DISCOVERING COMPACT TOPICAL DESCRIPTORS FOR WEB VIDEO RETRIEVAL

Fang Zhao, Yongzhen Huang, Liang Wang, Tieniu Tan

Center for Research on Intelligent Perception and Computing
National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing 100190, China
{fang.zhao, yzhuang, wangliang, tnt}@nlpr.ia.ac.cn

ABSTRACT

Describing videos efficiently is an important task for content based web video retrieval. To solve this problem, we propose an unsupervised approach based on an undirected topic model to learn a compact topical descriptor upon the bag-of-words (BoW) video representation. In our method, words in a BoW are assumed to have different topic features, and the topical descriptor of an entire video is obtained by aggregating those features, which makes the descriptor contain information about relative strength of topics. To improve the descriptor interpretability, an L_1 penalty is used to control the topical sparsity. Furthermore, efficient learning and inference algorithms are presented. We evaluate the proposed descriptor on the Columbia Consumer Video dataset. Experimental results demonstrate that compared with the BoW and other topical representations, the proposed compact descriptor has better performance in web video retrieval.

Index Terms— Web video retrieval, compact topical descriptor, undirected topic model, sparse representation

1. INTRODUCTION

Digital videos which are uploaded to online media websites (e.g., YouTube) have been growing explosively in recent years and most of them have very little textual annotation. Thus automatically learning compact video representations is an important task for content based web video retrieval, especially in mobile visual search system. However, this task is usually very challenging because such web videos are captured in uncontrolled conditions and generally contain huge intra-class variations.

The bag-of-words (BoW) representation [1] has been used well for visual search [1, 2] and web video content analysis [3]. But for large scale databases, the BoW representation is time- and memory-consuming [4]. Research efforts have been devoted to compact representations without serious loss of discriminability, e.g., the dimensionality reduction of local feature descriptors [5] and the compression of image/frame-level signatures [4, 6, 7].

However, to the best of our knowledge, little effort has been taken on the compactness of video-level representations. In this paper, we propose an unsupervised approach based on an undirected topic model to discover a compact topical descriptor from the BoW video representation for web video retrieval. Inspired by [6], we leverage the statistics of video corpus instead of a single video to achieve the descriptor's compactness.

In our method, an undirected topic model (i.e., Replicated Softmax [8]) is chosen for describing latent topics of videos. Compared with the directed models, it is much more efficient to estimate model parameters and calculate the posterior distribution over the latent topics, which is beneficial for modeling large scale web video databases. Meanwhile, through aggregating different topic features of words in a BoW, we generalize binary topic units in the undirected topic model to nonnegative real-valued units which are more expressive about relative strength of topics. Also, to make the descriptor more interpretable, we impose an L_1 regularizer on hidden unit activations to control the topical sparsity. In addition, learning video descriptors upon the BoW makes our descriptor suitable for various local features and vocabularies, which is clearly validated in subsequent experiments.

Our main contributions include: 1) We present a novel compact video descriptor based on an undirected topic model. 2) By using nonnegative real-valued units and an L_1 penalty, we extend the topic model to improve the descriptor discriminability. 3) A stochastic gradient descent method is introduced to efficiently learn the extended model.

The rest of this paper is organized as follows. In Section 2, we model web videos using an undirected topic model. Section 3 describes the learning procedure for compact topical video descriptor. Section 4 presents experimental evaluations. Finally, the paper is concluded in Section 5.

2. MODELING WEB VIDEOS BASED ON UNDIRECTED TOPIC MODEL

In this section, we start by reviewing the Replicated Softmax model, give its advantages compared to the directed topic model, and then formulate the model for web videos.

