Contracting Auto-Encoders

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot and Yoshua Bengio
Dept. IRO, U. Montreal

Introduction. We propose a novel approach for training deterministic auto-encoders. We show that by adding a well chosen penalty term to the classical reconstruction cost function, we can achieve results that equal or surpass those attained by other regularized auto-encoders as well as denoising auto-encoders on a range of data sets. This penalty term corresponds to the Frobenius norm of the Jacobian matrix of the encoder activations with respect to the input. We show that this penalty term results in a localized space contraction which in turn yields robust features on the activation layer. We find empirically that this penalty helps to carve a representation that better captures the local directions of variation dictated by the data, corresponding to a lower-dimensional non-linear manifold, while being more invariant to the vast majority of directions orthogonal to the manifold. Finally, we show that by using the learned features to initialize a MLP, we achieve state of the art classification error on a range of data sets, surpassing other methods of pre-training.

Contracting auto-encoder (CAE). The standard encoderRumelhart *et al.* (1986); Baldi and Hornik (1989) is a function f that maps an input $x \in \mathbb{R}^{d_x}$ to hidden representation $h(x) \in \mathbb{R}^{d_h}$. It has the form $h = f(x) = s_f(Wx+b_h)$, where s_f is a nonlinear *activation function*, typically a logistic $\mathrm{sigmoid}(z) = \frac{1}{1+e^{-z}}$. The encoder is parametrized by a $d_h \times d_x$ weight matrix W, and a bias vector $b_h \in \mathbb{R}^{d_h}$.

The decoder function g maps hidden representation h back to a reconstruction y: $y = g(h) = s_g(W'h + b_y)$, where s_g is the decoder's activation function, typically either the identity (yielding linear reconstruction) or a sigmoid. The decoder's parameters are a bias vector $b_y \in \mathbb{R}^{d_x}$, and matrix W'. In this work we only explore the tied weights case, in which $W' = W^T$.

Auto-encoder training consists in finding parameters $\theta = \{W, b_h, b_y\}$ that minimize the reconstruction error on a training set of examples D_n.

From the motivation of robustness to small perturbations around the training points, we propose a regularization term that corresponds to the Jacobian of the hidden representation with respect to the input $J_f(x) = \frac{\partial h}{\partial x}(x)$ which favors mappings that are more strongly contracting at the training samples. The resulting contracting auto-encoder (CAE) can then be expressed as:

$$\mathcal{J}_{\mathrm{CAE}}(\theta) = \sum_{x \in D_n} L(x, g(f(x))) + \lambda \|J_f(x)\|_F^2, \qquad (1)$$

where L is the reconstruction error. Typical choices include the squared error $L(x, y) = \|x - y\|^2$ used in cases of linear reconstruction and the cross-entropy loss when s_g is the sigmoid (and inputs are in $[0, 1]$): $L(x, y) = -\sum_{i=1}^{d_x} x_i \log(y_i) + (1 - x_i) \log(1 - y_i)$.

Experiments and results

Classification. We used the weights learned by the CAE to initialize a multi-layer neural network. We then compared the classification results obtained with other methods for initializing deep networks on a range of data sets Larochelle *et al.* (2007)[1]. In most case, we achieved state of the art results as can be seen in the table below

Geometric contraction. We compared the contraction effect of the CAE to other auto-encoder variants by two different methods. We first calculated for each example of the validation set the singular value decomposition of the Jacobian to obtain the average spectrum 1. We also measured the ratio of contraction of different models used for unsupervised training, by generating artificial samples around the data set points and calculating the ratio of their distance in the input space and in the feature space.

Conclusion. In this paper, we attempt to answer the following question: *what makes a good representation?*. Besides being useful for a particular task, which we can measure, or towards which we can train a representation, this paper highlights the advantages for representations to be *locally invariant in many directions* of change of the raw input. This idea is implemented by a penalty on the Frobenius norm of the Jacobian matrix of the encoder mapping, which computes the representation. The paper also introduces

[1]Data sets available at `http://www.iro.umontreal.ca/~lisa/icml2007`.

