Pattern Recognition 80 (2018) 94–108

Contents lists available at ScienceDirect

Pattern Recognition

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

Curriculum learning of visual attribute clusters for multi-task classification

Nikolaos Sarafianos[a,*], Theodoros Giannakopoulos[b], Christophoros Nikou[c], Ioannis A. Kakadiaris[a]

[a] Computational Biomedicine Lab, Department of Computer Science, University of Houston, 4800 Calhoun Rd. Houston, TX 77004, USA
[b] National Center of Scientific Research Demokritos, Athens GR 15310, Greece
[c] Department of Computer Science and Engineering, University of Ioannina, Ioannina GR 45110, Greece

ARTICLE INFO

Article history:
Received 14 September 2017
Revised 19 January 2018
Accepted 25 February 2018
Available online 5 March 2018

Keywords:
Curriculum learning
Multi-task classification
Visual attributes

ABSTRACT

Visual attributes, from simple objects (e.g., backpacks, hats) to soft-biometrics (e.g., gender, height, clothing) have proven to be a powerful representational approach for many applications such as image description and human identification. In this paper, we introduce a novel method to combine the advantages of both multi-task and curriculum learning in a visual attribute classification framework. Individual tasks are grouped after performing hierarchical clustering based on their correlation. The clusters of tasks are learned in a curriculum learning setup by transferring knowledge between clusters. The learning process within each cluster is performed in a multi-task classification setup. By leveraging the acquired knowledge, we speed-up the process and improve performance. We demonstrate the effectiveness of our method via ablation studies and a detailed analysis of the covariates, on a variety of publicly available datasets of humans standing with their full-body visible. Extensive experimentation has proven that the proposed approach boosts the performance by 4%–10%.

1. Introduction

Vision as reception. Vision as reflection. Vision as projection.
–Bill Viola, note 1986

When we are interested in providing a description of an object or a human, we tend to use visual attributes to accomplish this task. For example, a laptop can have a wide screen, a silver color, and a brand logo, whereas a human can be tall, female, wearing a blue t-shirt and carrying a backpack. Visual attributes in computer vision are equivalent to the adjectives in our speech. We rely on visual attributes since (i) they enhance our understanding by creating an image in our head of what this object or human looks like; (ii) they narrow down the possible related results when we want to search for a product online or when we want to provide a suspect description; (iii) they can be composed in different ways to create descriptions; (iv) they generalize well as with some fine-tuning they can be applied to recognize objects for different tasks; and (v) they are a meaningful semantic representation of objects or humans that can be understood by both computers and humans. However, effectively predicting the corresponding visual attributes of a human given an image remains a challenging task [1]. In real-life scenarios, images might be of low-resolution, humans might be partially occluded in cluttered scenes, or there might be significant pose variations.

Estimating the visual attributes of humans is an important computer vision problem with applications ranging from finding missing children to virtual reality. When a child goes missing or the police is looking for a suspect, a short description is usually provided that comprises such attributes (for example, tall white male, with a black shirt wearing a hat and carrying a backpack). Thus, if we could efficiently identify which images or videos contain images of humans with such characteristics we could potentially reduce dramatically the labor and the time required to identify them [2]. Another interesting application is the 3D reconstruction of the human body in virtual reality [3]. If we have such attribute information we can facilitate the reconstruction by providing the necessary priors. For example it is easier to reconstruct accurately the body shape of a human if we already know that it is a tall male with shorts and sunglasses than if no information is provided.

In this work, we introduce CILICIA (CurrIculum Learning multItask ClassIfication Attributes) to address the problem of visual attribute classification from images of standing humans. Instead of using low-level representations, which would require extracting hand-crafted features, we propose a deep learning method to

* Corresponding author.
E-mail address: nsarafia@central.uh.edu (N. Sarafianos).

https://doi.org/10.1016/j.patcog.2018.02.028
0031-3203/

solve multiple binary classification tasks. CILICIA differentiates itself from the literature as: (i) it performs end-to-end learning by feeding a single ConvNet with the entire image of a human without making any assumptions about predefined connection between body parts and image regions; and (ii) it exploits the advantages of both multi-task and curriculum learning. Tasks are split into groups based on their labels' cross-correlation using hierarchical agglomerative clustering. The groups of tasks are learned in a curriculum learning scenario, starting with the one with the highest within-group cross-correlation and moving to the less correlated ones by transferring knowledge from the former to the latter. The tasks in each group are learned in a typical multi-task classification setup. Parts of this publication appear in our previous work [4]. However, in this work we have:

- Proposed an effective method to obtain the groups of tasks using hierarchical agglomerative clustering, which can be of any number and not just two groups (strongly/weakly correlated).
- Conducted additional experiments to analyze the covariates of the proposed approach.
- Benchmarked our method in an additional dataset.
- Demonstrated the efficacy and robustness of our method by performing ablation studies in Section 5.