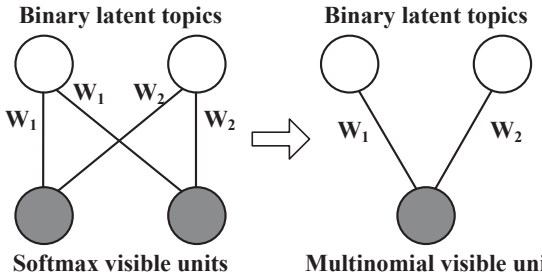


Fig. 1. Replicated Softmax model for a document containing two words.

2.1. Replicated Softmax model

The Replicated Softmax model is an undirected probabilistic topic model, which has been successfully applied in document analysis [8]. It can be used to automatically extract latent semantic topics from large unstructured datasets.

As shown in Fig. 1 (left), the Replicated Softmax model is a two-layer undirected graphical model. The bottom layer represents softmax visible units and the top layer represents binary hidden topics. Its main idea is that using K softmax units with identical weights (K is the document size) to model documents of different length. This is equivalent to using a single multinomial unit sampled K times (Fig. 1 (right)).

The undirected topic model is more appropriate for our task than the directed model (e.g., Latent Dirichlet Allocation (LDA) [9]). First, this model can be trained in an online manner, which is convenient to deal with web videos increasing continually. Second, the inference algorithm of latent topics is simple and easy to be implemented in parallel, which is valuable for analyzing large scale web video data. Third, it produces distributed topical representations of web videos, which can make more precise predictions for words using the intersection of the distributions predicted by individual topics.

2.2. Model formulation

Let a vocabulary with N words be indexed by $\{1,...,N\}$. We represent the video dataset as $D=\{\mathbf{v}_m\}_{m=1}^{M}$, where each video is represented as a vector $\mathbf{v}\in\mathbb{N}^N$ and each entry v_i (i = 1,…, N) of $\mathbf{v}$ denotes the number of times that word i appears in the video. Our goal is to map the high dimensional BoW vector $\mathbf{v}$ to an F-dimensional topical descriptor $\mathbf{c}$ ($F \ll N$) while preserving the discriminative power of the descriptor.

Consider modeling the video dataset D using the Replicated Softmax model. Let the BoW vector $\mathbf{v}\in\mathbb{N}^N$ be visible units, and $\mathbf{h}\in\{0,1\}^F$ be binary hidden topic features. The energy function of the state $\{\mathbf{v}, \mathbf{h}\}$ can be defined as follows:

$$E(\mathbf{v},\mathbf{h})=-\sum_{i=1}^{N}\sum_{j=1}^{F}W_{ij}v_i h_j-\sum_{i=1}^{N}a_i v_i-K\sum_{j=1}^{F}b_j h_j, \qquad (1)$$

where W_{ij} is the weight connected with v_i and h_j, a_i and b_j are the bias term of visible and hidden units respectively, and

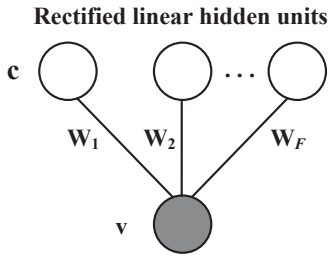


Fig. 2. Extended topic model for a video containing K words. The top layer represents the topical descriptor $\mathbf{c}$ and the bottom layer represents the BoW vector $\mathbf{v}$.

$K=\sum_i v_i$ is the total number of words in the video. The probability that the model assigns to the vector $\mathbf{v}$ is:

$$P(\mathbf{v})=\frac{1}{Z}\sum_{\mathbf{h}}\exp(-E(\mathbf{v},\mathbf{h})),\quad Z=\sum_{\mathbf{v}}\sum_{\mathbf{h}}\exp(-E(\mathbf{v},\mathbf{h})), \quad (2)$$

where Z is the normalizing constant. The conditional distributions are given by:

$$P(\mathbf{v}\mid\mathbf{h})=\mathrm{Mult}\left(K;\frac{\exp(a_i+\sum_{j=1}^{F}W_{ij}h_j)}{\sum_{i'=1}^{N}\exp(a_{i'}+\sum_{j=1}^{F}W_{i'j}h_j)},i=1,...,N\right), \qquad (3)$$

$$P(h_j=1\mid\mathbf{v})=\sigma\left(Kb_j+\sum_{i=1}^{N}W_{ij}v_i\right), \qquad (4)$$

where $\mathrm{Mult}(K;.)$ represents the multinomial distribution with K times sampling and $\sigma(x)=1/(1+\exp(-x))$ is the logistic function.

