Computer Vision and Image Understanding 136 (2015) 3–13

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Visual Topic Network: Building better image representations for images in social media

Zhenxing Niu [a], Gang Hua [b,*], Qi Tian [c], Xinbo Gao [a]

[a] The School of Electronic Engineering, Xidian University, China
[b] The Department of Computer Science, Stevens Institute of Technology, USA
[c] The Department of Computer Science, University of Texas at San Antonio, USA

ARTICLE INFO

Article history:
Received 5 March 2014
Accepted 28 January 2015

Keywords:
Image representation
Topic model
Visual Topic Network
Social media

ABSTRACT

Topic models have demonstrated to be effective on building image representations for general images. Recently, how to build better image representations for images in social media becomes an interesting problem, where one key issue is how to leverage images' social contextual cues, *e.g.,* user tags associated with images. Nevertheless, most previous methods either just exploited image content and neglect user tags, or assumed there are exact correspondences between image content and tags, *i.e.,* tags are closely related to image content. Thus, they cannot be applied to the realistic scenarios where the images are only weakly annotated with tags, *i.e.,* tags are only loosely related to image content as already manifested in real-world social media data. In this paper, we address the problem of building better image representations in social media, where the images are weakly annotated with user tags. In particular, we organize a collection of images as an image network where the relations between images are modeled by user tags. To model such image network and build image representations, we further propose a network structured topic model, namely *Visual Topic Network* (VTN), where the image content and their relations are simultaneously modeled. In this way, the weakly annotated tags can be effectively leveraged as building image representations. The proposed VTN model is inspired by the Relational Topic Model (RTM) recently introduced in the document analysis literature. Different from the binary article relations in RTM, the proposed VTN can model the multiple-level image relations. Our extensive experiments on two social media datasets demonstrated the advantage of the proposed VTN model.

1. Introduction

With the advancement of research on local image features such as SIFT [1], many works focus on building better image representations based on image local features. For example, the bag-of-words (BoW) representation [2–6] have been very popular. Typically, once local features are extracted and quantized, their overall distribution in an image is represented in terms of bag-of-words histograms.

Although the BoW representation is simple to build, the visual words are difficult to interpret, *i.e.,* the 'polysemy' and 'synonymy' issue [7]: a visual word may represent different image visual content and several visual words may represent the same image visual content. To alleviate such an issue, many works try to build a hierarchical image representation, where the intermediate level representation is more interpretable than the low level visual words. For example, an scene image can be described by a three-level hierarchical structure: *a scene* (*e.g.,* 'street') – *scene elements* (*e.g.,* 'sky', 'road') – *image patches*. The scene elements are more interpretable than the image patches.

To exploit such multi-level hierarchy, *topic models* have been introduced for image representation and recognition [8–12]. In topic models, each image is represented as a mixture of 'topics', and each topic is described by a distribution over visual words. Through topic modeling, some topics could be learnt from a collection of images, and usually they are easily interpretable, *i.e.,* the visual words representing the same image content are usually 'clustered' into one topic. Therefore, the mixture of topics can be regard as an intermediate level image representation, which is not only more interpretable but also has lower dimension compared to the BoW representation. The typical topic models introduced for visual recognition are the probabilistic Latent Semantic Analysis (pLSA) [13] and the Latent Dirichlet Allocation (LDA) [14]. For example, Fei-Fei and Bosch [9,8] exploited LDA and pLSA

* Corresponding author.
E-mail addresses: zhenxingniu@gmail.com (Z. Niu), ghua@stevens.edu (G. Hua), qitian@cs.utsa.edu (Q. Tian), xbgao@ieee.org (X. Gao).

http://dx.doi.org/10.1016/j.cviu.2015.01.010
1077-3142/

for scene recognition. Sudderth et al. [10] presented a hierarchical topic model for part-based object and scene category recognition.

Recently, with the popularity of social networks (*e.g.,* Facebook) and content-sharing websites (*e.g.,* Flickr and YouTube), images on social networks or content-sharing websites, called *social images*, become the dominating multimedia objects on the Internet. And how to build better image representations for such social images becomes more and more important for social media analysis. Different from general images, social images are often accompanied by various forms of contextual information like keywords, tags, comments, as well as surrounding texts. Such contextual cues are usually beneficial for understanding the image content. Thus, one key issue for social image understanding is how to leverage those contextual cues. However, the aforementioned works focused on modeling image content and neglecting such contextual cues.

One kind of the most leveraged contextual cues is the keyword associated with images. To leverage the associated keywords for social image understanding, some methods focused on modeling the joint distribution of visual features and keywords [15–19]. In [15,16], the process of building the relationship between the visual features and the keywords was analogous to a language translation. Further, Corr-LDA was proposed to extend this approach through a hierarchical probabilistic mixture model [17,18]. In [20], tr-mmLDA was proposed to capture correlations between images and annotation texts, where the correlations between the two data modalities were modeled with a linear Gaussian regression module. Recently, image recognition and annotation are simultaneously considered in [19] by modeling the joint distribution of image content, annotation keywords, and class labels.

