ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Interaction and User Integration in Machine Learning for Information Visualisation

Bruno Dumas[1], Benoît Frénay[1] and John A. Lee[2] *

1- NADI Institute - PReCISE Research Center
Université de Namur - Faculty of Computer Science
Rue Grandgagnage 21 - 5000 Namur, Belgium

2- Université catholique de Louvain - SST/ICTEAM & SSS/IREC
Place du Levant 3, 1348 Louvain-la-Neuve - Belgium

Abstract. Many methods have been developed in machine learning (ML) for information visualisation (infovis). For example, PCA, MDS, t-SNE and improvements are standard tools to reduce the dimensionality of high dimensional datasets for visualisation purposes. However, multiple other means are regularly used in the field of infovis when tackling datasets with high dimensionality. Letting the user manipulate the visualisation is one of these means, either through selection, navigation or filtering. Introducing manipulation of the visualisation also integrates the user as a core aspect of a given system. In the context of machine learning, beyond the informational and exploratory use of infovis, users' feedback can for example be highly informational to drive the dimensionality reduction process.

This special session of the ESANN conference is a followup of the special session on "Information Visualisation and Machine Learning: Techniques, Validation and Integration" at ESANN 2016. It aims to gather researchers that integrate users in the core of ML methods for infovis. New algorithms and frameworks are welcome, as well as experimental use cases that bring new insight in the integration of interaction and user integration in ML for infovis. This special session aims to provide practitioners from both communities a common forum of discussion where issues at the crossroads of machine learning and information visualisation could be discussed.

1 Introduction

Machine learning (ML) and information visualisation (infovis) are two opposite perspectives to look at data.

Machine learning is inspired by mathematics, statistics, and, to some extents, neurosciences. Machine learning develops mathematically rigourous and statistically sound algorithms to solve data-intensive problems, where the task to carry out can often be quantitatively assessed, be it supervised or unsupervised learning. This approach leaves little room for the user, who is generally just interested in the problem definition, its algorithmic solution, and how it impacts figures of merit. Data interpretation and semantics are then often overlooked.

On the other hand, information visualistion explores the other end of the spectrum. The user occupies a central and privileged place. Graphical display

*J.A.Lee is a Senior Research Associate with the Belgian F.R.S.-FNRS.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

and interfaces, interaction with the system, feedback loops are thoroughly studied, developed, and tested under several scenarios and use cases. Infovis is thus much more focused than ML on optimizing the user's experience of exploratory data analysis, in order to come to grips with the most subtle data intricacies. The potential shortcoming of infovis is thus to grant the user full responsibility and thus to neglect or avoid complex and automated steps of data processing, which decrease direct interpretability at first.

These two perspectives are somehow complementary but the dialog across communities is actually not so straightforward. Several initiatives have been recently undertaken to bridge this gap. For instance, two Dagstuhl seminars 12081 *Information Visualization, Visual Data Mining and Machine Learning* and 15101 *Bridging Information Visualization with Machine Learning* were held in 2012 and 2015, gathering researchers from both ML and infovis. These resulted in a few few joint works and publications (e.g., [1, 2, 3]). A previous ESANN special session on the topic *Information Visualisation and Machine Learning: Techniques, Validation and Integration* took place in 2016. After several years, though, the feeling remains that much work is still to be accomplished, justifying this followup. In this tutorial, in particular, we try to review some of the latest develoments on the middle ground. Moreover, we picked a few examples that we further describe and analyse. Finally we summarise in a few words the contributions presented in this special session.

The rest of this tutorial paper is organised as follows. In Section 2, we review current trends at the intersection of ML and infovis. In Section 3, we illustrate these trends with a few recent examples picked from the literature. In Section 4, we briefly summarize the contribution presented at this ESANN special session. Finally, Section 5 draws the conclusions and sketches some perspectives.

