Neurocomputing ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Neurocomputing

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

Annotation modification for fine-grained visual recognition

Changzhi Luo [a], Zhijun Meng [b,*], Jiashi Feng [c], Bingbing Ni [d], Meng Wang [a]

[a] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China
[b] School of Aeronautic Science and Engineering, Beihang University, Beijing 100191, China
[c] Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576, Singapore
[d] Department of Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 15 March 2016
Received in revised form
12 May 2016
Accepted 17 May 2016

Keywords:
Query modification
Annotation modification
Fine-grained visual recognition
Active set
Ranking loss

ABSTRACT

Query modification is an intensively studied and widely used technique in information retrieval, for it helps better understand the intention of the users. In this work, we introduce this idea into fine-grained visual recognition, which is important to ambiguous queries in image retrieval task. Unlike most existing works, which incorporate information about object bounding boxes or parts for extracting discriminative local features, we propose a novel approach from a new viewpoint to solve the fine-grained recognition problem, namely annotation modification. The proposed approach fully exploits the inter-class ambiguity (which is generally regarded as noise) to form active sets of annotations for boosting the fine-grained visual recognition. Specifically, it first obtains some most confusing classes of each image through an easy-to-evaluate classifier, and then modify the annotation of each image using the active set of annotations. To handle the modified annotations, a novel ranking based loss function is further designed to learn effective classification models. We evaluate the proposed approach on three popular fine-grained image datasets (i.e., Oxford-IIIT Pets, Flower-102 and CUB200-2011), and the experimental results clearly demonstrate its effectiveness.

1. Introduction

With the development of information retrieval technology, more and more tough query issues have emerged. For example, a user who wants to search a specified species of bird may get a list of birds with similar appearance but belong to different species. Such fine-grained query issue poses greater challenge for the image retrieval systems. Query modification [1–5] is an effective way to bridge the gap between the users' intention and the image content in the content-based image retrieval systems [6–13]. However, for the systems which adopt the input of images, query modification cannot be utilized. In this work, we adapt the idea of query modification, and proposed an approach named annotation modification for solving the fine-grained query issue.

Fine-grained query issue is also known as fine-grained visual recognition, which aims at distinguishing categories with subtle differences (e.g., recognizing different species of birds). It has attracted increasing research interest during the past few years [14–22]. However, recognizing fine-grained objects is still very challenging even for the cutting-edge classification techniques, e.g., Convolutional Neural Networks (CNN) [23–25]. The major difficulty arguably lies in the high similarity among samples from different fine-grained categories as shown in Fig. 1. In addition, large variance of object poses and viewpoints within the same category further increases the difficulty of such a task.

A popular pipeline for solving the fine-grained recognition problem is to localize the foreground objects as well as their discriminative parts at first, and then classify the objects based on the visual features from foreground and parts. For example, to classify birds of different species, one can first detect the birds in the images; and then localize their discriminative parts (such as heads, bodies and legs); finally combine these cues to conduct classification. However, such kind of methods heavily relies on the availability of annotations with extensive details, e.g., object bounding boxes and parts, whose collection however is usually very tedious and expensive. Requiring so much manually labeled information also hinders their application in the real world.

Benefiting from the rapid progress of object detection techniques (e.g., selective search [26], edge boxes [27], BING [28]), one can more efficiently discover and localize informative object parts in an image. As demonstrated in [18,22], using the cues from object proposals indeed improves the performance of fine-grained visual recognition. However, object proposals can only provide coarse localization of objects, and the noise from background or inaccurate localization will inevitably harm the recognition performance. Therefore, only carefully selected object proposals can

* Corresponding author.
E-mail address: mjz.beihang@gmail.com (Z. Meng).