When Vapnik and Vashist introduced the learning using privileged information (LUPI) paradigm [5], they drew inspiration from human learning. They observed how significant the role of an intelligent teacher was in the learning process of a student, and proposed a machine learning framework to imitate this process. Employing privileged information from an intelligent teacher at training time has recently received significant attention from the scientific community with remarkable results in areas ranging from object recognition [6–9] to biometrics [10–12].

Our work also draws inspiration from the way students learn in class. First, students find it difficult to learn all tasks at once. It is usually easier for them to acquire some basic knowledge first, and then build on top of that, by learning more complicated concepts. This can be achieved by learning in a hierarchical manner, which is commonly employed in the literature [13–15], or with a curriculum strategy. Curriculum learning [16–18] (presenting easier examples before more complicated and learning tasks sequentially, instead of all at the same time) imitates this learning process. It has the advantage of exploiting prior knowledge to improve subsequent classification tasks but it cannot scale up to many tasks since each subsequent task has to be learned individually. However, to maximize students' understanding a curriculum might not be sufficient by itself. Students also need a teaching paradigm that can guide their learning process, especially when the task to be learned is challenging. The teaching paradigm in our method is the split of visual attribute classification tasks that need to be learned by performing hierarchical agglomerative clustering. In that way, we exploit the advantages of both multi-task and curriculum learning. First, the ConvNet learns the group of tasks with the strongest intra cross-correlation in a multi-task learning setup, and once this process is completed, the weights of the respective tasks are used as an initialization for the more diverse tasks. During the training of the more diverse tasks, the prior knowledge obtained is leveraged to improve the classification performance. An illustrative example of our method is depicted in Fig. 1. Note that the proposed learning paradigm is not tied visual attribute classification domain and can be extended to other applications such as object recognition or [19] and domain adaptation [20].

In summary, this paper has the following contributions. First, we introduce CILICIA, a novel method of exploiting the advantages of both multi-task and curriculum learning by splitting tasks into groups by performing hierarchical agglomerative clustering. The tasks of each subgroup are learned in a joint manner. Thus, the proposed method learns better than learning all the tasks in a typical multi-task learning setup and converges faster than learning tasks one at a time. Second, we propose a scheme of transferring knowledge between the groups of tasks which speeds up the convergence and increases the performance. We performed extensive evaluations in three datasets of humans standing and achieved state-of-the-art results in all three of them.

The remainder of the paper is organized as follows: in Section 2, a review of the related work in visual attributes, curriculum learning, and transfer learning is presented. Section 3 presents CILICIA, the proposed curriculum learning approach for multi-task classification of clusters of visual attributes. In Section 4, experimental results are reported, a detailed analysis of covariates is provided, and a discussion about the performance and the limitations of the proposed approach is offered. Finally, conclusions are drawn in Section 6.

2. Related work

Visual attributes classification: The first to investigate the power of visual attributes were Ferrari and Zisserman [21]. They used low-level features and a probabilistic generative model to learn attributes of different types (e.g., appearance, shape, patterns) and segment them in an image. Kumar et al. [22] proposed an automatic method to perform face verification and image search. They first extracted and compared "high-level" visual features, or traits, of a face image that are insensitive to pose, illumination, expression, and other imaging conditions, and then trained classifiers for describable facial visual attributes (e.g., gender, race, and eyewear). A verification classifier on these outputs is finally trained to perform face verification. In the work of Scheirer et al. [23], raw attribute scores are calibrated to a multi-attribute space where each normalized value approximates the probability of that attribute appearing in the input image. This normalized multi-attribute space allows a uniform interpretation of the attributes to perform tasks such as face retrieval or attribute-based similarity search. Finally, attribute selection approaches have been introduced [24–26] which select attributes based on specific criteria (e.g., entropy). Zheng et al. [26] formulated attribute selection as a submodular optimization problem [27] and defined a novel submodular objective function.

Following the deep learning renaissance in 2012, several papers [28–32] have addressed the visual attribute classification problem using ConvNets. Part-based methods decompose the image to parts and train separate networks which are then combined at a feature level before the classification step. They tend to perform well since they take advantage of spatial information (e.g., patches that correspond to the upper body can better predict the t-shirt color than others that correspond to other body parts). Zhang et al. [33] proposed an attribute classification method which combines part-based models in the form of poselets [34], and deep learning by training pose-normalized ConvNets. Gkioxari et al. [35] proposed a deep version of poselets to detect human body parts which were then employed to perform action and attribute classification. Zhu et al. [36] introduced a method for pedestrian attribute classification. They proposed a ConvNet architecture comprising 15 separate subnetworks (i.e., one for each task) which are fed with images of different body parts to learn jointly the visual attributes. However, their method assumes that there is a pre-defined connection between parts and attributes and that all tasks depend on each other and thus, learning them jointly will be beneficial. Additionally, they trained the whole ConvNet end-to-end despite the fact that the size of the training dataset used was only 632 images. Based on our experiments, the only way to avoid heavy overfitting in datasets of that size is by employing a pre-trained network along with fine-tuning of some layers. Recycling pre-trained deep