Computer Vision and Image Understanding xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Structured deep hashing with convolutional neural networks for fast person re-identification☆

Lin Wu[a], Yang Wang[b,c,*], Zongyuan Ge[d], Qichang Hu[e,f], Xue Li[a]

[a] Information Technology and Electrical Engineering, The University of Queensland, QLD 4072, Australia
[b] Dalian University of Technology, China
[c] The University of New South Wales, NSW 2052, Australia
[d] IBM, Australia
[e] The University of Adelaide, SA 5005, Australia
[f] Data61, Canberra, ACT 2601, Australia

ARTICLE INFO

Keywords:
Person re-identification
Convolutional neural networks
Deep hashing
Structured embedding

ABSTRACT

Given a pedestrian image as a query, the purpose of person re-identification is to identify the correct match from a large collection of gallery images depicting the same person captured by disjoint camera views. The critical challenge is how to construct a robust yet discriminative feature representation to capture the compounded variations in pedestrian appearance. To this end, deep learning methods have been proposed to extract hierarchical features against extreme variability of appearance. However, existing methods in this category generally neglect the efficiency in the matching stage whereas the searching speed of a re-identification system is crucial in real-world applications. In this paper, we present a novel deep hashing framework with Convolutional Neural Networks (CNNs) for fast person re-identification. Technically, we simultaneously learn both CNN features and hash functions to get robust yet discriminative features and similarity-preserving hash codes. Thereby, person re-identification can be resolved by efficiently computing and ranking the Hamming distances between images. A structured loss function defined over positive pairs and hard negatives is proposed to formulate a novel optimization problem so that fast convergence and more stable optimized solution can be attained. Extensive experiments on two benchmarks CUHK03 (Li et al., 2014) and Market-1501 (Zheng et al., 2015) show that the proposed deep architecture is efficacy over state-of-the-arts.

1. Introduction

Re-identification is a task of matching persons observed from non-overlapping camera views based on visual appearance. It has gained considerable popularity in video surveillance, multimedia, and security system by its prospect of searching a person of interest from a large amount of video sequences (Sunderrajan and Manjunath, 2016; Wang et al., 2016; Ye et al., 2016). The major challenge arises from the variations in human appearances, poses, viewpoints and background cluster across camera views. Some examples are shown in Fig. 1. Towards this end, many approaches (Farenzena et al., 2010; Li et al., 2013; Paisitkriangkrai et al., 2015; Pedagadi et al., 2013; Zhao et al., 2014) are developed based on a combination of low-level features (including color histogram (Gray and Tao, 2008), spatial co-occurrence representation (Wang et al., 2007), LBP (Xiong et al., 2014) and color SIFT (Zhao et al., 2013)) against variations (*e.g.,* poses and illumination) in pedestrian images. However, these hand-crafted features are still not discriminative and reliable under such severe variations and misalignment across camera views.

Recently, deep learning methods (Ahmed et al., 2015; Chen et al., 2016; Li et al., 2014; Wu et al., 2016; 2017; 2018a; 2018b; Xiao et al., 2016) have been proposed to address the problem of person re-identification by learning deeply discriminative Convolutional Neural Network (CNN) features in a *feed-forward* and *back-propagation* manner. It extracts hierarchical CNN features from pedestrian images; the subsequent metric-cost part compares the CNN features with a chosen metric encoded by specific loss functions, *e.g.,*contrastive (pair-wise) (Ahmed et al., 2015; Li et al., 2014; Wu et al., 2016) or triplet (Chen et al., 2016; Yi et al., 2014) loss functions. However, such typical deep learning methods are not efficient in real-time scenario, due to the less-

☆ The name of the Editor in chief Dr. Herve Jegou

* Corresponding author.
E-mail addresses: lin.wu@uq.edu.au (L. Wu), wangy@cse.unsw.edu.au (Y. Wang), zongyuan@au1.ibm.com (Z. Ge), qichang.hu@adelaide.edu.au (Q. Hu), xueli@itee.uq.edu.au (X. Li).

https://doi.org/10.1016/j.cviu.2017.11.009
Received 16 December 2016; Received in revised form 2 October 2017; Accepted 25 November 2017
1077-3142/ © 2017 Elsevier Inc. All rights reserved.

Fig. 1. Typical samples of pedestrian images in person re-identification from CUHK03 data set (Li et al., 2014). Each column shows two images of the same individual observed by two different camera views.

efficiency of matching two pedestrian images by extracting and comparing hierarchical CNN features. In fact, the excellent recognition accuracy in neural network-based architectures comes at expense of high computational cost caused by the sequential updates in gradient decent of optimizing an objective function with a contrastive or triplet loss. As a result, the learning is slow by taking pairwise or triplet units as input, and the main computational expense for these deep models comes from many weight updates using stochastic gradient descent. This is, in our view, is mostly due to the less effective data sampling strategy underpinned by the simplified objective function without any high-order information. And directly matching these CNN features to obtain similarity values is not fast enough to be applicable in real-world applications. In this paper, we aim to reduce the computational burden of person re-identification by developing a fast re-identification framework whereby discriminative feature representations and hashing codes are learned effectively under a structured loss function which can automatically mine out meaningful hard examples, and exert boosting effect on parameter update during the back-propagation, leading to fast convergence and improved performance.

