Challenging Euclidean Topological Autoencoders

Michael Moor
Dept. Biosystems (D-BSSE)
ETH Zurich & Swiss Institute
of Bioinformatics, Switzerland
michael.moor@bsse.ethz.ch

Max Horn
Dept. Biosystems (D-BSSE)
ETH Zurich & Swiss Institute
of Bioinformatics, Switzerland
max.horn@bsse.ethz.ch

Karsten Borgwardt
Dept. Biosystems (D-BSSE)
ETH Zurich & Swiss Institute
of Bioinformatics, Switzerland
karsten.borgwardt@bsse.ethz.ch

Bastian Rieck
Dept. Biosystems (D-BSSE)
ETH Zurich & Swiss Institute
of Bioinformatics, Switzerland
bastian.rieck@bsse.ethz.ch

Abstract

Topological autoencoders (TopoAE) have demonstrated their capabilities for performing dimensionality reduction while at the same time preserving topological information of the input space. In its original formulation, this method relies on a Vietoris–Rips filtration of the data space, using the Euclidean metric as the base distance. It is commonly assumed that this distance is not sufficiently powerful to capture salient features of image data sets. We therefore investigate *alternative* choices of distances in the data space, which are generally considered to be more faithful for image data in comparison to the pixel distance. In our experiments on real-world image datasets, we find that the Euclidean formulation of TopoAE is surprisingly competitive with more elaborate, perceptually-inspired image distances.

1 Introduction

Topological autoencoders [6] were recently introduced as a novel dimensionality reduction method that satisfies topological constraints. Briefly put, their underlying concept is a loss term that aims to harmonise the topology of the input space and the topology of the learnt latent space. The loss term requires information about the topological features of each batch, provided in the form of *persistence diagrams*, calculated from a Vietoris–Rips complex [9]. Figure 1 provides an overview of the architecture.

In their experiments, Moor et al. [6] discussed the benefits of topology-constrained latent visualisations by means of various examples. While their method exhibits good performance on some datasets, large-scale image datasets such as CIFAR-10 do not result in directly-interpretable latent embeddings, even though quality metrics indicate that the quality of the topologically-constrained latent embeddings is higher than for other embedding methods.

The choice of a pixel-based Euclidean distance, which is unable to gauge the *perceived* similarity between images [2], might not be ideal. The goal of this paper is therefore to assess how changing the distance metric in the data space impacts the quality of the latent embeddings, measured using various dimensionality reduction quality metrics. In particular, we will investigate alternative formulations of topological autoencoders by employing distances that are assumed to be more faithful for the analysis of image data.

Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

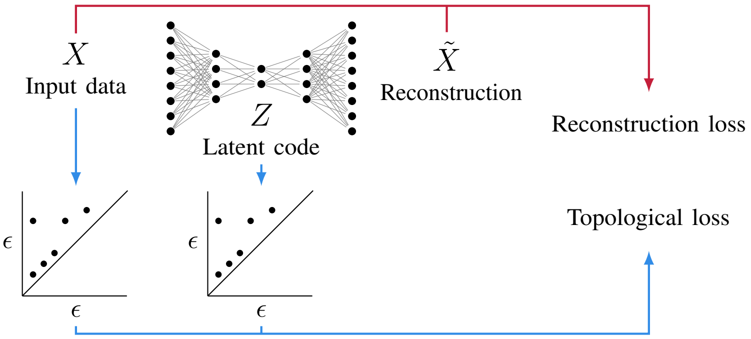


Figure 1: A schematic overview of the topological autoencoder calculation process. Using a mini-batch X of the data space $\mathcal{X}$, an autoencoder is trained to reconstruct X, resulting in a reconstructed mini-batch $\tilde{X}$. The crucial novel ingredient is a topological loss, which is calculated based on the topological differences between persistence diagrams (obtained from a mini-batch X and its corresponding latent code Z). The objective of the topological loss term is to impose a constraint on the autoencoder so that it is incentivised to preserve topological features of the data space within the respective latent representations. We refer the reader to Moor et al. [6] for more details.

