Multimedia Feature Mapping and Correlation Learning for Cross-Modal Retrieval

Xu Yuan, School of Software Technology, Dalian University of Technology, China
Hua Zhong, School of Software Technology, Dalian University of Technology, China
Zhikui Chen, School of Software Technology, Dalian University of Technology, China
Fangming Zhong, School of Software Technology, Dalian University of Technology, China
Yueming Hu, College of Natural Resources and Environment, South China Agricultural University, China

ABSTRACT

This article describes how with the rapid increasing of multimedia content on the Internet, the need for effective cross-modal retrieval has attracted much attention recently. Many related works ignore the latent semantic correlations of modalities in the non-linear space and the extraction of high-level modality features, which only focuses on the semantic mapping of modalities in linear space and the use of low-level artificial features as modality feature representation. To solve these issues, the authors first utilizes convolutional neural networks and topic modal to obtain a high-level semantic feature of various modalities. Sequentially, they propose a supervised learning algorithm based on a kernel with partial least squares that can capture semantic correlations across modalities. Finally, the joint model of different modalities is learnt by the training set. Extensive experiments are conducted on three benchmark datasets that include Wikipedia, Pascal and MIRFlickr. The results show that the proposed approach achieves better retrieval performance over several state-of-the-art approaches.

KEYWORDS
Correlation Learning, Cross-Modal Retrieval, Multimedia, Semantic Feature, Text and Image

INTRODUCTION

Cross-modal multimedia retrieval has become a widespread concern over the last few years owing to the explosion growth of multimedia information over the Internet. Multimedia data that is typical multimodal is derived from different channels, and data of different modalities can be represented by the same semantic type. Specifically, texts are used as the semantic representation of associated images or videos. The massive collections of images, texts and videos pose several challenges to multimedia retrieval. However, most of the conventional systems are only applied to the retrieval of single modal data, such as search engines (Google or Yahoo), resulting in the limited use of multimodal data. How to sustainably use these multimodal data for smart retrieval remains a challenge.

DOI: 10.4018/IJGHPC.2018070103

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
The key step of cross-modal retrieval task that the image or video can be found by text query is to reduce the semantic gap across modalities. A number of cross-modal retrieval approaches (Chen, Wang, Wang, & Zhang, 2012; Rasiwasia et al., 2010; Tang, Deng, & Gao, 2015; Zhang, Zhong, Yang, Chen, & Bu, 2016; Wang, Yang, & Meinel, 2015; Wang et al., 2016; Yu, Cong, Qin, & Wan, 2012; Zhuang, Wang, Wu, Zhang, & Lu, 2013) have been devoted to address the issue of semantic gap in the recent past. In our work, the semantic gap between image and text is mainly concerned.

Recently, the academic community has explored some models to bridge the semantic gap. The most popular technique may be canonical correlation analysis (CCA) (Rasiwasia et al., 2010), aiming to obtain a common space by maximizing the correlations between feature vectors of different modalities. Another typical approach is partial least squares (PLS) (Sharma & Jacobs, 2011), which also has attracted much attention. Besides CCA and PLS, some other methods are proposed to reduce the semantic gap. Yu et al. (Yu et al., 2012) used statistical correlation based on the topic model for image and text query. Zhai et al. (Zhai, Peng, & Xiao, 2012) proposed a joint model to exploit negative and positive correlation for cross-modal retrieval. Wang et al. (Wang, He, Wang, & Tan, 2013) applied penalty to projection matrices, and mapped multimodal data into a common latent subspace for feature matching.

The above methods only pay attention to the semantic mapping of modalities in linear space, while neglecting the latent semantic correlations of modalities in the highly non-linear space, as well as the high-level semantic feature in non-linear space. However, there may be non-linear correlations across modalities. Non-linear space may be more appropriate to mine the semantic correlations of different modalities than the linear space. If the multimodal correlation model is directly used in non-linear space, there exists a series of problems. Such as the selection of non-linear mapping functions and the curse of dimensionality in high-dimensional feature space. Additionally, the low-level artificial features utilized in these methods cannot contain enough semantic information that results in weakness semantic representation, such as scale invariant feature transformation (SIFT) or GIST used for image representation. Hence, constructing a joint high-level semantic model is crucial for cross-modal retrieval.

To solve these issues, we propose a novel supervised learning algorithm based on kernel partial least squares (KPLS) (Rosipal & Trejo, 2001), where the semantic correlations of different modalities can be effectively learnt for cross-modal retrieval. On the one hand, we capture the intra-modal correlations between image and text in the non-linear space, non-linear space has stronger capability to reduce the semantic gap than the linear model. On the other hand, the kernel trick is exploited to solve the curse of dimensionality, and reduce the computational cost effectively. Besides, we design a joint model for semantic representation. The model is inspired by the successful application of convolutional neural networks (CNN) in image representation (Wang et al., 2015) and topic model in textual representation (Rasiwasia et al., 2010). In our work, CNN is utilized to learn high-level semantic representation of image modality which can improve the performance of our approach. As for the text representation, topic model is used for text feature extraction from given document.

The contributions of our work are summarized as follows:

- Our supervised correlation learning algorithm is proposed to obtain semantic correlations across modalities, which can reduce the computational cost effectively.
- Different from previous works, image feature and text feature are extracted by CNN and topic model respectively, and our joint model explores the high-level semantic representation for cross-modal retrieval.
- Extensive experiments are conducted on three benchmark datasets to evaluate the proposed approach. Furthermore, the experimental results confirm that our approach outperforms several existing works.
Fednets: P2P Cooperation of Personal Networks Access Control and Management Framework
www.igi-global.com/chapter/fednets-p2p-cooperation-personal-networks/40835?camid=4v1a

Interflow Network Coding Approaches for Data Heterogeneous Unicast Sessions in Wireless Networks
www.igi-global.com/article/interflow-network-coding-approaches-for-data-heterogeneous-unicast-sessions-in-wireless-networks/149912?camid=4v1a