Hierarchical Nearest Neighbor Graph Embedding for Efficient Dimensionality Reduction

Supplementary Material

A. Datasets description

We used 9 datasets ranging from 1440 to 11 million samples in 28 to 16384 dimensions. Table 1 in the main paper provides a good overview. Here we include more details for each dataset:

Higgs [1]: Higgs bosons montecarlo simulations of kinematic properties measured by the particle detectors in the accelerator. The Higgs dataset has 11 million samples in 28 dimensions.

Google News [7]: is a dataset of 3 million words and phrases derived from a sample of Google News documents and embedded into a 300 dimensional space via word2vec. It is an unlabelled dataset and therefore we can not compute metrics on it.

COIL 20 [8]: is a set of 1440 greyscale images consisting of 20 objects under 72 different rotations spanning 360 degrees. Each image has size 128x128 pixels and is treated as a single 16384 dimensional vector for the purposes of computing distance between images.

CIFAR-10 [4]: 32×32 pixels RGB images of 10 object classes. We treat each image as a $32 \times 32 \times 3 = 3072$ dimensional pixel vector.

Fashion MNIST [9]: or F-MNIST is a dataset of 28x28 pixel grayscale images of fashion items (clothing, footwear and bags). There are 10 classes and 70000 images in total. Each image is treated as a ($28 \times 28 = 784$) dimensional pixel vector.

ImageNet [3]: The ILSVRC2012 ImageNet dataset. 1.2 million images belonging to 1000 classes. Each image is represented by a 2048 dimensional feature vector extracted using a trained ResNet-50 network.

BBT (season 1, episodes 1 to 6) and Buffy (season 5, episodes 1 to 6) are challenging video face identification/clustering datasets. They are generated based on the videos of the sitcoms *The Big Bang Theory* and *Buffy the Vampire Slayer* on small cast lists, for BBT: 5 main casts, and for Buffy: 6 main casts. The data comprises of detected faces in video frames represented by a trained (on the VGG-Face dataset) ResNet-50 model. BBT has a total of 199346 frames and Buffy has 206254 frames. The extracted ResNet-50 feature vectors are 2048 dimensional. The data for BBT and Buffy are provided by [2]. Since the feature vectors are obtained from a trained CNN model on face dataset, one should expect to see 5 main clusters (for BBT) and 6 clusters (for Buffy) in the embedding space.

MNIST & MNIST-8M [5,6]: is a dataset of 28x28 pixel grayscale images of handwritten digits. There are 10 digit classes (0 through 9). We use two variants of MNIST. MNIST 70000 (train + test) images and MNIST-8M [6] 8.1 million total images obtained by applying random transformations to each MNIST image. Each image is treated as a pixel concatenated 784 dimensional vector.

	1-NN ACC	Trustworthiness	CTA	Runtime
UMAP	0.238	0.752	0.595	1min 47s
FIt-SNE	0.526	0.933	0.637	1min 8s
h-NNE (ours)	0.518	0.937	0.618	19s

Table 1. New points projection: Performance and time comparison of new points projection using ImageNet validation set (50K samples) as new points and h-NNE, UMAP and FIt-SNE built on Imagenet Train set.

B. Projecting new points

As mentioned at the end of section 3.3 of the main paper, we can easily project new points by following the original algorithm. Here we perform a comparison with two of the other methods which also provide the option to project new points, namely FIt-SNE and UMAP. To create a realistic scenario, we create an embedding of the training part of the ImageNet dataset and based on the structure learned on it we project the test part of ImageNet. In Table 1, we can see the performance of different projections of ImageNet validation set and the corresponding time required to project the 50000 vectors of dimension 2048.

As we can see, both the performance and speed of our method is preserved on the new points projection.

