Signal Processing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Multimodal video classification with stacked contractive autoencoders

Yanan Liu *, Xiaoqing Feng, Zhiguang Zhou

Zhejiang University of Finance & Economics, Hangzhou, PR China

ARTICLE INFO

Article history:
Received 7 September 2014
Received in revised form
26 November 2014
Accepted 1 January 2015

Keywords:
Multimodal
Video classification
Deep learning
Stacked contractive autoencoder

ABSTRACT

In this paper we propose a multimodal feature learning mechanism based on deep networks (i.e., stacked contractive autoencoders) for video classification. Considering the three modalities in video, i.e., image, audio and text, we first build one Stacked Contractive Autoencoder (SCAE) for each single modality, whose outputs will be joint together and fed into another Multimodal Stacked Contractive Autoencoder (MSCAE). The first stage preserves intra-modality semantic relations and the second stage discovers inter-modality semantic correlations. Experiments on real world dataset demonstrate that the proposed approach achieves better performance compared with the state-of-the-art methods.

1. Introduction

With rapid progress of storage devices, Internet and social network, a large amount of video data are generated. How to index and search for these videos effectively is an increasingly active research issue in the multimedia community. To bridge the semantic gap between low-level features and high-level semantics, automatic video annotation and classification have emerged as important techniques for efficient video retrieval [1–3].

Typical approaches to accomplishing video annotation and classification are to apply machine learning methods only using image features from keyframes of video clips. As a matter of fact, video consists of three modalities, namely image, audio and text. Image features of keyframes just express visual aspects, whereas auditory and textual features are equivalently significant for video semantics understanding. A great deal of research has been focused on utilizing multimodal features for better understanding of video semantics [4,5]. Thus multimodal integration in video may compensate the limitations of learning from any single modality.

There are also many other multimodal learning strategies. One group focuses on multi-modal or cross-modal retrieval that learn to map the high-dimensional heterogeneous features into a common low-dimensional latent space [6–9]. Another group is composed of graph-based models, which generate geometric descriptors from multi-channel or multi-sensor to improve image or video analysis [10–16]. However, these methods are discriminative by supervised setting, which require a large amount of labeled data and waste abundant unlabeled data. Collecting labeled data is time-consuming and labor intensive. Thus discovering good representations of data that make it easier to extract useful information when building classifiers with only unsupervised learning has become a big challenge.

Recently, deep learning methods have tremendously attracted researchers interests. The breakthrough in deep learning was initiated by Hinton and quickly followed up in the same year [17–19], and many more later. A central idea, referred to as greedy layerwise unsupervised pre-training, was to learn a hierarchy of features one level at a

* Corresponding author.
E-mail addresses: liuyn@zju.edu.cn (Y. Liu), fxq_snake@163.com (X. Feng), zhouzhiguang@zjucadcg.cn (Z. Zhou).

http://dx.doi.org/10.1016/j.sigpro.2015.01.001
0165-1684/

time, using unsupervised feature learning to learn a new transformation at each level to be composed with the previously learned transformations; essentially, each iteration of unsupervised feature learning adds one layer of weights to a deep neural network. Finally, the set of layers could be combined to initialize a deep supervised predictor [20]. Methods that have been considered include Deep Belief Networks (DBNs) [17] with Restricted Boltzmann Machine (RBM), autoencoders [18], Convolutional Neural Network (CNN)[19], and so on.

Deep learning has been successfully applied to unsupervised feature learning not only for single modality, but also for multiple modalities [21–24]. However, these approaches just learn deep networks with two modalities, e.g., image-text or audio-image pairs. This paper explores useful feature representation by fusing three modalities of video into a joint representation that reflects the intrinsic semantics that the video data corresponds to. Specifically, we first build one Stacked Contractive Autoencoders (SCAE) for each single modality, whose outputs will be joint together and fed into another Multimodal Stacked Contractive Autoencoders (MSCAE). The Contractive Autoencoder (CAE) is a representation-learning algorithm which captures local manifold structure and has the potential for non-local generalization [25]. It is much more appropriate for multimedia data, such as video and image, which are intrinsically on low-dimensional manifolds. The proposed method has two stages. The first stage preserves intra-modality semantic relations and the following stage discovers inter-modality semantic correlations. Compared with existing supervised learners, our method requires minimum amount of prior knowledge of the training data.

