Label-consistent sparse auto-encoders

Thomas Rolland
IRIT, UPS, CNRS, Toulouse, France
INESC-ID, Lisbon, Portugal
Email: thomas.rolland91@orange.fr

Adrian Basarab
IRIT, Université Paul Sabatier, CNRS
Toulouse, France
Email: adrian.basarab@irit.fr

Thomas Pellegrini
IRIT, Université Paul Sabatier, CNRS
Toulouse, France
Email: thomas.pellegrini@irit.fr

Auto-encoders (AE) is a particular type of unsupervised neural networks that aim at providing a compact representation of a signal or an image [1]. Such AEs are useful for data compression but most of the time the representations they provide are not appropriate as is for a downstream classification task. This is due to the fact that they are trained to minimize a reconstruction error and not a classification loss. Classification attempts with AEs have already been proposed such as contractive AEs [2], correspondence AEs [3] and stacked similarity-aware AEs [4], for instance. Inspired by label-consistent K-SVD (LC-KSVD) [5], we propose a novel supervised version of AEs that integrates class information within the encoded representations.

I. Label-consistent sparse coding (LC-KSVD)

Sparse Coding (SC) and (sparse) AEs share a similar objective of providing compact data representations. LC-KSVD consists in adding to the standard SC reconstruction error objective: i) a label consistency constraint (a ”discriminative sparse-code error”), ii) a classification error term. It results in a unified objective function that can be solved with the standard K-SVD algorithm. To do this, Jiang *et al* [5] defined the following objective function:

$$(D, A, W, \gamma) = \underset{D,\gamma,A,W}{\arg\min} \|X - D\gamma\|_2^2 + \lambda\|\gamma\|_0 + \underbrace{\mu\|Q - A\gamma\|_2^2}_{label-consistent\ term} + \underbrace{\beta\|H - W\gamma\|_2^2}_{consistency\ term} \quad (1)$$

where D and γ are respectively the dictionary and sparse codes to be estimated, Q is a matrix of discriminative sparse codes of the input signals X, A a linear transformation matrix, W a linear classifier and H the labels associated to X. Q arbitrarily associates to an input signal a number of dictionary atoms, with non-zero values occurring when signal i and atom ki share the same label (see Fig 1). Q is arbitrarily defined by the user with the possibility to let some atoms ”empty” by not assigning them any class (in white in Fig. 1).

II. Proposed label-consistent Sparse autoencoders (LC-SAE)

Fig. 2 shows the architecture of the proposed LC-SAE comprised of a standard sparse convolutional AE central part, completed with a ”H branch” the consistency terms from (2). These branch is a fully-connected layer with softmax activation layer. The AE was trained with the cross-entropy cost function, the categorical variant for the classification H-branch.
Inspired by (1) we proposed the following loss function for our LC-SAE training:

$$L_{LC-SAE} = L_{AE} + \lambda L_{Sparse} + \beta L_{label} \quad (2)$$

Where L_{AE} is the mean squared error between the reconstructed signal and the original signal. L_{Sparse} is the L1 regularization (i.e. Sparse regularization) with an arbitrary weight λ weight. Finally, L_{Label} is the categorical cross-entropy between the output of the classification branch and the actual label, with β arbitrary weight.

III. Experiments

We compare the feature representation methods on MNIST. After extracting the sparse discriminative representations with each method, we train and test k-means and SVM with radial kernel (RBF) on the training and evaluation subsets of MNIST comprised of 50k and 10k images, respectively. SVM and k-means allow to compare the discriminative power of the representations in supervised and unsupervised settings.

The hyperparameters were tuned for classification. For the sparse coding approaches (standard SC and LC-KSVD), we used 1024 (about twice the dimension of the images $d = 728$) atoms for the dictionaries and $\lambda = 1.2/\sqrt{728}$ as suggested in [6]. For LC-KSVD, we used $\mu = 5.0$ and $\beta = 2.0$, which are large values to promote discriminative power over reconstruction [5].

