EVALUATING DIMENSIONALITY REDUCTION OF 2D HISTOGRAM DATA FROM TRUCK ON-BOARD SENSORS

Evaldas Vaiciukynas, Matej Ulicny, Sepideh Pashami & Sławomir Nowaczyk
Center for Applied Intelligent Systems Research (CAISR), Halmstad University
evavaic@ktu.lt, ulinm@tcd.ie, {seppas,slanow}@hh.se

ABSTRACT

This work presents evaluation of several approaches for unsupervised mapping of raw sensor data from Volvo trucks into low-dimensional representation. The overall goal is to extract general features which are suitable for more than one task. Comparison of techniques based on t-distributed stochastic neighbor embedding (t-SNE) and convolutional autoencoders (CAE) is performed in a supervised fashion over 74 different 1-vs-Rest tasks using random forest. Multiple distance metrics for t-SNE and multiple architectures for CAE were considered. The results show that t-SNE is most effective for 2D and 3D, while CAE could be recommended for 10D representations. Fine-tuning the best convolutional architecture improved low-dimensional representation to the point where it slightly outperformed the original data representation.

1 INTRODUCTION

High dimensionality and large amounts of data pose a challenge for data analysis in many fields. Variables which are often redundant can typically be compressed into more compact representations without a significant loss of information. While many modern methods try to overcome the "curse of dimensionality" and enable machine learning with high-dimensional data, good feature selection and extraction still can lead to significant improvements in many cases. The biggest challenge with representation learning methods is to ensure that the resulting features are sufficiently generic, i.e., that they are applicable to many different tasks. This work presents an early evaluation of t-SNE and CAE-based methods on real-world data from the automotive domain.

Multidimensional histograms, especially bivariate ones, can be encountered in diverse areas, for example, as image texture descriptors (Haralick et al., 1973; Johnson & Hebert, 1999; Lazebnik et al., 2005), in chromatic analysis (Reyes-Aldasoro et al., 2011), in radiometric measurements (Clamme & Deniz, 2005; Rautiainen et al., 2008), or in malware detection (Saxe & Berlin, 2015).

This research investigates an unsupervised non-linear mapping of high-dimensional data from on-board truck sensors, collected in the form of bivariate histograms, to a low-dimensional (2D, 3D or 10D) representation. Solutions based on flattening out the histogram bins into the feature vector, as well as those that preserve spatial proximity of the bins, are evaluated. Since the aim is to obtain a generic representation, the investigated methods are assessed by measuring the performance over a multitude of supervised machine learning tasks.

2 METHODS

Overall, 23 configurations of non-linear mapping were considered for each low-dimensional space, where 7 were based on t-SNE and 16 on CAE.

t-SNE (van der Maaten & Hinton, 2008) uses distance matrix as an input and computes data coordinates in low-dimensional space in a way that tries to maintain the neighborhood relationships, not necessarily the actual distance values. Distance matrix was computed using several common bin-to-bin, – Euclidean, cosine, correlation, and Spearman, – and cross-bin distances – Earth mover's (Ling & Okada, 2007) and diffusion (Ling & Okada, 2006).

Convolutional autoencoder is an unsupervised learning algorithm that trains weights of a neural network so that the computed output is as similar as possible to the provided input. Stacked layers have symmetric sizes and the smallest, middle-most layer, known as bottleneck, can be exploited for

dimensionality reduction. CAE allows for learning of local features and promotes weight sharing, where hidden layers are a result of convolving the input with a filter mask.

External evaluation was performed using random forest (Breiman, 2001), composed of 1000 unprunned CART-type trees. Detector's votes for the out-of-bag (OOB) data were converted to a soft decision through a normalized difference between class probabilities. The average cost of log-likelihood-ratio C_{llr} (Brümmer & de Villiers, 2013) over 74 different detection tasks was used to measure how good the investigated representation is.

