ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Insight-SNE: Understanding t-SNE Embeddings through Interactive Explanation

Sacha Corbugy[1], Thibaut Septon[2], Bruno Dumas and Benoit Frenay

University of Namur - NaDI - Faculty of Computer Science
rue Grandgagnage 21, B-5000 Namur - Belgium

Abstract. Non-linear dimensionality reduction techniques offer insights into complex datasets, yet interpreting them poses challenges. While some papers provide methods for explaining DR, and others focus on interactively exploring embeddings, there are currently no works that seamlessly combine both aspects. Our contributions, *Insight-SNE*, propose an interactive tool that allows exploring t-SNE embeddings and their related gradient-based explanations, as well as its evaluation with expert users.

1 Introduction

Non-linear Dimensionality Reduction (DR) techniques are widely used for data exploration. While they enable the visualization of data in two or three dimensions, interpreting them presents challenges due to their non-linearity. However, their ability to preserve local structure helps to understand low dimensions locally, revealing intricate patterns and insights within the data.

Different methods exist to explain non-linear DR. Some of them take inspiration from supervised machine learning [1, 2] and tend to generate explanations in the form of feature importance scores or local approximations of the underlying model's behavior. However, these approaches often fall short of providing concrete insights, as they merely compute explanations without facilitating their practical application. Other works focus on letting the user manipulate the DR graph in order to interpret the data in a more interactive manner. These methods prioritize user interaction with the DR visualization, allowing for exploration and discovery. While they offer a more engaging user experience, they lack the depth of explanatory power found in other methods.

Recently, the first gradient-based method for explaining non-linear DR was introduced [3]. The idea is inspired by saliency maps for neural networks. While this work introduces the methodology and the resulting explanation, it lacks of user experiments to evaluate how they understand those explanations. Furthermore, such explanations are not easy to use in practice, as they need to be generated depending on what the user is trying to achieve. As such explanations are designed to let the user understand and interpret the data, there is a need to let them explore efficiently the t-SNE embeddings and their explanations. Thus, this paper introduces *Insight-SNE* - an interactive tool that allows exploring t-SNE embeddings and their related gradient-based explanations - and its preliminary evaluation with expert users. Section 2 presents the

[1]Supported by the Walloon region, with a Ph.D. grant from FRIA (F.R.S.-FNRS).
[2]Supported by the Walloon region through the Pole MecaTech fund OPTIMIS (nb. 8564).

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

existing approaches for explaining DR and interacting with DR. It also briefly introduces the gradient-based explanation method used in our tool. Section 3 describes *Insight-SNE* and how to use it. Section 4 presents the results of a user evaluation. Finally, Section 5 discusses the outcomes and further works.

2 Related Work

Several works have proposed interpretation techniques for DR. Some focus on linear DR methods, which are easier to understand because the low-dimensional data are a linear combination of the high-dimensional data. However, linear techniques are less effective when dealing with complex data, hence the need for non-linear DR techniques. While non-linear DR methods produce better results, they are also more difficult to interpret. Some works have proposed approaches to generate explanations of their results, with many methods inspired by supervised machine learning. In [1], Bibal et al. adapted LIME to explain t-SNE locally by generating instance specific explanations. Since this method had limitations, Lambert et al. [4] proposed to improve the LIME approach by providing globally local and fast explanations of t-SNE embeddings. Another approach is [2] which was inspired by SHAP [5], an explanation method based on shapley values. While these methods provide ways to generate explanations of non-linear DR embeddings, they do not provide, by themselves, a way to interact and better understand those explanations. Another approach to understand non-linear DR visualizations is to use interactive tools. Some works proposed interactive exploration tools to interpret non-linear DR. t-viSNE [6] proposes different techniques to get insights about t-SNE embeddings. In their work, Stahnke et al. [7] introduced probing projections, an interactive framework for interpreting arrangements and errors in DR. DMT-EV [8] is an explainable deep network for DR. Zang et al. [8] provide a visual interface that helps to achieve better DR performance and explainability. All these methods employ statistical metrics to offer insights, but they do not generate explanatory outputs.