For the proposed LC-SAEs, the encoder part is comprised of three 3×3 convolution layers (16-10-10 filters respectively) with a rectifier-linear unit activation function, each followed by a 2×2 max-pooling layer for sub-sampling. The encoder output representations are 160-d vectors. Six variants of the proposed model are compared:

- Label-Consistent AE, without the sparse regularization with three different β values (0,1,2)
- Label-consistent Sparse AE with the ℓ_1-norm sparse regularization coefficient was tuned to 1e-7. Using the same three β values (0, 1, 2).

IV. Results

Fig. 3 shows examples of nine digit images from the MNIST eval subset with the original, reconstructed images on the first and second rows. The third row shows the sparse and the discriminative representations obtained with our method. These correspond to vector outputted by the AE that we reshaped in 2-d images for illustration. As can be seen, the reconstructed images are close to their original counterparts. Regarding the encoded activations shown in the third row, one can identify patterns similar between two instances of the same digit. Indeed, figure 4 represents the mean square error between all element corresponding to their classes. We can see that the minimum error is achieved for between all elements from the same class. That confirms our hypothesis that the extracted features are discriminative.

Table I gives a performance comparison between the different methods when using k-means (purity) and SVM (accuracy). For AEs, we always score the representation outputted by the encoder part of the model. As can be seen, SC and the sparse AE are not successful in providing discriminative representations that work with both clustering and SVM since purity values are close to chance (10%). Adding label-consistency constraints with, either to SC or AEs, drastically improve the representation separability, with 78%

purity for LC-KSVD, and 97-98% purity for LC-SAEs with $\beta > 0$. Finally, the LC-SAEs give the best results with the SVM classifier, showing that with only 16 atoms (i.e filter in the first layer) per class instead of about 100 with LC-KSVD, these models provide very discriminative encoded representations.

We showed in this work that the proposed LC-SAEs are effective in providing representations that allow for satisfactory image reconstructions and that embed discriminative information about the image classes. Ongoing experiments on other datasets are being conducted, such as tiny-imagenet and sound recordings (ESC-10), and similar trends are observed. It could also be interesting to add a constraint term α to the reconstruction loss L_{AE} in order to control perfectly the reconstruction power and discriminative power of our method.

ACKNOWLEDGMENT

This work was partly supported by the LABEX AMIES-UMS 3458, within the PEPS REP4SPEECH project, and the Agence Nationale de la Recherche LUDAU (Lightly-supervised and Unsupervised Discovery of Audio Units using Deep Learning) project (ANR-18-CE23-0005-01).

REFERENCES

[1] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1," D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds. Cambridge, MA, USA: MIT Press, 1986, ch. Learning Internal Representations by Error Propagation, pp. 318–362. [Online]. Available: http://dl.acm.org/citation.cfm?id=104279.104293

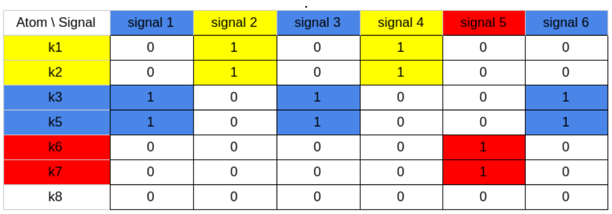
[2] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Omnipress, 2011, pp. 833–840.

[3] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Proc. ICASSP*. IEEE, 2015, pp. 5818–5822.

[4] W. Chu and D. Cai, "Stacked similarity-aware autoencoders." in *IJCAI*, 2017, pp. 1561–1567.

[5] Z. Jiang, Z. Lin, and L. S. Davis, "Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2651–2664, Nov 2013.

[6] Mairal, Bach, Ponce, and Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 689–696. [Online]. Available: http://doi.acm.org/10.1145/1553374.1553463



Atom \ Signal	signal 1	signal 2	signal 3	signal 4	signal 5	signal 6
k1	0	1	0	1	0	0
k2	0	1	0	1	0	0
k3	1	0	1	0	0	1
k5	1	0	1	0	0	1
k6	0	0	0	0	1	0
k7	0	0	0	0	1	0
k8	0	0	0	0	0	0

Fig. 1. Example for the user-defined Q matrix , each color corresponds to a class. In this example, signals 1, 3 and 6 belong to class 1; signals 2 and 5 to class 2 and signal 5 to class 3. Atom k8 is unassigned.

