Human Guided Dimensionality Reduction using Landmark Movement

Christian Raue, Frederic Sadrieh

Index Terms—Dimensionality Reduction, Human Guided Dimensionality Reduction

1 Introduction

Dimensionality reduction allows humans to grasp high dimensional relationships between data points, but dimensionality reduction is one-way, meaning that humans cannot influence how the high-dimensional data points are reduced [6]. We incorporate human interaction into the dimensionality reduction process. This allows the user to gain a better understanding of the dimensionality reduction result and use their domain knowledge.

Human interaction is achieved by modifying Landmark Multi-Dimensional Scaling (LMDS) [17] to allow the user to move the landmarks. After the movement, the remaining points are reduced relative to the landmark positions. We create an iterative process where the user can move the landmarks, create a new dimensionality reduction based on the movement and evaluate its quality through metrics. The human can learn with the dimensionality reduction, feel the consequences of his actions and create an appropriate dimensionality reduction.

2 Sources

2.1 Dimensionality Reduction

LMDS [17] is a variant of Multi-Dimensional Scaling (MDS) (as described in [4]). The classical MDS algorithm computes the pairwise squared distances between the high-dimensional points. It creates a *mean-centered inner-product matrix* from which the largest eigenvalues determine the position of the points in the low-dimensional space [4]. This deterministic algorithm scales in complexity with the number of points. Therefore, LMDS applies the MDS algorithm only to a select number of points—so-called landmarks—to reduce the complexity. The remaining points are embedded using the position of the landmarks and the distance to them [17].

For our project, LMDS is crucial because we can use landmarks to show the user a few points to interact with. LMDS is the ideal testbed for our research, since the landmark functionality is already implemented and the dimensionality reduction is easy to manipulate, but it is rather archaic. Today, t-SNE [18] and UMAP [13] are the most used [7].

t-SNE is a variant of Stochastic Neighbor Embedding [18]. t-SNE computes the pairwise similarities between the data points in the high-dimensional space and tries to construct an embedding that matches these similarities in the low-dimensional space. The difference between t-SNE and classical SNE is the use of joint probability distributions to describe the similarity and the Student t-distribution in the low-dimensional space [18]. t-SNE is good at preserving the local structure, which is beneficial for human interaction.

We have modified LMDS to embed the landmarks with t-SNE, which gives the user a better landmark placement. The method is limited by the unmodified second phase of LMDS.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

UMAP stands for Uniform Manifold Approximation and Projection and is considered to be faster and better at preserving global structure than t-SNE [13]. UMAP constructs a "fuzzy simplical complex" [13] in high-dimensional space. The complex is similar to a weighted graph, where each point is connected to its k-nearest neighbors. With this graph, it is possible to construct a lower dimensional graph by "iteratively applying attractive and repulsive forces at each edge and vertex" [13].

We did not use UMAP in our experiments because we already tested a global dimensionality reduction technique with MDS, and the problem of our t-SNE-LMDS approach lies within the second LMDS phase, which is unlikely to be improved by a UMAP-LMDS approach.

2.2 Label Propagation

Deep learning models require labeled data for supervised fine-tuning, but labeling data is a costly endeavor. Label propagation is a method to train a model with less labeled data using a semi-supervised approach [11]. A classifier is trained with labeled data. The classifier then predicts pseudo-labels for the unlabeled data. The quality of these labels may be inferior because the classifier has seen only a few examples. The pseudo-labeled data is used to train the model, which is fine-tuned with the labeled data [11]. Label propagation improves the performance of the model compared to training only on the labeled samples [11].

The authors of [2] organize all data, labeled or not, in a weighted graph. The weights are the distances between the data points in the feature space. The labeled samples are the roots, and the Optimum-Path Forest (OPF) minimizes the path between all points and them. Each unlabeled data point is given a pseudo-label belonging to the same class as the nearest root [2]. This technique can be used to propagate labels between labeled and unlabeled samples [2]. The authors in [2] use these new labels to train a CNN. It is first pre-trained with the pseudo-labels created by OPF, and then fine-tuned with the labeled data.

OPFSemi is used in [3], but dimensionality reduction is applied first. A deep neural network is trained with a small expert labeled dataset. The features of the last max-pooling layer from this model are projected into a 2D space using t-SNE [3]. In this 2D space, OPFSemi provides new pseudo-labels for the next iteration of training. The authors hope to stabilize the training by looping the OPFSemi approach. We can transfer this idea to human-guided dimensionality reduction by replacing the model with the human. The human can go through many iterations of dimensionality reduction and improve the landmarks, thus improving thequality of the dimensionality reduction.

