DEEP DISCRIMINATIVE MANIFOLD LEARNING

Jen-Tzung Chien and Ching-Huai Chen

Department of Electrical and Computer Engineering
National Chiao Tung University, Hsinchu, Taiwan 30010, ROC

ABSTRACT

This paper presents a new non-linear dimensionality reduction with stochastic neighbor embedding. A deep neural network is developed for discriminative manifold learning where the class information in transformed low-dimensional space is preserved. Importantly, the objective function for deep manifold learning is formed as the Kullback-Leibler divergence between the probability measures of the *labeled* samples in high-dimensional and low-dimensional spaces. Different from conventional methods, the derived objective does not require the empirically-tuned parameter. This objective is optimized to attractive those samples from the same class to be close together and simultaneously impose those samples from different classes to be far apart. In the experiments on image and audio tasks, we illustrate the effectiveness of the proposed discriminative manifold learning in terms of visualization and classification performance.

Index Terms— Manifold learning, deep neural network, discriminative learning, pattern classification

1. INTRODUCTION

Representation learning aims to explore the meaningful modeling of signals which is crucial for signal processing and machine learning [1]. The primary assumption behind most learning methods is that the minimum number of factors needed to describe the variance of dataset is much smaller than the dimensionality in the original signals [2]. Basically, the algorithms for learning representation range from linear transformations, such as principal component analysis (PCA) and linear discriminant analysis (LDA), to the nonlinear mappings, such as locally linear embedding [3] and stochastic neighbor embedding (SNE) [4, 5, 6, 7] where many of them are *nonparametric* approaches and there is no explicit mapping function between high-dimensional signal and low-dimensional representation. Such nonparametric manifold learning [3, 4] suffers from the generalization problem for unseen samples. To tackle this problem, the parametric mapping was proposed to predict unseen samples [8]. In [9], the manifold learning using deep neural network (DNN) was developed to improve the unsupervised representation learning. The parametric t-distributed SNE was proposed to learn the parametric mapping based on a DNN such that the representation for new samples was available. The deep model using DNN improved the mapping function for manifold learning.

Considering the dimensionality reduction from a probabilistic perspective, the representation learning could be realized by using latent variables based on maximum *a posteriori* probability [10]. Several probabilistic latent variable models such as probabilistic PCA and LDA (PLDA) [11] have been proposed for parametric manifold learning. When estimating the model parameters in parametric approaches, the latent variables are typically assumed to be independent with Gaussian distributions so that the relation between observations and latent variables is arranged as a linear function and the resulting solution is computationally efficient. Nevertheless, the latent variable model could be improved by introducing the non-Gaussian priors. In general, most manifold learning methods were performed in unsupervised manner. In this paper, we build a supervised DNN for dimensionality reduction and pattern classification. A parametric mapping function using DNN is adopted to conduct a supervised nonlinear transformation for manifold learning. The class labels are treated as targets in parametric manifold learning to learn the neighbor embedding for low-dimensional representation. The objective for learning representation is formed as the generalized Kullback-Leibler (KL) divergence between the probability measures of labeled samples in original and transformed spaces. Experiments on different tasks illustrate the merit of proposed method in a sense that the class information is preserved in the space with the reduced dimension and model discrimination.

2. MANIFOLD AND DEEP LEARNING

SNE was developed as a nonlinear unsupervised manifold learning [4]. Suppose we are given a set of high-dimensional signals $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$. SNE attempts to find the low-dimensional representations $\mathcal{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ where $\mathbf{y}_n \in \mathbb{R}^d$ preserves the pairwise similarity to $\mathbf{x}_n \in \mathbb{R}^D$ and $d < D$. The conditional probability $p_{m|n}$ that $\mathbf{x}_m$ is a neighbor of $\mathbf{x}_n$ is expressed by

$$p_{m|n} = \frac{\exp\left(-\|\mathbf{x}_n - \mathbf{x}_m\|^2\right)}{\sum_{t \neq n} \exp\left(-\|\mathbf{x}_n - \mathbf{x}_t\|^2\right)}. \quad (1)$$

