Learning generative models of handwritten digit images

Jiann-Ming Wu and Zheng-Han Lin

Correspondence: Jiann-Ming Wu, Department of Applied Mathematics, National Donghwa University, Hualien,Taiwan, R.O.C. Tel. 8863-8662500 ext. 21126, FAX : 8863-8662532, email: jmwu@server.am.ndhu.edu.tw

Abstract

This work explores generative models of handwritten digit images using natural elastic nets. The analysis aims to discover global features as well as distributed local features of handwritten digits. These features are expected to form a significant basis for discriminant analysis of handwritten digits and related analysis of handwritten characters that are beyond digits but possess similar features.

Keywords: neural networks, generative models, cortical maps, elastic net, Potts model, self-organization, unsupervised learning, handwritten digits.

I. Introduction

The natural elastic net has been applied to the analysis of natural images and has successfully achieved the orientation, localization and band pass feature of natural images[7][8]. This work will further extend the natural elastic net to the analysis of handwritten digit images, aiming to extract essential features and explore underlying generative models, that are expected to constitute an effective basis for processing character images beyond the training set. Related numerical experiments use the handwritten digit images from the well know USPS database [5].

Parameters of a natural elastic net is composed of a set of local means ordered on a lattice and a common covariance matrix. They characterize a set of disjoint multivariate Gaussian distribution serving as a generative model to training patches within a training set. A patch may contain global informations or local informations of a handwritten digit image depending on the sampling method. If a training patch is simply a digit image, the learning process performs a task of clustering analysis, locating each cortical point at the center of a cluster of digit images and ordering cortical points on a lattice under the measure of the Mahalanobis distance associated with the covariance matrix. If a training patch belongs the space spanned by the eigen-digits corresponding to significant eigenvalues, the learning process searches for generative models to digit images filtered by global features of principle component analysis. Neither an entire digit image nor its projection on significant eigen-digits helps decomposition of digit images as well as extraction of essential local features. Following the idea motivated by the visual system in neuroscience[1], we define a set of overlapping receptive fields, each composed of a set of nearby pixels centered at a particular position on the whole digit image and sample the digit image through all receptive fields simultaneously. After scanning all digit images, the sampling achieves multiple training sets, each containing tremendous subimages or patches captured through a particular receptive field. Then each set of training patches feed to a particular natural elastic net to attain its own generative model, which is composed of local means and covariance matrix as local features of digit images corresponding this receptive field. These distributed generative models constitute an associative memory with capability of reconstructing a distorted noisy digit image and a partial digit image, and provide essential local features and internal representations for incremental learning of new characters beyond digit images.

We will briefly review the natural elastic net in section II, and describe the applica-

tion to learning generative models of handwritten digit images as well as related numerical simulations in section III, and the conclusion is in the last section.

II. Natural Elastic Nets

The generative model is a flip-coin process directly operating on M^2 disjoint multivariate Gaussian distributions. Each time the process randomly selects one distribution according to the prior probabilities $\{\pi_k, 1 \leqq k \leqq M^2\}$, and then triggers the selected distribution to generate a training stimuli. Let π_k be $\frac{1}{K}$ for all k and $K = M^2$, indicating an equal selection among disjoint distributions. Assume that each individual distribution is as the following multivariate Gaussian distribution

$$\begin{aligned} P_k(x) &= P(x|y_k, A) \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} |A^{-1}|} \exp\left(\frac{(x-y_k)^t A (x-y_k)}{2}\right) \end{aligned} \tag{1}$$

,where y_k and A are the local mean and the covariance matrix respectively, $|A^{-1}|$ denotes the determinant of the inverse of the matrix A, and x^t denotes the transpose of vector x.

