Higher Order Contractive Auto-Encoder (CAE+H)

Salah Rifai, Grégoire Mesnil, Pascal Vincent, Xavier Muller,
Yoshua Bengio, Yann Dauphin, Xavier Glorot

Laboratoire d'Informatique
des Systèmes Adaptatifs
http://www.iro.umontreal.ca/~lisa

Université de Montréal

Overview of the presentation

The basics

- Auto-encoders learn efficient representations by trying to reconstruct the data
- Typical architecture is similar to a one layer MLP but where the output tries to be identical to the input

- **Encoder** : $h = f(x) = s(Wx + b_h)$
- **Decoder** : $y = g(h) = s(W'h + b_y)$

and s is a nonlinear *activation function*, typically a logistic function $\mathrm{sigmoid}(z) = \frac{1}{1+e^{-z}}$

Reconstruction error

The cost function usually corresponds to the reconstruction mean square error or cross-entropy:

- $L(x, y) = \|x - y\|^2$
- $L(x, y) = -\sum_{i=1}^{d_x} x_i \log(y_i) + (1 - x_i) \log(1 - y_i)$

Criterion

$$\mathcal{J}_{\mathrm{AE}}(\theta) = \sum_{x \in D_n} L(x, g(f(x))) \tag{1}$$

First Order Contractive Auto-Encoder

- Same as a regular auto-encoder but with an added penalty to the cost function
- Penalty corresponds to the Frobenius norm of the Jacobian of the hidden layer

Criterion

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

and

$$\|J_f(x)\|_F^2 = \sum_{i=1}^{d_h} (h_i(1-h_i))^2 \sum_{j=1}^{d_x} W_{ij}^2 \tag{3}$$

Higher Order CAE

- Computing parameters' gradient through higher orders derivatives of h is expensive.
- Instead we use a stochastic approximation of the Hessian Frobenius norm.

$$\|H_f(x)\|^2 = \lim_{\sigma \to 0} \frac{1}{\sigma^2} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)} \left[||J_f(x) - J_f(x+\epsilon)||^2 \right] \tag{4}$$

Criterion

$$\mathcal{J}_{\mathrm{CAE+H}}(\theta) = \mathcal{J}_{\mathrm{CAE}}(\theta) + \gamma \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)} \left[||J_f(x) - J_f(x+\epsilon)||^2 \right] \tag{5}$$

Why penalize the derivative's norm?

- Invariance: Encourages invariance of the hidden layer to small changes by contracting locally the input space.
- Locality: The projection in the feature space is locally contractive. Locality depends on the order of the derivative penalized.

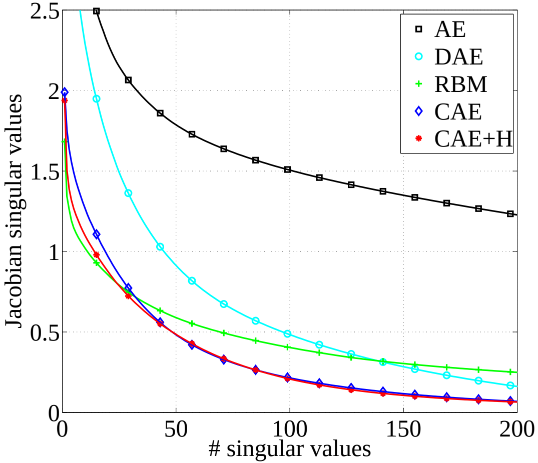
Geometric interpretation of the CAE+H

- Measure how contractive are the learnt features near sample points:
  - Locally: Spectrum of the jacobian
  - Globally: Contraction ratio as we move further away from sample points
- Compare the features of the CAE+H with other algorithms

Local space contraction

- Measure the spectrum of the singular values of the Jacobian at sample points.
- Average over many samples to see how the feature space has been contracted locally.
- This gives us an idea of the directions of contraction.

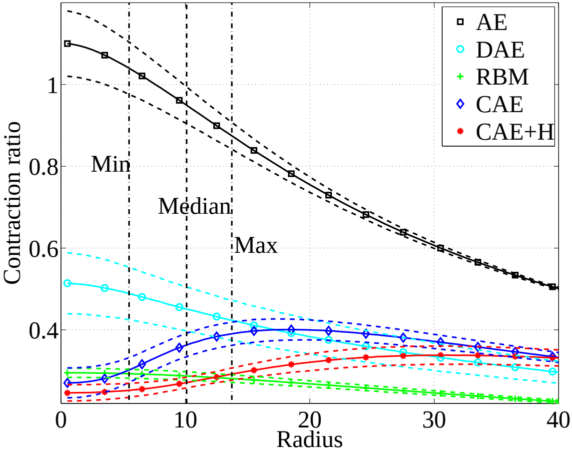
Local space contraction



Contraction Ratio

- We can estimate the isotropic contraction as a function of distance from sample points.
- Generate samples on a sphere of varying radius centered on an example.
- Measure the average distance of those points in the feature space as a ratio of the radius.
- This gives us an idea of how the space is deformed far from the samples.

