Self-Paced AutoEncoder

Tingzhao Yu, Chaoxu Guo, Lingfeng Wang, *Member, IEEE*, Shiming Xiang, *Member, IEEE*, and Chunhong Pan, *Member, IEEE*

***Abstract*—Autoencoder, which learns latent representations of samples in an unsupervised manner, has great potential in computer vision and signal processing. However, the diversity of samples makes learning a component autoencoder remaining a challenging task. This letter proposes a novel Self-Paced AutoEncoder (SPAE) for unsupervised feature extraction. The motivation behind this letter is to take samples gradually from simple to complex into consideration during training, which is similar to the mechanism of knowledge acquisition for humans. Under the unsupervised learning framework constructed on the autoencoder infrastructure, our SPAE first learns a weak autoencoder via samples with small losses and, then, elevates itself to a relatively strong autoencoder through samples with large losses. Then, the SPAE is generalized to a temporal domain, resulting to temporal SPAE (TSPAE), where the temporal information is explored and exploited to improve the performance. Typically, a TSPAE is capable of compressing temporal sequences into temporal-independent data. Experiments on the image classification and action recognition demonstrate the effectiveness of SPAE and TSPAE.**

***Index Terms*—Autoencoder (AE), self-paced learning (SPL), temporal encoding (TE), video analysis.**

I. Introduction

AUTOENCODER (AE) [1], [2] plays a fundamental role in unsupervised learning and signal processing. For a given input, an AE aims to learn a latent feature representation that can minimize the reconstruction loss. Recent years have witnessed a great deal of successes in this direction, e.g., word representation [3] and object detection [4]. The main advantage of AE lies in its high generality since it can automatically learn to extract salient patterns directly from the raw input, without any use of prior knowledge.

In spite of its great successes in many tasks, however, AE has difficulty in learning a latent representation directly, when the training samples disperse in a latent space. In fact, it is much easier to learn simpler concepts first and then build higher level ones on top of simpler ones [5]. Self-paced learning (SPL) [6] is exactly such a learning regime. It is inspired by the learning process of humans and animals that gradually proceeds from easy to complex samples in training. SPL inherits from curriculum learning (CL) [7], in which a curriculum determines a sequence of training samples that are ranked in ascending order of learning difficulty. SPL has many applications such as image and video segmentation [8], domain adaption [9], mixture of regression [10], multimedia retrieval [11], action recognition [12], saliency detection [13], [14], and face identification [15].

With the former illustration in mind, the key point of AE is how to deal with the relatively complicated samples, whereas fortunately, SPL provides a feasible paradigm for dealing with these samples. Therefore, this letter proposes to learn a capable AE under the framework of SPL [i.e., Self-Paced AE (SPAE)]. Specifically, SPAE learns a latent feature representation that is desired to reconstruct the input data. It assigns each sample a weight to reflect its learnability. The sample weight, corresponding to the reconstruction loss, and the reconstruction loss are jointly optimized. SPAE can also be extended to temporal domain, generating the temporal SPAE (TSPAE). Through which, a temporal sequence can be condensed in a temporal axis. The contributions of this letter are as follows.

1) An SPAE is proposed for feature extraction. Through which, a "young" model is first learned via easy samples, and then, it is aggregated to a relatively "mature" model via hard samples.
2) A surrogate Self-Paced regularizer (sSPr) is presented to deal with the diversity of samples. Both the rationality of sSPr and SPAE are thoroughly analyzed in theory.
3) A TSPAE is naturally proposed based upon SPAE. Within TSPAE, it can learn latent representations that are temporal independent.

II. Related Work and Motivation

SPL [6] defines a new machine learning paradigm based on CL [7]. There are also contributions taking the merit of both CL and SPL [13], [16]. Specifically, SPL learns easy samples first and, then, adds complex samples to train by gradually increasing the age parameter. Formally, suppose the training dataset is $\mathcal{X} = \{(\mathbf{x}^1, y^1), (\mathbf{x}^2, y^2), \ldots, (\mathbf{x}^n, y^n)\}$, where $\mathbf{x}^i \in \mathbb{R}^m$ denotes the ith observed sample and $y^i \in \{1, 2, \ldots, K\}$ represents the corresponding label. Then, the loss between the ground truth label y^i and the predicted label $g(\mathbf{x}^i; \boldsymbol{\theta}))$ can be defined as $\sum_i L(y^i, g(\mathbf{x}^i; \boldsymbol{\theta})))$, where $\boldsymbol{\theta}$ represents the parameter of model g. SPL jointly learns the model parameter $\boldsymbol{\theta}$ and a latent weight variable $\mathbf{v} = [v^1, v^2, \ldots, v^n]$ via