Moreover, the power of deep architecture used in the proposed algorithm has its own advantages than other shallow methods for video semantic analysis. For example, [26] used one hidden layer to learn an intermediate representation from video features. However, the single layer learning is limited, while deep network is able to discover more abstract and higher level semantic features for multimedia data. In addition, authors [27,28] have studied the fusion and adaption multiple features for Multimedia Event Detection (MED). But their work mostly focuses on various image features, ignoring the other modalities in video.

The remainder of this paper is organized as follows. We first provide some background to build our model in Section 2. The proposed framework is introduced in Section 3. Section 4 reports the experimental analysis. Finally, we summarize the conclusion and future work in Section 5.

2. Background

2.1. Autoencoder

An autoencoder is a special neural network consisting of three layers: the input layer, the hidden layer, and the reconstruction layer, which sets the target values to be equal to the input (as shown in Fig. 1). It is composed of two parts: (1) Encoder: a deterministic mapping f that transforms an input $x \in \mathbb{R}^{d_x}$ into hidden representation $y \in \mathbb{R}^{d_h}$:

$$y = f(x) = s_f(Wx + b_h) \tag{1}$$

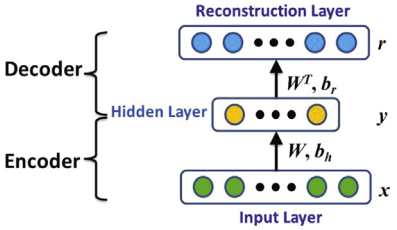

Fig. 1. Autoencoder.

(2) Decoder: the resulting hidden representation y is then mapped back to a reconstruction $r \in \mathbb{R}^{d_x}$ in input space by another mapping function g:

$$r = g(y) = s_g(W'h + b_r) \tag{2}$$

where s_f and s_g are the encoder and decoder activation functions, typically the sigmoid ($\mathrm{sigmoid}(x) = 1/(1+e^{-x})$) or hyperbolic tangent ($\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$) functions for nonlinearity, or the identity function if staying linear. The parameters $W \in \mathbb{R}^{d_h \times d_x}$ and $W' \in \mathbb{R}^{d_x \times d_h}$ are called encoder and decoder weight matrices, and $b_h \in \mathbb{R}^{d_h}$ and $b_r \in \mathbb{R}^{d_x}$ are the encoder and decoder bias vectors. Though autoencoder framework allows for a different matrix in the encoder and decoder, i.e., W and W', weight tying in which $W' = W^T$ may often be used in practice. In this paper we only explore this aforementioned case. Thus the parameter set $\theta = \{W, b_h, b_r\}$ are learned simultaneously on the task of reconstructing as similar as possible with the original input, i.e., aiming at the lowest reconstruction error on a training set of n examples $D_n = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\} \subset \mathbb{R}^{d_x}$, which corresponds to minimizing the following cost function:

$$\mathcal{J}_{AE}(\theta) = \sum_{x \in D_n} L(x, r) \tag{3}$$

where L is the reconstruction error, which could be squared Euclidean distance $L(x, r) = \|x - r\|^2$ in cases of linear reconstruction, or cross-entropy loss when s_g is sigmoid and inputs are in [0, 1]: $L(x, r) = -\sum_{i=1}^{d_x} x_i \log(r_i) + (1 - x_i)\log(1 - r_i)$.