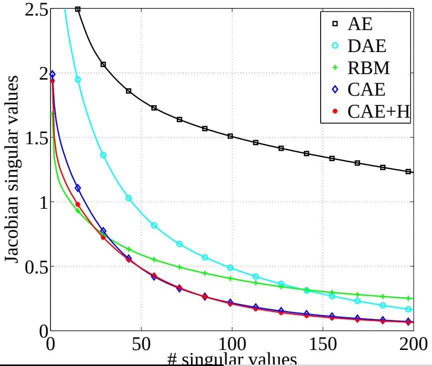
Contraction Ratio



Local contraction

Two observations from previous graphs:

- Highly localized contraction near sample points.
- However, a few directions are almost not contracted and there is a sharp dropoff in the singular values.



Reconstruction vs. Contraction penalty

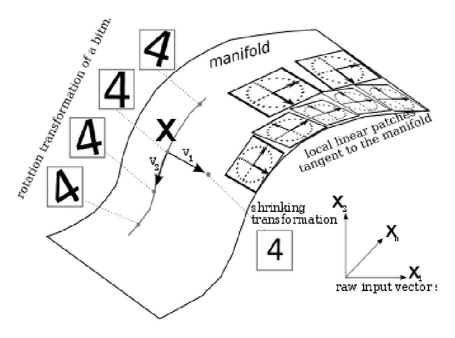
- The penalty is trying to make the features invariant in all directions near the samples by contracting isotropically
- The reconstruction cost is ensuring that the reconstruction is faithful by limiting the contraction in certain directions
- These directions correspond to the low dimensional manifold where the neighboring samples congregate

Approximating the manifold using the encoder's mapping

- We have no analytical parametrisation of the manifold.
- The contractive auto-encoder learns the directions of variation in the data.
- By looking at the directions and the magnitude of the principal singular vectors, we get an idea of the local dimensionality of the manifold (its local tangent)
- No prior knowledge is needed on these factors of variations as they are learned from the data

Manifold learning context

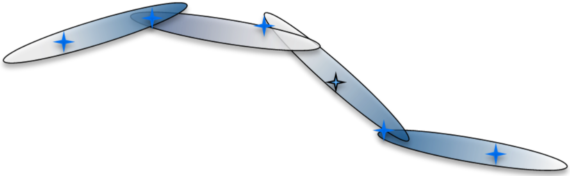
- Variations in the data correspond to dimensions parallel to the manifold,
- The orthogonal subspace to the manifold $\rightarrow$ unlikely data,
- Local directions are spanned by the PC of the Jacobian.



Local charts of the manifold

A linear local chart is a set of vectors associated to a datapoint.

- We can construct an atlas of the manifold using the union of local charts.
- Each local chart of this atlas is the low-dimensional tangent space to the manifold given by the first few singular values of the Jacobian



Local coordinates and saturation

Interpreting the hidden representation as a coordinate system.

- CAE+H yields highly saturated units (sparse representation) $\rightarrow$ null jacobian for these units
- Only non-saturated(linear) units are responsible for the high values in the spectrum of the Jacobian $\rightarrow$ directions of the local charts.

Non-saturated units $\Longrightarrow$ local coordinates.
Saturated units $\Longrightarrow$ global coordinates.

Formal definition of the atlas

- we define a local chart around x using the Singular Value Decomposition of $J^T(x) = U(x)S(x)V^T(x)$

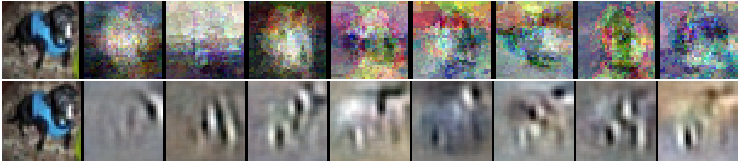
The tangent plane $\mathcal{H}_x$ at x is the span of the set of principal singular vectors B_x:

$$\mathcal{B}_x = \{U_{\cdot k}(x) | S_{kk}(x) > \epsilon\} \quad \text{and} \quad \mathcal{H}_x = \text{span}(\mathcal{B}_x),$$