2 Background

Given a point cloud $X := \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ and a distance $d\colon X \times X \to \mathbb{R}$, let $\mathbf{A} \in \mathbb{R}^{n \times n}$ refer to the distance matrix of X with $\mathbf{A}_{ij} = d(x_i, x_j)$. For $0 \leq \epsilon < \infty$, the Vietoris–Rips complex of X, denoted as $\mathfrak{R}(\mathbf{A})$, at scale ϵ contains all simplices of X whose elements $\{x_1, x_2, \ldots\}$ satisfy $d(x_i, x_j) \leq \epsilon$ for all i, j. Moor et al. [6] treat each mini-batch as a point cloud X paired with an encoded mini-batch Z. On top of the autoencoder architecture, which is illustrated in Figure 1, both point clouds X, Z are separately subjected to a Vietoris–Rips filtration to compute persistence diagrams. These diagrams approximate topological features of both the data and the latent spaces, which are then aggregated in a differentiable loss term. Even though it is common practice to use the Euclidean distance as base distance for the Vietoris–Rips complex, it has been previously shown that other distances may be used alternatively, and they do not necessarily have to satisfy the requirements of a metric [10].

3 Methods

While TopoAE employs the Euclidean distance $d(x_i, x_j) = ||x_i - x_j||_2$ in both the data and latent spaces, we investigate two alternative approaches for measuring distance in the data space, which are based on (i) random convolutions, and (ii) perceptual similarity . We note that there are alternative options of handling image data in practice, including specialised variants of convolutional neural networks [1, 4]. In this paper, we are primarily interested in performing an ablation study of the method proposed by Moor et al. [6], so we defer the comparison with computationally more involved architectures to future work.

3.1 Random Convolutions

Convolutional neural networks (CNNs) exhibit an inductive bias which is advantageous for the feature extraction and classification of images [5]. Notably, even randomly initialised convolutional layers can be used as rich feature extractors [7, 11]. We hypothesise that basing the distance of the *data* space on feature maps of untrained CNNs could preserve topological features of the input space. For this, let

$$d(x_i, x_j) = ||\mathrm{vec}(\mathcal{F}(x_i)) - \mathrm{vec}(\mathcal{F}(x_j))||_1, \tag{1}$$

where $\mathcal{F}$ denotes the output (feature maps) of a CNN employing 3 convolutional layers with ReLU activations, $\mathrm{vec}(\cdot)$ the vectorisation operation, and $||\cdot||_1$ the ℓ_1 norm, which is used to deal with potentially high-dimensional feature spaces.

3.2 Perceptual Similarity

In their work, Zhang et al. [12] showed that the feature representations of deep neural networks are well-suited to quantify the similarity of images similarly to human perception. Furthermore, their work shows that the derived similarity scores outperform (in terms of being aligned with human judgement) hand-engineered counterparts that rely on low-level features of the image or random projections using untrained neural networks. We thus also consider a learnt similarity score based on deep neural network features called Learned Perceptual Image Patch Similarity (LPIPS) [12]. The LPIPS distance is derived by training a deep neural network, such that the difference of its features is in line with human perceptual scores. In the dataset used to train LPIPS, humans were asked to judge which of two distorted images is closer to a reference image, and the neural network was trained to match these estimates. In particular, the distance between two images x_i and x_j was computed via their features $\hat{y}_i$ and $\hat{y}_j$

$$d\left(x_i, x_j\right) := \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{y}^l_{ihw} - \hat{y}^l_{jhw}\right)\right\|_2^2, \tag{2}$$

which corresponds to a per-channel scaled squared ℓ_2 distance (with w_l being the scale factor) of the features averaged over all spatial locations. Here, H_l and W_l denote the number of elements in the height direction and width direction, respectively. For details on how the distances were calibrated to match human perception, we refer to Zhang et al. [12]. In our experiments, we relied on the weights of the VGG deep convolutional neural network architecture [8] and the implementation of LPIPS provided with the publication, both of which are publicly available [13].