2 Trends at the Crossing of Machine Learning and Infovis

Two years ago, in the context of the special session "Information Visualisation and Machine Learning: Techniques, Validation and Integration" at ESANN 2016, we presented a classification of systems between information visualisation and machine learning taken from Bertini et Lalanne [4]. According to that classification, a system sharing features from both of these domaines will fall into one of three categories:

- *Systems towards the information visualisation perspective*: systems taking from this perspective will usually focus on the user side, with a heavy accent on the visualisation aspect. Machine learning techniques will be used as support for the generation of the visual analysis aspects, for example by preprocessing data to be displayed.
- *Systems towards the machine learning perspective*: with systems taking from this perspective, visualisation techniques are used typically as feedback with the goal of getting better information or improving the results of a given machine learning based approach.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

- *Systems seeking a fit between both of the domains* will integrate machine learning approaches and information visualisation approaches in equal measure to solve a given issue. Bertini et Lalanne remarked in their original paper that systems belonging to this category were in fewer numbers compare to the two other categories, for diverse reasons.

In the remainder of this article, we describe further a selection of recent works that fit into one of the categories described above.

3 Some illustrative examples

This section discusses a few methods fitting in the categories from Section 2. It is by no means exhaustive, but rather aims to gather inspiring examples to foster discussions and to illustrate the diversity of current research trends.

3.1 Improving deep learning interpretability at Facebook

Trained artificial neural networks are known to be difficult to analyze and interpret. This is even more true when they are deep [5], with many layers, and thus many architectural parameters. Especially in this case, users might want to inspect large, industry-size networks quickly and visually, in order to assess prototypes and improve their design. Researchers at Facebook have recently developed ActiVis [6], a visual analytics system that comes on top of Facebook's machine learning platform. The main features of ActiVis are:

- A novel visual representation that unifies instance- and subset-level inspections of neuron activation, facilitating comparison of activation patterns for multiple instances.
- An interface that tightly integrates an overview of graph-structured complex models and local inspection of neuron activations, allowing users to explore the model at different levels of abstraction.
- A deployed system scaling to large datasets and models.
- Case studies with Facebook engineers and data scientists that highlight how ActiVis helps them with their work.

ActiVis's multiple coordinated views supports exploration of complex deep neural network models, at both instance- (single datum) and subset-level (group of data, e.g., classes). ActiVis's interface is split in four main views or panels. In the first view, the user starts exploring the *model architecture*, through its computation graph overview. Selecting a data node displays its neuron activations. Next, the *neuron activation matrix view* shows the activations for instances and instance subsets; the projected view displays the 2D t-SNE embedding [7] of instance activations. From the *instance selection panel*, the user explores individual instances and their classification results. Adding instances to the *matrix*

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

view enables comparison of activation patterns across instances, subsets, and classes, revealing causes for misclassification.

Although ActiVis considerably eases the interpretation of complex neural networks, it remains a top layer that inspects networks as frozen objects. There is no feedback loop [1, 2]: the users cannot retropropagate their conclusions back to the network directly. They have to implement architectural changes outside ActiVis. As such, ActiVis belongs to the second category in Section 2.

3.2 Data-driven evaluation of visual quality measures

Evaluating the quality of a visualisation is not a travial task. Consequently, quality assessment of such artefacts has been a topic of research in recent years.

Sedlmair and Aupetit [8] learn to predict the quality of scatter plots based on data coming from both human and automated evaluation. First, they collect human judgment on a set of scatter plots. This labelling is used as a ground truth of *whether a human considers the scatter plot as relevant or not.* Scatter plots generated from 75 different datasets were labelled by humans from *not separable at all* to *nicely separable.* Then, for each scatter plot, a set of measures is automatically extracted from the positions of the points, based on distances within and between classes, distances to cluster centers, class entropies, pixels, etc. In their experiments, an AUC of 82.5% was achieved by the DSC measure (distance consistency) [9], but results were better with synthetic datasets than with real datasets. Sedlmair and Aupetit [8] conclude that there is *a lot of room for future improvements* [10] and found *a bias of current measures towards simple synthetic datasets.*

This paper is an example of approach were machine learning helps to automate information visualisation (selection of revelant representations, for example). There exist many other works in this line of research [11] that focus, e.g., on visualisation quality assessment [12, 13, 14, 15, 16, 17, 18, 19], modelling of human preferences [10, 20, 21, 11, 22] and links with information visualisation tasks [23, 24, 25, 26].