Fine-tuning. The encoding part of the best performing CAE was used to initialize weights of discriminative convolutional neural network. Two types of architectures for classification, connected to the bottleneck, were tested: simple (using a single densely connected layer with softmax activations, as in logit) and complex (2 layer perceptron with 100 rectified linear units in hidden layer and softmax activations in the output).

3 EXPERIMENTS

Data originates from 79974 unique Volvo trucks, and is recorded during a full year. The data of a single truck is represented with a bivariate histogram, where the axes correspond to a pair of sensors: turbocharger speed vs boost pressure. The original matrices with 9×10 dimensions were, for CAE experiments, zero-padded into 12×12. Discrete bin values of absolute frequency were converted into continuous values of relative frequency within the $[0-1]$ range through division by the overall sum of counts. Example of the resulting bivariate histogram for a typical truck is shown in Fig. 1 (**left**).

Validation sample of 7997 trucks was selected by a conditioned Latin hypercube method (Minasny & McBratney, 2006) from original data using age-based stratification. Example result of such sampling is shown in Fig 1 (**right**), visualized using t-SNE. The 71977 trucks remaining after this split were used for training of CAE-based methods. Notably, there is no need to for training data in case of methods based on t-SNE. Finally, the CAE-based fine-tuning was done using another, independent test sample of 7198 trucks, selected using the same hypercube method from this training set of 71977 trucks.

Supervised external evaluation and comparison of various methods was performed using a number of labels describing various truck configurations. The overall goal is not to find the best low-dimensional representation tailored to a very specific task, but rather to identify the method for learning a widely applicable representation. 74 different 1-vs-Rest detection tasks were devised from the following label groups (the number of categories within each group in parenthesis): engine (15), gearbox (9), chassis type (8), model (6), marketing type (6), fuel capacity (6), country of operation (5), model name (5), product class number (5), emission level (3), country (2), truck type (2), and brand type (2).

Experimental setup. For t-SNE, the perplexity parameter was set to 15 and the number of iterations was 1000. For CAE, robust internal representations were enforced by dropout (Srivastava et al., 2014) (rate = 0.5) and elastic net activity regularization (penalties $L_1 = L_2 = 0.00001$) in the bottleneck. Activation function was hyperbolic tangent, 8 and 32 filters of 3×3 mask size with one and two convolutional layers were investigated, as well as classical (Masci et al., 2011) and variational (Kingma & Welling, 2014) CAE architectures in Keras framework using NADAM (Dozat, 2016) optimizer.

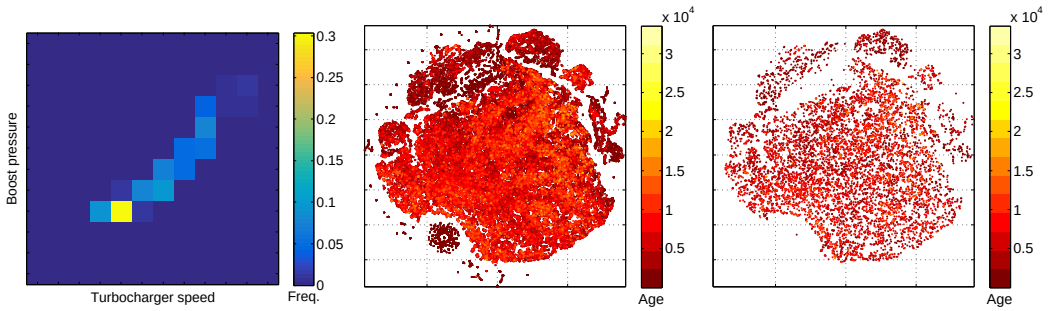

Figure 1: Truck sensor data, using original representation of relative frequency bivariate histogram (**left**). Data visualization by t-SNE: full (**middle**) and validation sample (**right**), where age is the total hours of individual truck usage (corresponding to the overall sum of bin counts).