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

Similarly, the conditional probability in low-dimensional representation is modeled by

$$q_{m|n} = \frac{\exp\left(-\|\mathbf{y}_n - \mathbf{y}_m\|^2\right)}{\sum_{t \neq n} \exp\left(-\|\mathbf{y}_n - \mathbf{y}_t\|^2\right)}. \tag{2}$$

$p_{n|n}$ and $q_{n|n}$ are set to zero. Intuitively, the difference between two sets of probability distributions $P_n = \{p_{m|n}\}_{m=1}^N$ and $Q_n = \{q_{m|n}\}_{m=1}^N$ can be measured by the Kullback-Leibler (KL) divergence $\mathcal{L}$. SNE is implemented to find low-dimensional representations $\mathcal{Y}$ from high-dimensional observations $\mathcal{X}$ by minimizing the objective function $\mathcal{L}$. Neighbor embedding of samples in two spaces is naturally preserved with this nonlinear and nonparametric transformation.

In a symmetric SNE (s-SNE), the pairwise similarities encoded in P_n and Q_n are measured by the joint probabilities

$$p_{nm} = \frac{p_{m|n} + p_{n|m}}{2N}, \; q_{nm} = \frac{\exp\left(-\|\mathbf{y}_n - \mathbf{y}_m\|^2\right)}{\sum_s \sum_{t, t \neq s} \exp\left(-\|\mathbf{y}_s - \mathbf{y}_t\|^2\right)}. \tag{3}$$

In [5], the t-distributed SNE (t-SNE) was implemented by calculating the pairwise similarity between two low-dimensional representations

$$q_{nm} = \frac{\left(1 + \|\mathbf{y}_n - \mathbf{y}_m\|^2/\nu\right)^{-\frac{\nu+1}{2}}}{\sum_s \sum_{t, t \neq s} \left(1 + \|\mathbf{y}_s - \mathbf{y}_t\|^2/\nu\right)^{-\frac{\nu+1}{2}}} \tag{4}$$

where ν is the degree of freedom in Student's t-distribution. The crowding problem in conventional SNE model is resolved accordingly. Such t-SNE can prevent attracting the low-dimensional representations mutually close together.

To deal with the unseen data problem, the parametric mapping function based on DNN can be incorporated for deep manifold learning. Generally, DNN consists of the connected neurons in many layers which receive the weighted outputs of the connected neurons in previous layer and pass their outputs to the connected neurons in next layer. The nonlinear activation function is applied in calculation of neuron output. A DNN is characterized by its layered structure and the connection weights. DNN model can be simply seen as a nonlinear function which maps between input space $\mathcal{X}$ and output space $\mathcal{Y}$, namely $f(\mathbf{w}, \mathbf{x}_n) = \mathbf{y}_n$, where $\mathbf{w}$ denotes the weight parameters and $\mathbf{x}_n$ and $\mathbf{y}_n$ are the samples in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Therefore, we would like to attain the desired outputs in the reduced-dimensional space by optimally estimating the weights of DNN from training data. This DNN is treated as a prediction function for unseen test data. The procedure of adjusting weights is referred to as DNN training. Recent works in [9, 12] showed that the improvement was obtained by applying the deep manifold learning with the pre-training procedure for the initial weights using the restricted Boltzmann machine (RBM) [13].

3. DEEP DISCRIMINATIVE MANIFOLD LEARNING

Different from previous works, this paper presents a new deep supervised manifold learning for pattern classification.

3.1. Supervised manifold learning

Suppose there are a set of high-dimensional data $\mathcal{X}$ and their corresponding labels $\mathcal{T} = \{t_1, \ldots, t_N\}$ collected for supervised manifold learning. We consider the assumption behind PLDA [11] that the members of the same class share the same latent variable $\mathbf{y}_c$ which is called the class variable. The point estimate of class variable $\hat{\mathbf{y}}_c$ is the variable that maximizes the posterior distribution, i.e. $\hat{\mathbf{y}}_c = \text{argmax}_{\mathbf{y}_n} p(\mathbf{y}_n|\mathbf{x}_n)$. Let $\mathbf{x}_n$ and $\mathbf{x}_m$ be two samples from the same class. The probability that $\mathbf{y}_n$ is identical to $\mathbf{y}_m$ equals to one if $\mathbf{x}_n$ and $\mathbf{x}_m$ belong to the same class or with the same target values $t_n = t_m$. To find the corresponding latent variable without using explicit probability model, we define $p_{nn} = 0$ and $p_{nm} = 1$ when $t_n = t_m$ and $p_{nm} = 0$ when $t_n \neq t_m$. The pre-assigned probabilities in high-dimensional space $P = \{p_{nm}\}$ are viewed as the desired probability values for latent variables given by the labeled samples. On the other hand, we define the joint probability of two samples in low-dimensional space as