3.3 Observation-level interaction with dimensionality reduction

Machine learning techniques for dimensionality reduction can be used to produce simple information visulisations like e.g. scatter plots. However, such visualisations may not be aligned with user needs. Since dimensionality reduction is often unable to react to user feedback, this can result in the rejection of these results. Several works have recently focused on addressing this lack of interactivity.

Endert et al. [27] enable observation-level interaction for several dimensionality reduction techniques. In their approach, users are allowed to move instances in the scatter plot. This provides a simple way to feed machine learning algorithms with rich knowledge coming from the user. Endert et al. envision two purposes for such interactions: *exploraty interactions* aim to *gain insight into the structure of the data by observing how other data reacts* [27] to such interactions, whereas *expressive interactions* aim to enforce constraints into the model. They

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

implemented user-guided versions of probabilistic principal component analysis (PPCA), multi-dimensional scaling (MDS) and generative topographic mapping (GTM), with specific heuristics to embed user feedback in each algorithm.

Interactive dimensionality reduction is not yet well represented in the literature [28], but solutions have been proposed to integrate user feedback through *must-link* and *cannot-link* constraints [29, 30], prior knowledge on the position of a few objects [31], hierarchical information [32, 33], quality metrics [34], etc.

3.4 ML techniques to enrich user experience in interactive systems

As an example of a system focusing on the information visualisation perspective, Cartograph [35] uses data from Wikipedia to create a visual cartography of knowledge. Semantic vectors based on the contents of Wikipedia pages are used with a standard t-SNE algorithm [7], and a 2D mapping of different topics is generated from there. Interestingly, if the text mining approach used is quite basic, the authors worked thoroughly on the presentation of the results, taking from the field of cartography to create a much more semantic-rich representation of the results. Pezzotti et al. [36] presented DeepEyes, a Progressive Visual Analytics system that supports the design of neural networks during training. Deepeyes presents to users perplexity histograms for the identification of stable layers. Then, an activation heatmap, an input map and a layer maps lets users explore further stable layers. DeepEyes thus supports DNN design decisions during training. A final interesting work is VizRec [37], a tool for recommending appropriate visualisation techniques considering a given set of data. Choosing the right graphical representation for some data is one of the core challenges in information visualisation, as the potential design space is huge. In the VizRec approach, its authors used a collaborative filtering based recommender system linked with a content-based recommender system to recommend appropriate visualisation to users. The selection of features used is of particular interest.

4 Contributions in this ESANN special session

This special session includes three contributions, on top of this tutorial paper.

Cutura et al. [38] address the issue of comparing several dimensionality reduction (DR) techniques. They propose VisCoDeR, a tool for *junior data scientists who seek to learn and better understand DR algorithms, as well as DR designers, who seek to evaluate and analyze DR methods* [38]. VisCoDeR provides textual information about dimensionality reduction algorithms, a view of the original data, a tool to identify distortions, a mechanism to interactively change algorithm metaparameters and a *meta-map* showing the results for different metaparameter values of a given dimensionality dimensionality technique.

Schlegel et al. [39] proposes G-Rap, an interactive, visual interface to design recurrent neural networks (RNNs), in particular for tasks where it is difficult to assess the output of the RNN. In their article, they present a tool doing text synthesis on rap song lyrics. The bigger interest of their approach lies in the seamless user interface allowing users to quickly compare the results of a set

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

of different RNNs, and strengthen the results of said neural networks through interaction and feedback loops.