Experiments using the methods outlined above were done with original data, as well as with a sparse matrix, capturing the deviation of an individual truck from the average operation of all vehicles, obtained from robust PCA (RPCA) (Aravkin et al., 2014).

3.1 Results of Experiments

Results of selected methods are compared in Fig. 2. The selection was based on evaluation using the validation sample for 74 detection tasks in each target dimensionality (2D, 3D and 10D), both for the original and RPCA-transformed data. The intervals around the average C_{llr} (or its rank) are such that two results being compared are significantly different if intervals are disjoint and are not significantly different if intervals overlap. None of the unsupervised methods was able to improve over the initial 90-dimensional data, but similar performance could be achieved after CAE-based complex fine-tuning.

The scatter plot on Fig. 2 (**right**) demonstrates how C_{llr} values vary for different detection tasks among the three selected methods. Each point corresponds to a single 1-vs-Rest labeling goodness-of-detection in minimal C_{llr} obtained using original data versus its low-dimensional representation: 3D from diffusion t-SNE (△), 10D from CAE (□), and 10D after complex fine-tuning (○).

To summarize, t-SNE was found to be suitable for 2D and 3D representations, but for 10D representation t-SNE was outperformed by CAE. Preprocessing by RPCA did not provide any significant improvement for methods analyzed. From various possible distances for t-SNE, Euclidean seems to be rather sufficient. Diffusion distance, although better than Earth mover's, provides similar performance to Euclidean. Classical CAE with 1 layer and 32 filters proved to be the best for 10D representation and complex fine-tuning allowed this representation to outperform the original slightly.

4 Conclusion and Future Work

Comparison of dimensionality reduction methods for bivariate histograms revealed that Euclidean or diffusion distance-based t-SNE is useful for visualization purposes (i.e. for producing 2D or 3D representations), while classical 1 layer 32 filters CAE is useful for learning a more generic representation. Low-dimensional CAE-based representation after supervised fine-tuning was able to outperform original representation slightly, but non-significantly, in various detection tasks. Bivariate histogram can be effectively compressed into universal low-dimensional representation, which can be further adapted to the supervised task at hand to achieve the discriminatory power of the original representation.

One promising direction of future research concerns using other pairs of sensors to obtain a more comprehensive comparison of the methods. Combination of multiple bivariate histograms could also be jointly compressed using CAE, if treating sensor pairs as separate channels, similarly to RGB in color images. Another idea is to exploit "repeated-measures" aspect of historical information due to regular reporting of on-board sensor data, which could help to find effective representations with regard to temporal evolution and not only a single snapshot.

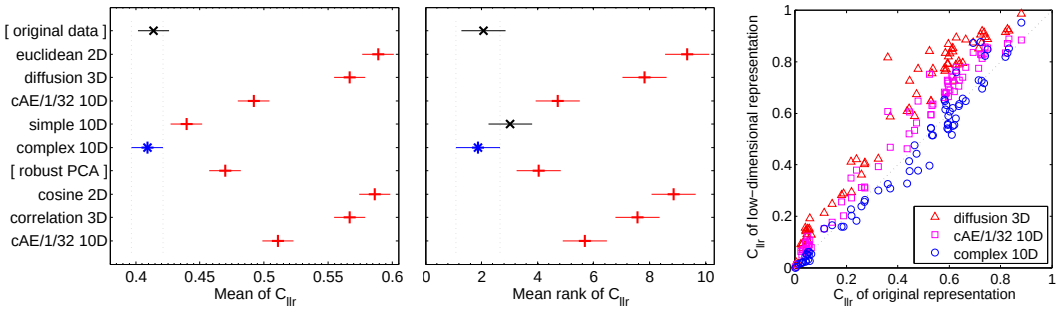

Figure 2: Results of the multiple comparisons procedure (using Tukey's HSD criterion with 95% confidence) for detection performance: parametric repeated-measures ANOVA (**left**) and the non-parametric Friedman test (**middle**). Of the 10 presented methods, the top 6 lines use original data, while bottom 4 lines use RPCA-transformed data. The best result is denoted by an asterisk (*), results similar to the best one by a cross (×) sign, and statistically significantly worse results are denoted by a plus (+) sign. Scatter plot (**right**) reveals performance for the 3 methods in each of 74 detection tasks.

