An improved human-in-the-loop model for fine-grained object recognition with batch-based question answering

by

Vyshnavi Gutta, Narendra Babu Unnam, P Krishna Reddy

in

CODS-COMAD

Report No: IIIT/TR/2020/-1

Centre for Data Engineering
International Institute of Information Technology
Hyderabad - 500 032, INDIA
January 2020

An improved human-in-the-loop model for fine-grained object recognition with batch-based question answering

Vyshnavi Gutta
IIIT Hyderabad, India
vyshnavi.gutta@research.iiit.ac.in

Narendra Babu Unnam
IIIT Hyderabad, India
narendra.unnam@research.iiit.ac.in

P.Krishna Reddy
IIIT Hyderabad, India
pkreddy@iiit.ac.in

Abstract

Fine-grained object recognition refers to a subordinate level of object the recognition such as recognition of bird species and car models. It has become crucial for recognition of previously unknown classes. While fine-grained object recognition has seen unprecedented progress with the advent of neural networks, many of the existing works are cost-sensitive as they are acutely picture-dependent and fail without the adequate number of quality pictures. Efforts have been made in the literature for a picture-independent recognition with hybrid human-computer recognition methods via single question answering with a human-in-the-loop. To this end, we propose an improved *batch-based question answering* method for making the recognition efficient and picture-independent. When pictures are unavailable, at each time-step, the proposed method mines N binary cluster-centric local questions to pose to a human-in-the-loop and incorporates the responses received into the model. After a preset number of time-steps, the most probable class of the target object is returned as the final prediction. When pictures are available, our model facilitates the plug-in of computer vision algorithms into the framework for better performance. Experiments on three challenging datasets show significant performance improvement with respect to accuracy and computation time as compared to the existing schemes.

CCS Concepts

• Classification → Fine grained recognition; **• Feature extraction →** Batch question answering.

Keywords

Fine-grained object recognition, human-in-the-loop, local attribute mining, batch-based question answering

ACM Reference Format:
Vyshnavi Gutta, Narendra Babu Unnam, and P.Krishna Reddy. 2018. An improved human-in-the-loop model for fine-grained object recognition with batch-based question answering. In *CODS-COMAD 2020: ACM India Joint International Conference on Data Science & Management of Data, 2020, Hyderabad, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

1 Introduction

While object recognition deals with the recognition of several objects belonging to broader entry levels like birds, humans, chairs, etc, fine-grained object recognition aims to distinguish the objects of subordinate categories that belong to the same entry-level object category, e.g., recognition of various species of birds. Its applications are numerous. Beyond simply being able to describe the world in more detail, fine-grained object recognition can be used for improved scene understanding, studying society and analyzing biodiversity.

Notably, fine-grained object recognition is a difficult task due to the small inter-class variance between the objects. Nevertheless, enormous progress has been made in fine-grained recognition in recent years. The existing fine-grained recognition methodologies can be categorized into three major recognition modes: Image-only based Recognition (IR), Question-answering based Recognition (QR), and Image and Question-answering based Recognition (IQR).

The most widely used recognition mode is *Image-only based Recognition (IR)*, wherein recognition is achieved by categorizing the image of the target object. Ever since the rise of Convolutional Neural Network (CNN) architectures for image classification, the accuracy of IR has dramatically improved and many CNN-based approaches have been proposed [6–8, 12, 16]. However, IR is highly resource-intensive as it requires a *large* number of images with *good* quality. As a result, IR is completely *picture-dependent*, i.e, without proper pictures, the performance of IR quickly collapses. Furthermore, it incurs huge monetary support for the required picture acquisition. (In this paper, we use the terms *image* and *picture* interchangeably.) This issue of monitory requirement becomes even more poignant in fine-grained domains due to their stricter picture requirements in terms of both quality and quantity. As a result, IR is *cost-sensitive* and falls short in producing *desirable* accuracies in *budget-restricted* domains. Moreover, the task of picture acquisition itself becomes very intractable in fine-grained domains.