For web videos with complex contents, binary hidden units in the Replicated Softmax are too limited to capture information about relative strength of topics. Similar to [10], here different words in a video are considered to have different binary topic features and the topical video descriptor $\mathbf{c}$ is obtained via an aggregation of those N topic features $\mathbf{h}$. Therefore, in our model, the original binary unit is replaced with a binomial unit which can be viewed as an aggregation of N separate binary units that share the same weights and bias. However, the binomial unit will make the model learning less stable. Instead we consider using the rectified linear units [11], the sampled value of which can be fast approximated by $\max(0,x+N(0,1))$ where $N(0,1)$ is Gaussian noise with zero mean and unit variance, to model the descriptor $\mathbf{c}$.

The extended model is shown in Fig. 2. Accordingly, the conditional distribution in (4) is rewritten as follows:

$$P(c_j\mid\mathbf{v})=\begin{cases}N(x,1),\ x=Kb_j+\sum_{i=1}^{N}W_{ij}v_i, & c_j\geq 0\\ 0, \quad \text{otherwise}\end{cases} \qquad (5)$$

where c_j is an entry of $\mathbf{c}$. It is worth nothing that the rectified linear unit variables which take nonnegative real values can represent not only the presence and absence of topics (like binary unit variables) but also the relative importance of topics.

3. LEARNING COMPACT TOPICAL VIDEO DESCRIPTOR

Based on the above probabilistic modeling, the learning process is to estimate the model parameters {**W**, **a**, **b**} through maximizing the probability that the model assigns to training videos. Meanwhile, in order to make the descriptor more interpretable and discriminative, we want hidden unit activations to be sparse. Different from [12] in which the sparsity constraint is only suited to binary units, here we impose an L_1 regularizer over the conditional expectation of the hidden units $\mathbf{c}$. Thus, given the video dataset D, we have the following optimization problem:

$$\min_{\mathbf{W},\mathbf{a},\mathbf{b}} -\sum_{m=1}^{M}\log P(\mathbf{v}_m)+\lambda\sum_{m=1}^{M}\left\|\mathbb{E}\left[\mathbf{c}_m \mid \mathbf{v}_m\right]\right\|_1 \tag{6}$$

where $\mathbb{E}[.]$ is the conditional expectation given the training data and λ is a coefficient controlling the sparsity degree. We can use stochastic gradient descent to solve this problem.

To calculate the gradient of the objective function, we consider the log-likelihood term and the regularizer in (6) respectively. The derivative of the log-likelihood with respect to parameters **W** is:

$$\frac{1}{M}\sum_{m=1}^{M}\frac{\partial \log P(\mathbf{v}_m)}{\partial W_{ij}} = <v_i c_j>_{data} - <v_i c_j>_{model} \tag{7}$$

where $<.>_{data}$ and $<.>_{model}$ denote expectations under the distributions specified by the data and model, respectively. In practice, to avoid expensive exact computation of the expectation $<v_i c_j>_{model}$, we use contrastive divergence [13] learning which gives an efficient approximation to the gradient of the log-likelihood. The update rule is then given by:

$$\Delta W_{ij}^{likelihood} = \alpha\left(<v_i c_j>_{data} - <v_i c_j>_{recon}\right) \tag{8}$$

where α is a learning rate and *recon* represents a "reconstruction" provided by one full Gibbs step using Equations (3) and (5). A similar update rule that using the values of individual units instead of pairwise products is applied to the biases **a** and **b**.