C. Preliminary projection: random initialization versus PCA

As described in section 3.1 of the main paper, We initialize with a preliminary projection using PCA. To reduce the computational complexity of PCA we proposed to use PCA on a reduced number of samples by using the centroids obtained on a predefined level of the h-NN graph. Here we provide further analysis and comparison on this initialization. We include an ablation using 6 medium scale datasets (up till 1 million samples). We show the impact on performance using 4 initialization methods:

1. Random init: We start with random uniformly distributed d-dimensional points.

2. Random Projection init: We project the original data in $\mathbb{R}^D$ to $\mathbb{R}^d$ with d random uniform vectors of

D-dimension each.

For random projections initialization we compute results over 5 runs and report the average.

3. `Full PCA init`: We use PCA on the full data to obtain the preliminary projection.

4. `PCA on centroids init`: proposed initialization used in the paper: We use PCA on the $\sim$ 1000 points/centroids of the full data from our built 1-NN hierarchy graph to obtain a faster preliminary projection.

In Table 2 we show a comparison on both local (Trustworthiness and KNN) and global (Centroid Triplet Accuracy (CTA)) structure preservation metrics. As seen the random initialization provides similar local structure preservation as the more time consuming PCA projection. However the random projections can not recover the global structure well in comparison (lower CTA scores). This is also depicted in a visual comparison of projections in Figure 2. Figure 2 shows our projections on the `BBT` dataset which has 5 classes. Both PCA and faster PCA on Centroids has very similar outputs whereas the projection based on random inits. shows noisier global structure (splitting the same class).

D. Impact of point cluster inflation for visualization

In the main paper near the end of section 3.3, we described the use of a single linear projection for all points can result to stretched point clusters when they are not well aligned to the global principal components. We add an option to inflate potentially squeezed point clusters using six local rotations with equally distanced angles in the interval $[0, \frac{\pi}{2}]$, followed by a scaling and the inverse rotation. This has no impact on the performance and only makes visualization more appealing. Here in Figure 3 we show a visual comparison on the `Fashion MNIST` dataset as an example.

E. KNN Accuracy Comparison

Figure 1 shows the KNN performance of methods on varying number of K on all datasets. As seen on all datasets h-NNE performs on par with the other methods on the whole range of k-neighbour values.

F. Clustering Properties

The method is built on the principle of grouping data points together in a hierarchical way which captures clustering properties. To demonstrate, we cluster the large scale datasets before and after the projection with k-means and compare to their groundtruth clusters. Table 3 shows the NMI scores of clustering in the original high-dim feature space (Original-dim) and in the 2-dim projection space of different methods. As seen, in comparison h-NNE maintains a high NMI score that shows its ability to preserve the clusters better.

	1-NN ACC	Trustworth.	CTA
COIL20			
Random	0.988	0.988	0.577
Random Proj	0.991	0.992	0.666
PCA-full	0.989	0.993	0.799
PCA-centroids	0.990	0.994	0.799
MNIST			
Random	0.946	0.970	0.630
Random Proj	0.960	0.982	0.715
PCA-full	0.962	0.984	0.752
PCA-centroids	0.965	0.983	0.671
Fashion MNIST			
Random	0.782	0.926	0.564
Random Proj	0.820	0.955	0.688
PCA-full	0.823	0.976	0.896
PCA-centroids	0.826	0.981	0.925
BBT			
Random	0.985	0.946	0.529
Random Proj	0.990	0.971	0.618
PCA-full	0.992	0.974	0.703
PCA-centroids	0.992	0.982	0.644
Buffy			
Random	0.967	0.951	0.621
Random Proj	0.976	0.968	0.516
PCA-full	0.982	0.975	0.867
PCA-centroids	0.975	0.976	0.857
ImageNet			
Random	0.436	0.857	0.589
Random Proj	0.560	0.931	0.624
PCA-full	0.567	0.933	0.654
PCA-centroids	0.557	0.928	0.604

Table 2. Ablation: Impact of preliminary projection.

data / GT clusters	Original-dim	FIt-SNE[18]	UMAP[20]	PaCMAP[31]	h-NNE
BBT / k=5	89.7	55.3	54.2	61.4	87.6
Buffy / k=6	76.1	53.2	48.1	41.5	74.7
ImageNet / k=1000	74.3	71.8	70.7	63.9	70.9
MNIST8M / k=10	61.3	55.5	-	57.7	59.6