In contrast to IR, *Question-answering based Recognition (QR)* does object recognition by utilizing the attribute (feature) information of the object collected through interaction with a user a.k.a. human-in-the-loop. (In this paper we use the terms *user* and *human-in-the-loop* interchangeably.) It achieves recognition by exploiting the notion that every visual object is characterized by its visual attributes. In QR, questions on the presence/absence of relevant attributes of a visible target object are posed to a user, whose perceived responses on the object to the questions collectively make recognition possible. But, as typical fine-grained objects have hundreds of such attributes, having to rely solely on the user for their identification is inefficient and cumbersome. Notably, QR does not use pictures for recognition. That is, unlike IR, QR is not cost-sensitive.

CODS-COMAD 2020, January 5-7, 2020, Hyderabad, India

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Research efforts have been made in the literature [3, 5, 9, 13, 17] to improve object recognition by exploiting the merits of IR and QR with a hybrid recognition which we refer to as *Image and Question-answering based Recognition (IQR)*. The idea behind IQR is to combinedly leverage a machine's extensive learning capability and a human's excellent visual capability for better recognition of a test object. The approaches proposed in [9, 17] incorporate IQR by utilizing the relative responses received for an attribute on multiple images to distinguish between the fine-grained objects. Likewise, an approach to pose discriminative features is proposed in [5] for recognition. The approach in [13] jointly leverages part-based click information and binary question answering for recognition. However, the above works [5, 9, 13, 17] are cost-sensitive as they are still picture-dependent.

In [2, 3], a picture-independent framework under IQR was introduced for object recognition in which a single question with binary [3] and multiple-choice [2] is posed to the user at each time-step. They use the information gain criterion along with the prior predictions of computer vision methods on the target image (if available) to decide the best question to pose to the user at each time-step. However, the approaches in [2, 3] are highly time-consuming as the system has to compute information gain for every attribute to select the best question at each time-step. As a result, the approaches would lead to high user waiting time.

In this paper, we propose an improved approach for fine-grained object recognition which we refer to as *Recognition via Image and Batch-based Local question answering (RIBQ)*. Instead of asking a single question at each time-step for object recognition as in existing approaches, there is an opportunity to improve the recognition performance by asking multiple questions at each time-step. In the approaches based on single question answering, only a single discriminative attribute is identified to distinguish the probable classes. Whereas in the case of the proposed batch-based question answering, multiple discriminative attributes are identified to distinguish the probable classes. As a result, object recognition is achieved quickly and effectively.

For extracting the potential batch of questions to be posed at each time-step, we propose a novel cluster-based local question mining method. At each time-step, the proposed method groups the probable classes into N non-overlapping clusters. From each cluster, a *cluster-centric local attribute*, which is the attribute whose presence/absence is exclusively pre-dominant in that cluster, is mined. The set of mined attributes are posed as questions to the user. After a preset number of time-steps, the most probable class based on the responses received is returned as the final prediction of the test object. When labelled images of target objects are available, we could plug-in vision's probabilistic class predictions on the object's test image into the proposed framework for more accurate recognition. We conducted an extensive performance study on three different datasets and demonstrate that the proposed approach improves both accuracy and computation time significantly w.r.t. the existing approaches.

Notably, with cluster-centric local question mining, the proposed approach mitigates the computational overhead involved in the information-gain based question mining in [2, 3] and is significantly faster. Since the proposed approach employs multiple discriminative questions at each time-step, it also improves the recognition quality. Furthermore, the proposed approach is flexible as it works as QR when pictures are unavailable and as IQR when pictures are available for better recognition. Thus the proposed method is also cost-effective as it aims to make the recognition effective regardless of the availability of pictures. Thus, the proposed method achieves *fast, accurate and cost-effective recognition* under *limited resource environments* (i.e., when pictures are unavailable.) With pictures, it facilitates the easy plug-in of vision's probabilistic class-estimates into the framework for more accurate recognition.