Diaz et al. [40] tackle one of the major issue of current dimensionality reduction techniques: scalability. Indeed, the computation cost of such procedures is usually an obstacle to interactive dimensionality reduction, since recomputation of instance projections should be performed in a fraction of second to achieve real-time interaction required by information visualisation tools. They first reduce the size of the dataset, compute the projection of the reduced dataset and then obtain out-of-sample projections using a fast neural network for regression. In the case of t-SNE, they tested their method with 4000 points and reduced the latency from 0.2 fps to 20 fps, achieving a speedup of two orders of magnitude.

These three contributions range from pure machine learning, in order to speed up information visualisation, to visualisation of machine learning results. They illustrate well the wide range of potential innovative approaches that appear when the fields of information visualisation and machine learning are meshed together, as underlined by Bertini et Lalanne [4].

5 Conclusions and Perspectives

This tutorial briefly reviewed some works at the crossing of machine learning and interactive information visualisation. Although many recent initiatives have fostered collaboration across fields, many more developments are still to come. Deeper feedback, not only to the visual analytics system itself, but also to the encapsulated machine learning techniques is certainly an open avenue. The industrialisation of machine learning is another forthcoming challenge, with ever bigger data, more complex systems of artificial intelligence, and stronger interaction with the public or customers.

Acknowledgement

Benoît Frénay would like to thank his research team, in particular Adrien Bibal and Minh Viet, for fruitful discussions about ML and information visualisation.

References

[1] D. Sacha, M. Sedlmair, L. Zhang, J.A. Lee, D. Weiskopf, S. North, and D. Keim. Human-centered machine learning through interactive visualization: Review and open challenges. In *Proc. ESANN 2016 - 24th European Symposium on Artificial Neural Networks2016, Pages 641-64624th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 641–646, Bruges, apr 2016.

[2] D. Sacha, M. Sedlmair, L. Zhang, J.A. Lee, J. Peltonen, D. Weiskopf, S.C. North, and D.A. Keim. What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing*, 268:164–175, 2017.

[3] D. Sacha, L. Zhang, M. Sedlmair, J.A. Lee, J. Peltonen, D. Weiskopf, S.C. North, and D.A. Keim. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):241–250, 2017.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

[4] E. Bertini and D. Lalanne. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *Acm Sigkdd Explorations Newsletter*, 11(2):9–18, 2010.

[5] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436 EP –, 2015.

[6] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. Chau. ActiVis: Vsual exploration of industry-scale deep neural network models. *IEEE Transactions on Visualization and Computer Graphics*, 24(1), 2018.

[7] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[8] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. *Computer Graphics Forum*, 34(3):201–210, 2015.

[9] Mike Sips, Boris Neubert, John P. Lewis, and Pat Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009.

[10] A. Tatu, P. Bak, E. Bertini, D. Keim, and J. Schneidewind. Visual quality metrics and human perception: An initial study on 2d projections of large multidimensional data. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '10, pages 49–56, New York, NY, USA, 2010. ACM.

[11] R. Etemadpour, R. Motta, J. G. d. S. Paiva, R. Minghim, M. C. F. de Oliveira, and L. Linsen. Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 21(1):81–94, Jan 2015.

[12] Samuel Kaski, Janne Nikkilä, Merja Oja, Jarkko Venna, Petri Törönen, and Eero Castrén. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC bioinformatics*, 4(1):48, 2003.

[13] M. Sips, B. Neubert, J.P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. In *Computer Graphics Forum*, volume 28, pages 831–838. Wiley Online Library, 2009.

[14] J.A. Lee and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7-9):1431–1443, 2009.

[15] J.A. Lee and M. Verleysen. Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters*, 31(14):2248–2257, 2010.

[16] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, 2011.

[17] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. In *Computer Graphics Forum*, volume 31, pages 1335–1344. Wiley Online Library, 2012.

[18] J. Lewis, M. Ackerman, and V. de Sa. Human cluster evaluation and formal quality measures: A comparative study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34, 2012.