Furthermore, by adding a *weight decay* term that helps to penalize large weights and avoid overfitting in Eq. (3), the simplest form of regularization is defined as follows:

$$\mathcal{J}_{AE-wd}(\theta) = \sum_{x \in D_n} L(x, r) + \frac{1}{2}\lambda \|W\|_2^2 \tag{4}$$

where the weight decay coefficient λ controls the relative importance of the regularization.

2.2. Contractive autoencoder

From the motivation of learning robust representations, the Contractive autoencoder (CAE) [25] is proposed with an alternative regularization yielding objective function:

$$\mathcal{J}_{CAE}(\theta) = \sum_{x \in D_n} L(x, r) + \frac{1}{2}\lambda \|J(x)\|_F^2 \tag{5}$$

where $J(x) = \partial f(x)/\partial x$ is the Jacobian matrix of the encoder f at x. Penalizing the Frobenius norm of the encoders Jacobian encourages the mapping to the feature space to be contractive in the neighborhood of the training data,

Please cite this article as: Y. Liu, et al., Multimodal video classification with stacked contractive autoencoders, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.01.001

i.e., the intermediate feature representation to be robust to small changes of the input. Rifai et al. [25] further provides empirical evidence that the trade-off between reconstruction error and the CAEs regularization term yield a representation that captures the local directions of variation dictated by the data, which often correspond to a lower-dimensional non-linear manifold, while being more invariant to the vast majority of directions orthogonal to the manifold. For a sigmoid encoder, i.e., $f(x)=1/(1+e^{-x})$, the contractive penalty term is easy to compute:

$$J_i(x)=f(x)_j(1-f(x)_j)W_j,$$

$$\|J(x)\|_F^2=\sum_{j=1}^{d_h}\left(f(x)_j(1-f(x)_j\right)^2\|W_j\|^2 \tag{6}$$

This has a similar computational cost as computing the linear reconstruction error (e.g. squared error is $L(x,r)=\|x-b_r-\sum_{i=1}^{d_h}f(x)_jW_j\|^2$). Thus computing the objective and the gradient update in a CAE is only about twice as expensive as in an ordinary AE; both have the same overall computational complexity of $O(d_hd_x)$.

2.3. Stacked autoencoders

The stacked autoencoders (SAE) is a neural network with multiple layers of autoencoders. There are h auto-encoders that are trained in a bottom-up and layer-wise manner, as illustrated in Fig. 2. The raw input vectors are fed to the bottom auto-encoder. After finishing training the bottom auto-encoder, the output hidden representations are wired to the subsequent layer. The same procedure is repeated until all the autoencoders are trained. After such a pre-training stage, the whole neural network is fine-tuned based on a pre-defined objective. The hidden layer of the top autoencoder is the output of the stacked autoencoders, which can be further fed into other applications, such as SVM for classification. The unsupervised pre-training can automatically exploit large amounts of unlabeled data to obtain a good weight initialization for the neural network than traditional random initialization [24].

3. Methodology

In this section, we introduce a framework for multimodal video classification. The two-stage training algorithm learns a set of parameters such that the mapped latent features capture both intra-modal semantics and inter-modal semantics well.

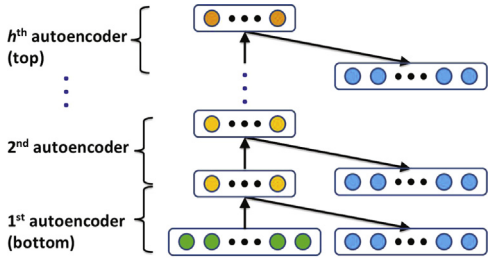

Fig. 2. Stacked autoencoders.

3.1. Single modal pre-training

In this stage, we first train a single SCAE for each modality. Given a set of n input feature vectors (i.e., image, audio and text features in a video shot respectively), the SCAE consisting of h contractive autoencoders is trained by minimizing the following objective function:

$$\mathcal{J}_{SCAE}(\theta)=\sum_{i=1}^{h}\sum_{x\in D_n}\left(L^{(i)}(x,r)+\frac{1}{2}\lambda\|J^{(i)}(x)\|_F^2\right) \tag{7}$$

The objective function can be considered as an extension of Eq. (5) to the stacked scenario. That is to say, a deep network is greedily built by feeding the latent representation of the CAE found on the layer below as input to the current layer. The unsupervised pre-training of such architecture is done one layer at a time. Each layer is trained as a CAE by minimizing the reconstruction of its input, as in Eq. (5).