The main contributions of this paper are three-fold:

(1) We introduce a batch-based question answering model (*RIBQ*) for fine-grained object recognition.
(2) We present an efficient dynamic cluster-centric local question mining approach.
(3) We have demonstrated that the proposed model improves both accuracy and computation time significantly w.r.t. the existing approaches by conducting extensive experiments on three different datasets.

The remainder of this paper is organized as follows. In the next section, we present the related work. In Section 3, we present the background. We present the proposed approach in Section 4. The experimental results and conclusions presented in Section 5 and Section 6 respectively.

2 Related Work

Regarding Image-only based Recognition (IR), several works [6, 7, 12, 16] utilize the visual attributes of objects to carry out the recognition process. In [6], attributes are used for semantic knowledge transfer between the known classes and the unseen classes through direct and indirect attribute predictions. In [12], inherent relativity between the attributes is exploited for mining information using deep neural networks, by adding a ranking layer. In [16], discriminative attributes are extracted from the class-specific discriminative patches. In [7], localized image features represented by attribute semantics are utilized for mining discriminative information.

Regarding Image and question answering based Recognition (IQR), research works [2–5, 9–11, 13, 17] integrate computer vision and human input to make the best of the both worlds. A picture-independent framework for object recognition is proposed in [3, 10], which asks an informative binary attribute as a binary question at each time-step. The attribute is extracted using the information gain criterion. If images are available, the vision's prior probabilistic class predictions on the object's test image are plugged into the framework for better recognition. In [2], an approach is proposed which poses a multiple-choice question at each time-step instead of posing a single binary question as in [3, 10]. However, the applicability of this approach in [2] is limited as the multiple-choice questions are not available for all the datasets and it takes extra human labour to prepare them. Note that, the proposed approach is different to that in [2] as the proposed approach poses multiple binary questions at each time-step rather than multiple-choice questions. In [13], part-based click information and binary question answering are jointly leveraged for recognition. In [4, 11], object annotation is carried out jointly by vision-based methods and human input by employing a Markov's decision process with the aid of reinforcement learning. In [5], discriminative bubbles/features are mined from the image using the user's responses. In [9], relative responses of the attributes are leveraged by learning a ranking function per attribute. In [17], local learning is done through image comparisons.

To summarize, few research efforts have been made to make the recognition picture-independent thereby cost-insensitive. As a part of this, picture-independent human-machine frameworks have been proposed which employ vision (if pictures are available) and binary [3] or multiple-choice [2] question answering for recognition. However, the information gain metric used in [2, 3] computes the information gain on all the attributes for selecting the best question to pose at each time-step. As a result, the interim user waiting period between successive time-steps is high. The proposed approach uses batch-based local question answering which facilitates faster and better knowledge acquisition than the preceding approaches.

3 Background

In this section, after explaining about attribute, object, class and class attribute vector, we discuss the picture-independent recognition framework used in [2, 3, 10].

3.1 Definitions

Definition 3.1. Attribute: A property of an object is called an attribute, if a human has the ability to decide whether the property is present or not for the object [6]. For example, *red color neck* is an attribute of an object *bird.* Its corresponding question to pose to the user would be *Is neck color red?.*

Definition 3.2. Object and *Class:* An object is a physical entity with pre-defined attributes. A class represents a collection of objects (instances) with the same attributes. For example, the class car represents all the cars (objects) in the real world.

