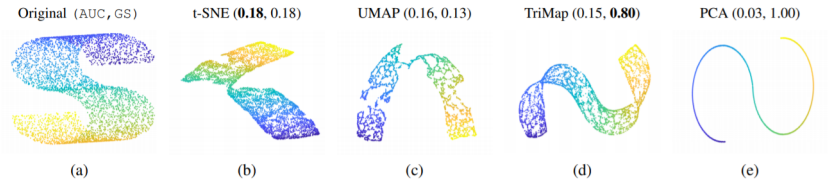
Abstract

The aim of our project will be to explore the TriMap method presented in the article by Ehsan Amid and Manfred K. Warmuth entitled "TriMap: Large-scale Dimensionality Reduction Using Triplets" and then making comparative visualizations with methods such as: t-SNE, LargeVis, UMap on 4 (or 5) data sets and verification of computational times, complexity, etc.

Introduction

A few words of introduction. Ehsan Amid and Manfred K. Warmuth have been written in their article: data visualization based on dimensionality reduction (DR) is a core problem in data analysis and machine learning. We know that DR is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension. DR methods are commonly divided into linear and non-linear approaches. One of the linear methods is PCA which is effective in preserving the global structure of the data. There are also methods that has been only on preserving the local neighborhood structure of each individual point. They belong to the nonlinear methods and they are well known to us: t-SNE, LargeVis and UMAP. The method we want to take a closer look at, TriMap (DR method) belongs to this group that focuses on preserving the global data structure while embedding.

The above dependencies can be seen in the example below, which is taken from the article by Amid and Warmuth. It shows a 2D visualization of the S curve dataset. In (a) the original 3D dataset is shown. In point (b) using the t-SNE method. In point (c) by the UMAP method. In point (d) using the TriMap method. And in the last point, the PCA method. What can be seen is that in the case of the PCA and TriMap methods, this global data structure is p goreserved, which cannot be seen in the case of t-SNE or UMap, which do not reflect the overall shape of the S curve. You can also see the GS values (global score) that it is high for TriMapa and PCA, while for t-SNE and UMapa it is low. AUC (The Area Under the Curve) is a dimensionality reduction performance measure called precision-recall.



The TriMap method

What exactly is TriMap? This is a dimensionality reduction method that uses triplet constraints to form a low-dimensional embedding of a set of points. The main idea in developing the TriMap method is to preserve a higher-order of structure in the data. In other words, we aim to reflect the relative (dis)similarities of triplets ofpoints, rather than pairs of points. Formally, a triplet is a constraint between three points i, j, and k, denoted as an ordered tuple (i, j, k), which indicates:

$$(i, j, k) \iff \text{point i is closer to point j than point k}$$

TriMap is initialized with the low dimensional PCA embedding, and this embedding is then modified using a set of carefully selected triplets from the high-dimensional representation. TriMap gives a better global view of the data than the aforementioned dimensionality reduction methods: t-SNE, LargeVis or UMap. TriMap compared to t-SNE is much faster, but provides comparable operating time with Umap and LargeVis, and at the same time scales much better to larger data sets.

Formal approach

1. TriMap starts by selecting a set of triples T = {(i, j, k)} such that Distance(i,j) < Distance(i,k) and assigns a weight to each triple $w_{ijk} \geq 0$. Note that a higher value of w_{ijk} means that the pair (i, k) is much further away than the pair (i, j).

2. Define the loss of the triplet (i, j, k) as:

$$l_{ijk} = w_{ijk} \frac{s(y_i, y_k}{s(y_i, y_j) + s(y_i, y_k} \tag{1}$$

where, $s(y_i, y_j) = (1 + ||y_i - y_j||^2)^{-1}$

It is also worth noting that the loss of the triple (i, j, k) approaches zero when $||y_i - y_j||$ decreases, and $||y_i - y_k||$ grows.

3. Development of a triple weighing scheme. In order to reflect the relative similarities in large dimensions, an unnormalized weight for three is defined:

$$w_{ijk}^{\sim} = exp(d_{ijk}^2 - d_{ij}^2) \geq 0 \tag{2}$$

where $d_{ij}^2 = \frac{||x_i - x_j||^2}{\sigma_{ij}}$

4. The final weight value is the result obtained by applying the log transformation

$$w_{ijk} = \zeta_\gamma(\frac{w_{ijk}^{\sim}}{W} + \delta) \tag{3}$$

where $\zeta_\gamma(u) = log(1 + \gamma u)$

and $W = max_{i',j',k' \in T} \tilde{w}_{i'j'k'}$

5. Initialization: initialize Y using PCA. As mentioned before, TriMap is initialized with PCA low-dimensional embedding. Such an initialization for TriMap enables faster convergence while retaining much of the global structure discovered by the PCA.

6. Finally, the loss function is computed as the sum of the loss of the triples collected in T.

Summarizing, TriMap is a dimensionality reduction method that uses triples to create a low dimensional point set embedding. It is a fast and efficient method that can be easily applied to large data sets. Moreover, it focuses on preserving the global data structure and provides a much better global view of the data than other dimensionality reduction methods such as t-SNE, LargeVis and UMAP. These, however, in turn, provide additional insight into the local neighborhood of individual points. In the next chapter, the comparison of these methods with respect to Trimap on various data sets will be presented

The datasets used to compare the methods

Five datasets were used for comparison: Mnist, Fashion Mnist, 20 News Group, Cifar-10, Forest Cover Type. The following methods were compared: PCA, t-SNE, UMAP, LargeVis, TriMap. First, some information about these data sets.

Mnist (Modified National Institute of Standards and Technology) contains a collection of 70,000, 28 x 28 images of handwritten digits from 0 to 9. This dataset is available in the sklearn.datasets library in Python.

Fashion Mnist is a dataset of Zalando's article images. It consists of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes. This dataset is also available in the sklearn.datasets library in Python.

The 20 News Groups dataset comprises around 18000 newsgroups posts on 20 topics. This dataset is also available in the sklearn.datasets library in Python.

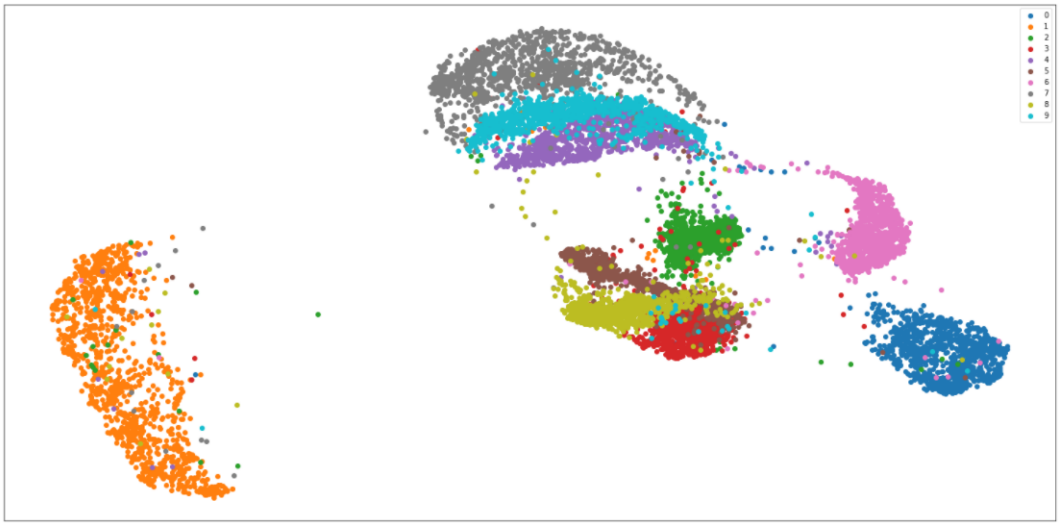
Cifar-10 is a dataset of 50,000 32x32 color training images and 10,000 test images, labeled over 10 categories. It is available in the tensorflow.keras.datasets library in Python.

Forest Cover Type is the actual forest cover type for a given observation (30 x 30 meter cell) was determined from US Forest Service Region 2 Resource Information System data. Independent variables were derived from data originally obtained from US Geological Survey and USFS data. Data is in raw form (not scaled) and contains binary columns of data for qualitative independent variables (wilderness areas and soil types). It is available at Machine Learning Repository.

Factors taken into account when comparing methods are: **Run Time, Average Cartesian Distance inside classes to Average Cartesian Distance for points from different classes Ratio** and **Global Score** only for TriMap.