1.1. Motivation

To cope with ever-growing amounts of visual data, deep learning based hashing methods have been proposed to simultaneously learn similarity-preserved hashing functions and discriminative image representation via a deep architecture (Lai et al., 2015; Zhang et al., 2015; Zhao et al., 2015). Simply delving existing deep hashing approaches into a person re-identification system is not trivial due to the difficulty of generalizing these pre-trained models to match pedestrian images in disjoint views. Fine-tuning is a plausible way to make pre-trained models suitable to re-identification, however, to suit their models, training images are commonly divided into mini-batches, where each mini-batch contains a set of *randomly* sampled positive/negative pairs or triplets. Thus, a contrastive or triplet loss is computed from each mini-batch, and the networks try to minimize the loss function and update the parameters through back-propagation by using Stochastic Gradient Decent (SGD) (Wilson and Martinez, 2003). We remark that randomly sampled pairs/triplets carry little helpful information to SGD. For instance, many triplet units can easily satisfy the relative comparison constraint in a triplet loss function (Eq. (3)), resulting into a slow convergence rate in the training stage since many of them easily satisfy the constraint well and give nearly zero loss. Worse still, mini-batches with random samples may fail to obtain a stable solution or collapse into a local optimum if a contrastive/triplet loss function is optimized (Song et al., 2016). To this end, a suitable loss function is highly demanded to work well with SGD over mini-batches, which has desirable property of augmenting meaningful hard examples.

In this paper, we propose a deep hashing scheme based on CNNs to efficiently address the problem of person re-identification. To mitigate the undesirable effects caused by contrastive/triplet loss function, we propose a structured loss function by *actively* adding hard negative samples into mini-batches, leading to a structured deep hashing framework. The proposed structured loss can guide sub-gradient computing in SGD to have correct directions, and thus achieves a fast convergence in training. Meanwhile, similarity-preserving hashing functions are jointly learned to enable a fast re-identification system.

1.2. Our approach

One may easily generate a straightforward two-stage deep hashing strategy by firstly extracting CNN features from a pre-trained model (Krizhevsky et al., 2012), followed by performing learned hash functions (separate projection and quantization step) to convert such CNN features into binary codes. However, due to the *independence* of two stages, such a strategy cannot obtain optimal binary codes in terms of the incapability of characterizing the supervised information from training data *i,e.,*intra-personal variation and inter-personal difference. Instead, the two stages can boost each other to achieve much better performance, that is, the learned binary codes can guide the learning of useful CNN features, meanwhile CNN features can be helpful in learning semantically similarity-preserving hash codes. Motivated by this, we present a structured deep hashing architecture to *jointly* learn feature representations and hash codes for person re-identification. The overall framework is illustrated in Fig. 2. In our architecture, mini-batches contain all positive pairs for a particular pedestrian, meanwhile each positive pair (has a query image and its correct match image from

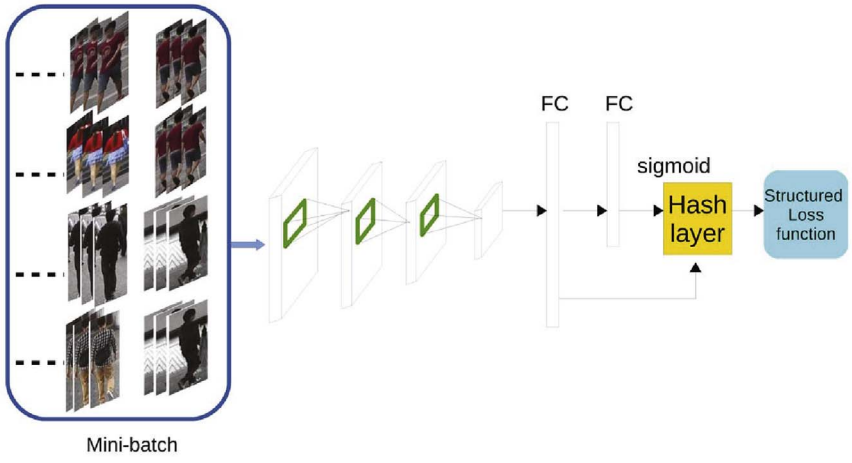

Fig. 2. Overview of our deep hashing framework for person re-identification. Our deep neural network takes a feed-forward, back-propagation strategy to learn features and hash codes simultaneously. During the feed-forward stage, the network performs inference from a mini-batch. The mini-batch is put through a stack of convolutional layers to generate nonlinear yet discriminative features, which are subsequently mapped to output feature vectors by fully-connected layers (FC). Meanwhile, a hash function layer is introduced atop of FC layer to learn hash codes that are optimized by a structured loss function to preserve their similarities/dissimilarities. In back-propagation, parameters are updated by computing their Stochastic Gradient Decent (SGD) w.r.t. the mini-batch.

Download English Version:

https://daneshyari.com/en/article/6937431

Download Persian Version:

https://daneshyari.com/article/6937431

Daneshyari.com