Definition 3.3. Class-attribute vector of class c_k ($CAV(c_k)$) and Class-attribute vectors (CAV): Class-attribute vector of a class depicts the relation between the class and all the attributes. Given a set $C = \{c_1, \ldots, c_m\}$ of all m classes and set $Q = \{q_1, q_2, \ldots q_n\}$ of all n attributes, we denote $CAV(c_k)$ as $< v(q_{1k}), v(q_{2k}), \ldots, v(q_{nk}) >$. Here, $v(q_{ik})$ is a real value from the interval [0-1] which indicates the degree of presence of the attribute q_i in the class c_k. For example, the value 0/1 for $v(q_{ik})$ indicates the complete absence/presence of q_i in c_k. CAV is the set of class-attribute vectors of all classes, i,e., CAV=$\{CAV(c_1), CAV(c_2), \ldots, CAV(c_m)\}$.

3.2 Picture-independent recognition framework

Given a domain's characteristic attributes and its class-attribute vectors, the recognition process of a test object is done in two stages. The stages repeat for a preset number of time-steps at the end of which the class with the highest probability estimate is returned as the final prediction of the target object. The details of each stage are as follows:

(1) **Question mining stage**: This stage mines the relevant visual attribute(s) to be posed as question(s) to a user at a time-step t based on the responses received till t and the class-attribute vectors. The approaches in [2, 3, 10] mine questions using information-gain. They compute information gain for each attribute and pose the question corresponding to the attribute with the maximum gain. The approaches in [3, 10] mine a binary question whereas the approach in [2] mines a multiple-choice question at t.
(2) **Response modelling stage**: This stage models the collected user response/s at time-step t to the questions as perceived by the user on the test object for processing in the next time-step. This includes re-computing the class-probability estimates of all the classes.

When pictures are available, the vision's probabilistic class prediction on the test image of the object is plugged into the framework for faster and more accurate recognition.

4 Proposed model

In this section, we first explain the basic idea and then present the proposed approach.

4.1 Basic idea

The basic idea of the proposed model is to improve the performance of object recognition by posing multiple discriminative questions in a batch at each time-step to the user. To this end, we employ a clustering-based approach to identify multiple discriminative attributes and then pose them as questions for distinguishing the probable classes. It can be noted that, given any collection of data points, the clustering process groups the data points into multiple clusters, such that members of the same cluster are similar and the members of the different clusters are dissimilar. So, by extracting the most predominant attribute from each cluster, it is possible to extract multiple discriminative attributes. As a result, object recognition can be realized effectively in terms of both accuracy and time.

So in the proposed approach, at each time-step, the probable classes are first grouped into N clusters. From each cluster, a potential attribute is mined using the proposed concepts of *Cluster-centric local attribute* and *locality degree of an attribute.* For a given cluster X, we use the term *Cluster-centric local attribute of* X to denote the attribute whose presence/absence is exclusively predominant in X. For an attribute q_i and a cluster X, we use the term *locality degree of* q_i *in* X for depicting the degree of q_i's exclusive presence/absence in X. Using these concepts, we mine the questions to pose to the user by computing the *locality degree of each attribute in each cluster* and identify those attributes which give the maximum locality degree in a particular cluster in comparison to other clusters and attributes as the potential cluster-centric local attributes. The mined attributes are then posed as questions to the user at each time-step whose responses to the questions collectively make recognition possible.

4.2 Proposed approach

The proposed approach uses picture-independent recognition framework (Section 3.2) for object recognition. We refer the proposed approach which is based on the concepts of cluster-centric local question mining and batch-based question answering as *Recognition via Image and Batch-based local question answering (RIBQ).* In this section, we first explain the overview of the proposed approach.