$$\phi(\boldsymbol{\theta}, \mathbf{v}) = \sum_{i=1}^{n} v^i L(y^i, g(\mathbf{x}^i; \boldsymbol{\theta})) + f(v^i, \lambda). \tag{1}$$

Manuscript received April 18, 2018; revised May 27, 2018; accepted May 27, 2018. Date of publication June 1, 2018; date of current version June 13, 2018. This work was supported by the National Natural Science Foundation of China under Grant 91646207, Grant 61573352, and Grant 61773377. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wei Li. *(Corresponding author: Tingzhao Yu.)*

T. Yu and C. Guo are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: tingzhao.yu@nlpr.ia.ac.cn; chaoxu.guo@nlpr.ia.ac.cn).

L. Wang, S. Xiang, and C. Pan are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: lfwang@nlpr.ia.ac.cn; smxiang@nlpr.ia.ac.cn; chpan@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2018.2843295

Herein, λ is the age parameter that controls the learning pace, and $f(v^i, \lambda)$ represents the self-paced regularizer (SPr). When λ is small, the model is young, and only easy samples with small losses are considered. As λ increases, more samples will be taken into consideration to construct a mature model. Equation (1) can be solved via an alternating method. Different from the former illustration, we mainly consider unsupervised feature representation where $y^i = \mathbf{x}^i$ in this letter.

Temporal encoding (TE), e.g., max pooling and mean pooling, is an essential part of temporal analysis. Specifically, for a given sample $\mathbf{x}^i_{1:t} \in \mathbb{R}^{m\times t}$ of temporal length t, TE is desired to learn a transformation function ρ that encodes $\mathbf{x}^i_{1:t}$ into $\hat{\mathbf{x}}^i \in \mathbb{R}^m$, where $\hat{\mathbf{x}}^i = \rho(\mathbf{x}^i_{1:t}; \mathbf{d})$ contains the compressed temporal information, and $\mathbf{d}$ is the temporal weight that defines ρ. Dynamic image [17] learns a novel TE strategy via minimizing

$$\psi(\mathbf{d}) = \frac{\delta}{2}\|\mathbf{d}\|^2 + \frac{2}{t(t-1)}\sum_{p>q}\max\{0, 1 - S(p|\mathbf{d}) + S(q|\mathbf{d})\}. \tag{2}$$

Herein, δ is a weight controller, S is a score function, and p, q are two temporal moments. In fact, the core of TE is to learn a temporal weight parameter to pool the temporal sequence $\mathbf{x}^i_{1:t}$ into a single $\hat{\mathbf{x}}^i$, thus, a learning-based strategy [18] can automatically learn the best pooling weight configuration jointly for each sample. Specifically, it can be formulated as follows:

$$\psi(\mathbf{d}, \boldsymbol{\theta}, b) = \sum_{i=1}^{n}(1 - y^i(\boldsymbol{\theta}^T\mathbf{x}^i_{1:t}\mathbf{d} + b))_+ + \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}. \tag{3}$$

Furthermore, motion image [19] generalizes TE utilizing a deep temporal convolution network, and this can be regarded as a deep nonlinear temporal weighting.

Our Motivation: With the former summarization of SPL and TE, our motivations are as follows. First, motivated by SPL, we aim to enhance the robustness of AE by learning easy samples first and then complex samples. Second, a qualified TE method should be reversible, and it holds that

$$\rho^{'}(\rho(\mathbf{x}^i_{1:t}; \mathbf{d}); \mathbf{d}^{'}) = \mathbf{x}^i_{1:t} \tag{4}$$

where $\rho^{'}$ is the reverse operation of ρ defined by parameter $\mathbf{d}^{'}$. We expect to design a temporal AE that is competent to TE. The rationality is that if a TE policy is reversible ((4) holds), then there will not be any loss of temporal relevance. Therefore, the condensed $\hat{\mathbf{x}}$ is capable of temporal related tasks.