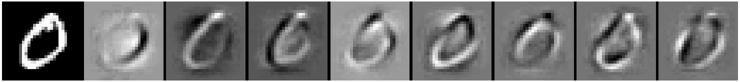
We can thus define an atlas $\mathcal{A}$ captured by h:

$$\mathcal{A} = \{(x, v) | x \in \mathcal{D}, v \in \mathcal{H}_x)\} \tag{6}$$

Visualizing the tangents



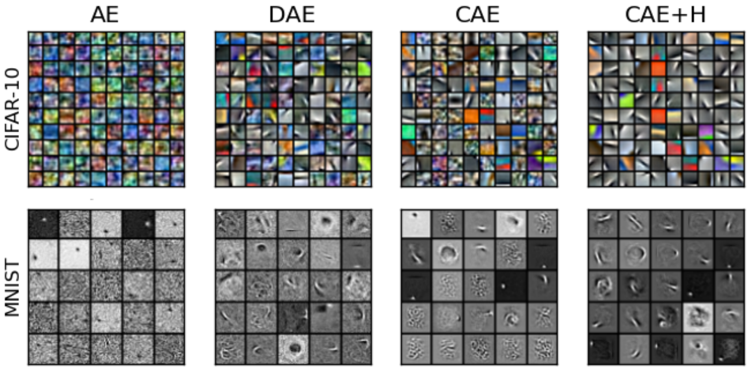
- Tangents learned on RCV1 and MNIST:



Trading & Markets	+gilt +yen +usda	-slow -term -debt	+matur +auction +treasur	-percent -sent -pressure	+bln +coupon +discount	-anti -predict -belgian	+interest +calcul +overnight	-sen -californ -introduc

Overcomplete representation

CAE+H benefits more from overcomplete representations.



CAE+H features

Model \ pretrain	AE	RBM	DAE	CAE	CAE+H
LogReg	2.17±0.29	2.04±0.28	2.05±0.28	1.82±0.26	**1.2**±0.21
MLP	1.78±0.26	1.3±0.22	1.18±0.21	1.14±0.21	**1.04**±0.20

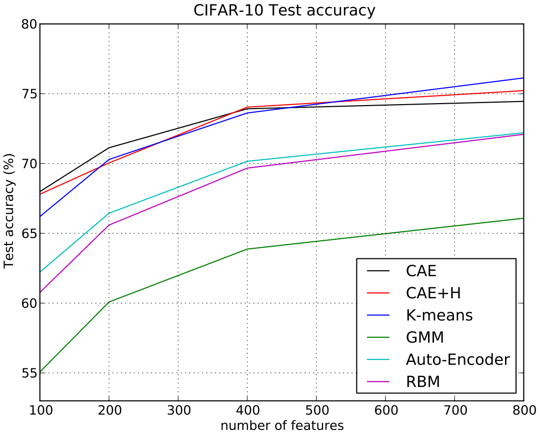
Table: Comparison of the quality of extracted features from different models when using them as the fixed inputs to a logistic regression (top row) or to initialize a MLP that is fine-tuned (bottom row).

CAE+H: What happens when we go deep?

Data Set	SVM_{rbf}	SAE-3	RBM-3	DAE-b-3	CAE-2	CAE+H-1	C AE+H-2
rot	11.11±0.28	10.30±0.27	10.30±0.27	**9.53**±0.26	**9.66**±0.26	10.9±0.27	**9 .2**±0.25
bg-img	22.61±0.379	23.00±0.37	16.31±0.32	16.68±0.33	15.50±0.32	15.9±0.32	**14.8** ±0.31
rect	2.15±0.13	2.41±0.13	2.60±0.14	1.99±0.12	1.21±0.10	**0.7**±0.07	**0.45**±0.06

Table: Comparison of stacked second order contractive auto-encoders with 1 and 2 layers (CAE+H-1 and CAE+H-2) with other 3-layer stacked models and baseline SVM.

CIFAR10 performance



We achieved a test error of 78.5% on CIFAR-10

Future Work

- Extending our definition of the local chart to a higher order approximation (curvy surfaces),
- Sampling new data points moving along the manifold approximation.
- Supervised learning algorithms taking advantage of the atlas extracted by the CAE+H (to appear in NIP2011)

Thanks to ...

theano

CRSNG
NSERC