21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Cross-Modal Topic Correlations for Multimedia Retrieval

Jing Yu† Yonghui Cong† Zengchang Qin† Tao Wan‡
† *Intelligent Computing and Machine Learning Lab, School of ASEE*
Beihang University, Beijing, China
‡ *School of Medicine, Boston University, USA*
E-mails: emy_yu@asee.buaa.edu.cn, zcqin@buaa.edu.cn, taowan@bu.edu

Abstract

In this paper, we propose a novel approach for cross-modal multimedia retrieval by jointly modeling the text and image components of multimedia documents. In this model, the image component is represented by local SIFT descriptors based on the bag-of-feature model. The text component is represented by a topic distribution learned from latent topic models such as latent Dirichlet allocation (LDA). The latent semantic relations between texts and images can be reflected by correlations between the word topics and topics of image features. A statistical correlation model conditioned on category information is investigated. Experimental results on a benchmark Wikipedia dataset show that the newly proposed approach outperforms state-of-the-art cross-modal multimedia retrieval systems.

1. Introduction

Online multimedia information, including texts, images, videos, music, etc., is increasing with a surprising speed. Much research has been done to develop information processing techniques and personalized interfaces to access such large scale multimedia information. The predominant information retrieval systems, e.g. Google and Yahoo, are still text-based that require loads of manual information annotation. So researchers are focused on building models that can bridge different information modalities. One important research direction is automatic image annotation [2]. However, most of previously proposed models are uni-modal only considering keywords, captions or category labels.

Extended from the uni-modal approaches, multi-modal retrieval systems have achieved significant progress [9]. One modality can be used to improve the retrieval performance by providing more correlated information from another modality. One can also project the image and text to the same semantic space and measure the similarity of these two modalities in such space [6]. But the main challenge still remains for further progress of how to build an effective joint model for representing multiple modalities, which is beneficial for addressing the "semantic gap" problem across different content modalities. The cross-modal retrieval system can take advantages of the whole structure of multimedia documents, not just the keywords, but the articles accompanied with relevant images, music or videos, which are widely existing on the pages of news, social network, electronic commerce, etc. This is much closer to what humans do than matching visual features and words separately.

Recently, Rasiwasia *et al.* [7] has demonstrated the advantages of jointly modeling texts and images by learning their correlations via canonical correlation analysis. Some other researchers employ probabilistic latent variable models to learn the shared hidden topics of the two data modalities [4]. Different from their work, we aim to design a cross-modal multimedia retrieval system by considering the statistical correlations between these two components. As a general cross-modal retrieval system, our model has the following two basic functions: (1) Given a query text, retrieve relevant images, and (2) given a query image, retrieve relevant texts. Our work mainly concentrates on the joint modeling between different modalities by using given category information.

2. Cross-modal Correlation

How to use appropriate content (image and text) representation is a key issue in a large and heterogeneous collection of images accompanied by unstructured noisy text. In content based image modeling, one of the most popular approaches is the bag-of-features (BoF) model [8]. Previous research has shown that the BoF model is effective in object and scene classifica-

tion, image retrieval [10], and video retrieval tasks. It captures the invariant aspects of local keypoint features. Among image features, SIFT features [3] are invariant to rotation, scaling, translation and small distortions. It has been proven to be robust with respect to different geometrical changes. In this research, images are represented by BoF with SIFT features.

In modeling text corpora, statistical methods have become increasingly popular and attracted more attention than classical syntactic rule-based natural language processing (NLP) techniques. Topic models are such hierarchical Bayesian models for modeling discrete data collections by clustering words based on co-occurrence information. Latent Dirichlet analysis (LDA) [1] is one of topic models that has been widely used in different NLP tasks including text classification, and question answering system [5]. In this paper, the text component in a document is described as a distribution over pretrained topics by using the standard LDA [1].

We can see the basic ideas of the above two models are similar. Local low-level features of image (or text) are not used directly in modeling. Mid-level representations (visual words or topics) are constructed for modeling the content information in a two-level hierarchical structure. In this paper, we use the term "topics of features" to represent visual words of the BoF model, because what we are interested in is the correlations between the topics of words and the topics of features. In other words, we concentrate on the topic correlations between different modalities.