As a result of the special bipartite structure of the model, the hidden unit variables are conditionally independent of each other given the values of the visible units **v**. Thus, according to Equation (5), the regularizer can be expanded as:

$$\left\|\mathbb{E}\left[\mathbf{c} \mid \mathbf{v}\right]\right\|_1 = \sum_{j=1}^{F}\mathbb{E}\left[c_j \mid \mathbf{v}\right] = \sum_{j=1}^{F}\frac{1}{\sqrt{2\pi}}\int_0^{+\infty} c_j \exp\left(-\frac{\left(c_j - x\right)^2}{2}\right)dc_j \tag{9}$$

where $x = Kb_j + \sum_i W_{ij}v_i$. Then the derivatives of the regularizer with respect to parameters **W** and **b** are given by:

$$\frac{\partial\left\|\mathbb{E}\left[\mathbf{c} \mid \mathbf{v}\right]\right\|_1}{\partial W_{ij}} = v_i\left(\frac{1}{2}+\frac{1}{2}\mathrm{erf}\left(\frac{\sqrt{2}}{2}x\right)\right), \tag{10}$$

$$\frac{\partial\left\|\mathbb{E}\left[\mathbf{c} \mid \mathbf{v}\right]\right\|_1}{\partial b_j} = \frac{1}{2}+\frac{1}{2}\mathrm{erf}\left(\frac{\sqrt{2}}{2}x\right), \tag{11}$$

where erf(.) is Gauss error function. Algorithm 1 outlines the entire learning procedure. After the learning is completed, the inference of the descriptor $\mathbf{c}$ is very simple, which can be directly calculated by Equation (5).

Algorithm 1 Sparse undirected topic model learning for compact topical video descriptor

1. **Input:** Video dataset $D = \{\mathbf{v}_m\}_{m=1}^{M}$ subdivided into small mini-batches
2. Randomly initialize parameters $\mathbf{W}^0$, $\mathbf{a}^0$ and $\mathbf{b}^0$
3. **for** t = 1 to T **do**
4. Draw a mini-batch from D randomly
5. Update the parameters $\{\mathbf{W}^t, \mathbf{a}^t, \mathbf{b}^t\}$ using contrastive divergence learning rule:

$$W_{ij}^t \leftarrow W_{ij}^{t-1} + \alpha\left(<v_i c_j>_{data} - <v_i c_j>_{recon}\right)$$

$$a_i^t \leftarrow a_i^{t-1} + \alpha\left(<v_i>_{data} - <v_i>_{recon}\right)$$

$$b_j^t \leftarrow b_j^{t-1} + \alpha\left(<c_j>_{data} - <c_j>_{recon}\right)$$

6. Update the parameters $\{\mathbf{W}^t, \mathbf{b}^t\}$ using the gradients of the L_1 regularizer:

$$x = Kb_j^{t-1} + \sum_i W_{ij}^{t-1}v_i$$

$$W_{ij}^t \leftarrow W_{ij}^t - \alpha\lambda\left\langle v_i\left(\frac{1}{2}+\frac{1}{2}\mathrm{erf}\left(\frac{\sqrt{2}}{2}x\right)\right)\right\rangle_{data}$$

$$b_j^t \leftarrow b_j^t - \alpha\lambda\left\langle \frac{1}{2}+\frac{1}{2}\mathrm{erf}\left(\frac{\sqrt{2}}{2}x\right)\right\rangle_{data}$$