Algorithm 1. trainSCAE(h,X,d).

Input: h, number of layers of SCAE; $X=[x_1,x_2,\ldots,x_n]^T$, training data, one example per row; d, a sequence of dimensions for each layer

Output: $\theta=\{W,b_h,b_r\}=\{\theta_i\}_{i=1}^{h}$, parameters of SCAE

for $i=1$ to h **do**
  Randomly initialize θ_i with dimensions d_{i-1} and d_i
  $(X_i,\theta_i)=$ **trainCAE** (X_{i-1},θ_{i-1})
end for

The detailed procedure of training a single modal SCAE is listed in Algorithm 1. It mainly consists of a layer-wise training stage that trains the contractive autoencoders from bottom to top by calling **trainCAE** (Algorithm 2), which is the algorithm for training a single-layer contractive autoencoder.

Algorithm 2. trainCAE(X,θ).

Randomly initialize θ
repeat
  Forward propagation (fProp): compute y and r according to Eqs. (1) and (2)
  Optimize Eq. (5) by stochastic gradient descent
until convergence
return fProp(X,θ)

Fig. 3 shows the single modal SCAE for image, audio and text. After learning the SCAE, the output hidden features of the top layer can then be used as a new representation for the original data.

3.2. Multimodal fine-tuning

Single modal pre-training explores the intra-modal semantic relations for each modality. The generated hidden features preserve the intrinsic structure in the original input features. However, inter-modal correlations are not involved in the single modal pre-training, which could be important to the higher-level feature representations. To discover both intra-modal and inter-modal correlations in the hidden features, we combine all modalities together to

Please cite this article as: Y. Liu, et al., Multimodal video classification with stacked contractive autoencoders, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.01.001

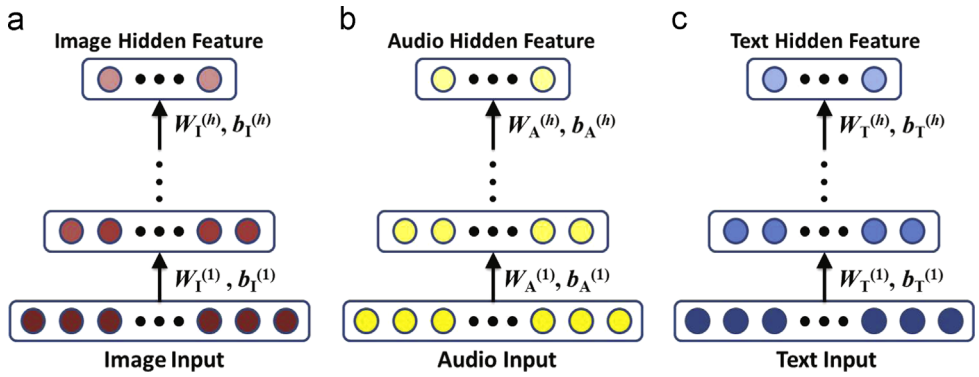

Fig. 3. The single modal SCAEs. (a) Image SCAE. (b) Audio SCAE. (c) Text SCAE.

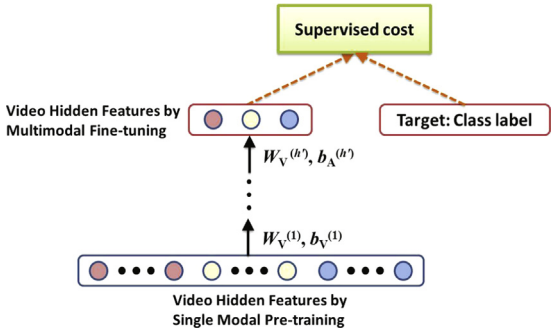

Fig. 4. The fine-tuning of MSCAE for classification.

learn a multimodal stacked contractive autoencoders (MSCAE) as shown in Fig. 4.