A novel explanation method for non-linear DR, inspired by saliency maps, has recently been introduced [3]. This method makes it possible to compute the derivatives of positions in t-SNE with respect to high-dimensional data. The resulting gradients provide a local explanation for a specific point within the embedding in terms of high-dimensional features. The underlying intuition behind these gradients is that they indicate how the point would move if the corresponding feature was changed. This explanation technique is the one used in this paper by the tool presented in Section 3. The gradients can be visualized in various ways. Their magnitude can be utilized to rank features by importance. The vectors can be plotted on the visualization to illustrate how changing features impacts the points position. Additionally, they can be aggregated to provide more global insights, even if each explanation is local (instance-specific).

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

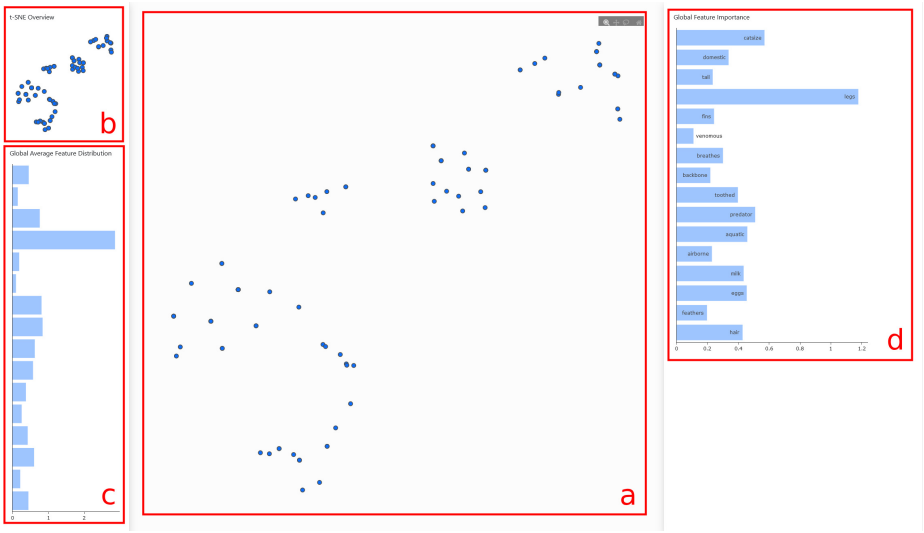


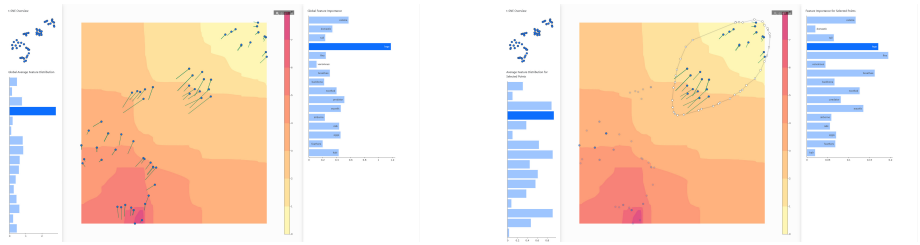
Fig. 1: Insight-SNE main interface, with (a) its t-SNE visualization, (b) t-SNE fixed overview, (c) *features distribution plot*, and (d) *features explanation plot*.