7. **end for**
8. **Return** $\mathbf{W}^T$, $\mathbf{a}^T$ and $\mathbf{b}^T$

4. EXPERIMENTAL RESULTS

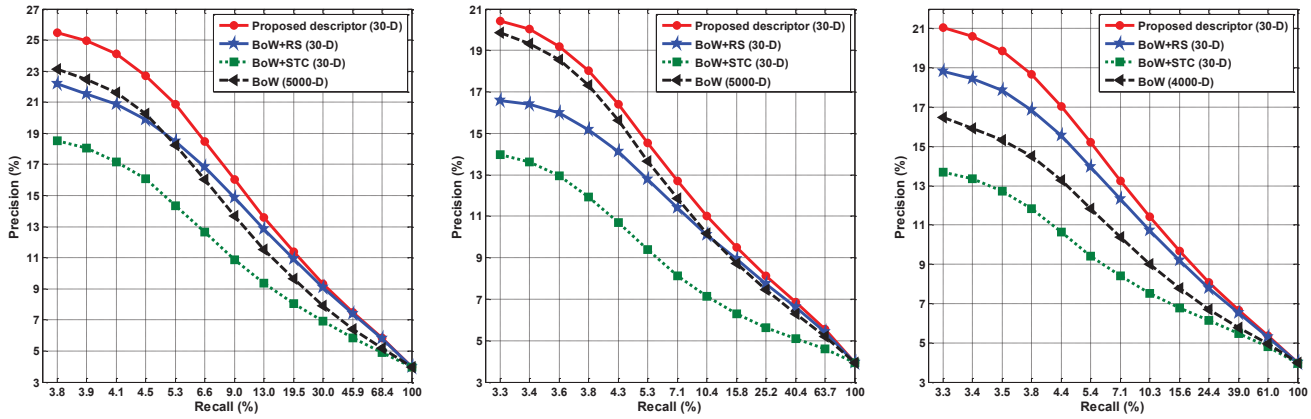
In this section we describe the used web video dataset, and present quantitative evaluations of the proposed descriptor and the comparison with BoW and two other topical representations based on the original Replicated Softmax (RS) and sparse topical coding (STC) [10] which has been successfully applied to compact image description [6].

4.1. Video dataset and representations

The Columbia Consumer Video (CCV) dataset [3], collected from the Internet, consists of 9,317 web videos. It contains 20 semantic categories, including events like "baseball" and "birthday", scenes like "playground", and objects like "bird".

According to [3], three visual and audio local keypoint features are extracted for each video: scale-invariant feature transform (SIFT) [14], spatial-temporal interest points (STIP) [15] and mel-frequency cepstral coefficients (MFCC) [16]. All the three features are then clustered into a single BoW using a soft-weighting scheme, respectively.

4.2. Retrieval results and comparisons



(a) SIFT (b) STIP (c) MFCC

Fig. 3. Precision-Recall curves for three local features and the comparison with BoW, BoW+RS and BoW+STC. The dimensionality of the BoW representations is 5000-D (for SIFT, STIP) and 4000-D (for MFCC). The proposed, RS-based and STC-based descriptors are all 30-D. A retrieval video is considered relevant to the query video if they have the same class label. Results are averaged over all 4,658 possible queries from the test set.

We evaluate the retrieval performance of the proposed descriptor on the CCV dataset for the SIFT, STIP and MFCC features. The sizes of vocabularies used to generate BoW histograms and the split for training and testing are the same as [3]. The regularizer coefficient λ is set to be 0.05 via cross-validation. The similarity is measured using the Cosine distance.

Fig. 3 shows that the proposed descriptor achieves the highest precision rate at a fixed recall rate for all the three local features. Especially, our descriptor still has better performance than BoW although its size is much smaller, which means it well captures the semantic topic structures of the web videos. In addition, our descriptor outperforms BoW +RS, which demonstrates that our extended model effectively improves the descriptor discriminability compared with the original Replicated Softmax. Note that BoW+STC does not achieve the expected performance as [6] in which STC-based image descriptor outperforms BoW because web videos contain much larger intra-class diversity than images.

