Deep Embedded SOM - Additional material

Florent Forest
April 29, 2019

This is some additional material for the two papers on the Deep Embedded Self-Organizing Map model (DESOM):

- F Forest, M Lebbah, H Azzag and J Lacaille. Deep Embedded SOM: Joint Representation Learning and Self-Organization. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2019. [1]
- F Forest, M Lebbah, H Azzag and J Lacaille. Deep Architectures for Joint Clustering and Visualization with Self-Organizing Maps. *Workshop on Learning Data Representations for Clustering (LDRC), PAKDD*, 2019. [2]

The first paper introduces the DESOM model as well as clustering and classification benchmarks on 3 datasets (short 6-page version). The second paper provides more details about the model and the training procedure, and focuses on its genericity by presenting possible architecture variants for structured data such as images and sequences.

1 Influence of loss weighting hyperparameter

The first additional experiment consists in evaluating the influence of the hyperparameter γ on the performance of a DESOM model with an 8×8 map. We evaluated the unsupervised clustering accuracy, purity, NMI and ARI on MNIST for $\gamma \in \{0.0001, 0.001, 0.01, 0.1, 1.0, 10.0\}$ by performing a subsequent k-means clustering with $k = 10$. The means and standard deviations on 10 runs as a function of γ are represented on figure 1. High values of γ lead to degenerate solutions for the autoencoder parameters, where all decoded map prototypes are identical and the performance is very low. This is due to the fact that the model tries hard to optimize the SOM loss, which is much easier to optimize than the reconstruction loss as we observed during our experiments, thus neglecting the code quality. As a consequence, the performance is better for values of the parameter that put a smaller weight on the SOM loss. The best performance in terms of accuracy is obtained for $\gamma = 0.001$. We did not cross-validate the γ parameter since we would like to remain in a completely unsupervised setting, and kept the value of 0.001 across all experiments and datasets. Moreover, the model is not very sensitive to the value of the γ as long as the hyperparameter stays in the right order of magnitude.

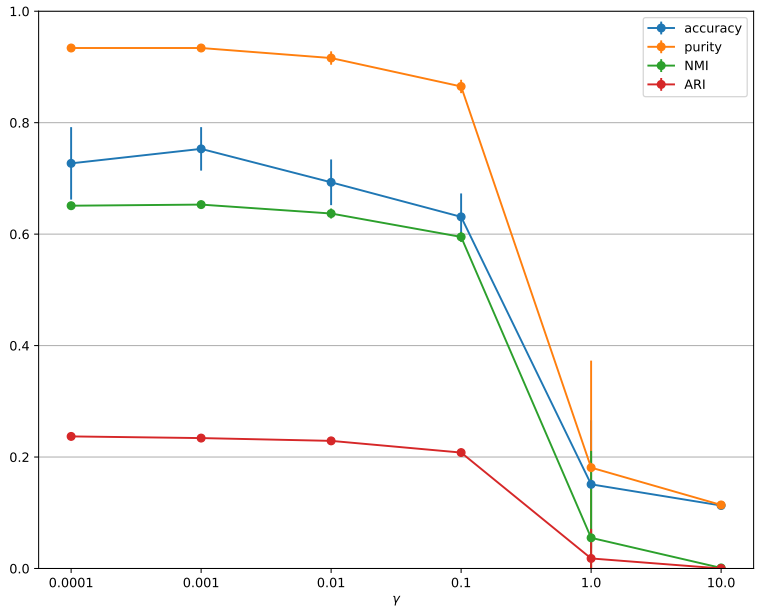

Fig. 1: Unsupervised clustering accuracy, purity, NMI and ARI evaluated on 10 runs on MNIST for different values of γ.