4 Experiments

4.1 Experimental Setup

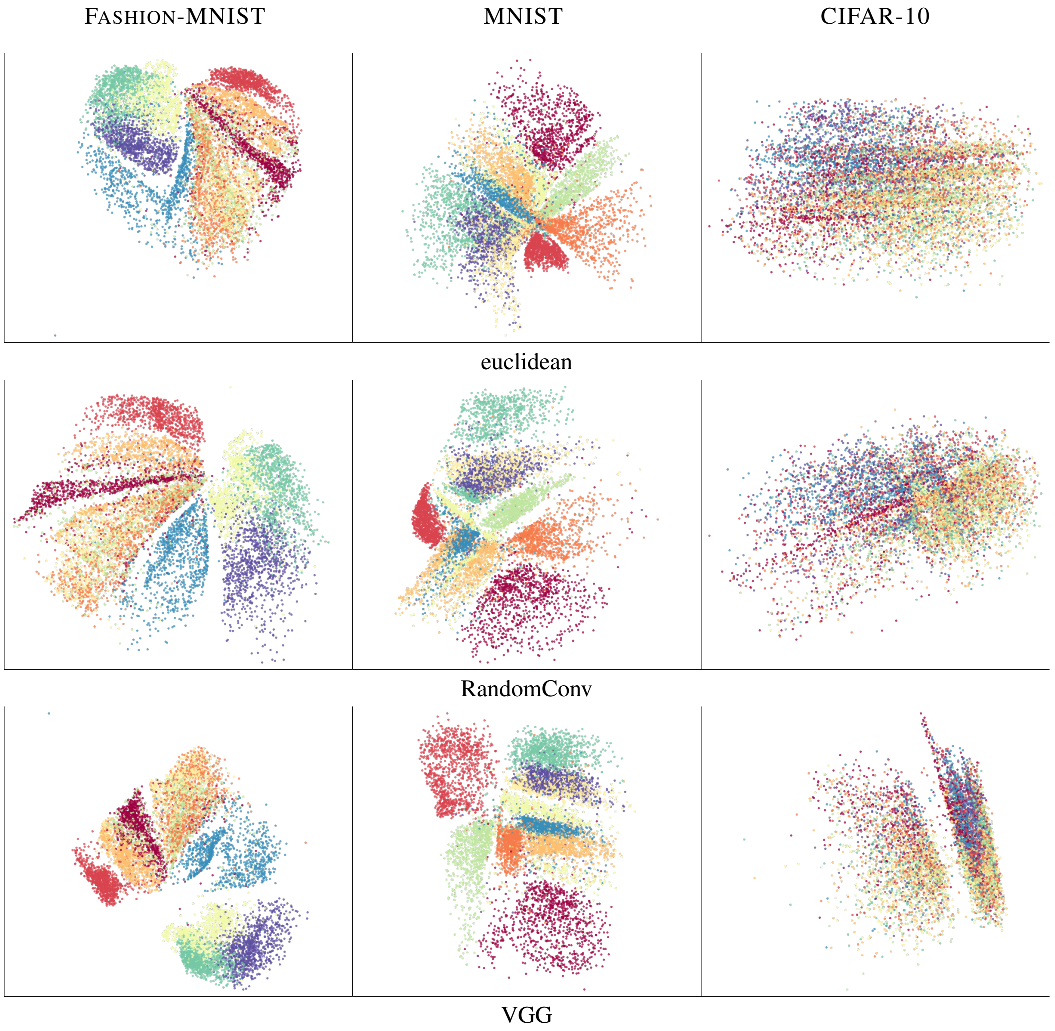
In our experiments, we consider three image datasets MNIST, FASHION-MNIST (abbreviated in the tables as F-MNIST), and CIFAR-10. We compare three variations of TopoAE by adapting the distance function used for the Vietoris–Rips filtration of the data space: (i) Euclidean (original formulation of Moor et al. [6]), (ii) RandomConv, which employs random convolutions, and (iii) VGG which uses the LPIPS distance with a pre-trained VGG network. In preliminary experiments, we identified that a strong topological regularisation parameter λ (see also Equation 1 of Moor et al. [6]) is essential for achieving good performance with TopoAE. Therefore, and also to ensure maximal comparability between the assessed methods, we used a fixed $\lambda = 2$, a mini-batch size of 64, a learning rate of 10^{-3} together with Adam, and a weight decay of 10^{-5}. Furthermore, we conform with the same MLP autoencoder architecture as used in Moor et al. [6] with $(1000-500-250)$ hidden units for the encoder, a two-dimensional bottleneck, and finally $(250 - 500 - 1000)$ hidden units for the decoder, which was inspired by DeepAE [3]. As for evaluation strategies, we follow Moor et al. [6] by visualising the low-dimensional embeddings as coloured by the class labels (which were not used for training, though), and report the same set of quantitative measures for dimensionality reduction, both evaluated on the predefined testing split. We make our code publicly available under https://github.com/BorgwardtLab/topo-ae-distances.

