Information Sciences 000 (2017) 1–13

Contents lists available at ScienceDirect

Information Sciences

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

Attention driven multi-modal similarity learning

Xinjian Gao[a], Tingting Mu[b,*], John Y. Goulermas[c], Meng Wang[a]

[a] School of Computer and Information, Hefei University of Technology, Hefei 230009, China
[b] School of Computer Science, University of Manchester, Manchester M1 7DN, United Kingdom
[c] Department of Computer Science, University of Liverpool, Liverpool L69 3BX, United Kingdom

ARTICLE INFO

Article history:
Received 13 December 2016
Revised 24 July 2017
Accepted 6 August 2017
Available online xxx

Keywords:
Multi-modal similarity
Attention mechanism
Representation learning
Multi-view
Neural network

ABSTRACT

To learn a function for measuring similarity or relevance between objects is an important machine learning task, referred to as similarity learning. Conventional methods are usually insufficient for processing complex patterns, while more sophisticated methods produce results supported by parameters and mathematical operations that are hard to interpret. To improve both model robustness and interpretability, we propose a novel attention driven multi-modal algorithm, which learns a distributed similarity score over different relation modalities and develops an interaction-oriented dynamic attention mechanism to selectively focus on salient patches of objects of interest. Neural networks are used to generate a set of high-level representation vectors for both the entire object and its segmented patches. Multi-view local neighboring structures between objects are encoded in the high-level object representation through an unsupervised pre-training procedure. By initializing the relation embeddings with object cluster centers, each relation modality can be reasonably interpreted as a semantic topic. A layer-wise training scheme based on a mixture of unsupervised and supervised training is proposed to improve generalization. The effectiveness of the proposed method and its superior performance compared against state-of-the-art algorithms are demonstrated through evaluations based on different image retrieval tasks.

1. Introduction

To learn a function that accurately calculates the similarity or relevance between objects is one of the most significant machine learning tasks, and is known as similarity learning. It is closely related to other fundamental machine learning paradigms, including clustering, ranking, classification and regression, and plays an important role in many real-world applications, such as image annotation and retrieval [48], intelligent recommendation systems [37] and knowledge graph completion [20]. Conventional similarity learning methods often learn a distance metric (e.g., Mahalanobis distance) [41] or use a kernel function [12] to measure the (dis)similarity between objects, where the metric (or kernel) formulation is adjusted by function parameters. These methods are mostly based on single modality. Although they are capable of measuring relevance in a standard environment, they may not be able to deal with tasks of more complex nature. For example, to retrieve images relevant to the query image of an apple fruit, images of apple juice (or the company Apple), which are related to the query in other relation types, can also be of interest. Therefore, this requires more sophisticated similarity learning models to encode multiple relation types.

[*] Corresponding author.
E-mail addresses: tingtingmu@me.com, tingting.mu@manchester.ac.uk (T. Mu).

http://dx.doi.org/10.1016/j.ins.2017.08.026
0020-0255/© 2017 Elsevier Inc. All rights reserved.

Multi-modal similarity learning takes into account multiple types of relevance patterns between objects. For example, image relevance reflected by their shape and colour appearance. Multi-modal extensions have been developed for conventional similarity learning based on distance metrics and kernel functions. For instance, multiple kernel similarity learning [39] is proposed to facilitate image ranking, where the multiple modalities of image connections are realized by multiple kernel functions and the overall similarity is computed as a weighted sum of these functions. Transfer distance metric learning [24] is developed to overcome the lack of available information in the target task and discovers multiple alternative connections between objects in relevant source tasks. These correspond to multiple modalities characterized by different base metrics combined to form the final metric. In general, the intermediate results of these methods, such as the parameters or learned relation types, are hard to interpret and the whole learning procedure is usually treated as a black box. Intelligent similarity learning methods that exhibit not only excellent performance but also good model interpretability are in demand.