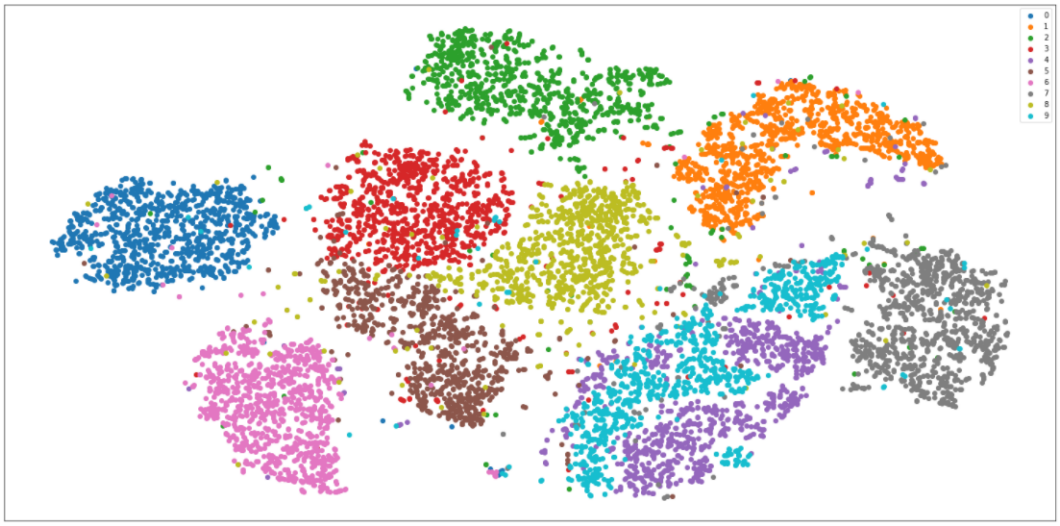
Comparison for the MNIST dataset

TriMap:

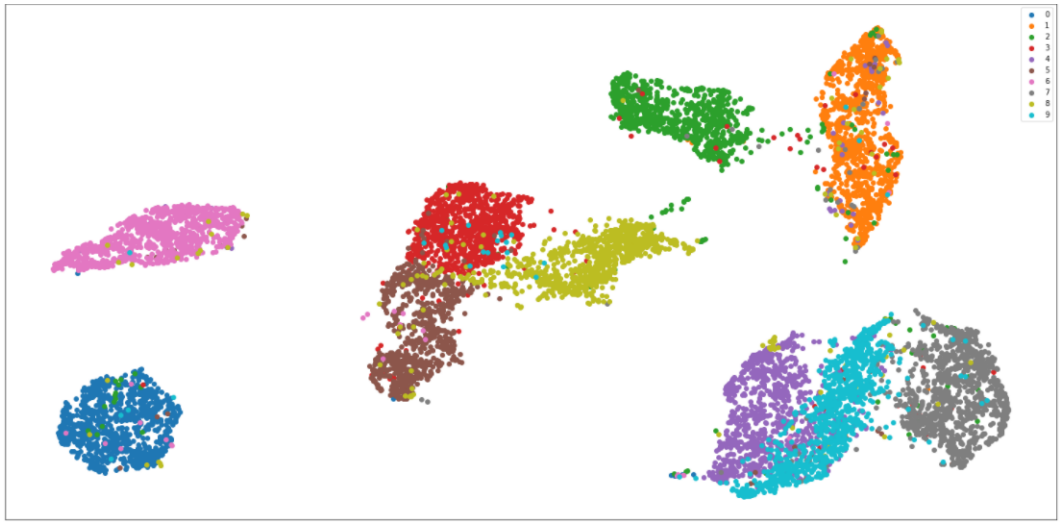


Global Score : 0.93

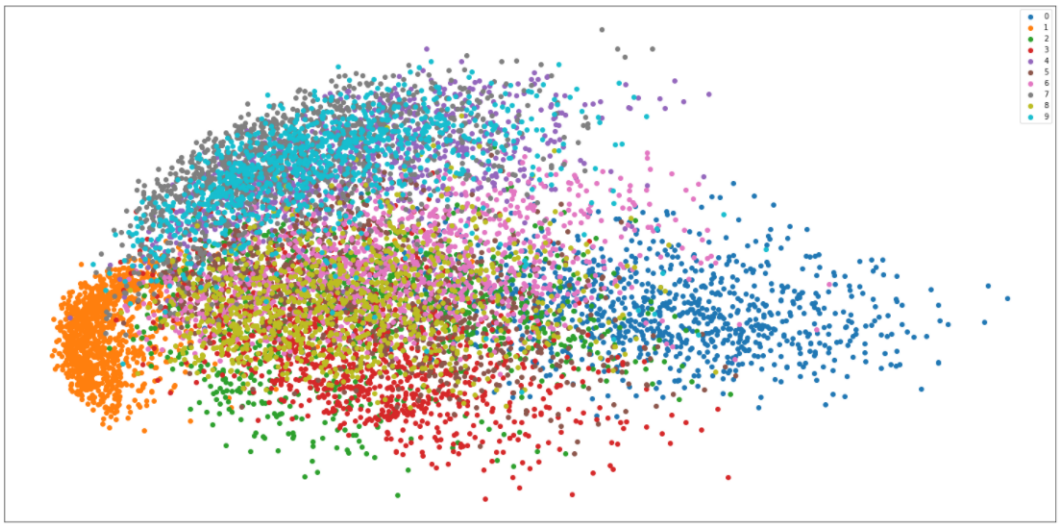
t-SNE:



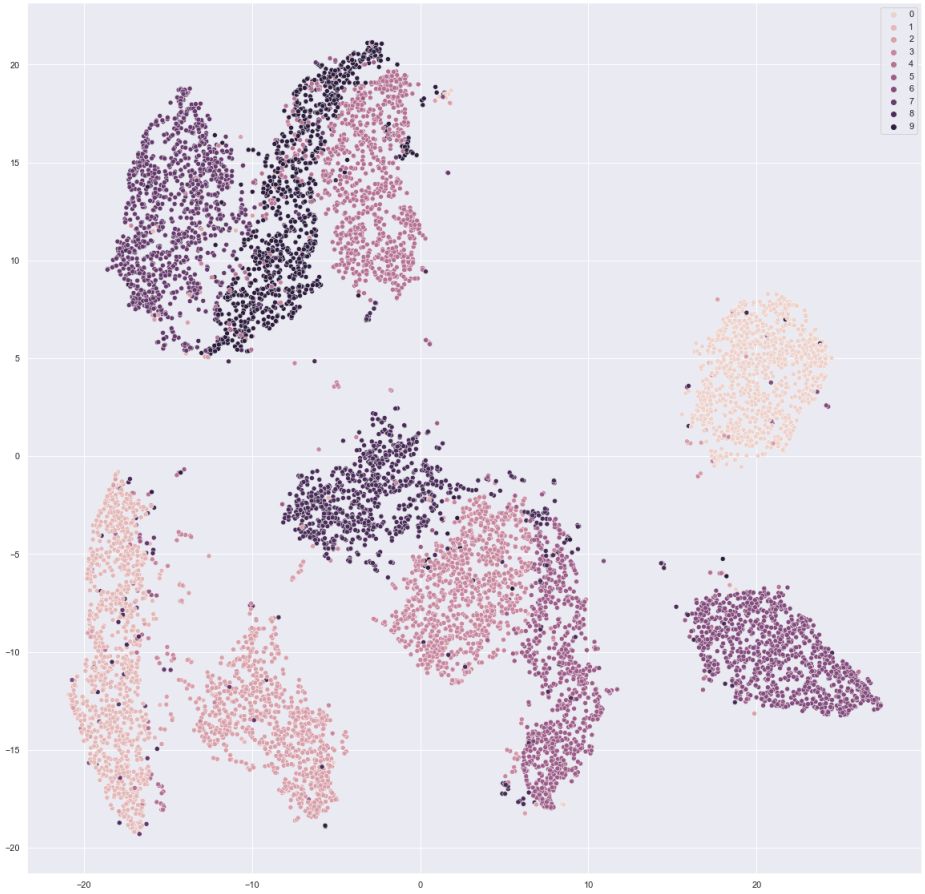
UMAP:



PCA:



LargeVis:



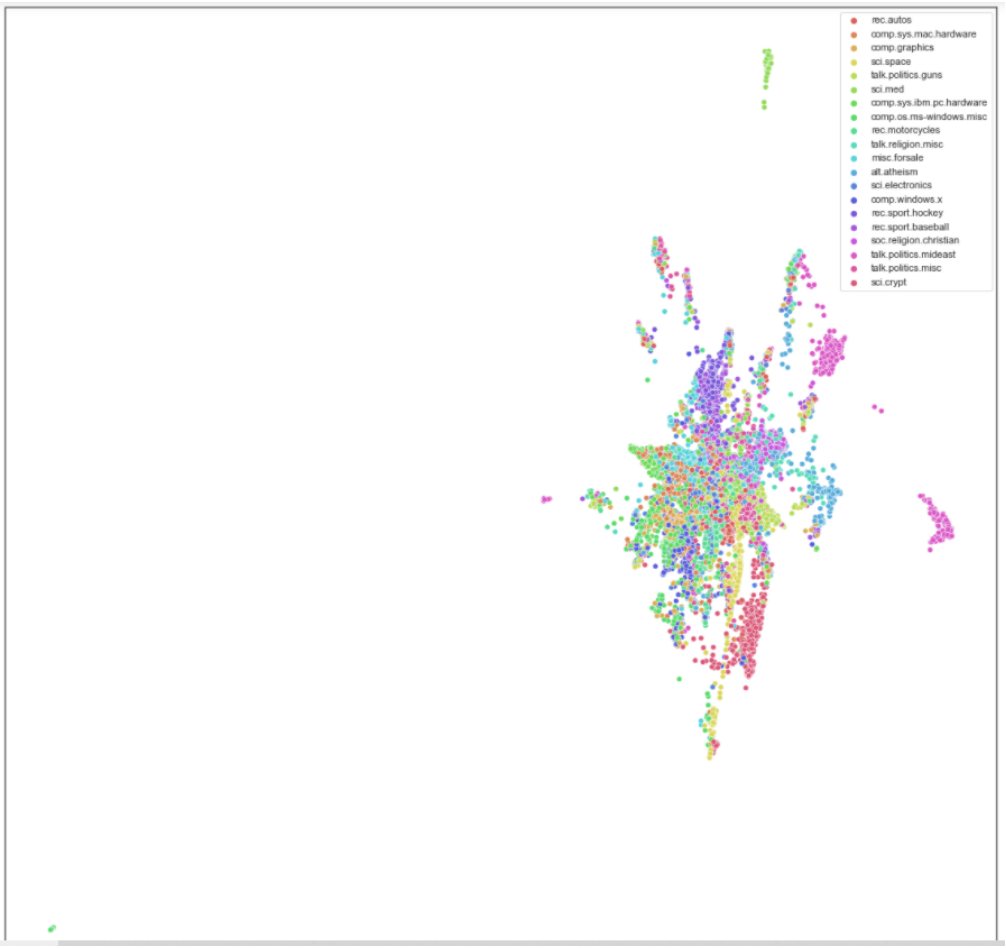
Results:

Method	TriMap	t-SNE	UMAP	LargeVis	PCA
Time (s)	33.23	45.51	31.98	425.17	0.18
Average distance inside/outside class ratio	0.25	0.34	0.23	0.27	0.57

Interpretation
For mnist dataset, the fastest method is PCA with a huge difference in comparison to other methods. Although, PCA has the worst average distance score. Methods with the best score were UMAP and TriMap. Every method except PCA classifies digits correctly.

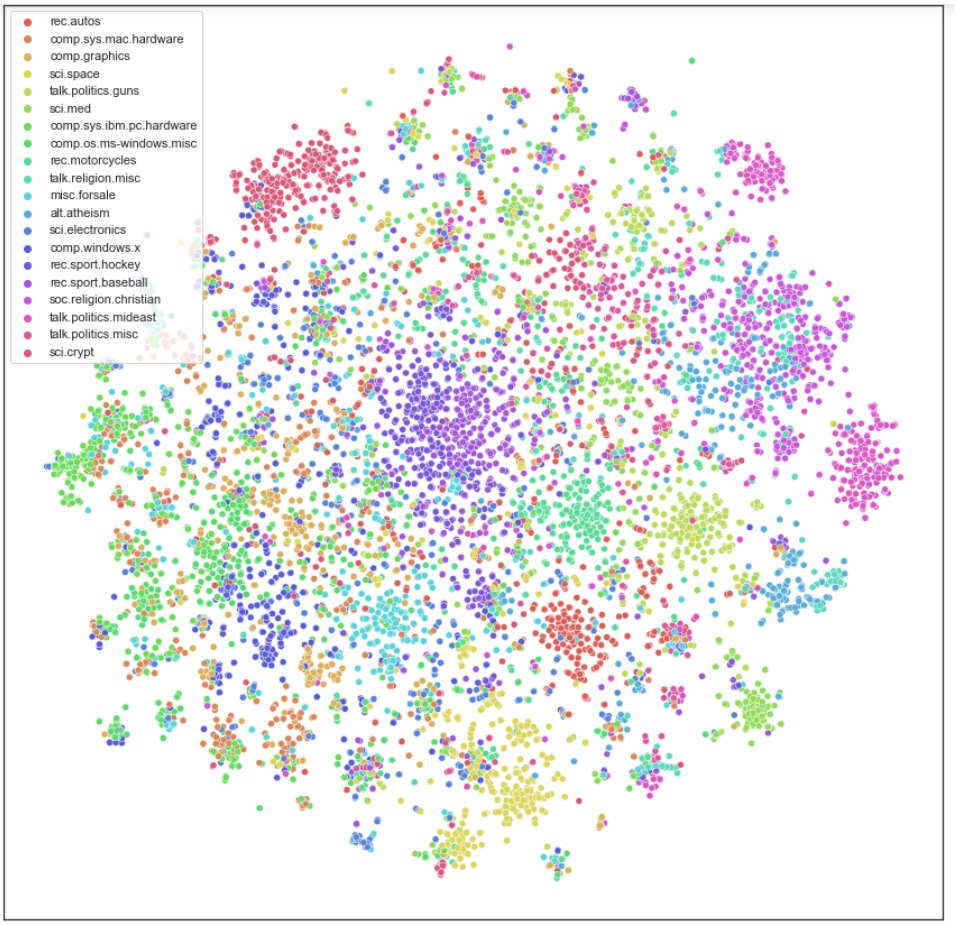
Comparison for the 20 News Group dataset

TriMap:

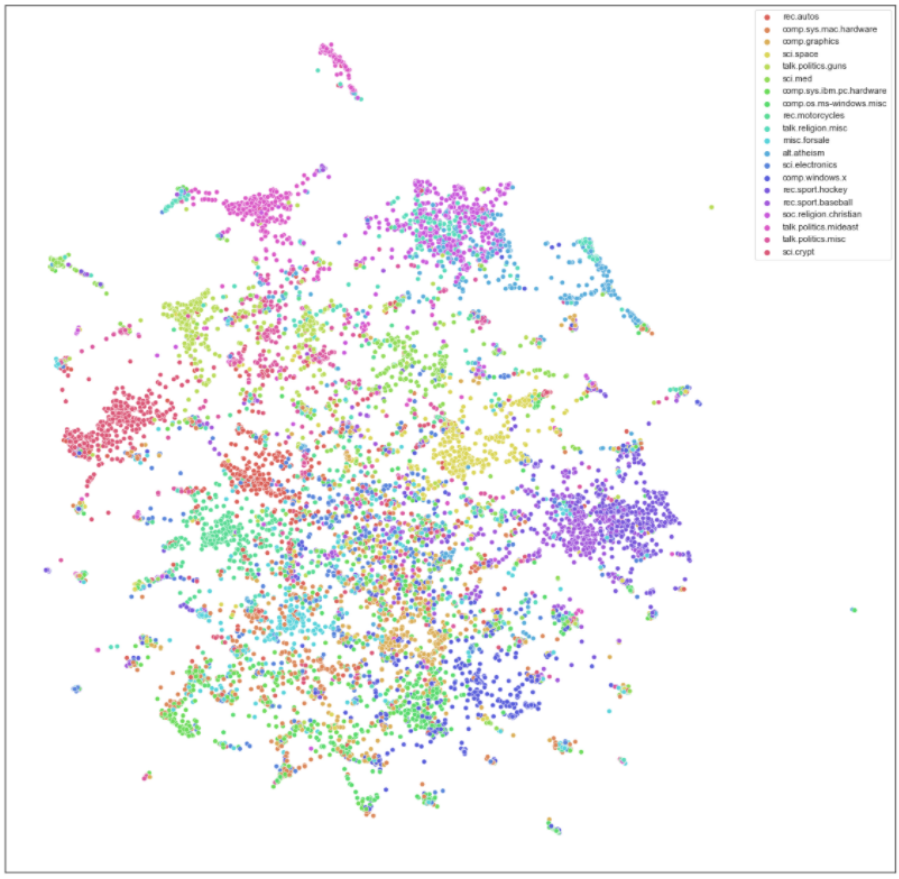


Global Score : 0.99

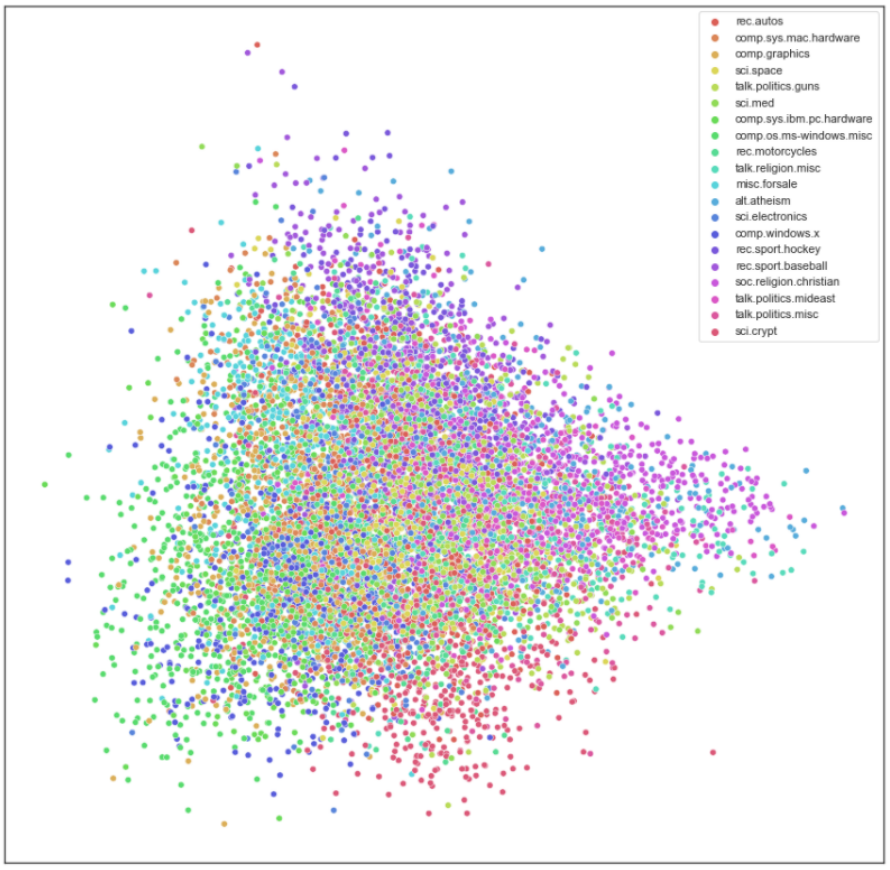
t-SNE:


rec.autos
comp.sys.mac.hardware
comp.graphics
sci.space
talk.politics.guns
sci.med
comp.sys.ibm.pc.hardware
comp.os.ms-windows.misc
rec.motorcycles
talk.religion.misc
misc.forsale
alt.atheism
sci.electronics
comp.windows.x
rec.sport.hockey
rec.sport.baseball
soc.religion.christian
talk.politics.mideast
talk.politics.misc
sci.crypt

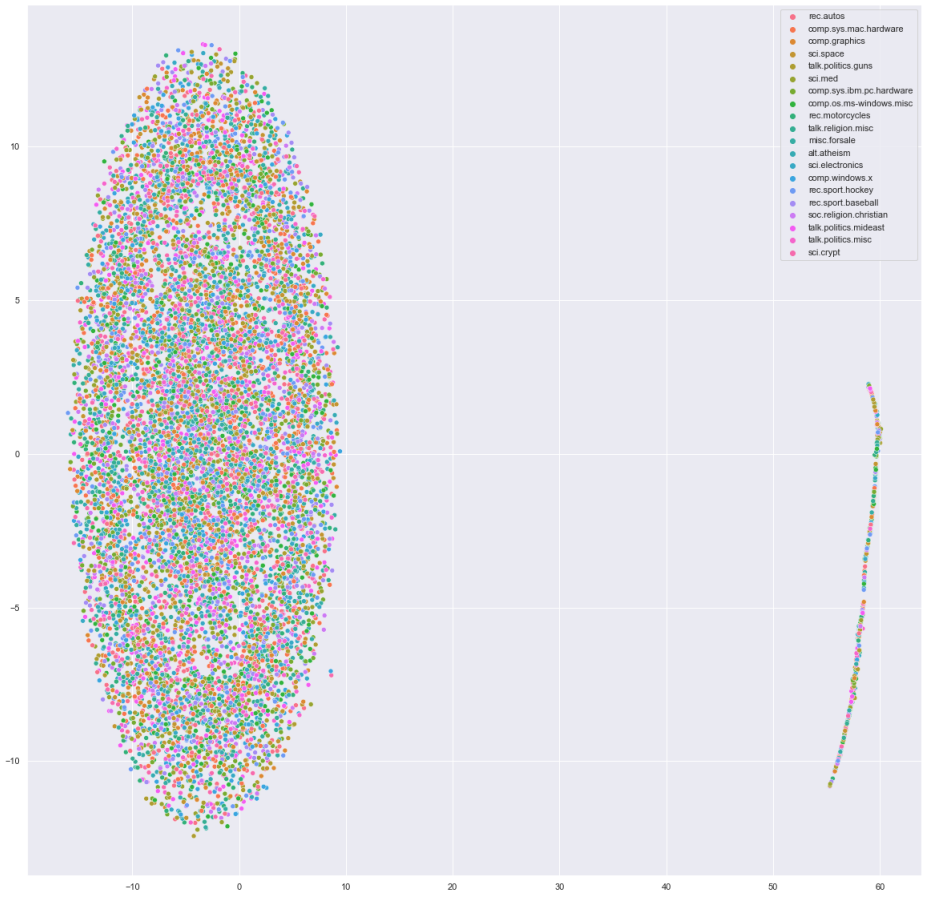
UMAP:


rec.autos
comp.sys.mac.hardware
comp.graphics
sci.space
talk.politics.guns
sci.med
comp.sys.ibm.pc.hardware
comp.os.ms-windows.misc
rec.motorcycles
talk.religion.misc
misc.forsale
alt.atheism
sci.electronics
comp.windows.x
rec.sport.hockey
rec.sport.baseball
soc.religion.christian
talk.politics.mideast
talk.politics.misc
sci.crypt

PCA:


rec.autos
comp.sys.mac.hardware
comp.graphics
sci.space
talk.politics.guns
sci.med
comp.sys.ibm.pc.hardware
comp.os.ms-windows.misc
rec.motorcycles
talk.religion.misc
misc.forsale
alt.atheism
sci.electronics
comp.windows.x
rec.sport.hockey
rec.sport.baseball
soc.religion.christian
talk.politics.mideast
talk.politics.misc
sci.crypt

LargeVis:



Results:

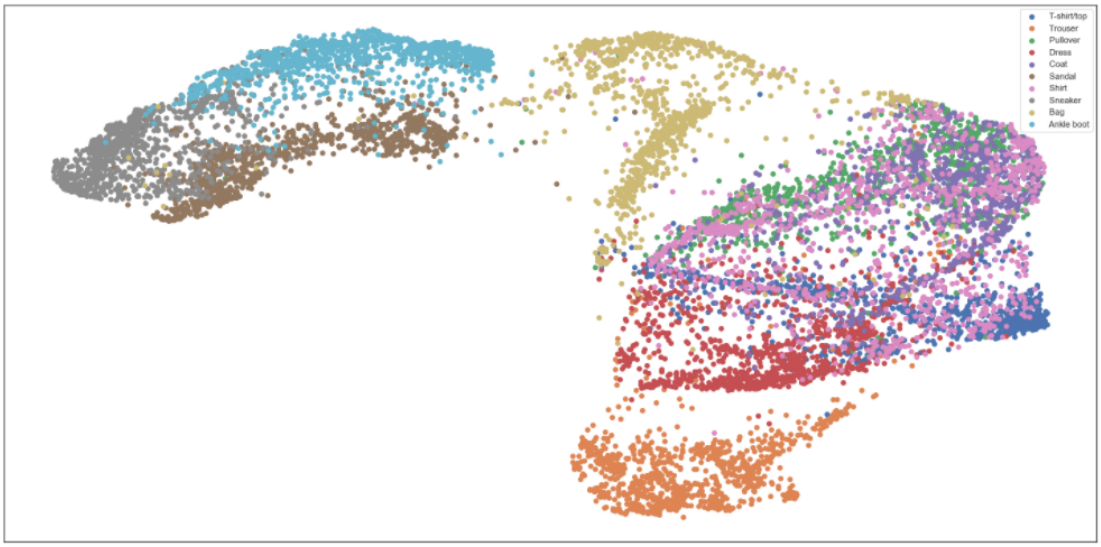
Method	TriMap	t-SNE	UMAP	PCA
Time (s)	38.09	57.50	64.97	1.08
Average distance inside/outside class ratio	0.67	0.64	0.63	0.74

Interpretation

On news dataset the used methods do not work well. Those methods are not good with text data preprocessing. Again, PCA method was the fastest but with worst score. The other methods have comparable scores but among them the best time achieve TriMap. Visualization of datasets was the best for TriMap, UMAP and t-SNE. PCA and LargeVis are unreadable.