A. The mathematical framework

According to the flip-coin process, each stimuli x_i is generated by one and only one individual distribution. Let the Potts neural variable $\delta_i = [\delta_{i1}, ..., \delta_{iK}]^t$ denote the membership of x_i with $\delta_{ik} \in \{0, 1\}$ for all k and $\sum_k \delta_{ik} = 1$. Therefore each δ_i belongs the set of $\{e^k, 1 \leqq k \leqq K\}$, where e^k is a standard unitary vector with the kth element one and the others zero. If δ_i is e^k, it is said that the ith training stimuli is generated by the kth individual distribution. The following local log likelihood function can then quantitatively measure the fitness of the individual distribution P_k to all of training stimulus whose membership vectors are e^k.

$$l_k = \log \prod_{\{i:\delta_i = e^k\}} P_k(x_i) = \sum_{\{i:\delta_i = e^k\}} \log P_k(x_i) \tag{2}$$

By summing up all l_k, we have the following log likelihood function

$$\begin{aligned} l &= \sum_k l_k \\ &= \sum_k \sum_{\{i:\delta_i = e^k\}} \log P_k(x_i) \\ &= -\frac{1}{2} \sum_i \sum_k \delta_{ik} (x_i - y_k)^t A (x_i - y_k) - \frac{N}{2} \log \left| A^{-1} \right| - \frac{Nd}{2} \log(2\pi) \end{aligned} \tag{3}$$

By neglecting the last constant term, reversing the sign and using the fact $\log|A^{-1}| = -\log|A|$, we obtain the first objective for the natural elastic net as follows

$$E_1 = \frac{1}{2} \sum_i \sum_k \delta_{ik} (x_i - y_k)^t A (x_i - y_k) - \frac{N}{2} \log|A| \tag{4}$$

The maximization of l is equivalent to the minimization of E_1 subject to the constraint of $\delta_i \in \{e^k, 1 \leqq k \leqq K\}$ for all i.

Following the minimal wiring principle, two nearby cortical points on the map should be as close as possible, which leads to the following minimal wiring criterion[2][3] for an effective dimensional-reduction mapping,

$$E_2 = \frac{1}{2} \sum_k \sum_{j \in NB(k)} \|y_k - y_j\|_A^2 \tag{5}$$

where $NB(k)$ is a set of all nodes connecting to the kth node on the lattice and $\|y\|_A^2$ denotes the Mahalanobis square length of a vector y, which is defined by $\|y\|_A^2 = y^t A y$.

Minimizing a weighted combination of E_1 and E_2 leads to the following mathematical framework for the natural elastic net.

Minimize

$$\begin{aligned} E &= E_1 + CE_2 \\ &= \frac{1}{2}\sum_i\sum_k \delta_{ik} \left\| x_i - y_k \right\|_A^2 - \frac{N}{2}\log|A| \\ &\quad + \frac{C}{2}\sum_k \sum_{j \in NB(k)} \left\| y_k - y_j \right\|_A^2 \end{aligned} \tag{6}$$

subject to

$$\sum_k \delta_{ik} = 1, \quad \forall i \tag{7}$$

where C is a weighting constant.

B. Dynamics for the Natural elastic net

The above mathematical framework is a mixed integer and linear programming. The optimization task involves with discrete combinatorial variables $\{\delta_i\}$ and continuous geometrical variables $\{y_k\}$, and the matrix A. Since the energy function E is not differentiable with respect to discrete variables, the gradient descent method can not be directly applied to the mathematical framework. By relating each membership vector δ_i to a Potts neural variable, the optimization task can be treated by a hybrid of the mean field annealing and the gradient descent, which has been successfully applied to the derivation of independent component analysis using Potts models[6].

The mean field equation can be derived from the following free energy function similar to the one proposed by Peterson and Söderberg[4]

$$\begin{aligned} \psi\left(A, Y, \langle\delta\rangle, u\right) &= E\left(A, Y, \langle\delta\rangle\right) + \sum_i\sum_m \langle\delta_{im}\rangle u_{im} \\ &\quad - \frac{1}{\beta}\sum_i \ln(\sum_m \exp\left(\beta u_{im}\right)) \end{aligned} \tag{8}$$

where u denote the set $\{u_i\}$ and each u_i is an auxiliary vector.