3 Interactive Explanation of t-SNE Embeddings

This Section details a novel interactive tool, *Insight-SNE**, designed to manipulate t-SNE representations while providing gradient-based explanations from [3]. An overview of its User Interface (UI) can be seen on Figure 1. When first using *Insight-SNE*, the user is prompted to choose a dataset and adapts the t-SNE parameters that best fit their needs. Then, the user is presented with the UI in Figure 1. It is divided into 4 parts. Part (a) hosts the t-SNE visualization. As it can be zoomed in and moved, Part (b) presents a fixed overview of the t-SNE representation. Part (c) shows the *average feature distribution plot*, illustrating the average value of each feature among the selected points. Meanwhile, Part (d) shows the *feature importance plot*, displaying the features importance (i.e., the derivatives of positions in t-SNE computed in [3]) among the selected points. If no points are selected, all points are considered in both parts. As the user can manipulate the plot and the explanations, the interface dynamically adjusts to represent user selection. Figure 2a shows the interface as the user selects a feature on the *feature importance plot*. As the UI gets updated, the heatmap reveals the selected feature's values across the embedding, with lower values depicted in yellow and higher values in red. The vectors drawn for each instance represent their gradient (i.e., the direction the instance would move towards, if its value for this feature increased) for the selected feature. When selecting a subset of points through the main plot (Figure 2b), the UI adapts accordingly: both the *feature importance plot* and the *feature distribution plot* change to accurately

*https://github.com/sady410/tsne_interactive_explanation

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.



(a) Selecting a feature. (b) Selecting a subset of points.

Fig. 2: User can select a subset of points with a lasso, and can select a feature (on the upper-right barplot) to get its corresponding explanations.

reflect features distribution and features importance for the selected subset.

4 Evaluation

As Section 3 introduced a tool aimed to facilitate interpretation of t-SNE embeddings through gradient explanations, it is important to evaluate users interests for such a tool and its usability with potential users. Therefore, this section describes an evaluation we conducted with 4 t-SNE expert users.

4.1 Participants

Participants were asked to answer three questions in order to assess their knowledge of DR and of t-SNE in particular. Answers to the questions were on a Likert scale ranging from 1 to 5. To the question "Do you use t-SNE regularly?", one participant answered 3, two answered 4 and one 5. For the question "How expert do you consider yourself to be with t-SNE?", two participants had an intermediate level (3), one felt somewhat expert (4) and one expert (5). And to the last question, "Are you familiar with explainability/interpretability in dimensionality reduction?", one participant described himself as having an intermediate knowledge (3), while three felt somewhat expert (4).

4.2 Setup

Study took place with the experiment conductor and each participant separately. First, Insight-SNE and its gradient-based explanations were presented through the *Iris* dataset use case. Then, participants were asked to explore and understand the *Zoo Animal*[†] dataset. Experiment conductor let the participant use the tool as they wish but asked a question if the participant felt stuck in order to ensure they explored and used all available features. Then, a semi guided interview was held, which aimed at answering the following three questions: $Q1$) "Do you find the tool supportive to better understand the t-SNE embeddings?", $Q2$)

[†]https://archive.ics.uci.edu/dataset/111/zoo

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

"Are the gradient-based explanations easy to understand?", $Q3$) "Is the tool's interaction easy to apprehend?".

4.3 Results

About $Q1$, all participants agreed that the tool was supportive, and interesting. $p2$ had an interesting remark, stating that as humans, we tend to interpret t-SNE results based on our knowledge of the data, while t-SNE might not have considered the same aspects. However, $p2$ and $p4$ were dubitative on how well the tool would support understanding a bigger dataset. For $Q2$, while $p1$ and $p2$ found it hard at first to apprehend the explanations, all of them found that the explanations were easy to use once you get the intuition behind them. Participants $p3$ added that the vectors (see Figure 2a) were redundant with the heatmap, suggesting that he did not fully understand the difference between the two visualizations. For $Q3$, participants $p1$, $p2$ and $p4$ found the tool had learning curve a bit hard to apprehend, but was easy to use when getting familiar with it. $p3$ found it easy to use directly. All agreed that the tool's user experience could be improved by improving some minor features currently missing.