[19] M. Aupetit and M. Sedlmair. Sepme: 2002 new visual separation measures. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pages 1–8, April 2016.

[20] R.A. Rensink and G. Baldridge. The perception of correlation in scatterplots. In *Computer Graphics Forum*, volume 29, pages 1203–1210. Wiley Online Library, 2010.

[21] J. Lewis, L. Van der Maaten, and V. de Sa. A behavioral investigation of dimensionality reduction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34, 2012.

[22] Adrien Bibal and Benoit Frénay. Learning interpretability for visualizations using adapted cox models through a user experiment. *arXiv preprint arXiv:1611.06175*, 2016.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

[23] M. Aupetit. Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing*, 70(7-9):1304–1330, 2007.

[24] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnork, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 59–66. IEEE, 2009.

[25] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11(Feb):451–490, 2010.

[26] S. Lespinats and M. Aupetit. Checkviz: Sanity check and topological clues for linear and non-linear mappings. In *Computer Graphics Forum*, volume 30, pages 113–125. Wiley Online Library, 2011.

[27] Alex Endert, Chao Han, Dipayan Maiti, Leanna House, and Chris North. Observation-level interaction with statistical models for visual analytics. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 121–130. IEEE, 2011.

[28] Enrico Bertini and Denis Lalanne. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*, pages 12–20. ACM, 2009.

[29] Daoqiang Zhang, Zhi-Hua Zhou, and Songcan Chen. *Semi-Supervised Dimensionality Reduction*, pages 629–634.

[30] I. Davidson. Knowledge driven dimension reduction for clustering. In *IJCAI*, pages 1034–1039, 2009.

[31] Xin Yang, Haoying Fu, Hongyuan Zha, and Jesse Barlow. Semi-supervised nonlinear dimensionality reduction. In *Proceedings of the 23rd international conference on Machine learning*, pages 1065–1072. ACM, 2006.

[32] Jing Yang, Matthew O Ward, and Elke A Rundensteiner. Visual hierarchical dimension reduction for exploration of high dimensional datasets. 2002.

[33] M. Williams and T. Munzner. Steerable, progressive multidimensional scaling. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 57–64. IEEE, 2004.

[34] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):993–1000, Nov 2009.

[35] Shilad Sen, Anja Beth Swoap, Qisheng Li, Brooke Boatman, Ilse Dippenaar, Rebecca Gold, Monica Ngo, Sarah Pujol, Bret Jackson, and Brent Hecht. Cartograph: Unlocking spatial visualization through semantic enhancement. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 179–190. ACM, 2017.

[36] N. Pezzotti, Th. Höllt, J. Van Gemert, B.P.F. Lelieveldt, E. Eisemann, and A. Vilanova. Deepeyes: Progressive visual analytics for designing deep neural networks. *IEEE transactions on visualization and computer graphics*, 24(1):98–108, 2018.

[37] B. Mutlu, E. Veas, and Ch. Trattner. Vizrec: Recommending personalized visualizations. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(4):31, 2016.

[38] Michaël Aupetit Michael Sedlmair Rene Cutura, Stefan Holzer. Viscoder: A tool for visually comparing dimensionality reduction algorithms. In *Proc. 26th Eur. Symp. Artificial Neural Networks*, Bruges, Belgium, April 2018.

[39] Juri Buchmüller Daniel Keim Udo Schlegel, Eren Cakmak. G-rap: interactive text synthesis using recurrent neural network suggestions. In *Proc. 26th Eur. Symp. Artificial Neural Networks*, Bruges, Belgium, April 2018.

[40] Abel A. Cuadrado Diego Garcia-Perez Dominguez Manuel Ignacio Diaz, Daniel Perez. Interactive dimensionality reduction of large datasets using interpolation. In *Proc. 26th Eur. Symp. Artificial Neural Networks*, Bruges, Belgium, April 2018.