Pattern Recognition 78 (2018) 331–343

Contents lists available at ScienceDirect

Pattern Recognition

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

Deep unsupervised learning of visual similarities

Artsiom Sanakoyeu*, Miguel A. Bautista, Björn Ommer

Heidelberg Collaboratory for Image Processing and Interdisciplinary Center for Scientific Computing, Heidelberg University, Germany

ARTICLE INFO

Article history:
Received 3 February 2017
Revised 3 October 2017
Accepted 23 January 2018
Available online 31 January 2018

Keywords:
Visual similarity learning
Deep learning
Self-supervised learning
Human pose analysis
Object retrieval

ABSTRACT

Exemplar learning of visual similarities in an unsupervised manner is a problem of paramount importance to computer vision. In this context, however, the recent breakthrough in deep learning could not yet unfold its full potential. With only a single positive sample, a great imbalance between one positive and many negatives, and unreliable relationships between most samples, training of Convolutional Neural networks is impaired. In this paper we use weak estimates of local similarities and propose a single optimization problem to extract batches of samples with mutually consistent relations. Conflicting relations are distributed over different batches and similar samples are grouped into compact groups. Learning visual similarities is then framed as a sequence of categorization tasks. The CNN then consolidates transitivity relations within and between groups and learns a single representation for all samples without the need for labels. The proposed unsupervised approach has shown competitive performance on detailed posture analysis and object classification.

1. Introduction

Learning similarities in the visual domain plays a central role for numerous computer vision tasks which range across different levels of abstraction, from low-level image processing to high-level object recognition or human pose estimation. Similarities have been usually obtained as a result of category-level recognition, where categories and the similarities of all their samples to other classes are jointly modeled. However, the large intra-class variability of visual categories has recently spurred exemplar methods [1,2], which split the category-level model into simpler subtasks for each sample. Therefore, separate exemplar classifiers are trained by learning the similarities of individual exemplars against a large set of negatives. This paradigm of exemplar learning has been applied with successful results in problems like object recognition [1,3], instance retrieval [4,5], and grouping [6]. Learning visual similarities has been also of particular importance for posture analysis [7] and video parsing [8], where exploiting both the appearance [9] and the temporal domain [10] has proven useful.

Throughout the numerous methods for learning visual similarities, supervised techniques have been of particular interest in the computer vision field. These supervised techniques have therefore followed different formulations either as ranking [11], regression [12], and classification [8] problems. Furthermore, with the recent advent of Convolutional Neural Networks (CNN), two stream architectures [13] and ranking losses [14] have shown great improvements over similarities learned using hand-crafted features. Nevertheless, these performance improvements obtained by CNNs come at the cost of requiring millions of samples of supervised training data or at least the fine-tuning [9] on large labeled datasets such as PASCAL VOC. Even though the amount of accessible image data is growing at an ever increasing rate, supervised labeling of image similarities is extremely costly. In addition to the difficulty of labeling a similarity metric, not only similarities between images are important, but also between objects and their parts. Annotating the fine-grained similarities between all these entities is hopelessly complex, in particular for the large datasets typically used for training CNNs.

Unsupervised deep learning of similarities that does not require any labels for pre-training or fine-tuning is, therefore, of great interest to the vision community. This way we can utilize large image datasets without being limited by the need for costly manual annotations. However, CNNs for exemplar-based learning have been rare [15] due to limitations resulting from the widely used cross-entropy loss. The learning task suffers from only a single positive instance, it is highly unbalanced with many more negatives, and the relationships between samples are unknown, cf. Section 2. Consequentially, stochastic gradient descend (SGD) gets corrupted and has a bias towards negatives, thus forfeiting the benefits of deep learning.

Our approach overcomes this shortcoming by updating similarities and CNN parameters. Normally, at the beginning, only a few

* Corresponding author.
E-mail addresses: artsiom.sanakoyeu@iwr.uni-heidelberg.de (A. Sanakoyeu), miguel.bautista@iwr.uni-heidelberg.de (M.A. Bautista), bjoern.ommer@iwr.uni-heidelberg.de (B. Ommer).

https://doi.org/10.1016/j.patcog.2018.01.036
0031-3203/

local estimates of similarities are easily available (i.e. pairs of samples that are highly similar (near duplicates) or that are very distant). Nevertheless, most of the initial similarities are unknown, or non-transitive, i.e. mutually contradicting. To nevertheless define balanced classification tasks suited for CNN training, we formulate an optimization problem that builds training batches for the CNN by selecting groups of compact cliques so that all cliques in a batch are mutually distant. Thus for all samples of a batch (dis-)similarity is defined—they either belong to the same compact clique or are far away and belong to different cliques. However, pairs of samples with no reliable similarities end up in different batches so they do not yield false training signal for SGD. Classifying if a sample belongs to a clique serves as a pretext task for learning exemplar similarity. Training the network then implicitly reconciles the transitivity relations between samples in different batches. Thus, the learned CNN representations impute similarities that were initially unavailable and generalize them to unseen data. Furthermore, to incorporate temporal context in our model, we introduce a Local Temporal Pooling strategy that models how similarities between exemplars change over short periods of time.

In the experimental evaluation, the proposed approach significantly improves over state-of-the-art approaches for posture analysis and retrieval by learning a general feature representation for a human pose that can be transferred across datasets.