5 Conclusion and Further Work

The evaluation presented in Section 4 highlights the interest of users for such a tool. Indeed, it eases the use of t-SNE and help users to better understand its embeddings through the different available explanations. While developing explanation methods for DR is a compelling area of research, it is imperative that these methods are not only effective but also designed to be seamlessly integrated into tools that end users can readily engage with. Without this integration, the potential value of these explanations may remain untapped. Therefore, there is a pressing need to focus not only on the development of explanation methods themselves but also on designing them to be user-friendly and accessible within interactive tools. By doing so, end users can truly find value and interest in the interpretability of DR results. While the evaluation demonstrates promising results, it is important to acknowledge several limitations. The experiment involved expert users, which might bias the findings towards individuals already familiar with the underlying concepts of t-SNE and its applications. Consequently, the tool's effectiveness with novice users remains uncertain and warrants further investigation. As an expert-oriented tool, users are expected to grasp the intuition behind the explanations before fully leveraging the tool's capabilities. This prerequisite understanding could hinder the tool's accessibility to a broader audience. Also, as the tool is currently a prototype, there is room for enhancing the user experience (UX). Addressing usability issues and refining the interface based on user feedback could significantly improve the tool's usability and UX. Furthermore, when asked about what improvements would be essential to enhance the tool, users made a few interesting suggestions. One idea was to arrange the bars in the feature importance plot in descending order, thereby highlighting the most important features. Another important improvement pertains to the *average feature importance plot*, which is currently not very

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

informative. Displaying the distribution for the selected feature amongst the selected points could be more informative. Other suggestions were to improve the tool's comparison ability. When interpreting t-SNE, users typically want to understand why one cluster differs from another. While the tool aids in this process, it could be made more user-friendly. For example, one suggestion was to enable users to create two selections and display the feature importance of both simultaneously for easy comparison. Another proposal was to provide several t-SNE plots with different explanations overlaid, to facilitate comparison.

While non-linear DR interpretation is being widely studied, there is a lack of tools combining interaction and explanation. Since explanation is designed for users there is a need to integrate their usage in interactive tools. In this regard, we presented *Insight-SNE*, a visual interactive tool to easily generate and manipulate t-SNE embeddings along with gradient-based explanations. As the tool is aimed towards users, we conducted a preliminary evaluation with t-SNE expert users to evaluate how the tool is perceived, and how it could be improved. Results show a need and an interest for such tool as it significantly enhances the quality of interpreting DR results. Making non-linear DR techniques accessible through user-friendly explanations within interactive tools is crucial. This ensures that end users can easily understand and benefit from the interpretability of their results, maximizing the technique's potential.

References

[1] Adrien Bibal, Viet Minh Vu, Géraldin Nanfack, and Benoît Frénay. Explaining t-sne embeddings locally by adapting lime. ESANN, 2020.

[2] Wilson E Marcilio-Jr and Danilo M Eler. Explaining dimensionality reduction results using shapley values. *Expert Systems with Applications*, 178:115020, 2021.

[3] Sacha Corbugy, Rebecca Marion, and Benoît Frénay. Gradient-based explanation for non-linear non-parametric dimensionality reduction. *Data Mining and Knowledge Discovery*, pages 1–29, 2024.

[4] Pierre Lambert, Rebecca Marion, Julien Albert, Emmanuel Jean, Sacha Corbugy, and Cyril de Bodt. Globally local and fast explanations of t-sne-like nonlinear embeddings. In *CIKM-WS*. CEUR Workshop Proceedings, 2022.

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[6] Angelos Chatzimparmpas, Rafael M Martins, and Andreas Kerren. t-visne: Interactive assessment and interpretation of t-sne projections. *IEEE Trans. Vis. Comput. Graph*, 26(8):2696–2714, 2020.

[7] Julian Stahnke, Marian Dörk, Boris Müller, and Andreas Thom. Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *IEEE Trans. Vis. Comput. Graph*, 22(1):629–638, 2015.

[8] Zelin Zang, Shenghui Cheng, Hanchen Xia, Liangyu Li, Yaoting Sun, Yongjie Xu, Lei Shang, Baigui Sun, and Stan Z Li. Dmt-ev: An explainable deep network for dimension reduction. *IEEE Trans. Vis. Comput. Graph*, 30(3):1710–1727, 2022.