By setting

$$\frac{\partial \psi}{\partial \langle \delta_{im} \rangle} = 0 \; for \; all \; i, m \tag{9}$$

$$\frac{\partial \psi}{\partial u_{im}} = 0 \; for \; all \; i, m \tag{10}$$

we have the following mean field equation for evaluating mean activations of discrete neural variables

$$\begin{aligned} \mu_{im} &= -\frac{\partial E}{\partial \langle \delta_{im} \rangle} \\ &= -\frac{1}{2} (x_i - y_m)^t A (x_i - y_m) \end{aligned} \tag{11}$$

$$\langle \delta_{im} \rangle = \frac{\exp(\beta u_{im})}{\sum_k \exp(\beta u_{ik})} \tag{12}$$

The mean configuration satisfying the mean field equation (11) and (12) is a saddle point of the free energy (8) corresponding to a particular β value. Based on the mean configuration, we can apply the gradient descent method to derive the following updating rule for each y_m.

$$\begin{aligned} \Delta y_m &\propto -\frac{\partial E}{\partial y_m} \\ &= \frac{1}{2} \sum_i \langle \delta_{im} \rangle \left(A + A^t\right) (x_i - y_m) + \frac{C}{2} \sum_{n \in NB(m)} \left(A + A^t\right) (y_n - y_m) \end{aligned} \tag{13}$$

To zero gradient, such as $\Delta y_m = 0$, we have the following linear system.

$$(CN_m + \sum_i \langle \delta_{im} \rangle) y_{ma} - C \sum_{n \in NB(m)} y_{na} = \sum_i \langle \delta_{im} \rangle x_{ia}, \quad 1 \leqq m \leqq K, 1 \leqq a \leqq d$$

where N_m denotes the number of nodes in the set $NB(m)$. By solving the linear system, we have

$$[y_1, ..., y_K]^t = (C(H - G) + Z)^{-1} R \tag{14}$$

where both H and Z are $K \times K$ diagonal matrices with diagonal entries $H_{mm} = N_m$ and $Z_{mm} = \sum_i \langle \delta_{im} \rangle$, $1 \leqq m \leqq K$, respectively, G is a $K \times K$ adjacent matrix corresponding to the lattice with entries

$$\begin{aligned} G_{mn} &= 1 \text{ if the } m\text{th node and the } n\text{th node are connected,} \\ &= 0 \text{ otherwise,} \end{aligned}$$

, and the matrix R has entries

$$R_{ma} = \sum_i \langle \delta_{im} \rangle x_{ia}, \ 1 \leqq m \leqq K, \ 1 \leqq a \leqq d$$

The updating rule for each element A_{ab} in the covariance matrix can be derived as follows

$$\begin{aligned} \Delta A_{ab} &\propto -\frac{\partial E}{\partial A_{ab}} \\ &= -\frac{1}{2} \sum_i \sum_m \langle \delta_{im} \rangle (x_{ia} - y_{ma})(x_{ib} - y_{mb}) \qquad (15) \\ &\quad -C \sum_m \sum_{j \in NB(m)} (y_{ma} - y_{ja})(y_{mb} - y_{jb}) + \frac{N}{2} \left[\left(A^t \right)^{-1} \right]_{ab} \qquad (16) \end{aligned}$$

Again, when $\Delta A_{ab} = 0$, we have

$$A = \left(W^{-1} \right)^t \qquad (17)$$

where

$$\begin{aligned} W_{ab} &= \frac{1}{N} \sum_i \sum_m \langle \delta_{im} \rangle (x_{ia} - y_{ma})(x_{ib} - y_{mb}) \qquad (18) \\ &\quad + \frac{2C}{N} \sum_m \sum_{j \in NB(m)} (y_{ma} - y_{ja})(y_{mb} - y_{jb}) \end{aligned}$$

The following step-by-step statement is the learning process for the natural elastic net toward the minimum of the objective function (6).