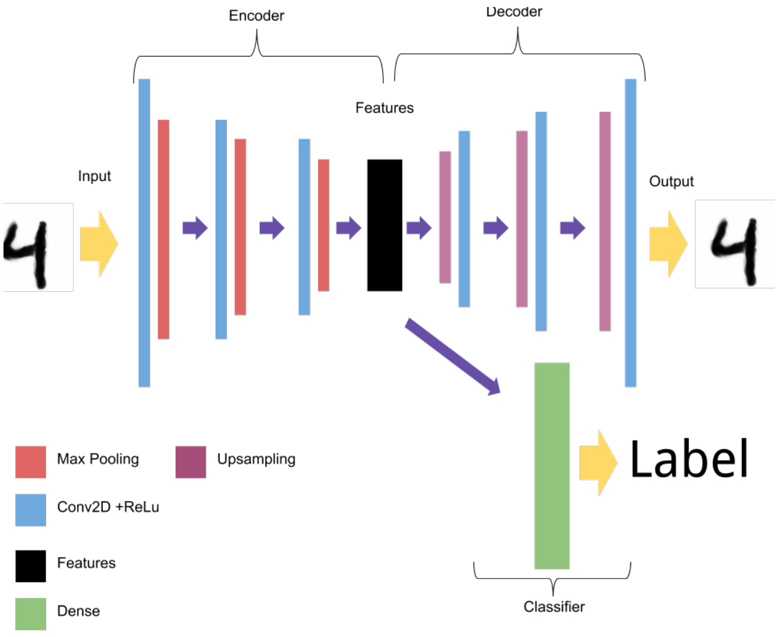


Fig. 2. The proposed LC-AE architecture.

Fig. 3. MNIST samples and representations obtained with a Sparse LC-AE: original images (top row), reconstructed images (second-top row), classification branch (bottom row)

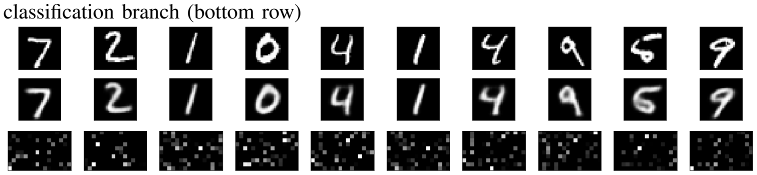
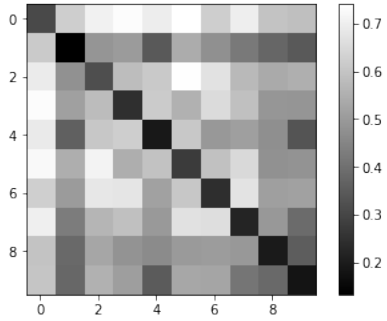


Fig. 4. Reconstruction error matrix for all elements of a class with all elements of the all other classes



Approach	K-means	SVM (RBF)	MSE loss
Sparse Coding	0.13	0.90	
LC-KSVD	0.78	0.91	
AE	0.19	0.96	
LC-AE ($\beta = 0$)	0.62	0.970	0.0155
LC-AE ($\beta = 1$)	0.95	0.990	0.0219
LC-AE ($\beta = 2$)	0.95	0.990	0.0237
LC-SAE ($\beta = 0$)	0.65	0.920	0.0173
LC-SAE ($\beta = 1$)	**0.98**	**0.992**	0.0237
LC-SAE ($\beta = 2$)	0.97	0.991	0.0232

TABLE I
PERFORMANCE COMPARISON IN TERMS OF PURITY FOR K-MEANS AND ACCURACY FOR SVM.