III. Self-Paced AE (SPAE)

A. SPAE Model

Different from the supervised strategy, in which the true label is given, AE is unsupervised. For a given AE g defined by $\boldsymbol{\theta}$, the essence is to minimize the reconstruction loss

$$\phi(\boldsymbol{\theta}) = \sum_{i=1}^{n} L(\mathbf{x}^i, g(\mathbf{x}^i; \boldsymbol{\theta})). \tag{5}$$

Within the framework of SPAE, the weighted loss is minimized together with an SPr $f(v^i, \lambda)$. Thus, the objective function of SPAE can be formulated as follows:

$$\phi(\boldsymbol{\theta}, \mathbf{v}) = \sum_{i=1}^{n} v^i L(\mathbf{x}^i, g(\mathbf{x}^i; \boldsymbol{\theta})) + f(v^i, \lambda). \tag{6}$$

Here, v^i is the sample loss and λ is the age parameter. For simplicity, the reconstruction loss $L(\mathbf{x}^i, g(\mathbf{x}^i; \boldsymbol{\theta}))$ is abbreviated as l^i. Then, (6) can be written as follows:

$$\phi(\boldsymbol{\theta}, \mathbf{v}) = \sum_{i=1}^{n} v^i l^i + f(v^i, \lambda). \tag{7}$$

Without loss of comprehension, the superscript i can be omitted for a specific sample $\mathbf{x}^i$, i.e., $l = l^i$ and $v = v^i$.

B. SPAE Regularizer

The SPr is vital to define a reasonable update policy of weight controller v. Basically, an eligible SPr is desired to match the requirements illustrated in Definition 1.

Definition 1 (SP-regularizer[11]): Suppose v, l, and λ are the corresponding sample weight, sample loss, and age parameter, respectively, $f(v; \lambda)$ is called an SPr, if following statements holds:

1) $f(v; \lambda)$ is convex with respect to $v \in [0, 1]$;
2) $v^*(l, \lambda)$ is monotonically decreasing with respect to l, and $\lim_{l\to 0} v^*(l, \lambda) = 1$, $\lim_{l\to\infty} v^*(l, \lambda) = 0$;
3) $v^*(l, \lambda)$ is monotonically increasing with respect to λ, and $\lim_{\lambda\to 0} v^*(l, \lambda) = 0$, $\lim_{\lambda\to\infty} v^*(l, \lambda) \leq 1$;

where $v^*(l, \lambda) = \arg\min_{v\in[0,1]} vl + f(v; \lambda)$

Under this definition, multiple SPrs have been constructed, e.g., hard SPr [6], linear SPr [11], log SPr [11], mixture SPr [20], and dynamic SPr [21]. Nevertheless, all of these SPrs are diluted as the age parameter is increasing. In fact, the purpose of constructing SPr is to depict sample weights more accurately. Therefore, a capable SPr should have a horizontal intercept when the training sample loss is small. For better dealing with the extreme easy and hard samples, this letter proposes a simple sSPr. The proposed SPr is defined as follows:

$$f(v; \lambda) = \frac{\lambda}{2\times 3}v^3 - \lambda v. \tag{8}$$

The rationality of (8) being an SPr can be proved as follows.

Proof: The first and second derivatives of (8) for v are

$$\begin{cases} \frac{\partial f(v,\lambda)}{\partial v} = \frac{\lambda}{2}v^2 - \lambda \\ \frac{\partial^2 f(v,\lambda)}{\partial^2 v} = \lambda v \end{cases}. \tag{9}$$

The second derivative of (8) is $\frac{\partial^2 f(v,\lambda)}{\partial^2 v} \geq 0$, and $f(v; \lambda)$ is convex with respect to $v \in [0, 1]$. Therefore, condition 1 of Definition 1 holds.

With respect to v^*, it can be derived via

$$v^*(l, \lambda) = \arg\min_{v\in[0,1]} vl + f(v; \lambda). \tag{10}$$

Taking the definition of $f(v; \lambda)$ [i.e., (8)] into (10), (10) can be written as follows:

$$v^*(l, \lambda) = \arg\min_{v\in[0,1]} vl + \frac{\lambda}{2\times 3}v^3 - \lambda v. \tag{11}$$

Suppose $\varphi = vl + \frac{\lambda}{2\times 3}v^3 - \lambda v$, then its first derivatives for v can be formulated as follows:

$$\frac{\partial\varphi}{\partial v} = l + \frac{\lambda}{2}v^2 - \lambda = 0. \tag{12}$$

Algorithm 1: Optimization of SPAE.