$$q_{nm} = \exp\left(-\|\mathbf{y}_n - \mathbf{y}_m\|^k\right) \tag{5}$$

and $q_{nn} = 0$. The supervision of training samples is correspondingly provided. According to the above definition, if $t_n = t_m$, $\mathbf{y}_n$ and $\mathbf{y}_m$ are imposed to be identical such that the probability equals to one in the latent space. In Eq. (5), there is a parameter $k > 0$ that controls the shape of an exponentially decay function. Smaller k gives a longer tail.

To pursue the latent variables satisfying the probability assumption for p_{nm} and q_{nm}, we consider the objective function for elastic embedding [14] and extend it for supervised manifold learning by minimizing the objective $\mathcal{L}$ given by

$$\sum_n \sum_m p_{nm} \|\mathbf{y}_n - \mathbf{y}_m\|^k + \lambda \sum_n \sum_m r_{nm} \exp\left(-\|\mathbf{y}_n - \mathbf{y}_m\|^k\right) \tag{6}$$

where $r_{nm} = 1 - p_{nm}$. Given the objective function in Eq. (6), if p_{nm} equals to one, then $\mathbf{y}_n$ and $\mathbf{y}_m$ shall affect the objective function through the first term. In other words, the first term forces the latent variables in the same class to be as close as possible. On the other hand, if $\mathbf{y}_n$ and $\mathbf{y}_m$ are not in the same class, the second term pushes them away. This circumstance becomes negligible when they have been far apart. Typically, both cases depend on the tuning parameters k and λ. The parameter λ governs the trade-off between the attraction in the first term and the repulsion in the second term. The objective function in Eq. (6) is rewritten in a form of generalized KL divergence or I divergence $\mathcal{D}_I(P\|Q) = \sum_n \sum_m \left(p_{nm} \log\left(p_{nm}/q_{nm}\right) - p_{nm} + q_{nm}\right)$ [15] as

$$\mathcal{L} = \mathcal{D}_I\left(P \| \lambda R \circ Q\right) + G(\lambda) \tag{7}$$

where $R = \{r_{nm}\}$, $Q = \{q_{nm}\}$, $\circ$ denotes the element-wise multiplication and $G(\lambda)$ is defined by

$$\sum_n \sum_m \left[p_{nm} \left(\log p_{nm} - \log \lambda - \log r_{nm} - 1\right)\right]. \tag{8}$$

Minimizing $\mathcal{L}$ over $\mathcal{Y}$ is equivalent to minimizing the objective $\mathcal{D}_I \left(P \| \lambda R \circ Q\right)$ over $\mathcal{Y}$ because $G(\lambda)$ is a constant.

3.2. Discriminative objective function

One issue in the objective function of Eq. (6) is the empirical trade-off parameter λ which should be determined beforehand. Here, we consider Eq. (7) and propose a new optimization objective

$$\min_{\mathcal{Y}} \mathcal{D}_{\text{KL}} \left(P \| R \circ Q\right) = \min_{\mathcal{Y}} \left[\min_{\lambda \geq 0} \mathcal{D}_I \left(P \| \lambda R \circ Q\right)\right] \tag{9}$$

where KL divergence is defined as $\mathcal{D}_{\text{KL}}(P\|Q) = \sum_n \sum_m \tilde{p}_{nm} \log\left(\tilde{p}_{nm}/\tilde{q}_{nm}\right)$ with $\tilde{p}_{nm} = p_{nm}/\sum_s \sum_t p_{st}$ and $\tilde{q}_{nm} = q_{nm}/\sum_s \sum_t q_{st}$. By expanding Eq. (9) and dropping off the terms irrelevant to $\mathcal{Y}$, the discriminative objective function based on SNE (disc-SNE) $\mathcal{L}_{\text{disc-SNE}}$ is derived as

$$\begin{aligned} &\sum_n \sum_m p_{nm} \|\mathbf{y}_n - \mathbf{y}_m\|^k + \left(\sum_s \sum_t p_{st}\right) \\ &\quad \times \log \sum_n \sum_m r_{nm} \exp\left(-\|\mathbf{y}_n - \mathbf{y}_m\|^k\right). \end{aligned} \tag{10}$$