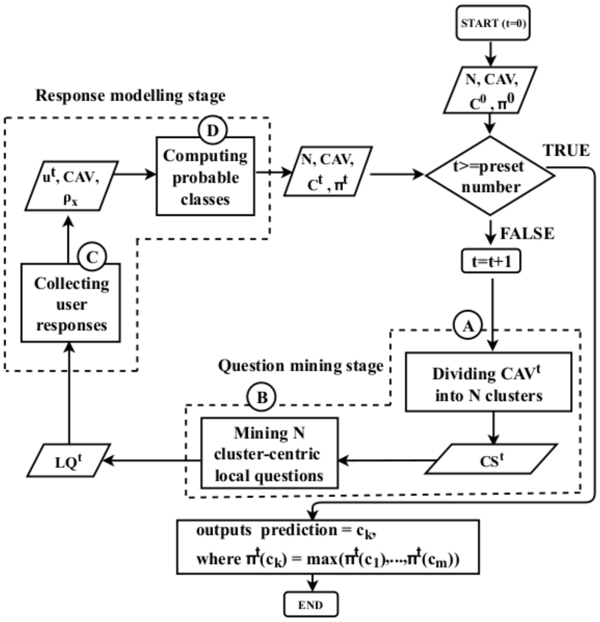

START (t=0)
N, CAV, C^0, π^0
Response modelling stage
D
u^t, CAV, ρ_x
Computing probable classes
N, CAV, C^t, π^t
t>=preset number
TRUE
FALSE
t=t+1
C
Collecting user responses
A
Question mining stage
Dividing CAV^t into N clusters
B
Mining N cluster-centric local questions
LQ^t
CS^t
outputs prediction = c_k, where $\pi^t(c_k) = \max(\pi^t(c_1),\dots,\pi^t(c_m))$
END

Figure 1: Recognition process in RIBQ.

Table 1: Commonly used terms and their notations

Notations	**Terms**
t	t^{th} time-step
N	Number of question in the batch
$Q = \{q_1, q_2,\dots, q_n\}$	Questions/attributes
Q^t	Un-asked question set till t
$C = \{c_1, c_2,\dots, c_m\}$	classes
u^t	Set of user responses collected at t
U^t	Set of all user responses collected till t
C^t	Probable classes after t
CS^t	Set of clusters at t
$\pi^t = \{\pi_1^t, \dots, \pi_m^t\}$	Class-probability estimates for $c_1(\pi_1^t)$, $\dots, c_m(\pi_m^t)$ given image x and U^t
$CAV = \{CAV(c_k) \forall k \in 1 \dots m\}$	Class-attribute vectors for $c_1(CAV(c_m)), \dots, c_m(CAV(c_k))$
$CAV^t = \{CAV^t(c_1) \forall k \in 1 \dots m\}$	Probabilistic class-attribute vectors at t for $c_1(CAV^t(c_m)), \dots, c_m(CAV^t(c_k))$
$\rho_x = p(c_k\|x) \forall k \in 1 \dots m$	Class-probability estimates for $c_1, \dots, c_m$ given image x
LQ^t	Set of cluster-centric local questions at t

Next, we present the details of question mining and response modeling stages in RIBQ.

4.2.1 *Overview*

Table 1 shows the commonly used terms and their notations. The recognition process of the proposed approach is depicted in Figure 1. The question mining stage consists of two steps: A) *Dividing CAV^t into N clusters* and B) *Mining N cluster centric local questions.* The response modelling stage consists of two steps: C) *Collecting user responses* and D) *Computing probable classes.*

At $t = 0$, the proposed approach takes the class-attribute vectors CAV, number of questions N in a batch, probable classes C^0 = C, and class-probability estimates π^0 as input. If pictures are not available, we consider all the classes as probable initially. So, the class probability estimates in π^0 are equal for all the classes. If pictures are available, class probability estimates in π^0 are initialized to probabilistic class predictions produced by a vision-based method. The steps at t are as follows: A) The probabilistic class-attribute vectors CAV^t at t are computed and then clustered into a set of clusters CS^t. B) From N clusters in CS^t, N cluster-centric local questions LQ^t are mined. C) A set u^t of user responses to the derived LQ^t are collected from the user. D) Using u^t, CAV and vision's predictions ρ_x (if pictures are available) class-probability estimates π^t and the probable classes C^t are computed which are input to the next time-step $t + 1$. The above steps are repeated until the preset number of time-steps. In the end, the class with the maximum probability estimate is returned as the final prediction.

