TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

ALEKSI HIETANEN

UNTITLED

Master's thesis

Examiner:

1. INTRODUCTION

- A major challenge in modern data analysis
- analysis of high dimensional data, examples
- dimensionality reduction
- curse of dimensionality
- linear and non linear
- problems with existing methods
- neural networks provide way for scalable learning of complex functions
- additionally enables parametric extension
- in this thesis...
- The structure of this thesis is as follows. Chapters /refch:vae and /refch:tsne cover the relevant background literature for the methods proposed in chapter /refch:methods, where chapter /refch:vae discusses the theory behind Variational Autoencoders and chapter /refch:tsne presents t-SNE, as well as its parametric extension. In chapter /refch:results several empirical results for the proposed method are presented, along with comparisons to other popular methods. The final chapters are reserved for discussion and conclusions.

2. VARIATIONAL AUTOENCODERS

3. T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING

4. METHODS

5. RESULTS

5.1 Evaluation metrics

The quantitative evaluation of unsupervised, non-linear dimensionality reduction methods is a challenging problem.

To evaluate the efficacy of the method proposed in this work we employ three different quantitative metrics.

Namely, the trustworthiness score, continuity and for labeled data sets 1-NN classification accuracy will be used.

5.1.1 Trustworthiness

5.1.2 Continuity

5.1.3 1-NN classifier

5.2 Data sets

5.2.1 MNIST

One of the most widely used data sets in current machine learning, being the most used

5.2.2 SVHN

The Google *Street View House Numbers* data set bares a great resemblance to MNIST yet being a considerably harder data set to learn.

Typically supervised feature extraction has been used to preprocess the original images into feature vectors, such as in ... where convolutional neural networks are applied to ...

In this work we directly use the cropped and uncropped street view house numbers set.

5.2.3 Mass cytometry data

6. DISCUSSION

[**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]

REFERENCES

[1] N. Aghaeepour, G. Finak, H. Hoos, T.R. Mosmann, R. Brinkman, R. Gottardo, R.H. Scheuermann, Critical assessment of automated flow cytometry data analysis techniques, Nature Methods, Vol. 10, Iss. 3, Feb. 2013, pp. 228–238.

[2] Amir El-ad David, Davis Kara L, Tadmor Michelle D, Simonds Erin F, Levine Jacob H, Bendall Sean C, Shenfeld Daniel K, Krishnaswamy Smita, Nolan Garry P, Pe'er Dana, viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia, Nature biotechnology, Vol. 31, Iss. 6, may, 2013, p. 545–552.

[3] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science, Vol. 315, Iss. 5814, 2007, pp. 972–976.

[4] A. Gisbrecht, A. Schulz, B. Hammer, Parametric nonlinear dimensionality reduction using kernel t-SNE, Neurocomputing, Vol. 147, Advances in Self-Organizing Maps Subtitle of the special issue: Selected Papers from the Workshop on Self-Organizing Maps 2012 (WSOM 2012), 2015, pp. 71–82.

[5] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, CoRR, Vol. abs/1502.03167, 2015.

[6] D.P. Kingma, T. Salimans, M. Welling, Improving variational inference with inverse autoregressive flow, CoRR, Vol. abs/1606.04934, 2016.

[7] D.P. Kingma, M. Welling, Auto-encoding variational bayes., CoRR, Vol. abs/1312.6114, 2013.

[8] Levine Jacob H., Simonds Erin F., Bendall Sean C., Davis Kara L., Amir El-ad D., Tadmor Michelle D., Litvin Oren, Fienberg Harris G., Jager Astraea, Zunder Eli R., Finck Rachel, Gedman Amanda L., Radtke Ina, Downing James R., Pe'er Dana, Nolan Garry P., Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis, Cell, Vol. 162, Iss. 1, doi: 10.1016/j.cell.2015.05.047, feb, 2018, pp. 184–197.

[9] L. Maaten, Learning a parametric embedding by preserving local structure, in: Dyk, D. van, Welling, M. (eds.), Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr, 2009, Proceedings of Machine Learning Research 5, PMLR, pp. 384–391.

[10] L. van der Maaten, Accelerating t-SNE using Tree-Based Algorithms, Journal of Machine Learning Research, Vol. 15, 2014, pp. 3221–3245.

[11] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research, Vol. 9, 2008, pp. 2579–2605.

[12] A. Nima, N. Radina, H. Hoos Holger, R. Brinkman Ryan, Rapid Cell Population Identification in Flow Cytometry Data, Cytometry. Part A : the journal of the International Society for Analytical Cytology, Vol. 79, Iss. 1, jan, 2011, p. 6–13.

[13] D. Rezende, S. Mohamed, Variational inference with normalizing flows, in: Bach, F., Blei, D. (eds.), Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 07–09 Jul, 2015, Proceedings of Machine Learning Research 37, PMLR, pp. 1530–1538.

[14] D.J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, in: Xing, E.P., Jebara, T. (eds.), Proceedings of the 31st International Conference on Machine Learning, Bejing, China, 22–24 Jun, 2014, Proceedings of Machine Learning Research 32, PMLR, pp. 1278–1286.

[15] J.M. Tomczak, M. Welling, Improving variational auto-encoders using householder flow, CoRR, Vol. abs/1611.09630, 2016.