To extract information from visual objects, primate visual systems employ attention mechanisms to dynamically focus on important information that is relevant to the current behavior or visual tasks [32]. Using the image retrieval task as an example, if the query is the image of a beach, the users could move their focus from the whole scene, to certain parts of the image, e.g., boat, people who swim, or sea. Recent advances in attention mechanisms use a set of dynamic attention weights to control the contribution of different parts [1]. Such techniques have been successful in tasks, such as machine translation [1] and image caption generation [42]. Taking the attention based image caption generation model [42] as an example, it works with high-level representations extracted from image patches using a convolutional neural network (CNN). The model learns the attention weights for each patch to construct a weighted context vector that represents relevant parts of an image based on which a long short-term memory (LSTM) network is used to generate text captions. Inspired by the recent success of attention learning in language and vision, we propose an interaction-oriented attention mechanism to improve the accuracy of similarity learning, and meanwhile show that the attention weights returned by the mechanism are able to improve the model interpretability.

In addition to multi-modal similarity analysis and attention mechanisms that can potentially improve the robustness and interpretability of a learning model, it is also important to improve the model performance. When dealing with complex real-world tasks, features that exhibit heterogeneous properties should be considered. For example, in image retrieval, shape feature is more important for measuring similarity between a brown bear and a polar bear, while the color feature is more important for examining brown bears in different poses. One representative work that deals with this problem is [6], which leverages shared knowledge from multiple related tasks to improve the performance of feature selection. Another commonly used technique for combining multi-view information is multi-view embedding. It aims at mixing and refining information provided by different types of features within a low-dimensional embedding space [28,36,43,50]. Recent developments on multi-view learning have shown that complementary information across different views has the potential of improving the performance of many machine learning tasks [10]. To further improve similarity learning, we take into account multi-view local structures in similarity formulation.

In summary, this work proposes a powerful similarity measure by exploring multiple hidden relationships between image objects that suit the multi-modal nature of real-world tasks. To improve model robustness and interpretability, dynamic attentions are incorporated to selectively capture salient parts that contribute to the object interactions. To deal with heterogeneous object properties, we encode multi-view information that improves the object representation. These result in proposing a novel attention-driven multi-modal similarity (AMoS) model possessing a multi-layered architecture. Neural networks are used to compute representation vectors of a given object and its corresponding patches. Different relation modalities are encoded as different hidden neurons in the relation layer. Dynamic attention weights are modeled as functions receiving the entire image for their patch representation, and multi-view information provided by different feature extraction methods are used to enhance the image representation in pre-training. The effectiveness of the proposed model is compared with various state-of-the-art methods evaluated through image retrieval tasks. The remaining paper is organized as follows. Section 2 briefly introduces related works. Section 3 delivers the proposed algorithm, while Section 4 contains experimental results and comparative analyses. Finally, Section 5 concludes the work.

2. Related works

2.1. Multi-modal similarity learning

Multi-modal similarity learning is a type of learning that relies on measuring the relevance between objects from multiple aspects. It has been shown to be effective in many real-world applications. One example is person identification over camera networks, using multiple Mahalanobis distance metrics designed to characterize different cameras that contain different types of noise [26]. These metrics are connected by enforcing joint regularization that reduce overfitting. Another work mines complementary information among features that exhibit heterogeneous properties by optimizing different distance metrics in different feature spaces [38] . To facilitate tasks such as inter-modal label transfer and zero-shot learning, multi-modal models are developed to formulate the relations between text and image features [31]. In social media network analysis, a latent semantic space is computed to encode multi-modal links, e.g., context and content links between the multimedia and context objects [30]. Another example work in data retrieval [19], develops multi-modal algorithms to achieve cross-modal hashing. For instance, linear subspace ranking hashing maps data from different modalities into a common binary space, so that the cross-modal similarity can be measured using Hamming distance, where different modalities

Please cite this article as: X. Gao et al., Attention driven multi-modal similarity learning, Information Sciences (2017), http://dx.doi.org/10.1016/j.ins.2017.08.026