The input of MSCAE is the concatenation of the output hidden features learned by single modal pre-training, i.e., image SCAE, audio SCAE and text SCAE. The objective function is similar with that of SCAE in Eq. (7), so **trainSCAE** can be called to perform the training task. Once thus trained, a logistic or softmax regression layer can be added on top of the encoders. The parameters of the whole network are then fine-tuned by performing backpropagation on a supervised objective appropriate for classification, i.e., minimizing the error in predicting the class labels in the training set.

3.3. Prediction and classification

After training three single modal SCAEs and a combined multimodal SCAE, the parameters are already well learned. For a test video shot, high-dimensional low-level features are extracted from each modality and first mapped into low-dimensional hidden space using the trained SAE, with the parameters $\{(W_I^{(1)}, b_I^{(1)}), \ldots, (W_I^{(h)}, b_I^{(h)})\}$ for image feature, $\{(W_A^{(1)}, b_A^{(1)}), \ldots, (W_A^{(h)}, b_A^{(h)})\}$ for audio feature and $\{(W_T^{(1)}, b_T^{(1)}), \ldots, (W_T^{(h)}, b_T^{(h)})\}$ for text feature as shown in Fig. 3. The hidden features are then connected to generate a uniform representation for the video data, and projected to a lower-dimensional hidden space similarly with parameters $\{(W_V^{(1)}, b_V^{(1)}), \ldots, (W_V^{(h')}, b_V^{(h')})\}$ in Fig. 4. Thus the obtained feature by forward propagation is fed into the trained logistic or softmax classifier, where its class label will be predicted.

4. Experiments

4.1. Dataset

The experimental data are mainly based on the collection of TREC video retrieval evaluation (TRECVID) [29] provided by the National Institute of Standards and Technology (NIST). We use TRECVID 2005 video dataset, which is composed of about 168 h multilingual digital video captured from LBC (Arabic), CCTV4, NTDTV (Chinese), and CNN, NBC, MSNBC (English). Due to the limitation of multi-languages, we select the videos broadcasted in English. We then partition the whole dataset into a training dataset of 50,000 shots, a validation dataset of 10,000 shots and a test dataset of 29,450 shots.

The low-level features we used are extracted and preprocessed as follows:

Image features: One key frame within each shot is obtained as a representative image. Image features are then extracted from the key frame. We use three different types of image features: color histograms, textures, and Canny edge.

Audio features: We sample audio signal every 20 msecs (512 windows at 44,100 Hz sampling rate). The audio features extracted from every 20 msecs include MFCC, spectral centroid, rolloff, flux and zero crossings. Because of the variable lengths of shots, we calculate the statistic (mean and variance) of audio features for each shot.

Text features: The source text in video is the ASR (Automatic Speech Recognition) transcript. The text of each shot is represented by a bag of words using TF-IDF features, with a vocabulary of terms which appear in the whole dataset.

In our experiments, we use 546-D image features, 38-D audio features and 1285-D text features. We compare 39 frequently-used semantic concepts (classes) defined in LSCOM [30]. These concepts cover a broad range of interesting topics in news video. The ground truth of the presence of each concept is assumed to be binary (either present or absent in a video shot).

4.2. Evaluation metric

We measure the video classification performance using average precision (*AP*) and mean average precision (*MAP*),

Please cite this article as: Y. Liu, et al., Multimodal video classification with stacked contractive autoencoders, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.01.001

Table 1
MAP comparison with different deep models.