Notably, the advantage of the objective in Eq. (10) over Eq. (6) is that there is no need of choosing λ. Parameter λ has been inherently merged during the optimization of $\mathcal{L}_{\text{disc-SNE}}$ with respect to $\mathcal{Y}$.

3.3. Optimization procedure

It is important that we adopt a DNN as the parametric manifold learner for dimensionality reduction over training samples as well as unseen new samples. The optimal network weights $\mathbf{w}$ in different layers are trained by minimizing the objective $\mathcal{L}_{\text{disc-SNE}}$ by using the training samples $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and their labels $\mathcal{T} = \{t_1, \ldots, t_N\}$. The RPROP algorithm [16] with weight backtracking is implemented for updating $\mathbf{w}$ in an optimization procedure where the gradients of objective function with respect to the weight parameters are calculated according to $\frac{\partial \mathcal{L}_{\text{disc-SNE}}}{\partial \mathbf{w}} = \sum_n \frac{\partial \mathcal{L}_{\text{disc-SNE}}}{\partial \mathbf{y}_n} \frac{\partial \mathbf{y}_n}{\partial \mathbf{w}}$ where $\frac{\partial \mathcal{L}_{\text{disc-SNE}}}{\partial \mathbf{y}_n}$ is yielded as

$$\begin{aligned} &\sum_m 2k \left(p_{nm} - \frac{\sum_s \sum_t p_{st}}{\sum_s \sum_t r_{st} q_{st}} r_{nm} q_{nm}\right) \\ &\quad \times \|\mathbf{y}_n - \mathbf{y}_m\|^{\frac{k-2}{2}} (\mathbf{y}_n - \mathbf{y}_m) \end{aligned} \tag{11}$$

and $\frac{\partial \mathbf{y}_n}{\partial \mathbf{w}}$ is estimated through the error back propagation algorithm. The key difference compared with conventional DNN is the construction of objective function $\mathcal{L}_{\text{disc-SNE}}$. Conventional DNN minimizes the sum-of-square-error function while our model minimizes the KL divergence for elastic embedding and dimensionality reduction.

3.4. Comparison with other methods

It is interesting to compare the objective functions in different methods. The proposed objective $\mathcal{D}_{\text{KL}} \left(P \| R \circ Q\right)$ in Eq. (10) is related to that of the weighted symmetric SNE (ws-SNE) $\mathcal{D}_{\text{KL}}(P \| M \circ Q)$ [15] where P and Q are defined in Eq. (3) and M is the weighting matrix which is imposed to force the centroids in different clusters repulsed mutually. This ws-SNE obtained better performance for manifold learning by alleviating the crowding problem. In our study, the definition of P and R is used to map the samples of the same class into a single low-dimensional representation while preventing the low-dimensional representations from other classes to be close each other. A *discriminative* SNE (also denoted as disc-SNE) is implemented. In [12], the idea of mapping those samples from the same class into a single representative sample was also incorporated in the deep metric learning by means of collapsing classes, which was named as the d-MCML where the objective function was proposed in a form of [12]