We also use mean Average Precision (mAP) to evaluate the retrieval performance under different compactness conditions (i.e., different sizes). As illustrated in Table 1, our descriptor outperforms other topical representations in terms of both the compactness and the ranking distortion, and still preserves discriminability comparable to BoW when the size of the descriptor is very small (e.g., 10 dimensions). It can be also observed that the descriptor using SIFT performs better than those using other two local features because scene and object information are usually more helpful for analyzing unconstrained web videos.

Table 1. mAP comparison with other topical descriptors under different compactness conditions.

Local Feature	Size	mAP (%)			
		BoW+ STC	BoW+ RS	Proposed descriptor	BoW
SIFT	10-D	8.73	10.00	**11.45**	11.18 (5000-D)
	20-D	8.98	11.25	**12.01**	
	30-D	9.14	11.94	**12.44**	
STIP	10-D	7.32	9.00	**9.67**	9.85 (5000-D)
	20-D	7.40	9.44	**10.00**	
	30-D	7.55	9.78	**10.33**	
MFCC	10-D	7.75	8.98	**9.78**	9.06 (4000-D)
	20-D	7.86	9.63	**10.25**	
	30-D	7.93	9.90	**10.28**	

5. CONCLUSION

This paper has developed a compact topical video descriptor based on an undirected topic model for web video retrieval. Nonnegative real-valued hidden units and an L_1 penalty are used to improve the descriptor discriminability. Stochastic gradient descent is performed to efficiently learn the extended topic model. Experimental results have demonstrated the effectiveness of the proposed descriptor in terms of the descriptor compactness and the retrieval accuracy. Currently, we have only considered the single-modal input and two-layer model. Future work will focus on multimodal and multilayer structured descriptor learning.

6. ACKNOWLEDGEMENT

This work is jointly supported by National Natural Science Foundation of China (61175003, 61135002, 61203252), Hundred Talents Program of CAS, National Basic Research Program of China (2012CB316300), National Key Technology R&D Program (2011BAH11B01), and Tsinghua National Laboratory for Information Science and Technology Cross-discipline Foundation.

7. REFERENCES

[1] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," *Proc. IEEE International Conference on Computer Vision*, pp. 1470-1477, Oct. 2003.

[2] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabulary and fast spatial matching," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, Jun. 2007.

[3] Y.G. Jiang, G. Ye, S.F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," *Proc. ACM International Conference on Multimedia Retrieval*, Apr. 2011.

[4] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704-1716, Sep. 2012.

[5] V. Chandrasekhar, G. Takacs, and D. Chen et. al., "Chog: Compressed histogram of gradients a low bit-rate feature descriptor," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2504-2511, Jun. 2009.

[6] R. Ji, L.Y. Duan, J. Chen, and W. Gao, "Towards compact topical descriptors," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2925-2932, Jun. 2012.

[7] M. Douze, H. Jégou, C. Schmid, and P. Pérez, "Compact video description for copy detection with precise temporal alignment," *Proc. European Conference on Computer Vision*, pp. 522-535, Sep. 2010.

[8] R. Salakhutdinov and G.E. Hinton, "Replicated softmax: an undirected topic model," *Advances in Neural Information Processing Systems*, pp. 1607-1614, 2009.

[9] D. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, Jan. 2003.

[10] Jun Zhu and Eric P. Xing, "Sparse topical coding," *Proc. Conference on Uncertainty in Artificial Intelligence*, 2011.

[11] V. Nair and G.E. Hinton, "Rectified Linear units improve restricted Boltzmann machines," *Proc. International Conference on Machine Learning*, 2010.

[12] H. Lee, C. Ekanadham, and A.Y. Ng, "Sparse deep belief net model for visual area V2," *Advances in Neural Information Processing Systems*, pp. 873-880, 2008.

[13] G.E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1711-1800, Aug. 2002.

[14] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 60: 91-110, 2004.

[15] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, 64:107-123, Sep. 2005.

[16] B. Logan, "Mel frequency cepstral coefficients for music modeling," *International Symposium on Music Information Retrieval*, 28:5, Oct. 2000.