Model		AE	AE+wd	DAE-g	RBM	MSCAE
Dimension of video hidden features	16-D	0.313	0.327	0.368	0.385	**0.359**
	32-D	0.309	0.321	0.354	0.362	**0.331**

which are adopted by NIST. Let S be the size of the test set and R the number of relevant shots returned. At any given index k, let R_k be the number of relevant shots in the top k shots detected. And let $I_k=1$ if the kth shot is relevant and 0 otherwise. Then AP is defined as

$$AP=\frac{1}{R}\sum_{k=1}^{S}\frac{R_k}{k}*I_k$$

And we average the AP over all the 40 concepts to create MAP, which is the overall evaluation result. The parameters of the algorithms are determined through a validation process according to their performances on the validation dataset. In addition, we also measure the precision at top 50 predictions (precision@50).

4.3. Experimental results

4.3.1. Comparisons of deep learning models

We compare the proposed MSCAE against other deep learning models as follows:

- AE: Basic autoencoder without any regularization;
- AE+wd: Autoencoder with weight decay regularization;
- DAE-g: Denoising autoencoder [31] with Gaussian noise;
- RBM-binary: Restricted Boltzmann Machine.

All autoencoder variants used tied weights, a sigmoid activation function for both encoder and decoder, and a cross-entropy reconstruction error. They were trained by optimizing their (regularized) objective function on the training set by stochastic gradient descent. As for RBMs, they were trained by Contrastive Divergence.

Moreover, to make a fair comparison, all models are stacked through single modal pre-training and multimodal fine-tuning as our MSCAE. In the first stage, models are configured with 2 layers, where the image features are mapped from 546-D to 128-D, audio features from 38-D to 20-D, and text features from 1285-D to 128-D. In the second stage, the concatenated joint feature are reduced from 276-D to 128-D, and finally to 16-D and 32-D with a 3-layer model, respectively. For each case, we selected the value of hyperparameters (such as the strength of regularization) that yielded, after supervised fine-tuning, the best classification performance on the validation set. Final classification accuracy was then computed on the test set.

The results are reported in Table 1. The proposed MSCAE achieves the best performance for both dimensions of final video hidden features. On the one hand, contractive autoencoder demonstrates better learning and reconstructing ability than other models. On the other hand, our two-stage method is less sensitive to the dimension of the output layer. The reason is that MSCAE has stronger representation power and can better avoid local optimality by a good unsupervised pre-training.

Table 2
MAP comparison with shallow models.

Model	SVM	LDA	MSCAE
MAP	0.321	0.314	**0.359**
Precision@50	0.672	0.687	**0.791**

Table 3
MAP comparison with shallow models.

Model	OMG-SSL	FWOT	MSCAE
MAP	0.348	0.353	**0.359**
Precision@50	0.758	0.763	**0.791**

4.3.2. Comparisons between deep and shallow models

In this section, we compare our deep model with traditional shallow models, i.e., Support Vector Machine (SVM) with a RBF kernel and Linear Discriminant Analysis (LDA). The SVM classifications (C-SVM) were obtained using LIBSVM [32]. Values for the C (cost) and γ (RBF) parameters were selected by 10-fold cross-validation.

From Table 2, it can be seen that our MSCAE outperforms the other shallow classifiers as well. MSCAE achieves a MAP of 0.359, compared to 0.321 and 0.314, achieved by SVM and LDA models. Besides, MSCAE achieves a precision@50 of 0.791, compared to 0.672 and 0.687 by SVM and LDA models, respectively. It proves that deep learning is able to learn higher level abstract concepts and understand the semantics better.

4.3.3. Comparisons with video annotation methods

In this section, we compare the proposed algorithm with two state-of-the-art video annotation approaches. One is optimized multigraph-based semi-supervised learning (OMG-SSL) [5], and the other is Feature Weighting via Optimal Thresholding (FWOT) [28]. Table 3 shows that the performances of the three methods are comparable, which proves the advantages of multimodal feature fusion as well. Specifically, our method still outperforms the others, revealing the powerful learning ability of deep architecture.