Input: $\mathcal{X} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$, $\mathbf{x}^i \in \mathbb{R}^m$.
Output: $\boldsymbol{\theta}$
Initialization: $\boldsymbol{\theta}$, $\mathbf{v}$, λ μ;
while not converge **do**
  Fix $\mathbf{v}$, update $\boldsymbol{\theta}$ via (14);
  Fix $\boldsymbol{\theta}$, update λ and $\mathbf{v}$ via (16) and (13);
end

The closed-form optimal solution for v can be obtained as follows:

$$v^* = \begin{cases} 1, & l < \frac{\lambda}{2} \\ \left(2 - \frac{2l}{\lambda}\right)^{\frac{1}{2}}, & \frac{\lambda}{2} \leq \quad l \leq \lambda \\ 0, & l > \lambda \end{cases}. \tag{13}$$

According to (13), the following conclusions can be drawn: $\lim_{l\to 0} v^*(l,\lambda) = 1$, $\lim_{l\to\infty} v^*(l,\lambda) = 0$, $\lim_{\lambda\to 0} v^*(l,\lambda) = 0$, and $\lim_{\lambda\to\infty} v^*(l,\lambda) \leq 1$. As a result, the conditions 2 and 3 of Definition 1 also hold.

Consequently, (8) is a rational SPr. ■

C. SPAE Optimization

There are two variables $\boldsymbol{\theta}$ and $\mathbf{v}$ in (7) need optimization; therefore, this letter adopts an alternating method to solve the SPAE model. There are two steps named $\boldsymbol{\theta}$-*step* and $\mathbf{v}$-*step*. Each step optimizes one parameter by keeping the other fixed. The optimization process is summarized in Algorithm 1.

$\boldsymbol{\theta}$-*step*: Solving $\boldsymbol{\theta}$ by fixed $\mathbf{v}$ equals optimize a weighted AE via

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}, \mathbf{v}) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} v^i L(\mathbf{x}^i, g(\mathbf{x}^i; \boldsymbol{\theta})). \tag{14}$$

Equation (14) can be solved via back propagation. Specifically, suppose $\phi_{\text{AE}}(\boldsymbol{\theta})$ and $\phi_{\text{SPAE}}(\boldsymbol{\theta})$ denote the typical loss of AE and the weighted loss of SPAE, respectively; then, the following formulation holds:

$$\begin{aligned} \Delta\boldsymbol{\theta} &= -\eta \frac{\partial \phi_{\text{SPAE}}}{\partial \boldsymbol{\theta}} \\ &= -\eta\, v\, \frac{\partial \phi_{\text{AE}}}{\partial \boldsymbol{\theta}}. \end{aligned} \tag{15}$$

Herein, η is the learning rate. The sample weight v indicates taking different samples into consideration gradually.

$\mathbf{v}$-*step*: The closed-form solution for $\mathbf{v}$ of a fixed $\boldsymbol{\theta}$ has been illustrated in (13). Generally, at each step, the age parameter λ should be given. For simplicity, this letter updates λ according to its original definition [21] via

$$\lambda = \mu\lambda \tag{16}$$

where μ is set to be 1.3 experimentally. By substituting (16) into (13), the $\mathbf{v}$-*step* can be solved. The following *Proof* demonstrates the convergence of SPAE.