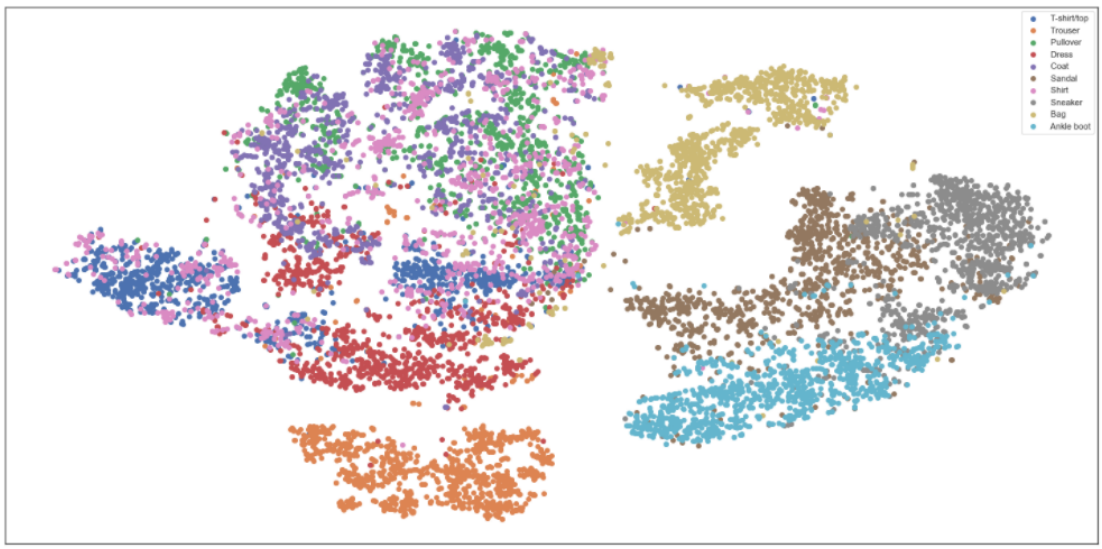
Comparison for the Fashion Mnist dataset

TriMap:



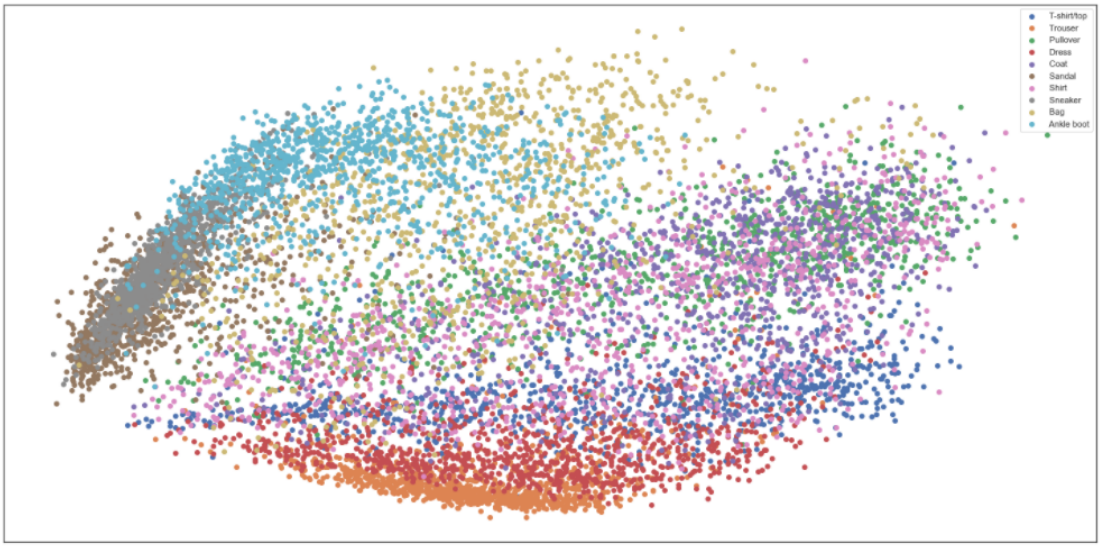
Global Score : 0.88

t-SNE:

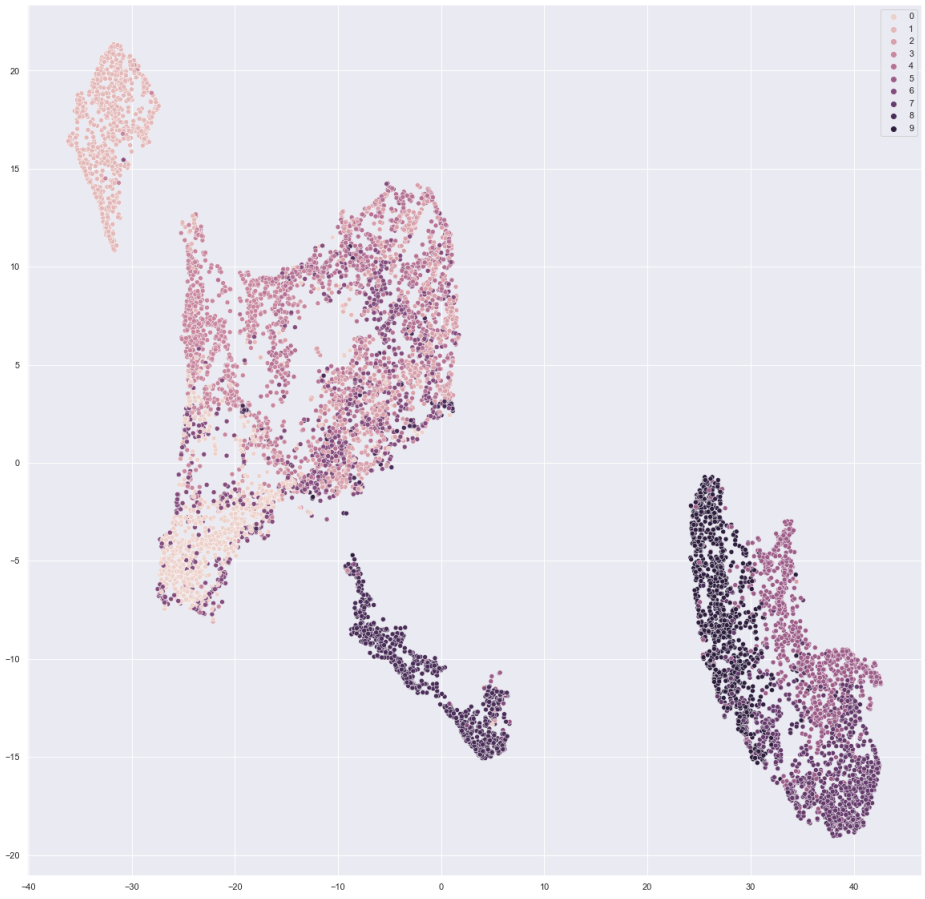


UMAP:

PCA:



LargeVis:



Global Score : 0.96

Results:

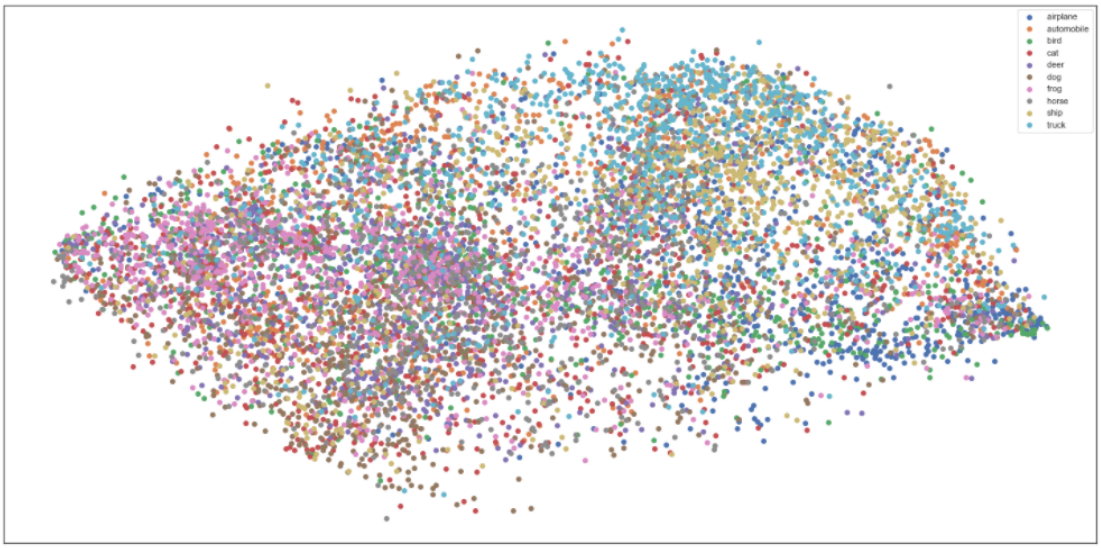
Method	TriMap	t-SNE	UMAP	LargeVis	PCA
Time (s)	49.84	52.99	95.78	461.62	0.47
Average distance inside/outside class ratio	0.31	0.39	0.25	0.23	0.51

Interpretation
Once again PCA method has the best time with the worst distances score. The best distances score achieve LargeVis but this method takes really long to run. Trimap method has good scores with one of the best times. Every

method separate each class properly but PCA is less readable than other methods.

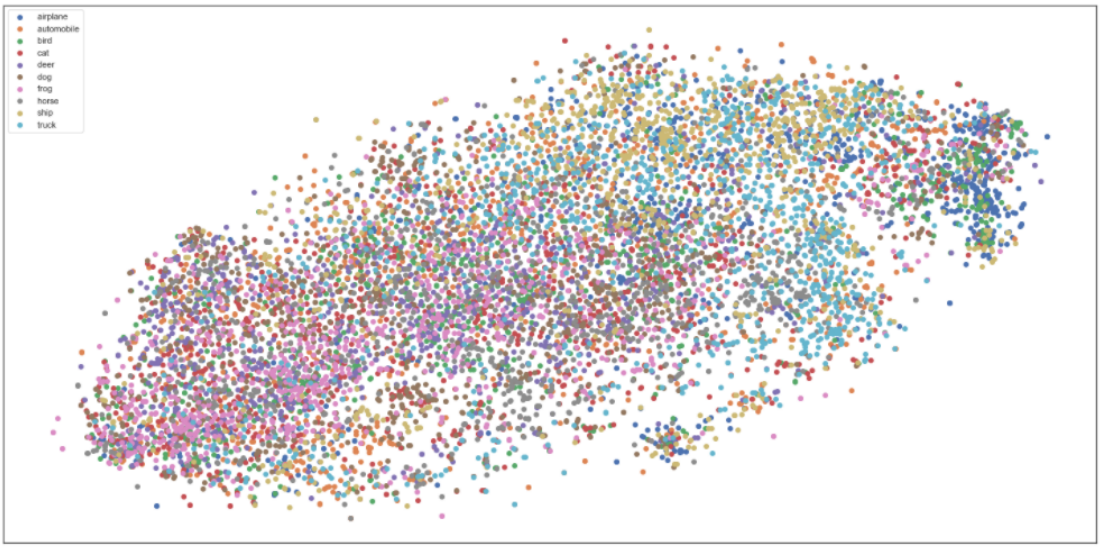
Comparison for the Cifar-10 dataset

TriMap:

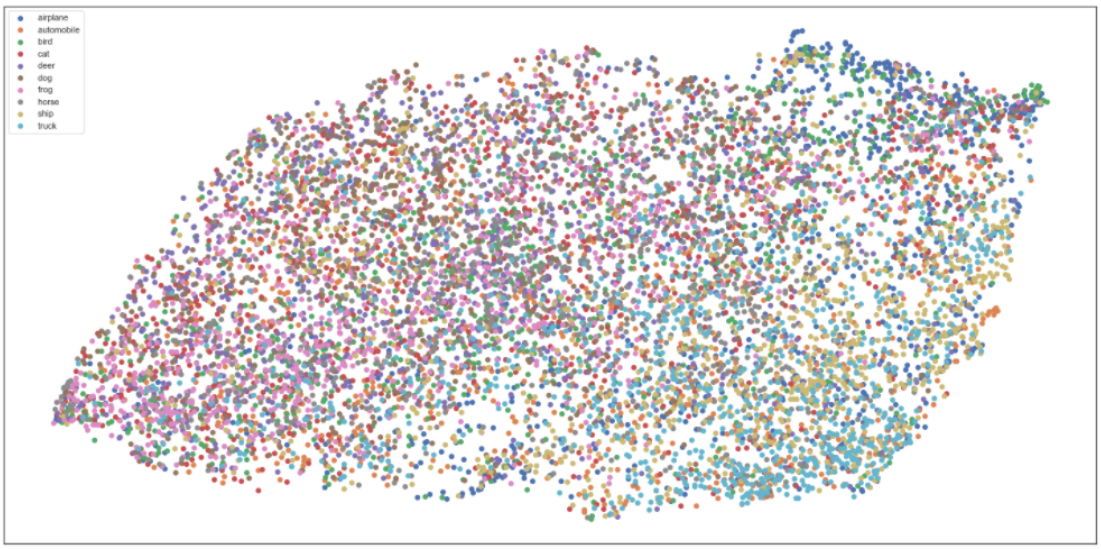


Global Score : 0.96

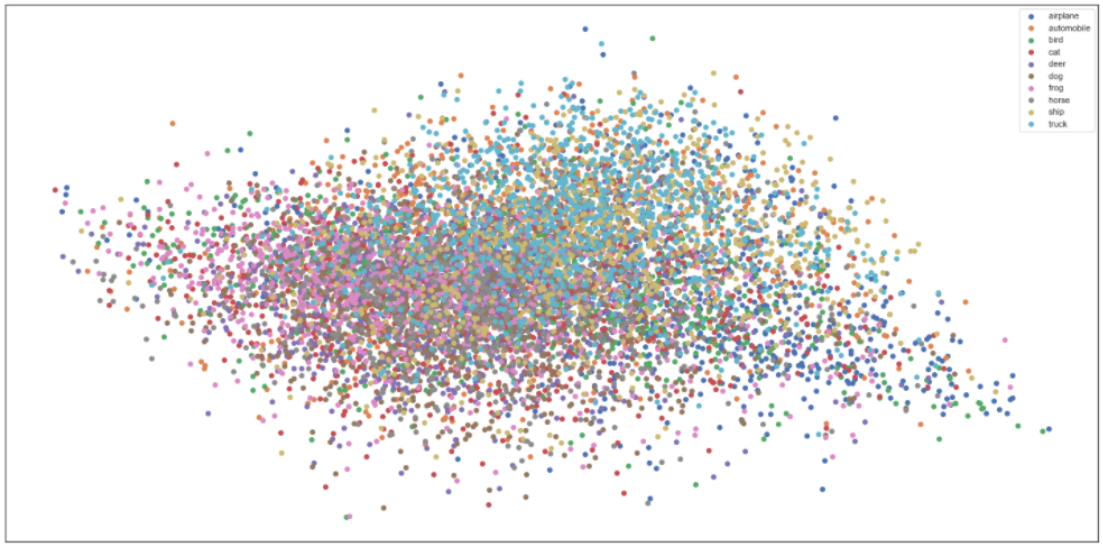
t-SNE:



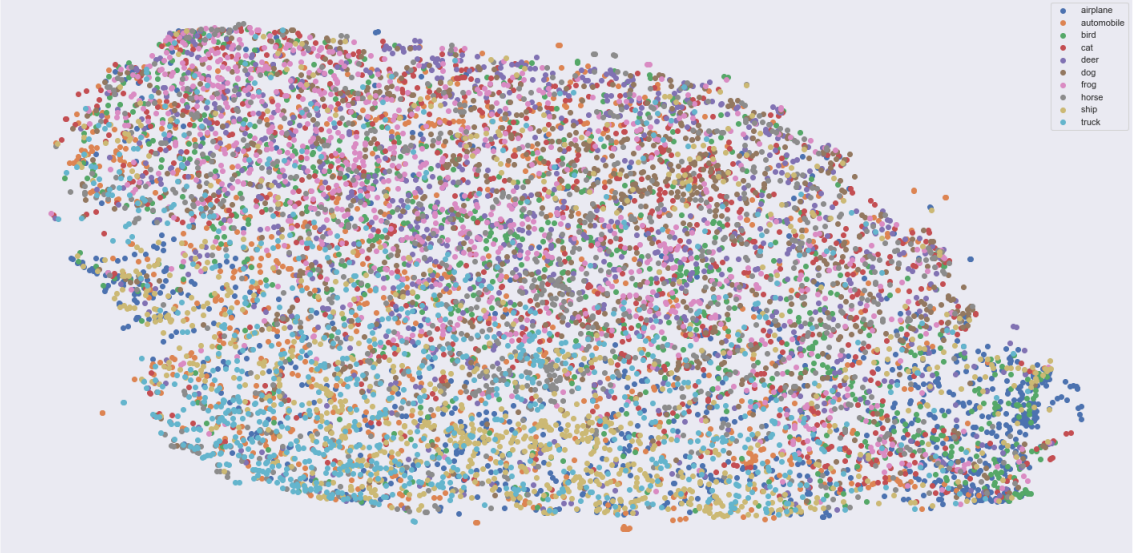
UMAP:



PCA:



LargeVis:



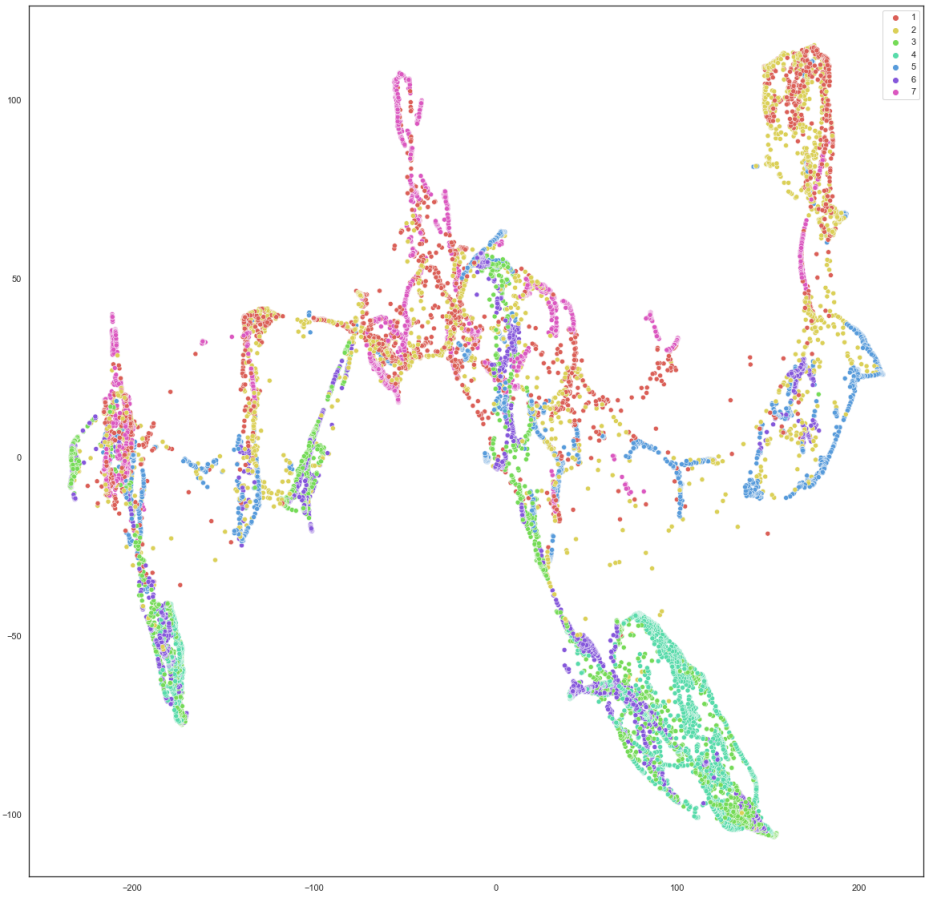
Results:

Method	TriMap	t-SNE	UMAP	LargeVis	PCA
Time (s)	45.10	65.03	74.83	528.48	3.36
Average distance inside/outside class ratio	0.89	0.89	0.89	0.89	0.91

Interpretation
Cifar10 was the hardest dataset for preprocessing with methods we used. Score of each method are unsatifaying. There is no possibility to outline groups of similiar observation.

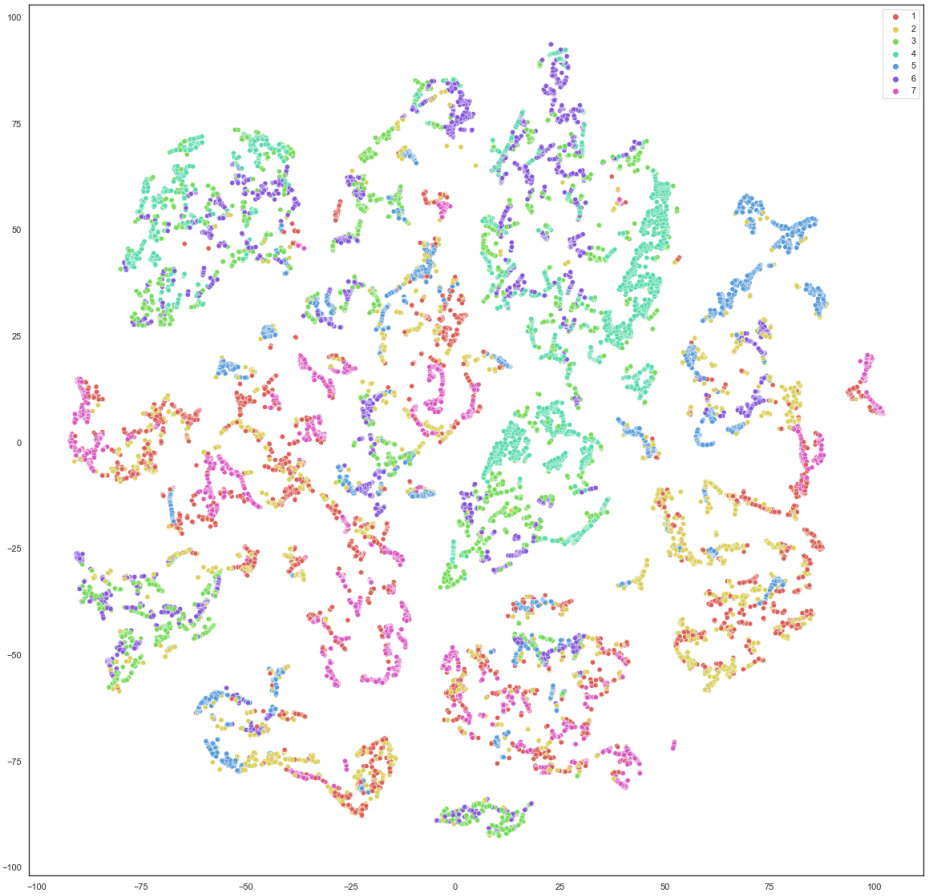
Comparison for the Covertype dataset

TriMap:

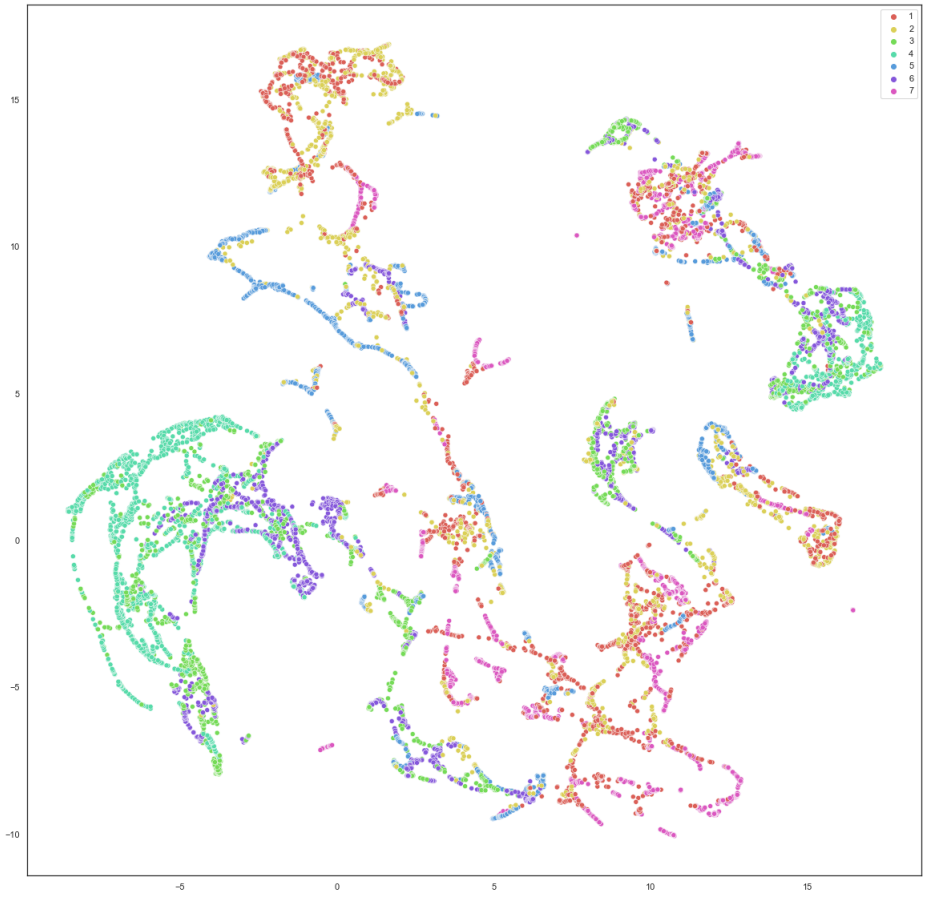


Global Score : 0.23

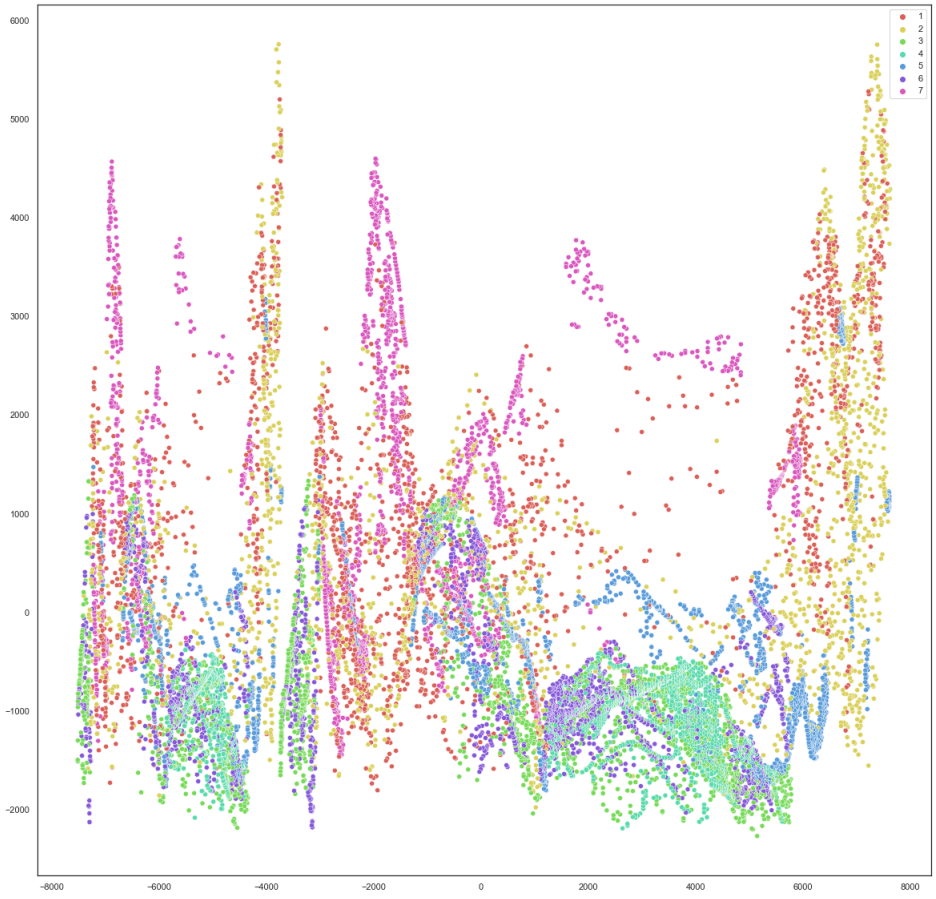
t-SNE:



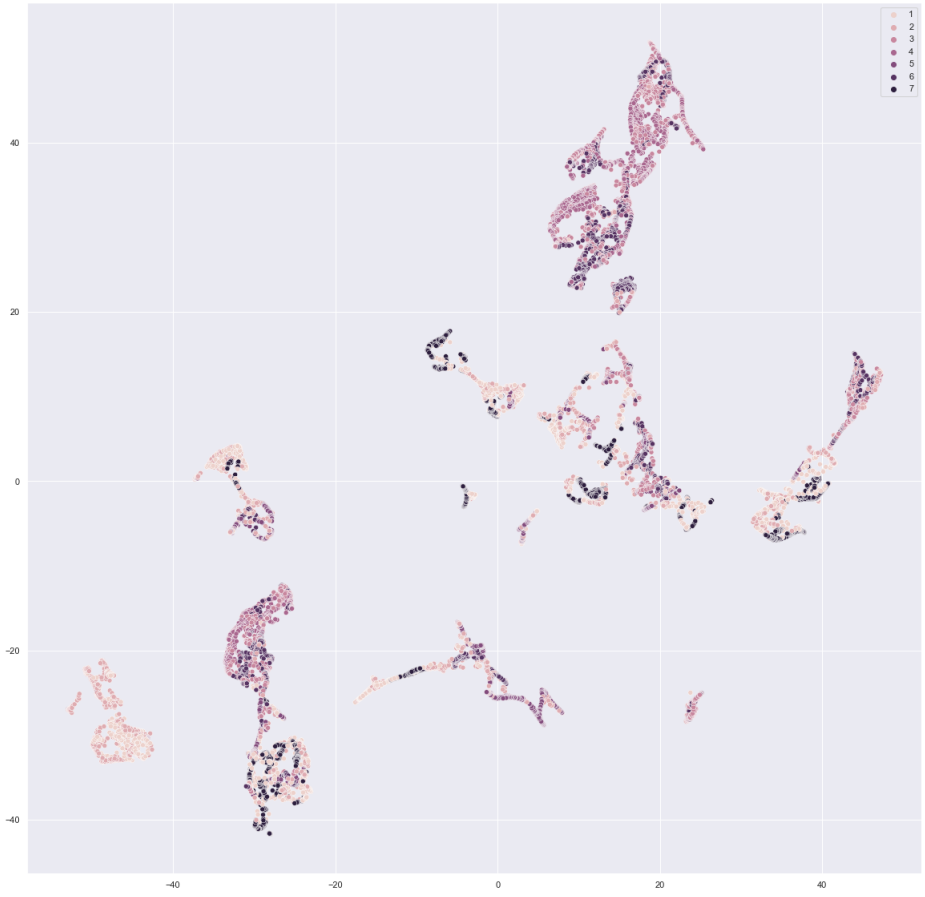
UMAP:



PCA:


6000
5000
4000
3000
2000
1000
0
−1000
−2000
−8000
−6000
−4000
−2000
0
2000
4000
6000
8000
1
2
3
4
5
6
7

LargeVis:



Results:

Method	TriMap	t-SNE	UMAP	LargeVis	PCA
Time (s)	65.4	65.19	29.88	439.68	1.31
Average distance inside/outside class ratio	0.90	0.89	0.88	0.83	0.89

Interpretation
Average distance is not the best measure for this dataset. As we can see results of this measure are not good but groups of similiar observations are easily visible. Each method has similiar score. As always PCA was the fastest. Best score for this dataset achieved LargeVis.

Conclusions
PCA is always the fastest in comparison to other methods but average distance and visualization are the worst. LargeVis is method which operation time is incomparably worse than methods such as UMAP or t-SNE. t-SNE, UMAP and TriMap works pretty well at this datasets - each of them achiving good score and good time. However, TriMap scales better for big datasets.