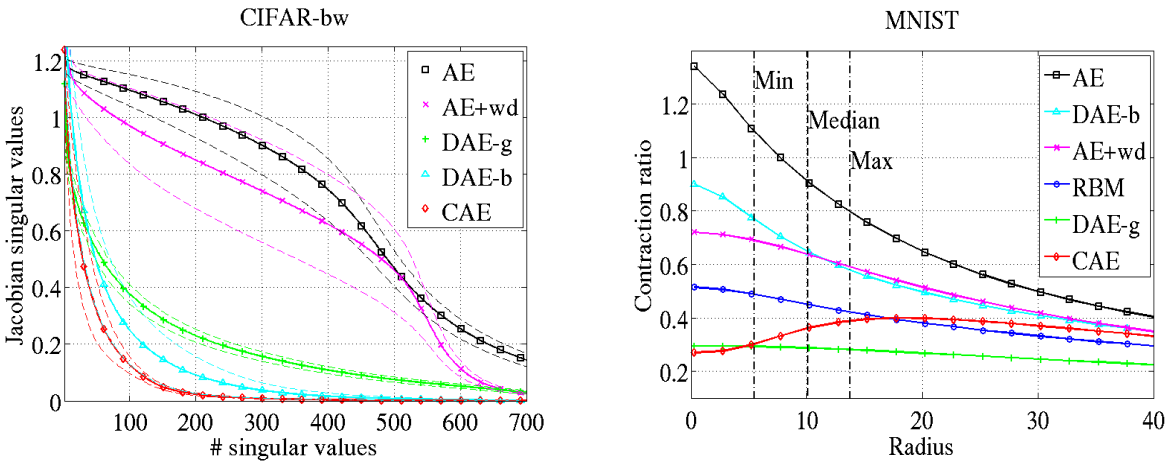

Figure 1: **Left**: Average spectrum of the encoder's Jacobian, for the CIFAR-bw dataset. Large singular values correspond to the local directions of "allowed" variation learned from the dataset. The CAE having fewer large singular values and a sharper decreasing spectrum suggests that it does a better job of characterizing a *low-dimensional manifold* near the training examples.**Right**: Contraction curves obtained with the considered models on MNIST show the CAE has a more localized contraction near the dataset points than other models. Furthermore, the contraction does not increase in a monotone fashion as with the other models, thus resulting in a more disentangled space.

Data Set	**SVM**$_{rbf}$	**SAE-3**	**RBM-3**	**DAE-b-3**	**CAE-1**	**CAE-2**
basic	3.03±0.15	3.46±0.16	3.11±0.15	2.84±0.15	2.83±0.15	**2.48**±0.14
rot	11.11±0.28	10.30±0.27	10.30±0.27	**9.53**±0.26	11.59±0.28	**9.66**±0.26
bg-rand	14.58±0.31	11.28±0.28	**6.73**±0.22	10.30±0.27	13.57±0.30	10.90 ±0.27
bg-img	22.61±0.379	23.00±0.37	16.31±0.32	16.68±0.33	16.70±0.33	**15.50**±0.32
bg-img-rot	55.18±0.44	51.93±0.44	47.39±0.44	**43.76**±0.43	48.10±0.44	45.23±0.44
rect	2.15 ±0.13	2.41±0.13	2.60±0.14	1.99±0.12	1.48±0.10	**1.21**±0.10
rect-img	24.04±0.37	24.05±0.37	22.50±0.37	**21.59**±0.36	**21.86**±0.36	**21.54**±0.36

Table 1: Comparison of stacked contracting auto-encoders with 1 and 2 layers (CAE-1 and CAE-2) with other 3-layer stacked models Vincent *et al.* (2008, 2010) and baseline SVM. Test error rate on all considered classification problems is reported together with a 95% confidence interval. Best performer is in bold, as well as those for which confidence intervals overlap. Clearly CAEs can be successfully used to build top-performing deep networks. 2 layers of CAE often outperformed 3 layers of other stacked models.

empirical measures of robustness and invariance, based on the contraction ratio of the learned mapping, at different distances and in different directions around the training examples. We hypothesize that this reveals the manifold structure learned by the model, and we find (by looking at the singular value spectrum of the mapping) that the Contracting Auto-Encoder discovers lower-dimensional manifolds. In addition, experiments on many data sets suggest that this penalty always helps an auto-encoder to perform better, and competes or improves upon the representations learned by Denoising Auto-Encoders or RBMs, in terms of classification error.

References

Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, **2**, 53–58.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. In Z. Ghahramani, editor, *ICML 2007*, pages 473–480. ACM.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, **323**, 533–536.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *ICML 2008*.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, **11**(3371–3408).