2.1 Cross-modal retrieval system

In multimedia information retrieval, we assume that each document contains contents in different modalities, e.g., texts and images. The problem is about how to retrieve semantically matched texts given a query image and vice-versa. Formally, given a set of documents denoted by $\mathbf{D} = [D_1, D_2, ..., D_K]$, each document contains an image and its related text. In fact, this combination can be multiple texts accompanied with more than one image or no image. For simplicity, we only consider a simplified case that there is only one-to-one mapping between the image and the text, i.e., $D_k = [I_k, TX_k]$, where $D_k \in \mathbf{D}$, I_k and TX_k is the related image and related text in D_k respectively. Our tasks are: given a query I_q (or TX_q), return a document $D_j \in \mathbf{D}$ that has the most semantically related text (image).

Given the above representation of two modalities, the key technique for cross-modal retrieval is to build a joint model to bridge the texts and images by modeling their correlations. For simplicity, we introduce a score function to evaluate the likelihood of an image I_k given a text query TX_q: given a query text, we seek for an image I_k, that is the most relevant in the dataset via computing the conditional probability of I_k given TX_q, denoted as $S(I_k) = P(I_k|TX_q)$. Similarly, the score function for an image query I_q is: $S(TX_k) = P(TX_k|I_q)$. These two scores will be used to rank a list of returned data in response to the query.

For a particular document, though its contents may be in different modalities, the underlying semantic meaning for these contents are similar. For instance, documents about history will most probably contain words like "war","army", and "weapon". Meanwhile, these documents may have high probabilities over specific visual words. In view of this, our approach is to make use of this relation to compute the conditional probability. We define that $\mathbf{V} = [V_1, V_2, ..., V_M]$ as a set of visual words in the codebook where M is the size of the codebook, and $\mathbf{T} = [T_1, T_2, ..., T_N]$ as a set of topics learned by topic model where N is a predefined number of topics. For a visual word V_i and a topic T_j, the underlying probabilistic relation can be computed based on the training document $\mathbf{D}$.

$$P(V_i|T_j) = \sum_{k=1}^{K} P(V_i|I_k)P(I_k|TX_k)P(TX_k|T_j) \quad (1)$$

where $P(V_i|I_k)$ is just the BoF distribution of the image I_k. Since the image I_k and the text TX_k appear in the same document d_k, then $P(I_k|TX_k) = P(TX_k|I_k) = 1$. For the third term $P(TX_k|T_j)$, according to the Bayes theorem, we can obtain

$$P(TX_k|T_j) = \frac{P(T_j|TX_k)P(TX_k)}{\sum_{k=1}^{K} P(T_j|TX_k)P(TX_k)} \quad (2)$$

where $P(TX_k)$ is the prior probability of the text component in a document $d_k \in \mathbf{D}$. Without any information on each document, we use the uniform distribution as the prior according to the principle of maximum entropy. Formally, $P(TX_k) = P(I_k) = \frac{1}{K}$. Based on topic model, TX_k is represented by a distribution over topics, $P(T_j|TX_k)$ is just the probability of the jth topic T_j for $j = 1, \ldots, N$. Similarly, the likelihood of topic T_j given a visual word V_i is evaluated by

$$P(T_j|V_i) = \sum_{k=1}^{K} P(T_j|TX_k)P(TX_k|I_k)P(I_k|V_i) \quad (3)$$

Based on the above correlation between the topics of words and the topics of features, we can calculate the connection between any images (texts) and a query text (image). The likelihood of being the image I_k given a query text TX_q can be evaluated by Eq. 4. Similarly, $P(TX_k|I_q)$ can be evaluated. We referred to this model

[1] Standard LDA code in C can be downloaded for free from: http://www.cs.princeton.edu/∼blei/lda-c/

as *Naive Topic Correlations* (NTC).

$$P(I_k|TX_q) = \sum_i \sum_j P(I_k|V_i)P(V_i|T_j)P(T_j|TX_q) \quad (4)$$

2.2. Correlation with category information

In this section, we will consider a local learning approach to find better matching patterns within one semantic category. Our approach is to consider the topic correlation within a category. Therefore, relations between documents and categories can be established via feature topics and word topics. Their relationships can be utilized by computing the conditional probability given each category.