5. Conclusion and future work

Multimodal integration plays important role in video semantics classification, based on which we propose a two-stage learning framework with contractive stacked autoencoders. By considering both intra-modal and inter-modal semantics, we learn a set of effective SCAEs for feature

Please cite this article as: Y. Liu, et al., Multimodal video classification with stacked contractive autoencoders, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.01.001

mapping from single modal pre-training to multimodal fine-tuning. Compared to other deep and shallow models, experimental results show the improvements of our approach in video classification accuracy.

In the recent years, the TRECVID related research has gradually shifted from controlled videos (e.g., news videos) to unconstrained videos (e.g., internet videos, such as the TRECVID MED data and the HMDB action video dataset). For example, some authors [26–28,33] have used the TRECVID MED dataset, and Shen et al. [34] has used the HMDB dataset. It is a new trend to focus on the research of more complex semantic analysis on video data.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 61100084, 61202197, 61303133), Zhejiang Province Department of Education Fund (No. Y201223321) and Zhejiang Provincial Natural Science Foundation of China (No. LQ13F020003).

References

[1] M. Wang, X.-S. Hua, J. Tang, R. Hong, Beyond distance measurement: constructing neighborhood similarity for video annotation, IEEE Trans. Multimed. 11 (3) (2009) 465–476.
[2] M. Wang, B. Ni, X.-S. Hua, T.-S. Chua, Assistive tagging: a survey of multimedia tagging with human–computer joint exploration, ACM Comput. Surv. 44 (4) (2012) 1–24.
[3] G. Li, M. Wang, Z. Lu, R. Hong, T.-S. Chua, In-video product annotation with web information mining, ACM Trans. Multimed. Comput. Commun. Appl. 8 (4) (2012) 1–19.
[4] Y. Yang, J. Song, Z. Huang, Z. Ma, N. Sebe, A. Hauptmann, Multi-feature fusion via hierarchical regression for multimedia analysis, IEEE Trans. Multimed. 15 (3) (2013) 572–581.
[5] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, Y. Song, Unified video annotation via multigraph learning, IEEE Trans. Circuits Syst. Video Technol. 19 (5) (2009) 733–746.
[6] M. Bronstein, A. Bronstein, F. Michel, N. Paragios, Data fusion through cross-modality metric learning using similarity-sensitive hashing, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 3594–3601.
[7] Y. Zhuang, Y. Yang, F. Wu, Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval, IEEE Trans. Multimed. 10 (2) (2008) 221–229.
[8] L. Zhang, Y. Gao, C. Hong, Y. Feng, J. Zhu, D. Cai, Feature correlation hypergraph: exploiting high-order potentials for multimodal recognition, IEEE Trans. Cybern. 44 (8) (2014) 1408–1419.
[9] M. Wang, H. Li, D. Tao, K. Lu, X. Wu, Multimodal graph-based reranking for web image search, IEEE Trans. Image Process. 21 (11) (2012) 4649–4661.
[10] L. Zhang, Y. Gao, R. Zimmermann, Q. Tian, X. Li, Fusion of multichannel local and global structural cues for photo aesthetics evaluation, IEEE Trans. Image Process. 23 (3) (2014) 1419–1429.
[11] L. Zhang, Y. Gao, Y. Xia, Q. Dai, X. Li, A fine-grained image categorization system by cellet-encoded spatial pyramid modeling, IEEE Trans. Ind. Electron. 99 (2014) 1.
[12] Y. Xia, X. Li, Z. Shan, Parallelized fusion on multisensor transportation data: a case study in cyberits, Int. J. Intell. Syst. 28 (6) (2013) 540–564.
[13] L. Zhang, M. Song, X. Liu, L. Sun, C. Chen, J. Bu, Recognizing architecture styles by hierarchical sparse coding of blocklets, Inf. Sci. 254 (0) (2014) 141–154.