However, these methods assume that there are explicit correspondences between the keywords and image content. Thus, they can only be applied to the case where all the *keywords* have a visual interpretation rather than the realistic scenarios where the images are weakly annotated with *tags*, *i.e., tags* are usually loosely related to the image content compared to *keywords* [21]. For example, a photo for the 'Lincoln Memorial' could have a tag 'National Mall' since the Lincoln Memorial is located at the National Mall street, but there does not exist a correspondence between the photo content and such a tag.

Recently, an image representation learning method [22,23] was proposed to leverage those loosely related tags in a new way. Specifically, a multimedia information network (MINets) was constructed with images and tags. And then the image similarity and the image-tag relationship were described by relational matrices. At last, the learning of image representation was formulated as a low-rank matrix approximation problem. Similarly, in [24] an image network was constructed with images and their social-network metadata, and a discriminative method was proposed to model such image network for image recognition. Although those social-network metadata (*e.g.,* user tags, the comments, the groups to which those images belong, the uploaders of those images, *etc.*) are loosely related to image content, the experimental results reveal that they can still be effectively leveraged for understanding image content through such a model. So, these works suggest that the loosely related tags or loosely related metadata can be effectively leveraged by organizing them as a network.

Inspired by it, in this paper we make use of these loosely related tags by constructing an image network. Moreover, by proposing a topic model to model such image network, we build a hierarchical image representation. As shown in Fig. 1, our method consists of the following steps: firstly, each image is represented as a visual document by using a bag-of-words representation, meanwhile user tags are leveraged to define the relations between each pair of images. As a result, a Visual document Network (VN) is constructed where the nodes represent images and links represent image relations. Secondly, a network structured topic model, namely Visual Topical Network (VTN), is proposed to model the VN and build an intermediate level image representation.

In this paper, instead of directly modeling these loosely related tags, we leverage them to describe image relations. For example, if two images share more common tags, we define that they have a *strong* relation which indicates they are more likely to have similar representations. Otherwise, we define that they have a *weak* relation which indicates they are not likely to have similar representations. In such a way, these tags can still be adequately explored although they are only loosely related to image content. For example, two photos about 'Lincoln Memorial' are both tagged with a common tag, *e.g.,* 'National Mall'. Although tag 'National Mall' is loosely related to the image content 'Lincoln Memorial', the positive image relation is still beneficial for understanding image content.

To model the constructed image network, a network structured topic model VTN is proposed in this paper. In particular, the proposed VTN is inspired by the idea of the Relational Topic Model (RTM) proposed by Chang and Blei [25]. The RTM is designed to model a text document network formed according to the citation relationships among scientific articles.

However, in [25] the document citation is described by a binary variable (*i.e., with* or *without* citation between two articles). But for our task, the relations among images are much more complicated. In particular, the image relations are defined as the correlation of the two tag sets, *e.g.,* how many tags are shared between two images. Hence it is better to describe such relations by multiple-valued variables. And subsequently, a new distribution is needed to describe such multiple-valued variables. These are the major differences between the RTM and the proposed VTN model.

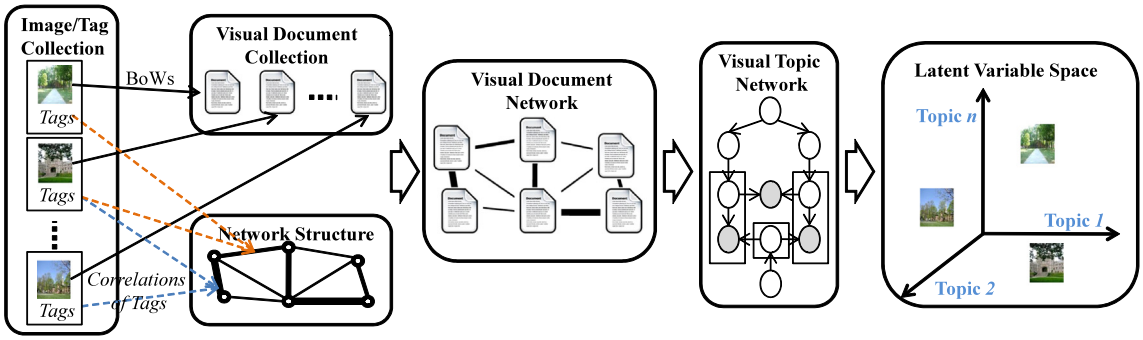

Fig. 1. Building a hierarchy image representation by leveraging image content and user tags. In particularly, the proposed method has two stages: first a network is constructed by leveraging image content (*e.g.,* induced from local features) and image relations (*e.g.,* induced from the user tags); and then a network structured topic model (*i.e.,* Visual Topic Network) is used to model such a network, which could build a hierarchical image representation.