1. Initialize β as a sufficiently low value, $A = 0.01 \times I$ (identity matrix), $y_k \approx \frac{1}{N} \sum_i x_i$, $\langle \delta_{ik} \rangle \approx \frac{1}{K}$

2. Update $\{\langle\delta_{im}\rangle\}$ by equations (11) and (12).

3. Update $\{y_m\}$ by equation (14).

4. Update A by equations (18) and (17).

5. If $\sum_i \sum_m \langle\delta_{im}\rangle^2 > \theta$ then halt, else $\beta \leftarrow \beta * \frac{1}{0.98}$, and *goto step* 2

where θ is a threshold , ex. $\theta = 0.98 * N$.

III. Numerical Simulations for learning generative models

In this work, we use the USPS handwritten digit database. This database consists of 9298 segmented numerals digitized from handwritten zip codes that appeared on real U.S. Mail passing through the Buffalo, N.Y. post office. Many different people, using a great variety of sizes, writing styles and instruments, wrote the digits. The training images consisted of 7291 handwritten digits and the remaining 2007 handwritten digits were used for future test.

Extraction of global features or distributed local features of the handwritten digit depends on the sampling method and the way of applying the natural elastic net.

A. Generative models for global features

For analysis of handwritten digits, a 20×20 natural elastic net is directly applied to the 7291 training images, each containing 16×16 grey-level pixels. As a result, the natural elastic net has 400 cortical points ordered on a 20×20 lattice and a matrix A as parameters of the generative model. The cortical points partition these training images into 400 non-overlapping clusters based on the Mahalanobis distance corresponding to the covariance matrix. The obtained cortical points essentially capture the local mean of handwritten digits. These cortical points are shown in figure 1.

As we apply the principle component analysis to these training images, we have a set of eigenvectors or eigen-digits, each corresponding to an eigenvalue. We select forty significant eigenvectors to form a basis for dimensionality reduction. The projection of 7291 training images on the forty eigen-digits leads to a new set of training samples, each composed of 40 elements. We then apply a 20×20 natural elastic net to learn the generative model of these training samples. As a result, we have a set of cortical points ordered on the lattice and a covariance matrix for the generative model. A training sample is exactly obtained from a training image, so it inherits the label of the training digit image. Assume that each training image has its own label denoting the digit type from zero to nine. We can assign each cortical point a label by looking up the label of its nearest training sample. With the labeling, the natural elastic net becomes a classifier that can recognize a digit image for testing. After projecting the 2007 testing images on the space spanned by the forty eigen-digits, we have 2007 testing samples, each also with its own label. Numerical simulations show that the classifier has a correct rate, 96.42%, for recognizing 2007 testing samples, which is comparable to the best known one, 97.5%[5].

B. Generative models for distributed local features

In neuroscience, the visual system receives visual signals from a set of localized receptive fields, transmitting the information from each receptive field to a corresponding population of visual neurons for further integration.

We use 49 natural elastic nets to emulate a visual system for the process of handwritten digits. The whole digit image, 16×16 pixels, is uniformly partitioned into 49 overlapping receptive fields, each containing 4×4 pixels. Then sampling the 7291 training images through these receptive fields produces 49 sets of training patches, where each patch contains 4×4