4.2.2 *Question Mining Stage*

Question mining stage consists of 2 steps A and B.

A) Dividing CAV^t into N clusters

The inputs to this step are the number of clusters N, class attribute vectors (Definition 3.3), probable classes C^{t-1} and class-probability estimates. Here, C^{t-1} and π^{t-1} are the outputs of response modeling stage at $t - 1$. After computing probabilistic class-attribute vectors CAV^t at t, N non-overlapping clusters CS^t are computed from CAV^t. The details are as follows.

(i) **Computing CAV^t:** For each class c_k in C^{t-1}, we multiply its class-probability estimate $\pi^{t-1}(c_k)$, and its class-attribute vector for obtaining the probabilistic class-attribute vector of c_k at t denoted by $CAV^t(c_k)$.

$$CAV^t(c_k) = \pi^{t-1}(c_k) * CAV(c_k), c_k \in C^{t-1} \tag{1}$$

Now, $CAV^t = \{CAV(c_k), \forall c_k \in C^{t-1}\}$

(ii) **Clustering CAV^t into N clusters:** We perform *centroid initialized K-means clustering* on the obtained CAV^t for getting the set of clusters CS^t. At a time-step t, in K-means clustering, instead of initializing centroids randomly, we initialize them to the final centroids of clusters from the previous time-step $t - 1$. So, the clustering algorithm converges faster [1] thereby reducing the computation time.

B) Mining N cluster-centric local questions

The input to this step is the set CS^t of clusters produced in the preceding step. The output is the set LQ^t of the mined questions to be posed.

Given the set Q^t of n previously unasked questions till t and the set CS^t of N clusters at t, the issue is to select a potential question from each cluster. Ideally, we should select a question with high discriminating power such that the user's answer to the question prunes most of the improbable classes.

To select such questions, we have developed the following method based on the observation that the classes within a cluster are similar and classes from different clusters are dissimilar. We extract an attribute for each cluster which is exclusively predominant in that cluster either by *presence* or *absence.* As a result, we get multiple discriminative attributes, which we refer to as ***Cluster-centric local attributes (LQ^t).*** The resultant attributes LQ^t are later posed as questions to the user. To find the exclusively predominant attribute for each cluster, we propose the concept of *locality degree.*

We present the details of the proposed question mining approach after explaining the notions of *compound cluster representative vector* and *locality degree of an attribute in a cluster.*

- Compound cluster representative vector of cluster CS_j^t ***(***CR_j^t***):*** CR_j^t is the probability vector $< CR_{1,j}^t, CR_{2,j}^t \ldots, CR_{n,j}^t >$ where, $CR_{i,j}^t$ is the probability of presence of q_i in CS_j^t. CR_j^t is obtained by the summation of probabilistic class-attribute vectors (Equation 1) of all the classes contained in cluster CS_j^t. We denote the collection of compound cluster representatives of all clusters in CS^t as CR^t.

Given a cluster CS_j^t, we compute CR_j^t as follows:

$$CR_j^t = \sum_{CAV^t(c_k) \in CS_j^t} CAV^t(c_k) \tag{2}$$

- Locality degree of an attribute q_i ***in cluster*** CS_j^t: $LD(i,j)$ depicts the *degree* of an attribute q_i's exclusive presence or absence in the cluster CS_j^t. $LD(i,j)$ indicates the expected probability of q_i occurring in CS_j^t and expected probability of q_i not occurring in every other cluster $CS_{j'}^t$. So, for a given CS_j^t, $LD(i,j)$ is computed by taking the mean of the differences of the probability of occurrence of q_i in CS_j^t to that in every other cluster. The computed mean's absolute value represents the locality degree q_i in CS_j^t and its sign represents q_i's dominance either by presence($+ve$) or by absence($-ve$). If there is only a single cluster CS_j^t, we compute $LD(i,j)$ as the maximum of the probability of q_i's presence and probability of q_i's absence in the cluster CS_j^t.