4.2 Results

Table 1 lists the quantitative results of our experiments. We observe that the standard Euclidean approach performs surprisingly well over various measures, foremost in terms of KL_1, ℓ-Cont and ℓ-RMSE. Overall, we find that VGG shows competitive performance with the Euclidean approach on MNIST and FASHION-MNIST, whereas RandomConv competes with the Euclidean formulation on CIFAR. We included ℓ-RMSE for the sake of completeness, but acknowledge that it serves foremost as a sanity check, since it computes a mean squared error of the Euclidean distance matrices of both spaces (which the Euclidean TopoAE is predisposed to minimise). Figure 2 shows the latent embeddings for all methods and datasets. Here, we observe a better separability of the classes for the two alternative approaches, in particular for MNIST and FASHION-MNIST, whereas CIFAR remains challenging for all investigated approaches.

Table 1: A summary of our quantitative evaluations. Please refer to Moor et al. [6] for details on the metrics.

Dataset	Metric	$\text{KL}_{0.01}$	$\text{KL}_{0.1}$	KL_{1}	ℓ-Cont	ℓ-MRRE	ℓ-Trust	ℓ-RMSE	Data MSE
	Model								
CIFAR	RandomConv	**0.573000**	0.026776	0.000519	0.884307	0.116578	**0.865882**	37.986596	**0.13567**
	VGG	0.747940	0.035598	0.000545	0.852621	0.134274	0.857791	38.790536	0.18757
	Euclidean	0.589208	**0.021210**	**0.000324**	**0.919586**	**0.107334**	0.851644	**37.628791**	0.14111
F-MNIST	RandomConv	0.421326	0.066843	0.001259	0.973635	0.025571	0.970463	22.048758	**0.10510**
	VGG	**0.380293**	**0.055905**	0.001052	0.977465	**0.023314**	**0.973091**	22.813647	0.10957
	Euclidean	0.391267	0.060878	**0.000981**	**0.980441**	0.025026	0.967511	**21.436654**	0.10919
MNIST	RandomConv	0.355687	0.130055	0.001506	0.921422	0.060333	0.930401	19.791551	0.14990
	VGG	0.674952	0.187962	0.001844	0.920575	**0.059505**	**0.931545**	20.174628	**0.14978**
	Euclidean	**0.343812**	**0.098164**	**0.000797**	**0.926969**	0.069833	0.906727	**18.976329**	0.15464



euclidean

RandomConv

VGG

Figure 2: Latent representations of the Fashion-MNIST (left column), MNIST (middle column), CIFAR-10 (right column) data sets. All methods used a fixed batch size of 64.

5 Conclusion

In this paper, we challenged the common assumption that Euclidean distances are ill-suited for measuring distances on image datasets. We found that even the use of elaborate distances based on perceptual similarity [12] does *not* result in marked improvements in terms of quality metrics—in fact, the Euclidean distance was competitive with all of these approaches. The resulting visualisations, however, appear to exhibit an improved separation when employing the alternative distances specific to the image domain.

For future work, we envision extending similar considerations to other domains, i.e. choosing domain-specific distance metrics for graphs, times series, and others. In the graph domain, for instance, it might be worthwhile to experiment with variants of the *graph edit distance* or other similarity measures.

References

[1] Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999, 2016.

[2] Dorin Comaniciu and Peter Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[3] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[4] A. Komarichev, Z. Zhong, and J. Hua. A-CNN: Annularly convolutional neural networks on point clouds. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7413–7422, 2019.

[5] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, 3361(10):255–258, 1995.

[6] Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7045–7054, 2020.

[7] Andrew M. Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y. Ng. On random weights and unsupervised feature learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1089–1096, 2011.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations*, 2015.

[9] Leopold Vietoris. Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. *Mathematische Annalen*, 97(1):454–472, 1927.

[10] Hubert Wagner and Paweł Dłotko. Towards topological analysis of high-dimensional feature spaces. *Computer Vision and Image Understanding*, 121:21–26, 2014.

[11] Zhenlin Xu, Deyi Liu, Junlin Yang, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. *arXiv preprint arXiv:2007.13003*, 2020.

[12] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[13] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. LPIPS code repository, 2018. URL https://github.com/richzhang/PerceptualSimilarity. Code version 0.1.