In this framework, we first train SVM [2] classifiers based on texts and images separately. Given a query text (image), text classifier (image classifier) is applied to compute its probability distribution over categories. We define C_i as the ith category given a set of categories $\mathbf{C} = [C_1, C_2, ..., C_n]$. Then the probability of the image I_k given a query text TX_q is computed by summing up the conditional probabilities across all the categories.

$$P(I_k|TX_q) = \sum_i P(I_k|C_i)P(C_i|TX_q) \quad (5)$$

Based on the Bayes theorem, we can obtain

$$P(I_k|C_i) = \frac{P(C_i|I_k)P(I_k)}{\sum_k P(C_i|I_k)P(I_k)} \quad (6)$$

Similarly, given an image query, the probability of a text component can also be computed. The values of $P(C_i|I_k)$ and $P(C_i|TX_k)$ can be obtained by the predictions from trained SVM classifiers. And these two values are not necessarily the same as the classifiers trained individually based on the contents of different modalities.

3 Experimental Studies

To evaluate the effectiveness of the newly proposed model, a document corpus with paired texts and images is required. We conduct a series of experiments by testing our algorithms on Wikipedia dataset. The Wikipedia corpus[3] is collected from the "Wikipedia featured articles" that contains 2866 paired images and texts divided into 10 categories, among which 2173 documents are for training and 693 documents for testing. In Wikipedia dataset, texts are of good quality and representative enough of their categories. However, images of each category are comparatively ambiguous. That's because categories in Wikipedia are more abstract and they have semantic overlap. This dataset represents the main retrieval application in practice when texts are more classable than images.

[2]The Support Vector Machine (SVM) code is free available online from: http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

[3]The Wikipedia corpus is free available online from: http://www.svcl.ucsd.edu/projects/crossmodal/

Table 1. Retrieval results in MAP.

Model	Image query	Text query	Average
$Baseline$[7]	.118	.118	.118
SCM[7]	.277	.226	.252
NTC	.142	.162	.152
$CMTC$	**.293**	**.232**	**.266**

We test our proposed correlations model on the wikipedia data for two tasks: (1) using a query image to retrieve related texts and (2) vice-versa. The mean average precision (MAP) and precision-recall (PR) curves are used to evaluate the retrieval performance. In our experiments, we set $M = N = 100$ such that 100 topics are used for LDA and 100 visual words are used for the bag-of-features model. For each image data, we extract 128-dimensional SIFT descriptors and quantize them into a 100-dimensional codebook to form a feature vector. Besides, we remove the stop-words and employ stemming methods to pre-process the text data and use the processed texts to train LDA model with the topic number of 100. In the retrieval procedure, the query image (text) is also described as the distribution over visual words (topics) at first, and then compute the score of each text (image) in the training data. Finally, the system returns a list of ranked texts (images) in descending order of the scores.

For the cross-modal topic correlation model with category information (CMTC for short), we train SVM classifiers on training texts and images' topic features respectively. Then texts and images can be represented as distributions over categories. Experimental results indicates that SVM classifiers with the linear kernel have the best performance in classification. So we use the SVM with linear kernel for category estimation and prediction. The score of each text (image) in the training dataset is evaluated and the system returns a ranked set of texts (images).