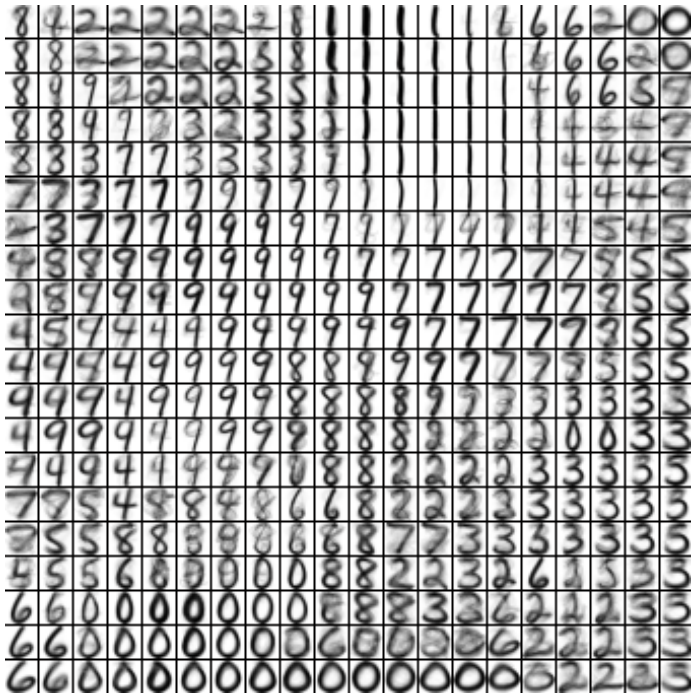

Figure 1: The cortical map of a 20x20 natural elastic net for handwritten digits.

pixels. Each set of training patches is associated with a 10×10 natural elastic net that emulates a population of visual neurons with the same receptive field. As a result, each natural elastic net has 100 cortical points ordered on the 10×10 lattice and a matrix A as parameters of the generative model for 7291 patches obtained through a corresponding receptive field. Figure 2 show all cortical points of 49 natural elastic nets in 7×7 blocks. Each block shows 10×10 cortical points of a natural elastic net and is located on a position relative to the center of the corresponding receptive field.

The generative model of distributed local features can be used to reconstruct an image that is beyond the training images. We select some from 2007 testing images, splitting each image into 49 overlapping patches through 49 receptive fields, then feeding each patch to a corresponding natural elastic net. Each natural elastic net generates a cortical point that is nearest to the patch. Combining 49 cortical points achieves a restored digit. 128 testing images are shown in the odd columns of figure 3, and the reconstructed digits are shown in