http://dx.doi.org/10.1016/j.neucom.2016.05.089
0925-2312/

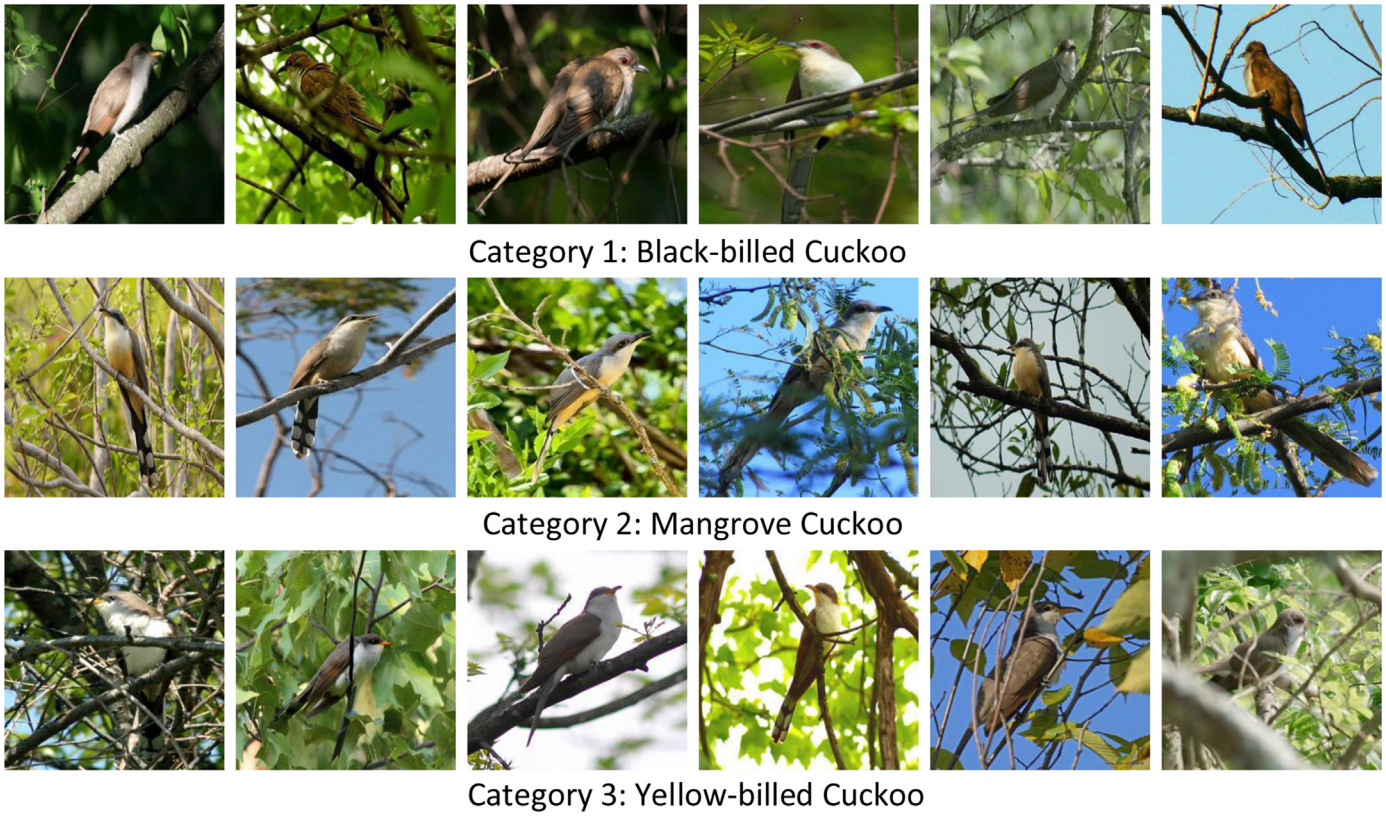

Fig. 1. Sample images from CUB200-2011 dataset. The major difficulty of fine-grained visual recognition lies in the high similarity among samples from different categories. In this figure, totally three categories are showed, however, even human beings cannot clearly distinguish them.

help localize the critical differences for different categories, which are rather difficult to obtain.

In this work, we provide an alternative approach to the fine-grained visual recognition task. Our approach is motivated by interactive search [29–32] in information retrieval, i.e., allowing for multiple queries instead of one will produce more promising search results. The intuition in this work is that, for fine-grained visual recognition, the challenge mainly lies in distinguishing those highly similar categories, thus when handling those categories, allowing the classification model to make multiple times of predictions is promising for increasing the probability of successfully predicting the correct category. Thus, different from existing works relying on side information from complicated object detectors, we propose to exploit useful yet commonly ignored information hidden in the category annotation space. To verify our assumption, we conduct an experimental study on the top-k accuracy of the VGG-Net [25] on the CUB200-2011 dataset and plot the results in Fig. 2. One can observe from the figure that the recognition accuracy indeed increases sharply with the expansion of the candidate predictions. This demonstrates that compared with predicting a single category, it will be easier for a deep network to hit the category of ground truth within top-k predictions.

Motivated by this observation, we propose our annotation modification approach, as our approach is built upon the networks with active sets of annotations, we named it NASA. This idea is similar to query modification [1–4], which is widely used in information retrieval. In a searching task, the search engine inevitably encounters ambiguous queries, and thus it needs relevance feedback or other query modification methods for returning promising search results. The ambiguous issue also lies in fine-grained visual recognition task, and we introduce the solution into this work. We first obtain an initial active set of annotations for each image

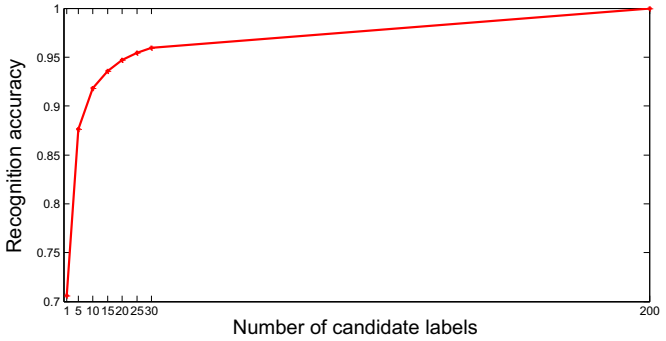

Fig. 2. Top-k accuracy using VGG-Net on the CUB200-2011 dataset. The plot shows how the accuracy of a deep network varies along with different numbers of candidate predictions. It can be seen that the accuracy increases sharply with the expansion of the candidate annotation set, which demonstrates that VGG-Net can predict the ground truth of most examples from the top-k candidates with $k < 30$.

through applying a multi-class linear SVM (or any other classifier that can be trained efficiently) on pre-extracted CNN features of the image. The active set is formed by the categories with top largest confidence scores, which indicates that the classes in the active set are confusing. Then, we propose to employ a ranking based loss function to train the network with the active sets. Specifically, NASA explicitly assigns greater penalty to the prediction on the category of ground truth if the rank of its confidence score is low and penalizes more on the confused classes (other classes in the active set) if their confidence scores rank wrongly high. By incorporating active sets and such a novel ranking loss function, NASA improves the performance on discriminating the fine-grained categories with subtle differences significantly.

In summary, we make the following contributions in this work in solving the fine-grained recognition problem. (1) We propose to

Please cite this article as: C. Luo, et al., Annotation modification for fine-grained visual recognition, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.05.089