Table 3. NMI scores of k-means clustering

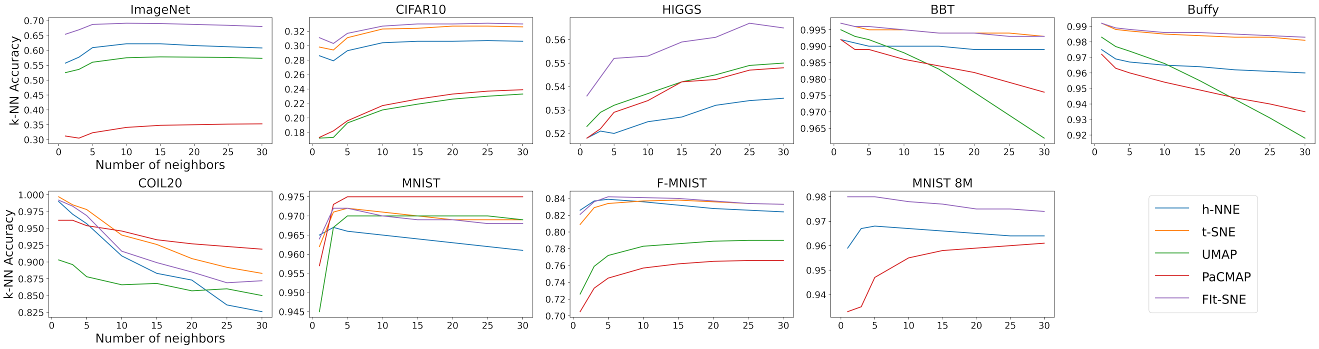

Figure 1. Local structure preservation: k-NN classifier accuracy on different datasets with increasing k neighbours.

G. Visual comparison

We end our supplementary with a visual comparison of the embeddings in two dimensions for a sample of real-world datasets which can be viewed in figure 4. In each plot the colors are based on the labels of each dataset with the exception of Google News for which no labels are available. Also note that the Google News plot is missing for the case of t-SNE due to it not finishing the projection of 3 million points in a reasonable time interval.

References

[1] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):1–9, 2014. 1

[2] Martin Bäuml, Makarand Tapaswi, and Rainer Stiefelhagen. Semi-supervised Learning with Constraints for Person Identification in Multimedia Data. In *CVPR*, 2013. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 1998. 1

[6] Gaëlle Loosli, Stéphane Canu, and Léon Bottou. Training invariant support vector machines using selective sampling. *Large scale kernel machines*, 2007. 1

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 1

[8] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). 1996. 1

[9] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 1

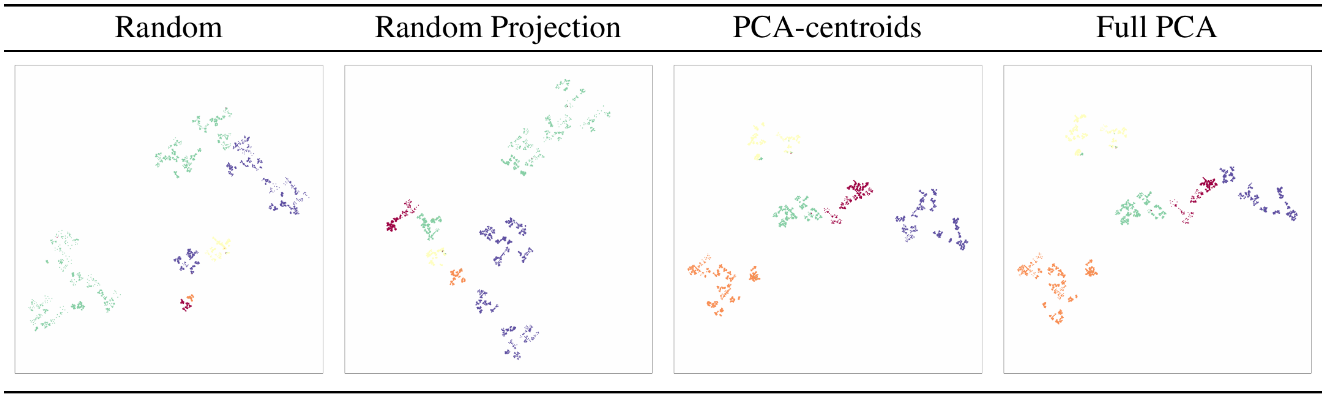

Figure 2. Impact of preliminary projection - Random initialization versus PCA on the BBT dataset