Proof: As is illustrated in early research works [11], [21], (14) equals quadratic programming. Thus, the solution to $\boldsymbol{\theta}$ is the global optimum, i.e.,

$$\phi(\boldsymbol{\theta}^e, \mathbf{v}^{e-1}) \leq \phi(\boldsymbol{\theta}^{e-1}, \mathbf{v}^{e-1}) \tag{17}$$

where e indicates the number of epoch. As proofed in Section III-B, (8) is convex to v; thus, (6) is convex to $\mathbf{v}$. Therefore, it holds that

$$\phi(\boldsymbol{\theta}^e, \mathbf{v}^e) \leq \phi(\boldsymbol{\theta}^e, \mathbf{v}^{e-1}). \tag{18}$$

Consequently, SPAE converges to a stationary solution as follows:

$$\phi(\boldsymbol{\theta}^e, \mathbf{v}^e) \leq \phi(\boldsymbol{\theta}^{e-1}, \mathbf{v}^{e-1}). \tag{19}$$

■

D. Temporal SPAE (TSPAE)

TSPAE is built on long short term memory (LSTM) AE [22], through which, a temporal sequence $\mathbf{x}_{1:t} \in \mathbb{R}^{m\times t}$ can be encoded into a fixed length feature vector $\hat{\mathbf{x}} \in \mathbb{R}^m$. The basic formulation of TSPAE can also be summarized as in (6). In addition, the major difference between SPAE and TSPAE lies in the encoding–decoding strategy. To be specific, SPAE utilizes spatial convolution or fully connected layers as g for spatial reconstruction, whereas TSPAE employs LSTM AE for temporal reconstruction.

IV. Evaluation

A. MNIST Example

For demonstrating the effectiveness of the proposed SPAE, this section evaluates its performance on several variants of the MNIST dataset [23]. Except for the basic MNIST dataset (mnist-basic), the variants includes MNIST with rotation (mnist-rot), MNIST with random background (mnist-back-rand), MNIST with image background (mnist-back-image), and MNIST with both image background and rotation (mnist-rot-back-image). All of these datasets consists of 12 000 training samples and 50 000 testing samples.

Table I presents the testing accuracy obtained by SVM, DBN, NNet, SAA [23], AE, SDAE+Dropout [24], DDL [25], and SPAE. For AE, SDAE, DDL, and SPAE, the encoded features are obtained in an unsupervised manner, and the testing accuracy are based upon these features utilizing a fully connected network. Basically, SPAE outperforms these baseline methods and SPAE is also superior to AE owing to the SPL paradigm.

Fig. 1 presents a comparison of linear SPr, logarithmic soft SPr, mixture SPr, and the proposed sSPr on MNIST dataset and its variants. Linear, logarithmic soft, and mixture SPr assign low weights to some easy samples at the early stage, whereas there is a cutoff in sSPr, which enables sSPr to assign weights large enough to sample with loss less than the threshold (13).

B. Action Recognition

In order to illustrate the performance of TSPAE, this section conducts experiments on video-based action recognition. The datasets in comparison includes *Penn* [26] and *HMDB-51* [27]. Penn[1] contains 2326 video sequences of 15 action classes, whereas HMDB-51[2] is a more challenging dataset with 6849 videos divided into 51 human action classes.

Fig. 2 visualizes the embedded features of mean pooling, max pooling, min pooling, random pooling, AE-LSTM [22],

[1] http://dreamdragon.github.io/PennAction/
[2] http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-data

TABLE I
TESTING ERROR ON MNIST AND ITS VARIANTS

Datasets	mnist-basic	mnist-rot	mnist-back-rand	mnist-back-image	mnist-rot-back-image
SVM-rbf	**3.03 ± 0.15**	11.11 ± 0.28	14.58 ± 0.15	22.61 ± 0.37	55.18 ± 0.44
SVM-poly	3.69 ± 0.17	15.42 ± 0.32	16.62 ± 0.15	24.01 ± 0.37	56.41 ± 0.43
DBN	3.94 ± 0.17	14.69 ± 0.31	9.80 ± 0.15	16.15 ± 0.32	52.21 ± 0.44
NNet	4.69 ± 0.19	18.11 ± 0.34	11.28 ± 0.15	27.41 ± 0.39	62.16 ± 0.43
SAA-3	3.46 ± 0.16	10.30 ± 0.27	20.04 ± 0.15	23.00 ± 0.37	51.93 ± 0.44
AE	4.08 ± 0.04	16.93 ± 0.07	9.26 ± 0.07	15.96 ± 0.07	47.65 ± 0.05
SDAE+*Dropout*	3.71 ± 0.06	10.92 ± 0.07	28.78 ± 0.11	18.47 ± 0.12	55.06 ± 0.08
DDL	4.20 ± 0.10	13.17 ± 0.04	10.67 ± 0.08	19.02 ± 0.09	44.23 ± 0.25
SPAE	3.32 ± 0.04	**10.26 ± 0.07**	**9.01 ± 0.02**	**13.24 ± 0.02**	**44.15 ± 0.05**

Note: AE and SPAE are unsupervised; thus, the testing accuracy is slightly lower than supervised methods. When they are fine tuned (+ *fine tune*), they achieve the best performance among methods in comparison.
The boldface values indicate the best performance.