1.1. Related work

The Exemplar Support Vector Machine (Exemplar-SVM) has been one of the driving methods for exemplar-based learning [1]. Each Exemplar-SVM classifier is defined by a single positive instance and a large set of negatives. To improve performance, Exemplar-SVMs require several rounds of hard negative mining, increasing greatly the computational cost of this approach. To circumvent this high computational cost, [6] proposes to train Linear Discriminant Analysis (LDA) over Histogram of Gradient (HOG) features [6]. LDA whitened HOG features with the common covariance matrix estimated for all the exemplars removes correlations between the HOG features, which tend to amplify the background of the image.

Recently, several CNN approaches have been proposed for supervised similarity learning using either pairs [13], or triplets [14] of images. However, supervised formulations for learning similarities require that the supervisory information scales quadratically for pairs of images, or cubically for triplets. This results in very large training times.

The literature on exemplar-based learning in CNNs is very scarce. In [15] the authors of Exemplar-CNN tackle the problem of unsupervised feature learning. A patch-based categorization problem is designed by randomly extracting patch for each image in the training set and defining it as a surrogate class. Hence, since this approach does not take into account (dis-)similarities between exemplars, it fails to model their transitivity relationships, resulting in poor performances (see Section 3.1).

Furthermore, recent works [9,10,16,17] showed that temporal information in videos and spatial context information in images can be utilized as a convenient supervisory signal for learning feature representation with CNNs. However, the computational cost of the training algorithm is enormous since the approach in [9] needs to tackle all possible pair-wise image relationships requiring a training set that scales quadratically with the number of samples. On [16] authors leverage time-contrastive loss to learn representations leveraging the temporal structure of the data. However, this approach is limited to video sequences without repetitions since the method is based on the assumption of mutual independence of time segments. In contrast, our approach leverages the relationship information between compact cliques, framing similarity learning as a multi-class classification problem. As each training batch contains mutually distinct cliques the computational cost of the training algorithm is greatly decreased.

2. Methodology

In this section, we show how a CNN can be employed for learning similarities between all pairs of a large number of exemplars. In particular, the idiosyncrasies of exemplar learning have made it difficult to unravel its full capabilities in CNNs. First, deep learning is extremely data hungry, which conflicts with having a single positive exemplar for training, we now abbreviate this setup as 1-sample CNN. This 1-sample setup then faces several issues. *(i)* The within-class variance of an individual exemplar cannot be modeled. *(ii)* The ratio of one exemplar and many negatives is highly imbalanced so that the cross-entropy loss over SGD batches overfits against the negatives. *(iii)* An SGD batch for training a CNN on multiple exemplars can contain arbitrarily similar samples with different label (the different exemplars may be similar or dissimilar), resulting in label inconsistencies. *(iv)* Provided the single training sample, exemplar learning cannot exploit the temporal context of training data, if available.

The methodology proposed in this paper overcomes this issues as follows. In Section 2.2 we discuss why simply appending an exemplar with its nearest neighbors and data augmentation (similar in spirit to the Clustered Exemplar-SVM [18], which we abbreviate as NN-CNN) is not sufficient to address *(i)*. Section 2.3 deals with *(ii)* and *(iii)* by generating batches of cliques that maximize the intra-clique similarity while minimizing inter-clique similarity. In addition, Section 2.5 shows how to exploit temporal information to further impose structure on the learned similarities by using a temporal average pooling.

To show the effectiveness of the proposed method we give empirical proof by training CNNs following both 1-sample CNN and NN-CNN training protocols. Fig. 1(a) shows the average ROC curve for posture retrieval in the Olympic Sports dataset [19] (refer to Section 3.1 for further details) for 1-sample CNN, NN-CNN and the proposed method, which clearly outperforms both exemplar based strategies. In addition, Fig. 1(b)–(d) shows an excerpt of the similarity matrix learned for each method. It becomes evident that the proposed approach captures more detailed similarity structures, e.g., the diagonal structures correspond to repetitions of the same gait cycle within a long jump.

2.1. Initialization

In the previous section, we have shown the shortcomings of exemplar-based training of CNNs. The key obstacle is the discrepancy between the single positive sample used in exemplar learning and the large amounts of data needed to train deep CNNs. Therefore, given a single exemplar $\mathbf{d}_i$ we attempt to find an initial number of related samples to enable the training of a CNN which further improves the similarities between exemplars. To obtain this initial group of related samples we employ LDA whitened HOG [6], which is a fundamental and computationally efficient approach to estimate similarities s_{ij} between large numbers of samples. Moreover, since they constitute a view-based approach, HOG features are viewpoint and rotation variant, which is therefore beneficial for pose estimation in 2D. We define $s_{ij} = s(\phi(\mathbf{d}_i), \phi(\mathbf{d}_j)) = \phi(\mathbf{d}_i)^\top \phi(\mathbf{d}_j)$, where $\phi(\mathbf{d}_i)$ is the whitened HOG descriptor of the exemplar and $\mathbf{S}' = (s_{ij}) \in \mathbb{R}^{N \times N}$ is the resulting kernel matrix. The nearest neighbor of the sample i is the sample j which maximizes s_{ij}.

As can be seen from Fig. 4(b) most of these similarities are evidently unreliable and, thus, the majority of samples cannot be