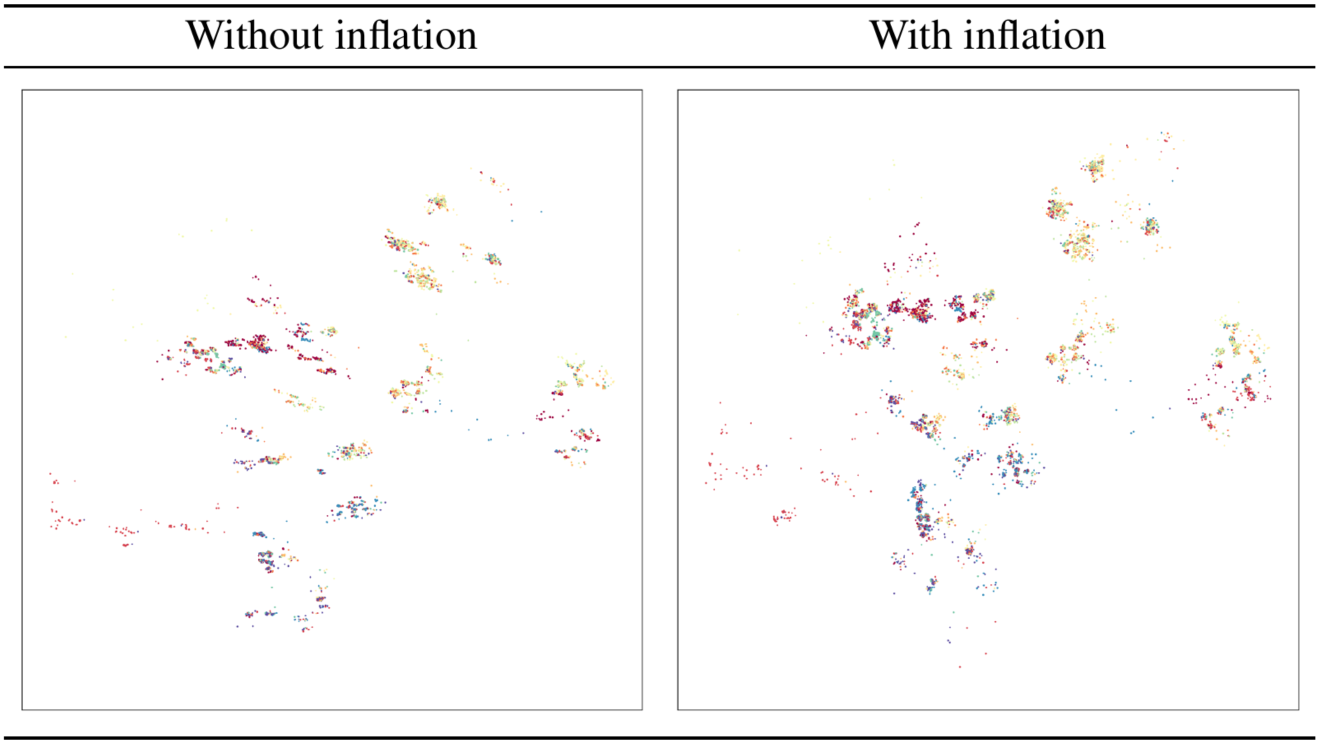

Figure 3. Impact of Point Cluster Inflation for visualization purposes on the Cifar10 dataset