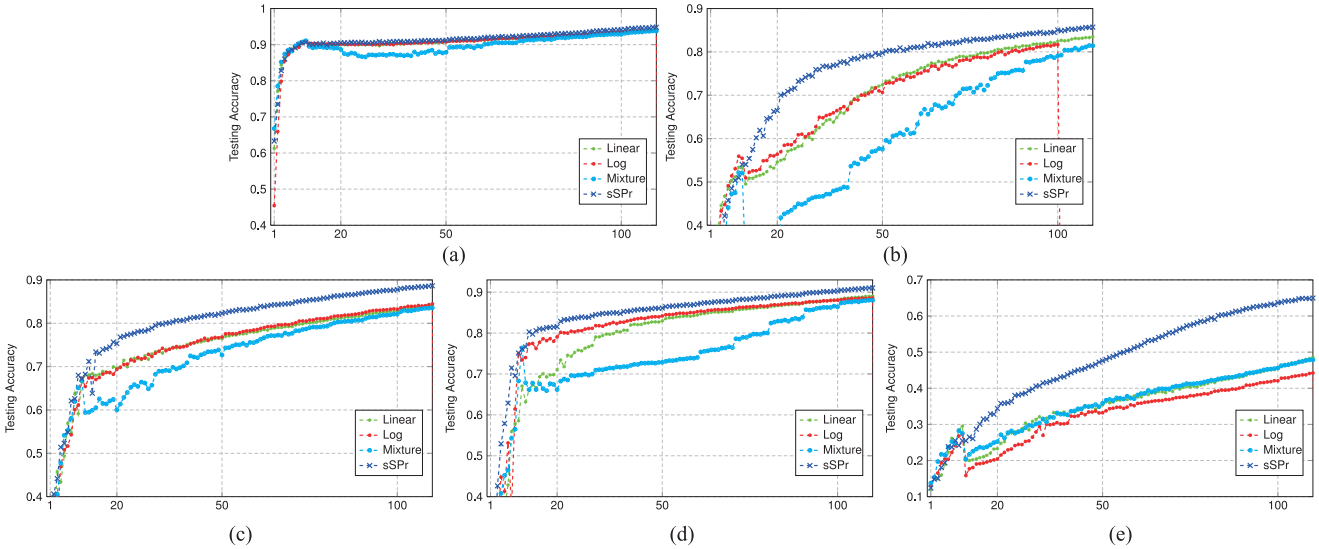

Fig. 1. Testing accuracy versus epoch of MNIST and its variants with regard to various SPrs. The proposed sSPr is superior to other SPrs. (a) mnist-basic. (b) mnist-rot. (c) mnist-back-image. (d) mnist-back-rand. (d) mnist-rot-back-image.

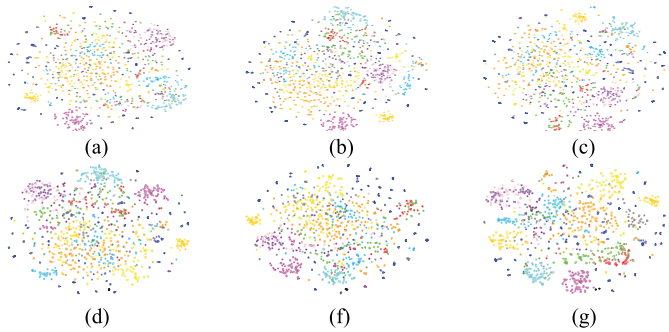

Fig. 2. Visualization of feature embedding on Penn dataset. (a) Mean—72.26%. (b) Max—73.12%. (c) Min—69.09%. (d) Rand—71.42%. (e) AE-LSTM—73.04%. (f) TSPAE—73.79%.