[14] Y. Xia, J. Hu, M.D. Fontaine, A cyber-its framework for massive traffic data analysis using cyber infrastructure, Sci. World J., http://dx.doi.org/10.1155/2013/462846.
[15] L. Zhang, Y. Han, Y. Yang, M. Song, S. Yan, Q. Tian, Discovering discriminative graphlets for aerial image categories recognition, IEEE Trans. Image Process. 22 (12) (2013) 5071–5084.
[16] L. Zhang, M. Song, X. Liu, J. Bu, C. Chen, Fast multi-view segment graph kernel for object classification, Signal Process. 93 (6) (2013) 1597–1607.
[17] G. E. Hinton, S. Osindero, A fast learning algorithm for deep belief nets, Neural Comput. 18 (2006) 2006.
[18] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, U.D. Montral, M. Qubec, Greedy layer-wise training of deep networks, Adv. Neural Inf. Process. Syst. (NIPS) (2007) 153–160.
[19] M. Ranzato, C. Poultney, S. Chopra, Y. Lecun, Efficient learning of sparse representations with an energy-based model, Adv. Neural Inf. Process. Syst. (NIPS) (2006) 1137–1144.
[20] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[21] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proceedings of the 28th International Conference on Machine Learning (ICML), 2011.
[22] N. Srivastava, R. Salakhutdinov, Multimodal learning with deep Boltzmann machines, J. Mach. Learn. Res. 15 (2014) 2949–2980.
[23] A. Frome, G.S. Corrado, J. Shlens, S. Bengio, J. Dean, M.A. Ranzato, T. Mikolov, Devise: a deep visual-semantic embedding model, Adv. Neural Inf. Process. Syst. (NIPS) (2013) 2121–2129.
[24] W. Wang, B.C. Ooi, X. Yang, D. Zhang, Y. Zhuang, Effective multi-modal retrieval based on stacked auto-encoders, in: Proceedings of International Conference on Very Large Data Bases (VLDB), 2014.
[25] S. Rifai, P. Vincent, X. Muller, X. Glorot, Y. Bengio, Contractive auto-encoders: explicit invariance during feature extraction, in: L. Getoor, T. Scheffer (Eds.), Proceedings of the International Conference on Machine Learning (ICML), Omnipress, 2011, pp. 833–840.
[26] Z. Ma, Y. Yang, N. Sebe, K. Zheng, A. Hauptmann, Multimedia event detection using a classifier-specific intermediate representation, IEEE Trans. Multimed. 15 (7) (2013) 1628–1637.
[27] Z. Ma, Y. Yang, N. Sebe, A. Hauptmann, Knowledge adaptation with partially shared features for event detection using few exemplars, IEEE Trans. Pattern Anal. Mach. Intell. 36 (9) (2014) 1789–1802.
[28] Z. Xu, Y. Yang, I. Tsang, N. Sebe, A. Hauptmann, Feature weighting via optimal thresholding for video analysis, in: Proceedings of the International Conference on Computer Vision (ICCV), 2013, pp. 3440–3447.
[29] Trevid. ⟨http://www-nlpir.nist.gov/projects/trevid/⟩.
[30] C.G.M. Snoek, M. Worring, J.C.V. Gemert, J. Mark Geusebroek, A.W.M. Smeulders, The challenge problem for automated detection of 101 semantic concepts in multimedia, in: Proceedings of the ACM International Conference on Multimedia, 2006, pp. 421–430.
[31] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.
[32] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27.
[33] Y. Yang, Z. Ma, Z. Xu, S. Yan, A. Hauptmann, How related exemplars help complex event detection in web videos? in: Proceedings of the International Conference on Computer Vision (ICCV), 2013, pp. 2104–2111.
[34] H. Shen, Y. Yan, S. Xu, N. Ballas, W. Chen, Evaluation of semi-supervised learning method on action recognition, Multimed. Tools Appl. (2014) 1–20.

Please cite this article as: Y. Liu, et al., Multimodal video classification with stacked contractive autoencoders, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.01.001