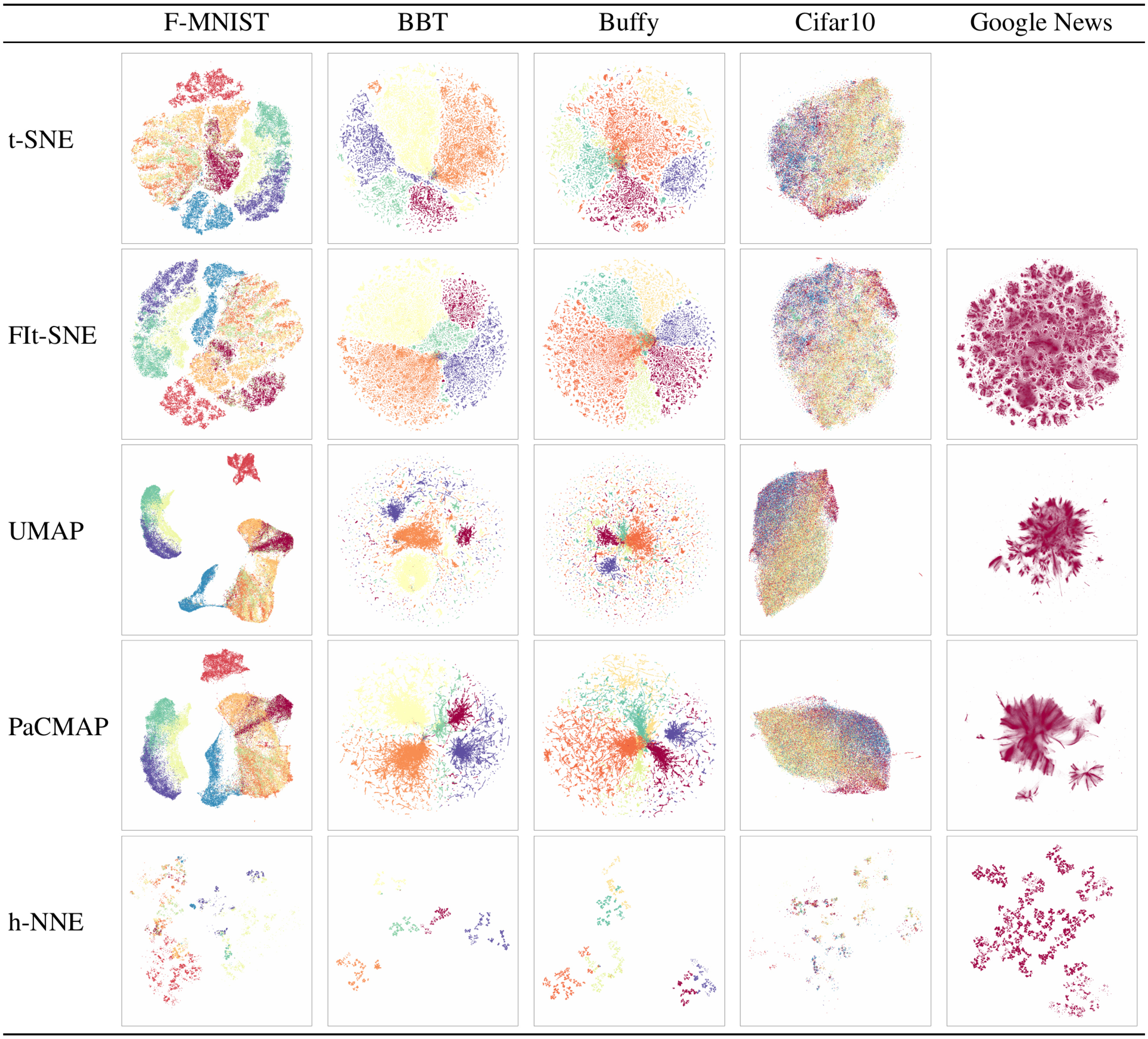

Figure 4. Visual comparison between h-NNE, t-SNE, FIt-SNE, UMAP and PaCMAP projections in 2D.