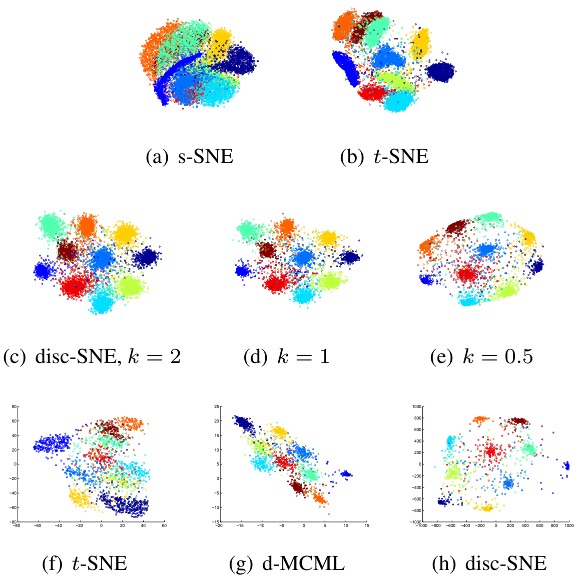
$$\begin{aligned} \mathcal{L}_{\text{d-MCML}} &\propto \sum_n \sum_m p_{nm} \|\mathbf{y}_n - \mathbf{y}_m\|^2 \\ &+ \sum_n \sum_m \log \left(\sum_s \sum_{t, t \neq s} \exp\left(-\|\mathbf{y}_s - \mathbf{y}_t\|^2\right)\right). \end{aligned} \tag{12}$$

Here, the first term aims to map the samples into a single representative sample while the second term would like to repulse low-dimensional representations to be apart from each other. There are two issues in this objective function. The first one is that the term in the brackets of the second term is shared for different samples $\mathbf{y}_n$ and $\mathbf{y}_m$ in different classes. The force of repulsion is seen as a fixed value. The second issue is that the physical meaning of the first term and the second term are possibly conflicting for those samples in the same class. However, such issues do not happen in the proposed objective Eq. (10) where the effect of the second term is individually caused by each pair of samples $\mathbf{y}_n$ and $\mathbf{y}_m$. Either the first term or the second term is activating for each data pair $\{\mathbf{y}_n, \mathbf{y}_m\}$. Namely, the proposed manifold learning aims to move all samples of the same class in reduced dimension space toward the class centroid and also move the samples of different classes far apart mutually.

4. EXPERIMENTS

4.1. Experimental setup

We conducted the experiments on the MNIST and USPS handwritten digit datasets and also on the NIST i-vector speaker recognition challenge by following the experimental setups in [9, 12, 17]. For MNIST and USPS, we implemented the DNN supervised manifold learning which mapped an image into a two-dimensional sample vector for visualization using the topology D-500-500-2000-2. For speaker verification task, the DNN topology 600-300-300-d was applied to reduce the dimension of i-vector D=600 to dimension d=300. The equal error rate (EER) was examined for speaker verification and the classification error was measured for image recognition. RBM pre-training was applied. The RPROP algorithm with mini-batch size of 100 was implemented. For comparison, we carried out the s-SNE, t-SNE [5], d-MCML [12] and the proposed disc-SNE with different k. The 1-nearest neighbor classifier was applied for image recognition.


(a) s-SNE (b) t-SNE

(c) disc-SNE, $k = 2$ (d) $k = 1$ (e) $k = 0.5$

(f) t-SNE (g) d-MCML (h) disc-SNE

Fig. 1. Two-dimensional visualization of test images of MNIST ((a)-(e)) and USPS ((f)-(h)) using different methods.

4.2. Experimental result

Figures 1(a)-(e) demonstrate the visualization of $10,000$ test samples in MNIST dataset by s-SNE, t-SNE and disc-SNE under different k. We can see that the crowding problem is serious by using s-SNE. The heavy-tail of t-distribution in t-SNE alleviates such problem in s-SNE. However, using the class information in SNE algorithm is feasible to move the test samples of the same class with tendency of closeness in the low-dimensional space. In case of $k = 2$, disc-SNE obtains clear separation between classes. When decreasing k, a heavy-tailed condition is increasing. The separation between classes is further enhanced but the shape of a class is changed. Figures 1(f)-(h) show the visualization of 2007 test images in USPS dataset using t-SNE, d-MCML and dis-SNE with $k = 0.5$. Disc-SNE visualizes better than the other methods.

	MNIST (2)	MNIST (10)	USPS (2)	USPS (10)
s-SNE	38.7%	5.7%	36.4%	7.6%
t-SNE	11.6%	5.0%	21.9%	7.5%
d-MCML	4.4%	1.9%	13.3%	6.0%
disc-SNE	4.0%	1.6%	10.3%	5.3%

Table 1. Comparison of classification error rates. The number in brackets indicates the reduced dimension d.

Baseline	PCA	LDA	s-SNE	t-SNE	d-MCML	disc-SNE
7.26 %	7.92 %	6.08 %	7.15%	6.85%	5.91%	5.85%

Table 2. Comparison of EERs for speaker verification.