3.1. Results and discussions

The MAP for the two proposed cross-modal retrieval systems as well as the semantic correlation matching (SCM)[7] model are shown in Table 1. As is illustrated in Table 1, we can obtain the following conclu-

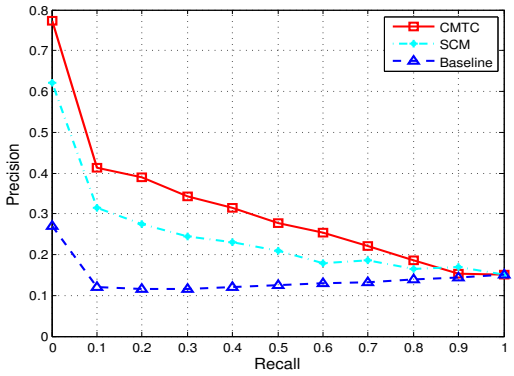

Figure 1. Text Query

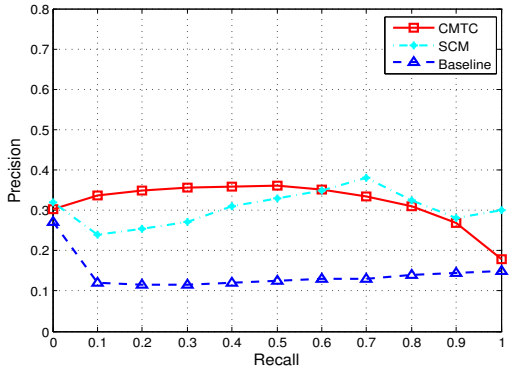

Figure 2. Image Query

sions: First, CMTC results have much improvement over random retrieval as baseline. The average MAP of CMTC is higher than double performance of random case. Second, the CMTC model outperforms SCM in both image retrieving given text query and vice-versa. Compared to the SCM model, the image query result is improved by 1.6% and text query result is improved by 0.6%. Also, the CMTC model doesn't need to map texts and images to a correlation space, that is more simple and timesaving than SCM. Compared to the NTC model, the category information helps to improve the retrieval performance to a great extent. Fig.1 and Fig.2 show PR curves of cross-modal retrieval results with CMTC, SCM, and the random baseline on Wikipedia. They show that our CMTC improves the precision of the two retrieval tasks at all levels of recall compared to the random baseline. For text query, CMTC outperforms SCM at all levels of recall. For image query, CMTC has higher precision than SCM at first five levels of recall and has comparable precision with SCM at three levels of recall. It indicates that more related retrieval texts via CMTC rank closer to the top than those results via SCM , which is more applicable in practice because users mainly concern the top retrieval results.

4. Conclusions and future work

In this paper, we proposed a novel approach for cross-modal multimedia retrieval. We have investigated two models of correlation between mid-level topics in different modalities. Conditional probabilities are used to measure the correlation between the image (text) query and the documents for retrieval. Experimental studies on a benchmark Wikipedia dataset show that the category-based correlation model achieves the best performance compared to other state-of-the-art cross-modal retrieval systems. The future work will be focused on building a more reasonable generative model to investigate the text-image correlation. In such a framework, both images and texts can be considered as observed variables generated by some hidden semantic concepts. The observable images and texts can be used to estimate hyper-parameters or the latent structure of the model.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *JMLR*, Vol. 3: pp. 993-1022, 2003.

[2] D. Blei and M. Jordan. Modeling annotated data. In *ACM SIGIR*, pp. 127-134, 2003.

[3] D.G. Lowe. Object recognition from local scale-invariant features, In *ICCV*, pp. 1150-1157, 1999.

[4] D. Putthividhya, H. Attias, and S. Nagarajan. Topic-regression multimodal latent dirichlet allocation for image annotation. In *CVPR*, 2010, pp. 3408C3415.

[5] Z. Qin, M. Thint and Z. Huan. Ranking answers by hierarchical topic models. In *IEA/AIE*, LNAI 5579, pp. 103-112.

[6] N. Rasiwasia, P. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. In *IEEE Transactions on Multimedia*, 9(5):923-938, 2007.

[7] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM Multimedia*, 2010.

[8] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470-1477, 2003.

[9] T. Westerveld. Probabilistic multimedia retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference*, page 438, 2002.

[10] X. Yuan, J. Yu, Z. Qin and T. Wan, A SIFT-LBP retrieval model based on bag-of features. In *ICIP*, pp. 1061-1064, 2011.