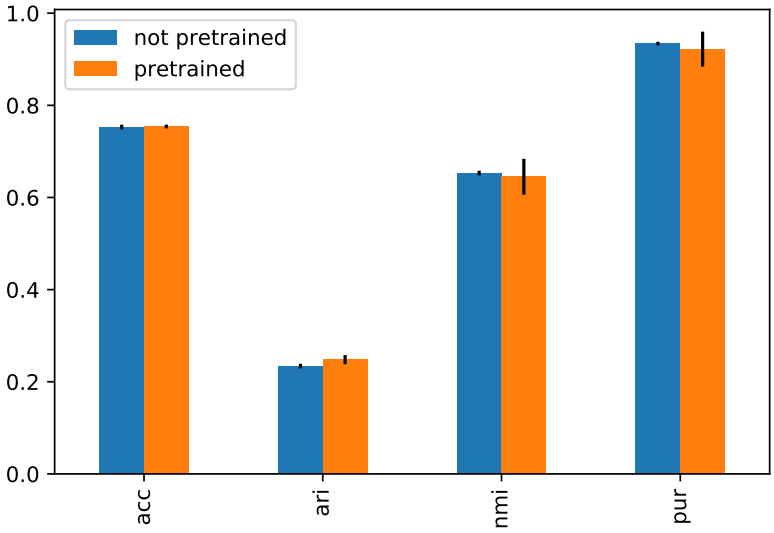

Fig. 2: Influence of autoencoder pretraining on the 4 metrics evaluated on 10 runs on MNIST.

2 Influence of autoencoder pre-training

The second experiment studies the influence of pre-training the autoencoder only with the reconstruction loss before performing the joint task. Usually, pre-training an autoencoder helps avoiding local minima by initializing the autoencoder weights at a good initial solution. There are several ways to pre-train an autoencoder: traditional end-to-end training, stacked denoising autoencoders [3] (also called layer-wise pre-training), RBM pre-training [4] etc. End-to-end training can be problematic because it could learn the identity function (but not an issue in the case of an under-complete autoencoder), and is less robust and prone to overfitting. Pre-training is used in most of deep clustering approaches, either layer-wise [5–7], RBM [8] or end-to-end [9], and improves results. We pre-trained the autoencoder in a simple end-to-end fashion for 200 epochs on MNIST, and compared the performance of a DESOM model trained for 10000 iterations first with random intialization of all weights, and then by using the pre-trained autoencoder weights (10 runs for meaningful means and standard deviations). The conclusion is that pre-training does not lead to a significant performance improvement (see figure 2).

Our explanation is that at the very beginning of the training, the SOM loss is very high which leads to strong gradients that completely disturb the encoder

weights and thus cancel out the pre-training (the reconstruction loss quickly increases, before decreasing again during the joint training). Even without pre-training, the learned representations are of good quality and sufficient for SOM visualization. As a conclusion, pre-training is not needed and learning of meaningful representations and self-organization is achieved jointly. This has the advantage of reducing the training time of our model.

References

[1] Florent Forest, Mustapha Lebbah, Hanane Azzag, and Jérôme Lacaille. Deep Embedded SOM: Joint Representation Learning and Self-Organization. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2019)*, pages 1–6, 2019.

[2] Florent Forest, Mustapha Lebbah, Hanane Azzag, and Jérôme Lacaille. Deep Architectures for Joint Clustering and Visualization with Self-Organizing Maps. In *Workshop on Learning Data Representations for Clustering (LDRC), PAKDD 2019*, pages 1–12, 2019.

[3] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

[4] Geoffrey E Hinton and Ruslan Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(July):504–507, 2006.

[5] J Xie, R Girshick, and A Farhadi. Unsupervised Deep Embedding for Clustering Analysis. In *ICML*, volume 48, 2015.

[6] B Yang, X Fu, N D Sidiropoulos, and M Hong. Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering. In *ICML*, 2016.

[7] Z Jiang, Y Zheng, H Tan, B Tang, and H Zhou. Variational Deep Embedding : An Unsupervised and Generative Approach to Clustering. In *IJCAI*, pages 1965–1972, 2017.

[8] C Song, Y Huang, F Liu, Z Wang, and L Wang. Deep auto-encoder based clustering. *Intelligent Data Analysis*, 18(6):S65–S76, 2014.

[9] Maziar Moradi Fard, Thibaut Thonet, and Eric Gaussier. Deep k-Means: Jointly Clustering with k-Means and Learning Representations. 2018.