Table 1 reports the classification error rates of test images by using different dimensionality reduction methods with the reduced dimensions d=2 and 10. MNIST and USPS datasets are used. This table shows that the supervision in manifold learning does improve the feature discrimination and accordingly reduce the recognition error in different conditions. Classification performance is improved by increasing d. The supervised learning using d-MCML and disc-SNE with $k = 0.5$ performs better than unsupervised learning using s-SNE and t-SNE. The lowest classification error is achieved by using disc-SNE. On the other hand, the comparison of EER (%) using different approaches to reduce the dimensionality of i-vector is shown in Table 2. The cosine distance scoring is performed for speaker recognition. Dimensionality reduction using PCA and LDA is implemented for comparison. PCA, s-SNE and t-SNE correspond to unsupervised methods while LDA, d-MCML and disc-SNE are seen as supervised method. Different dimensionality reduction methods are superior to baseline system with i-vectors. The supervised methods perform better than unsupervised methods. The proposed disc-SNE obtains the lowest EER among different methods.

5. CONCLUSIONS

This paper presented a supervised and parametric manifold learning method based on the stochastic neighbor embedding. The proposed method considers the condition that the samples from the same class share the same latent representation. Using a DNN as the mapping function, the proposed objective is optimized to transform the samples of the same class into a single representative centroid and simultaneously map those samples from different classes to be far apart. A meaningful objective is realized for discriminative manifold learning. The experiments on image and audio tasks show that the proposed manifold learning reflects the clustering structure of the classes in low-dimensional visualization, achieves the goal of extracting the discriminative features, and successfully improves the performance of pattern recognition. This framework could be further extended by building a hybrid transformation and classification deep neural network.

6. REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[2] E. Levina and P. J. Bickel, “Maximum likelihood estimation of intrinsic dimension,” in *Advances in Neural Information Processing Systems*, 2004, pp. 777–784.

[3] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[4] G. E. Hinton and S. T. Roweis, “Stochastic neighbor embedding,” in *Advances in Neural Information Processing Systems*, 2002, pp. 857–864.

[5] L. Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov. 2008.

[6] Z. Yang, I. King, Z. Xu, and E. Oja, “Heavy-tailed symmetric stochastic neighbor embedding,” in *Advances in Neural Information Processing Systems*, 2009, pp. 2169–2177.

[7] K. Bunte, S. Haase, M. Biehl, and T. Villmann, “Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences,” *Neurocomputing*, vol. 90, pp. 23–45, 2012.

[8] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, “Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering,” in *Advances in Neural Information Processing Systems*, 2004, vol. 16, pp. 177–184.

[9] L. Maaten, “Learning a parametric embedding by preserving local structure,” in *Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009, pp. 384–391.

[10] S. Watanabe and J.-T. Chien, *Bayesian Speech and Language Processing*, Cambridge University Press, 2015.

[11] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.

[12] R. Min, L. Maaten, Z. Yuan, A. Bonner, and Z. Zhang, “Deep supervised *t*-distributed embedding,” in *Proc. of International Conference on Machine Learning (ICML)*, 2010, pp. 791–798.

[13] G. E. Hinton, “A practical guide to training restricted Boltzmann machines,” in *Neural Networks: Tricks of the Trade (2nd ed.)*, pp. 599–619. 2012.

[14] M. A. Carreira-Perpinan, “The elastic embedding algorithm for dimensionality reduction,” in *Proc. of International Conference on Machine Learning (ICML)*, 2010, pp. 167–174.

[15] Z. Yang, J. Peltonen, and S. Kaski, “Optimization equivalence of divergences improves neighbor embedding,” in *Proc. of International Conference on Machine Learning (ICML)*, 2014, pp. 460–468.

[16] M. Riedmiller and H. Braun, “A direct adaptive method for faster backpropagation learning: The RPROP algorithm,” in *Proc. of IEEE International Conference on Neural Networks*, 1993, pp. 586–591.

[17] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, “The NIST 2014 speaker recognition i-vector machine learning challenge,” in *Proc. of Odyssey: The Speaker and Language Recognition Workshop*, 2014.