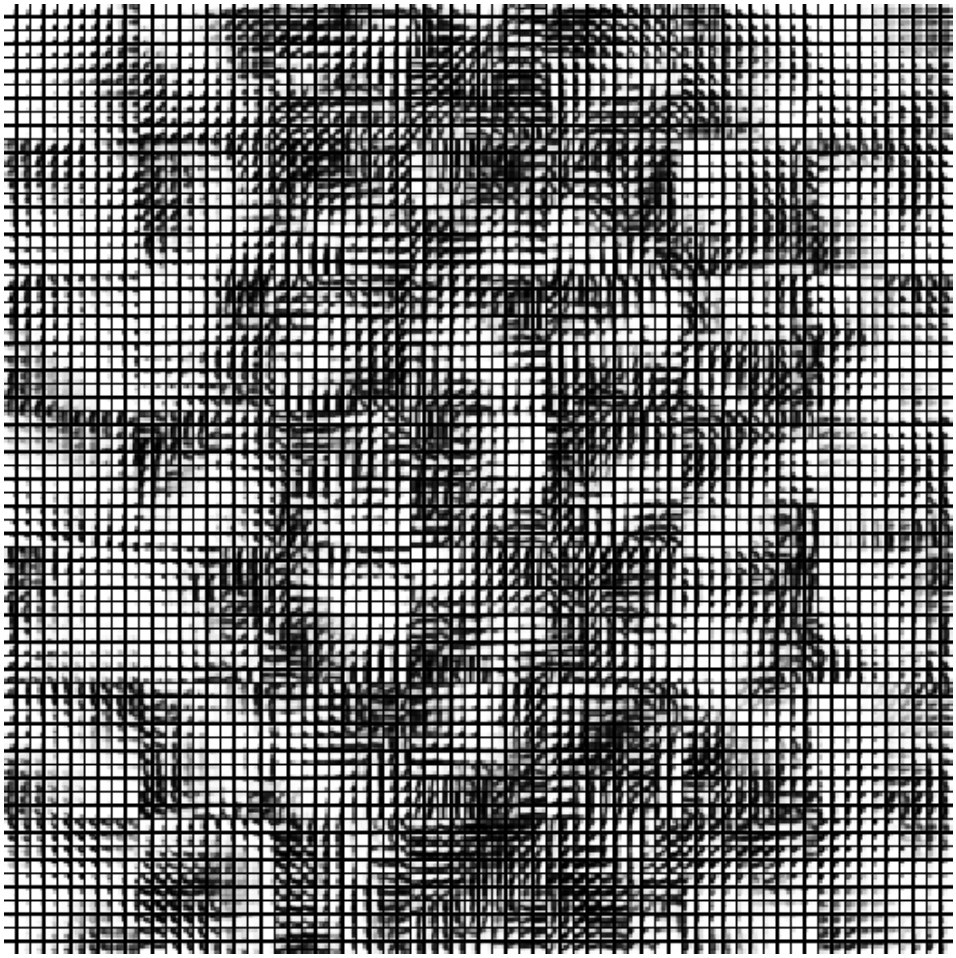

Figure 2: all cortical points of 49 natural elastic nets in 7×7 blocks.

Figure 3: Some testing images and the corresponding reconstructions are respectively shown in the odd column and the even column.

the even columns of figure 3.

IV. Conclusions

We have derived generative models of handwritten digits for the extraction of global features and distributed local features using natural elastic nets. Toward global feature extraction without using dimensionality reduction, the natural elastic net performs the task of clustering analysis, locating a set of local means at centers of digit clusters partitioned by the Mahalanobis distance based on a covariance matrix. When operating on the space spanned by significant eigen-digits, the labeled natural elastic net serves as a classifier, successfully discriminating 2007 testing images up to a correct rate of 96.42%. Based on distributed local features, a group of natural elastic nets, each emulating a population of visual neurons with the same receptive field, constitute an associative memory with capability of digit image reconstruction. Each natural elastic net extracts local features from training images via its

own receptive field, achieving internal representations in terms of a set of local means and a covariance matrix. All of internal representations obtained by distributed natural elastic nets may serve as a basis for an incremental learning process of characters that are not digits but possess similar local features. The relation of the distributed local features to the task of discriminant analysis and dimensionality reduction will be explore in near future.

Reference

References

[1] Peter Dayan and L. F. Abbott, (2001), *Theoretical Neuroscience : Computational and Mathematical Modeling of Neural Systems*.

[2] Durbin, R,.and Mitchison, G., (1990). *A dimension reduction framework for cortical maps*, Nature 343, 644-647.

[3] Durbin, R., and Willshaw, D., (1987). *An Analogue Approach to the Travelling Salesman Problem Using an Elastic Net Method*, Nature 326, 689-691.

[4] Peterson, C., and Söderberg, B., (1989). *A new method for mapping optimization problems onto neural network*, Int. J. Neural Syst. 1,3.

[5] Patrice Simard, Yann Le Cun, John Denker (1993), "*Efficient Pattern Recognition Using a New Transformation Distance*", Advances in Neural Information Processing System 5, 50-58.

[6] Wu, J.M., and Chiu, S.J., (2001). *Independent component analysis using Potts models*, IEEE Trans. on Neural Networks, Vol. 12, No. 2, March.

[7] J.M. Wu and Z.H. Lin (2002), "*Learning generative models of natural images*", Neural Networks,. Vol. 15(3), 337-347.

[8] J.M. Wu and Z.H. Lin (2000), "*Natural Elastic Nets for faithful representations*", International Computer Symposium, Taiwan.