Regularization of Contractive Autoencoders using the Schatten Norm

Adam Maus
Department of Computer Sciences
University of Wisconsin - Madison
Madison, WI 53706
maus@cs.wisc.edu

Brian Nixon
Department of Computer Sciences
University of Wisconsin - Madison
Madison, WI 53706
nixon@cs.wisc.edu

Abstract

Contractive autoencoders have been used in unsupervised learning to learn a useful representation of data that minimizes reconstruction error and captures manifold structure through regularization of the hidden layer by the Frobenius norm. We study a more general contractive autoencoder by replacing the Frobenius norm penalty term with the Schatten norm. The Schatten norm is a matrix norm that is equivalent to the p-norm of the singular values of a matrix with the Forbenius norm being a special case of the Schatten Norm with p = 2. We study the effect of the choice of p in classification when the autoencoder is used to project data into a overcomplete feature space. We also look at the autoencoder's sensitivity to changes in the input data and learned manifold structure. Lastly, we study how classification can be improved by using the autoencoder to perform dimensionality reduction.

1 Introduction

Autoencoders are unsupervised artificial neural networks used to learn new representations of data [Hinton 1995, Hinton 2006] with applications in hand-written character recognition, image recognition [Tan 2010], novelty assessment [Thompson 2002], and data visualization [Nadjarpoorsiyahkaly 2011]. Traditionally, autoencoders have been used to learn low-dimensional codes of data but new types of autoencoders have been introduced that learn useful high-dimensional representations of data with applications in deep learning [Bengio 2009], manifold learning [Tenenbaum 2000, Roweis 2000], initialization of state-of-art classifiers [Bengio 2007], Deep Belief Networks [Hinton 2006 - Neural Comp], and Deep Boltzmann Machines [Salakhutdinov 2009]. Autoencoders of this nature focus more on learning a sparse network structure and invariance to noisy data. To encourage sparsity and weight decay, a penalty term is added to the loss function that requires the sum of the weight values to be as small as possible [Kavukcuoglu 2009, Lee 2008]. Denoising autoencoders encourage invariance by first corrupting the input with noise and then try to learn the uncorrupted form of the data [Vincent 2010].

The basic autoencoder maps data of dimension d into an h dimensional space. The autoencoder encodes data, x, using $e(x) = f_1(Wx + b_1)$ and decodes it to the original input space with $d(x) = f_2(W^T e(x) + b_2)$, where the activation functions $f_1(x)$, $f_2(x)$ take the form of the sigmoid function $\frac{1}{1+e^{-x}}$. $f_2(x)$ may also be a linear activation function depending on the application. W is a matrix of real numbers of size $h \times d$, b_1 and b_2 are bias vectors of real numbers of size h and d, respectively. The autoencoder learns an accurate reconstruction of data by randomly initializing the parameters W, b_1, and b_2 and is trained through gradient descent while trying to minimize the objective function $L(W, b_1, b_2) = \sum_i^n ||x_i - d(e(x_i))||^2$. Autoencoders that encourage weight decay add the following term $\lambda \sum_{i,j} W_{ij}^2$.

Table 1: Results of KNN classifiers on first 1,000 samples of MNIST

CAE Encoding	Error
Schatten $p = 1$	23.50%
Schatten $p = 2$	0.0%
Schatten $p = \infty$	0.0%
KNN with no encoding	27.25%

2 Contractive Autoencoders (CAE)

Contractive autoencoders introduced by Rifai et al. [Rifai 2011] are similar in spirit to autoencoders with weight decay however instead of trying to minimize the weights, the contractive autoencoder objective function is $L(W, b_1, b_2) = \sum_i^n ||x_i - d(e(x_i))||^2 + \lambda ||J_e(x_i)||_{S_p}$ where $||Y||_{S_p}$ represents the Schatten norm of Y with a value of p. Rifai et al. use the Schatten norm with $p = 2$ also known as the Frobenius norm of the Jacobian of the nonlinear encoding function, $J_e(x_i) = \partial e(x_i)/\partial x_i$. Through this regularization, contractive autoencoders encourage a sparse representation of the data that is also locally invariant and maximizes contraction around each of the training points. The contraction is maximized in directions orthogonal to the manifold so representations change very little in these directions while parallel directions have the most change in representation.

In addition to the regularization, CAEs learn deep representations of the data by using an overcomplete representation of the data. An overcomplete representation is produced by choosing more hidden nodes than dimensions allowing the autoencoder to have multiple perfect reconstructions. By also including the regularization, the autoencoder must choose reconstructions that are robust to small changes to the data with the added benefit that the CAE learns the manifold's tangent directions. It should be noted, that this regularization is not limited to autoencoders and other models could also benefit by maximizing the contraction around each data point.

3 Experiments

Training contractive autoencoders involves minimizing the $L(W, b_1, b_2)$. For these results, a gradient descent method was used to train the weights, W, b_1, and b_2.

We compare the performance of the CAE with varying p using the basic MNIST dataset. After training each CAE with 1,000 hidden nodes on the first 400 samples taken from the dataset, the CAE was tested and compared by using a k nearest neighbors classifier on the MNIST test data. The classifier was set up such that the trained CAE encoded each training point and each test point. For each encoded test point, the Euclidean distance was computed between it and each training point. The k closest neighbors voted on the label for the test point with ties were broken randomly using a uniform distribution. Results can be seen in Table **??**.

One way to measure the invariance of in training data is to study how the encoding changes as you vary the data. Figure **??** shows the contraction curves for each of the CAEs with 1,000 hidden nodes trained on 400 samples from the MNIST data set.

Additionally, a contraction curve can be calculated after each epoch during the CAE training.

Lastly, the average singular values for each trained CAE can be plotted as shown in Figure **??**. The most sensitive directions have high singular values with values that quickly decrease due to the contractive nature of the autoencoder.

4 Conclusions

4.1 Headings: second level

Second level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 10. One line space before the second level heading and 1/2 line space after the second level heading.

4.1.1 Headings: third level

Third level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 10. One line space before the third level heading and 1/2 line space after the third level heading.

5 Citations, figures, tables, references

These instructions apply to everyone, regardless of the formatter being used.

5.1 Citations within the text

Citations within the text should be numbered consecutively. The corresponding number is to appear enclosed in square brackets, such as [1] or [2]-[5]. The corresponding references are to be listed in the same order at the end of the paper, in the **References** section. (Note: the standard BIBTEX style `unsrt` produces this.) As to the format of the references themselves, any style is acceptable as long as it is used consistently.

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]", not "In our previous work [4]". If you cite your other papers that are not widely available (e.g. a journal paper under review), use anonymous author names in the citation, e.g. an author of the form "A. Anonymous".

5.2 Footnotes

Indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).[2]

5.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction; art work should not be hand-drawn. The figure number and caption always appear after the figure. Place one line space before the figure caption, and one line space after the figure. The figure caption is lower case (except for first word and proper nouns); figures are numbered consecutively.

Make sure the figure caption does not get separated from the figure. Leave sufficient space to avoid splitting the figure and figure caption.

You may use color figures. However, it is best for the figure captions and the paper body to make sense if the paper is printed either in black/white or in color.

5.4 Tables

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See Table **??**.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

[1] Sample of the first footnote

[2] Sample of the second footnote

Figure 1: Sample figure caption.

Table 2: Sample table title

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

6 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

References follow the acknowledgments. Use unnumbered third level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to 'small' (9-point) when listing the references. **Remember that this year you can use a ninth page as long as it contains *only* cited references.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609-616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer-Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.