REFERENCES

Aleksandr Aravkin, Stephen Becker, Volkan Cevher, and Peder Olsen. A variational approach to stable principal component pursuit. In Nevin L. Zhang and Jin Tian (eds.), *30th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 32–41, Quebec City, Quebec, Canada, July 23-27 2014. AUAI Press, Corvallis, Oregon.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, October 2001. doi:10.1023/A:1010933404324.

Niko Brümmer and Edward de Villiers. The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF. *arXiv*, 1304(2865v1):1–23, April 2013. Presented at the NIST SRE'11 Analysis Workshop, Atlanta, December 2011. `http://sites.google.com/site/bosaristoolkit`.

Jean-Pierre Clamme and Ashok A. Deniz. Three-color single-molecule fluorescence resonance energy transfer. *ChemPhysChem*, 6(1):74–77, 2005. ISSN 1439-7641. doi:10.1002/cphc.200400261.

Timothy Dozat. Incorporating Nesterov momentum into Adam. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, pp. 1–4, San Juan, Puerto Rico, May 2 2016.

Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, November 1973. ISSN 0018-9472. doi:10.1109/TSMC.1973.4309314.

Andrew E. Johnson and Martial Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):433–449, May 1999. ISSN 0162-8828. doi:10.1109/34.765655.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, volume 1312, pp. 1–14, Banff, Canada, April 15 2014.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, August 2005. ISSN 0162-8828. doi:10.1109/TPAMI.2005.151.

Haibin Ling and Kazunori Okada. Diffusion distance for histogram comparison. In Andrew Fitzgibbon, Camillo J. Taylor, and Yann LeCun (eds.), *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 246–253. IEEE Computer Society, June 17-22 2006. doi:10.1109/CVPR.2006.99.

Haibin Ling and Kazunori Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(5): 840–853, May 2007. ISSN 0162-8828. doi:10.1109/TPAMI.2007.1058.

Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In Timo Honkela, Włodzisław Duch, Mark Girolami, and Samuel Kaski (eds.), *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN)*, volume Part I, pp. 52–59, Espoo, Finland, June 14-17 2011. Springer Berlin Heidelberg. ISBN 978-3-642-21735-7. doi:10.1007/978-3-642-21735-7˙7.

Budiman Minasny and Alex B. McBratney. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, 32(9):1378–1388, 2006. ISSN 0098-3004. doi:10.1016/j.cageo.2005.12.009.

Kimmo Rautiainen, Juha Kainulainen, Tuomo Auer, Jörgen Pihlflyckt, Jani Kettunen, and Martti T. Hallikainen. Helsinki university of technology l-band airborne synthetic aperture radiometer. *IEEE Transactions on Geoscience and Remote Sensing*, 46(3):717–726, March 2008. ISSN 0196-2892. doi:10.1109/TGRS.2007.914805.

Constantino Carlos Reyes-Aldasoro, Leigh J. Williams, Simon Akerman, Chryso Kanthou, and Gillian M. Tozer. An automatic algorithm for the segmentation and morphological analysis of microvessels in immunostained histological tumour sections. *Journal of Microscopy*, 242(3): 262–278, June 2011. ISSN 1365-2818. doi:10.1111/j.1365-2818.2010.03464.x.

Joshua Saxe and Konstantin Berlin. Deep neural network based malware detection using two dimensional binary program features. In *10th International Conference on Malicious and Unwanted Software (MALWARE)*, pp. 11–20, October 2015. doi:10.1109/MALWARE.2015.7413680.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, June 2014. ISSN 1532-4435.

Laurens van der Maaten and Geoffrey Hinton. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, November 2008.