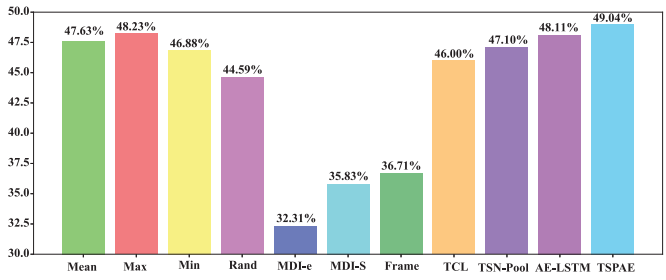

Fig. 3. Testing accuracy on HMDB dataset with regard to various TE techniques. MDI-e is multiple dynamic images with end-to-end training, whereas MDI-S is multiple dynamic images with SVM training. TSPAE performs the best among the encoding methods in comparison.

and TSPAE. TSPAE is semantically separable than the other TE techniques.

For better demonstrating the effectiveness of TSPAE, Fig. 3 gives a more intuitive comparison on HMDB. Specifically, multiple dynamic image (MDI) [17], TSN-Pool [28], and temporal convolution layer [29] are compared, and TSPAE is superior to other TE strategies.

V. CONCLUSION

This letter has proposed a new SPAE, which is inspired from the learning paradigm of humans. A simple sSPr is introduced, and its effectiveness has been demonstrated both theoretically and experimentally. In addition, SPAE is extended to temporal domain (i.e., TSPAE) for TE. Experiments on both toy tasks, e.g., mnist and its variants, and complex tasks, e.g., action recognition, illustrated the superiority of SPAE and TSPAE.

REFERENCES

[1] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. W. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, Apr. 2017.

[2] J. Deng, Z. Zhang, F. Eyben, and B. W. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1068–1072, Sep. 2014.

[3] S. Chandar *et al.*, "An autoencoder approach to learning bilingual word representations," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 1853–1861.

[4] J. Li, C. Xia, and X. Chen, "A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 349–364, Jan. 2018.

[5] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[6] P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. Neural Inf. Process. Syst.*, 2010, pp. 1189–1197.

[7] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[8] D. Zhang, L. Yang, D. Meng, D. Xu, and J. Han, "SPFTN: A self-paced fine-tuning network for segmenting objects in weakly labelled videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5340–5348.

[9] K. Tang, V. Ramanathan, F. Li, and D. Koller, "Shifting weights: Adapting object detectors from image to video," in *Proc. Neural Inf. Process. Syst.*, 2012, pp. 647–655.

[10] L. Han *et al.*, "Self-paced mixture of regressions," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 1816–1822.

[11] L. Jiang, D. Meng, T. Mitamura, and A. Hauptmann, "Easy samples first: Self-paced reranking for zero-example multimedia search," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 547–556.

[12] L. Jiang, D. Meng, S. Yu, Z. Lan, S. Shan, and A. G. Hauptmann, "Self-paced learning with diversity," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 2078–2086.

[13] D. Zhang, D. Meng, L. Zhao, and J. Han, "Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 3538–3544.

[14] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

[15] L. Lin, K. Wang, D. Meng, W. Zuo, and L. Zhang, "Active self-paced learning for cost-effective and progressive face identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 7–19, Jan. 2018.

[16] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, "Self-paced curriculum learning," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2694–2700.

[17] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3034–3042.

[18] L. Wang, C. Gao, J. Liu, and D. Meng, "A novel learning-based frame pooling method for event detection," *Signal Process.*, vol. 140, pp. 45–52, 2017.

[19] T. Yu, H. Gu, L. Wang, S. Xiang, and C. Pan, "Cascaded temporal spatial features for action recognition," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3034–3042.

[20] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, and A. Hauptmann, "Self-paced learning for matrix factorization," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3196–3202.

[21] H. Li and M. Gong, "Self-paced convolutional neural networks," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2110–2116.

[22] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 843–852.

[23] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 473–480.

[24] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[25] A. Majumdar and V. Singhal, "Noisy deep dictionary learning: Application to Alzheimer's disease classification," in *Proc. Int. Joint Conf. Neural Netw.*, 2017, pp. 2679–2683.

[26] W. Zhang, M. Zhu, and K. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2248–2255.

[27] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2556–2563.

[28] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 33–44.

[29] L. Sun, K. Jia, D. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4597–4605.