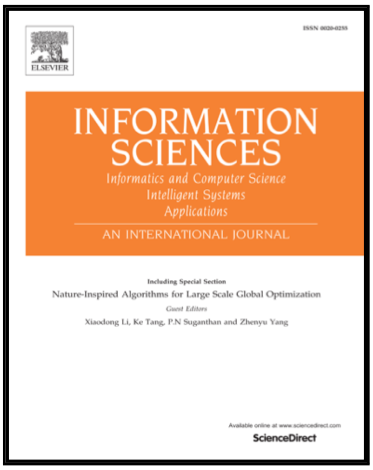
Accepted Manuscript



Manifold Regularized Cross-Modal Embedding for Zero-Shot Learning

Zhong Ji , Yunlong Yu , Yanwei Pang , Jichang Guo , Zhongfei Zhang

PII: S0020-0255(16)31236-1
DOI: 10.1016/j.ins.2016.10.025
Reference: INS 12584

To appear in: *Information Sciences*

Received date: 15 May 2016
Revised date: 2 October 2016
Accepted date: 9 October 2016

Please cite this article as: Zhong Ji , Yunlong Yu , Yanwei Pang , Jichang Guo , Zhongfei Zhang , Manifold Regularized Cross-Modal Embedding for Zero-Shot Learning, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.10.025

Manifold Regularized Cross-Modal Embedding for Zero-Shot Learning

Zhong Ji[a], jizhong@tju.edu.cn, Yunlong Yu[a], yuyunlong@tju.edu.cn, Yanwei Pang[a], pyw@tju.edu.cn, Jichang Guo[a], jcguo@tju.edu.cn, Zhongfei Zhang[b], zhongfei@cs.binghamton.edu

[a]School of Electronic Information Engineering, Tianjin University, Tianjin, 300072, China
[b]Department of Computer Science, State University of New York, Binghamton, NY 13902, USA

Abstract
Zero-Shot Learning (ZSL) aims at classifying previously unseen class samples and has gained its popularity in applications where samples of some categories are scarce for training. The basic idea to address this issue is transferring knowledge from the seen classes to the unseen classes through mapping the visual feature to an embedding space spanned by class semantic information. The class semantic information can be obtained from human-labeled attributes or text corpus in an unsupervised fashion. Therefore, the embedding function from visual space to the embedding space is extremely important. However, the existing embedding approaches to ZSL mainly focus on aligning pairwise semantic consistency from heterogeneous spaces but ignore the intrinsic structure of the locally homogeneous isomorph. In order to preserve the locally visual structure in the embedding process, this paper proposes a Manifold regularized Cross-Modal Embedding (MCME) approach for ZSL by formulating the manifold constraint for intrinsic structure of the visual features as well as aligning pairwise consistency. The linear, closed-form solution makes MCME efficient to compute. Furthermore, rather than applying the embedding function learned from the seen classes directly, we also propose a new domain adaptation strategy to overcome the domain-shift problem during the knowledge transfer process. The MCME with the domain adaptation method is called MCME-DA. Extensive experiments on the benchmark datasets of AwA and CUB validate the superiority and promise of MCME and MCME-DA.

Keywords
Zero-shot learning, image classification, cross-modal embedding, manifold, domain adaptation.

1 Introduction

Image classification is a major research area in computer vision [17][26][29][35][43]. Tremendous improvements have made in recent years mainly due to the prosperous progress of Convolutional Neural Networks (CNN) [1] and large scale annotated datasets, e.g., ImageNet. Amazing results have even been reported surpassing human-level performance on the 1000-class ImageNet dataset [19]. However, labeling all the classes for training is still expensive, time-consuming, and impractical. In addition, the labeled data are not always available for some classes in real-world settings.

Thus, it is imperative to introduce Zero-Shot Learning (ZSL) to improve the scalability of conventional image classification methods [3][8][11][27][38]. ZSL aims at learning classification models for the novel classes with no labeled data for training. To address this challenging task, existing methods solve the problem by transferring knowledge from the seen classes (i.e., classes with labeled data for training) to the unseen ones (i.e., classes with no labeled training data). Actually, ZSL is inspired by the inferential capability of humans. We human beings have the capability to organize information and infer what unknown things might be. A practical example is as follows. Imaging that a child has never seen a zebra, however, he/she has heard that the zebra looks like a horse but with black and white stripes. When this child sees a zebra in the first time, he/she may recognize the zebra with his/her own inferential capability. To this end, two key issues should be considered for ZSL: (1) Collect auxiliary information for each class, including seen classes and unseen classes, and (2) Connect the visual samples and the class auxiliary information. Specifically, this connection is often achieved by constructing a semantic embedding space. In this way, each class is associated with a vector in this embedding space, so that the knowledge can be transferred from the visual space to class label.

Most work utilizes attributes as the semantic embedding space [7][27][28][30]. Attributes define a few properties of an object, such as the color, the shape, and the presence or absence of a certain body part. They are shared across both seen and unseen categories. In this way, both the seen classes and unseen classes are associated with semantic vectors in the class embed-