2.3 Confident Learning

[14] shifts the focus from model quality to data quality. They estimate the joint distribution between the given and hypothetical correct labels [14]. With this technique, they can train better models and find flaws in common datasets. [14] serves as motivation: The labels given by the user are not necessarily correct, and we need to incorporate uncertainty. A potential pipeline could allow the user to label landmarks along with their uncertainty. Using both could improve the dimensionality reduction. In the end, we did not follow this path, as we find landmark movement to be more intuitive than labeling with uncertainty.

2.4 Inverse MDS

To incorporate human-landmark-movement into existing dimensionality reduction techniques, we need to invert the dimensionality reduction, specifically the moved LMDS landmarks. To inverse MDS, we considered two approaches. The first approach uses the unmodified low-dimensional distance matrix as the new high-dimensional one [9]. This method is quite simple and easy to implement, but it ignores the differences between the spaces. [9] use this approach to generate high-dimensional embeddings from users placing objects in a 2D space. In our project, we found that this approach is not well suited for inversing MDS.

The second approach is to train a small neural network to do the inversion (following [6]). One can interpret the dimensionality reduction and its inverse conceptually as an autoencoder. The encoder part is realized by the dimensionality reduction algorithm itself. The decoder is a neural network trained to reconstruct the high-dimensional data points from the low-dimensional ones. [6] presents the feasibility of using a neural network for decoding, and uses it to show users in a finished dimensionality reduction what high-dimensional features the "empty parts" would have. We use this method to invert MDS and t-SNE.

2.5 Quali-quantative metrics

The survey [7] benchmarks several dimensionality reduction techniques. We focus on the metrics to better evaluate our technique. They use:

- trustworthiness: measuring whether the k-nearest neighbors in the high-dimensional space, are close neighbors in the projection. The metric first takes all high-dimensional k-nearest neighbors that are not in the low-dimensional k-nearest neighbors. For these points, it uses the low dimensional neighbor rank to determine how close they are in the projection [7].
- continuity: measuring whether the k-nearest neighbors in the projection are close neighbors in the high-dimensional space. The metric is calculated similarly to trustworthiness, but taking the opposite set difference [7].
- normalized stress: measuring whether the point-pairwise distance from the high-dimensional space to the projection is preserved [7].
- neighborhood hit: measuring the average proportion of neighbors in the low-dimensional space that have the same label as the point [7].
- average local error: measuring for each point the average normalized distance to all other points [7].

We have implemented these metrics to give the user an overview of how well the dimensionality reduction is working.

[8] extends trustworthiness and continuity to the cluster level. Steadiness measures whether the clusters in the high-dimensional space are clusters in the projection and, continuity measures if the clusters in the projection are clusters in the high-dimensional space [8]. This is done by randomly selecting points and their neighbors as clusters and calculating their distance in the other space [8]. We have not implemented these metrics because we found 5 metrics to be sufficient for the user.

2.6 Research parameters

We use a subsample of 2,000 examples out of the 25,000 labeled training examples from the IMDb dataset [12], which contains an equal split of strongly positive and negative movie reviews. The dataset is downloaded using the hugging face dataset library [10][1]. This dataset has been shown to be challenging for dimensionality reduction techniques [7].

In addition to the binary classification task of IMDb, we use the emotion dataset [16], which has six labels. The dataset contains English tweets and assigns each one of six emotions("sadness, disgust, anger, joy, surprise, and fear") [16]. We generate a sample of 2,000 examples from the 16,000 train split of the dataset downloaded from hugging face datasets [10][2].

We only train on the IMDb and emotion datasets. As a hold-out set, we use the mnli dataset [20], which is part of the glue benchmark [19]. The dataset contains a premise and a hypothesis sentence, which we concatenate with a semicolon (i.e., `premise;hypothesis`). The label indicates whether the premise entails the hypothesis, is neutral, or contradicts it [20]. We have generated a sample of 2,000 examples out of the 393,000 train split of the dataset downloaded from hugging face datasets [10][3].

To embed the textual input, we use the all-mpnet-base-v2 [1] Sentence-Transformer-model [15] from hugging face. This model is pre-trained on sentence similarity: It takes two sentences as input and outputs their similarity. The model generates a 768 dimensional embedding that includes the semantics of the text [1].

For the training pipeline, we used parts of a template for fast ML training and iteration [5].

[1] https://huggingface.co/datasets/imdb

[2] https://huggingface.co/datasets/dair-ai/emotion

[3] https://huggingface.co/datasets/glue/viewer/mnli

References

[1] T. Aarsen, O. Espejel, and N. Reimers. all-mpnet-base-v2. Technical report. 2

[2] W. P. Amorim, G. H. Rosa, R. Thomazella, J. E. C. Castanho, F. R. L. Dotto, O. P. R. Júnior, A. N. Marana, and J. P. Papa. Semi-supervised learning with connectivity-driven convolutional neural networks. *Pattern Recognition Letters*, 128:16–22, 2019. doi: 10.1016/j.patrec.2019.08.012 1

[3] B. C. Benato, J. F. Gomes, A. C. Telea, and A. X. Falcão. Semi-supervised Deep Learning Based on Label Propagation in a 2D Embedded Space. In J. M. R. S. Tavares, J. P. Papa, and M. González Hidalgo, eds., *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 371–381. Springer International Publishing, Cham, 2021. 1

[4] V. De Silva and J. B. Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, technical report, Stanford University, 2004. 1

[5] K. Dobler, M. Schall, and O. Zimmermann. nlp-research-template, Oct. 2023. 2

[6] M. Espadoto, G. Appleby, A. Suh, D. Cashman, M. Li, C. Scheidegger, E. W. Anderson, R. Chang, and A. C. Telea. UnProjection: Leveraging inverse-projections for visual analytics of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 29(2):1559–1572, 2023. doi: 10.1109/TVCG.2021.3125576 1, 2

[7] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea. Toward a Quantitative Survey of Dimension Reduction Techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173, 2021. doi: 10.1109/TVCG.2019.2944182 1, 2

[8] H. Jeon, H.-K. Ko, J. Jo, Y. Kim, and J. Seo. Measuring and Explaining the Inter-Cluster Reliability of Multidimensional Projections. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):551–561, 2022. doi: 10.1109/TVCG.2021.3114833 2

[9] N. Kriegeskorte and M. Mur. Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, 3, 2012. doi: 10.3389/fpsyg.2012.00245 2

[10] Q. Lhoest, A. Villanova del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, J. Davison, M. Šaško, G. Chhablani, B. Malik, S. Brandeis, T. Le Scao, V. Sanh, C. Xu, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, C. Delangue, T. Matussière, L. Debut, S. Bekman, P. Cistac, T. Goehringer, V. Mustar, F. Lagunas, A. Rush, and T. Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 conference on empirical methods in natural language processing: System demonstrations*, pp. 175–184. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, Nov. 2021. arXiv: 2109.02846 [cs.CL]. 2

[11] Z. Li, B. Ko, and H.-J. Choi. Naive semi-supervised deep learning using pseudo-label. *Peer-to-Peer Networking and Applications*, 12(5):1358–1368, Sept. 2019. doi: 10.1007/s12083-018-0702-9 1

[12] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150. Association for Computational

Linguistics, Portland, Oregon, USA, June 2011. doi: 10.5555/2002472.2002491 2

[13] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 1

[14] C. Northcutt, L. Jiang, and I. Chuang. Confident Learning: Estimating Uncertainty in Dataset Labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, Apr. 2021. doi: 10.1613/jair.1.12125 1

[15] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing*. Association for Computational Linguistics, Nov. 2019. 2

[16] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 3687–3697. Association for Computational Linguistics, Brussels, Belgium, 2018. doi: 10.18653/v1/D18-1404 2

[17] V. Silva and J. Tenenbaum. Global Versus Local Methods in Nonlinear Dimensionality Reduction. In S. Becker, S. Thrun, and K. Obermayer, eds., *Advances in Neural Information Processing Systems*, vol. 15. MIT Press, 2002. doi: 10.5555/2968618.2968708 1

[18] L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008. 1

[19] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In T. Linzen, G. Chrupała, and A. Alishahi, eds., *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355. Association for Computational Linguistics, Brussels, Belgium, Nov. 2018. doi: 10.18653/v1/W18-5446 2

[20] A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In M. Walker, H. Ji, and A. Stent, eds., *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)*, pp. 1112–1122. Association for Computational Linguistics, New Orleans, Louisiana, June 2018. doi: 10.18653/v1/N18-1101 2