$$LD(i,j) = \left\{ \begin{array}{ll} abs\left(\frac{\sum_{j \neq j'}(CR_{i,j}^t - CR_{i,j'}^t)}{N-1}\right) & \text{if } size(CR^t) \geq 2 \\ max(CR_{i,j}^t, 1 - CR_{i,j}^t) & \text{if } size(CR^t) = 1 \end{array} \right\} \tag{3}$$

where $j, j' \in 1..size(CR^t)$.

Example 1., Consider three representative vectors of CR^t to be [0.01,0.6,0.1], [0.4,0.2,0.8], and [0.5,0.1,0.02].

Then $LD(0,1) = abs(\frac{(0.01-0.4)+(0.01-0.5)}{2}) = 0.44$.

The steps for computing cluster-centric local questions is as follows. Algorithm 1 shows the pseudocode.

(1) Compute compound cluster representatives of all clusters in CS^t (Line 2).

(2) For each attribute q_i in Q^t,

(2.1) Compute q_i's *locality cluster* $LC(i)$ which is the cluster in which locality degree of q_i is the highest (Line 7). Let $LC(i)$ be CS_J^t in which q_i's locality degree is $LD(i,J)$.

(2.2) If local attribute information field for CS_J^t ($LAI(J)$) is either empty or if $LD(i,J)$ is greater than an existing locality degree in $LAI[J]$, update $LAI(J)$ (Lines 8-9).

(3) Add the questions stored in the local attribute information fields of clusters to LQ^t and remove them from Q^t (Line 11).

(4) Repeat steps 2,3 till each cluster gets its question by removing the clusters which have mined a local question along with their representatives from the respective sets (Line 12).

Algorithm 1 Cluster-centric local question mining at t

Input: CS^t: set of N clusters at t.
Output: LQ^t: set of N cluster-centric local questions to be posed at t.

```
1:  $LQ^t$={}. If $t = 0$, $Q^t = \{q_1, \ldots, q_n\}$, else $Q^t = Q^{t-1}$
2:  Compute $CR^t = \{CR_1^t, \ldots, CR_N^t\}$ using Equation 2.
3:  while $CS^t$ is not empty do
4:      for $j \in 1..size(CS^t)$ do
5:          $LAI[j] = [False, \emptyset, \emptyset]$
6:      for all the attributes in $Q^t$ do
7:          $LC(i) = CS_J^t = argmax(LD(i,J)), J \in \{argmax(CR_{i,j}^t), argmin(CR_{i,j}^t)\}, j \in 1 \ldots size(CR^t)$
8:          if $LAI[J][0] == False$ OR $LAI[J][2] \leq LD(i,J)$ then
9:              $LAI(J)$= $[True, q_i, LD(i,J)]$.
10:     for all $CS_j^t \in CS^t$ with $LAI[j][0] == True$ do
11:         $LQ^t$.add($LAI[j][1]$), $Q^t$.remove($LAI[j][1]$)
12:         $CS^t$.remove($CS_j^t$), $CR^t$.remove($CR_j^t$),
```

4.2.3 *Response modelling stage*

Response modelling stage consists of 2 steps C and D.

C) Collecting user responses

The input to this step is the set of cluster-centric local questions to be posed LQ^t. The output is the set of user responses to LQ^t.

The user response set to the N questions in LQ^t is denoted by $u^t = \{u_1^t, \ldots, u_N^t\}$. Here, u_i^t denotes the user's response to i^{th} question in LQ^t. Each response u_i^t=(a_i,r_i), where a_i denotes the answer to the i^{th} question, i.e, $a_i \in$ {yes/no} (as the questions are binary) and r_i denotes the user's confidence in his answer, $r_i \in$ {guessing, probably, definitely}.

D) Computing probable classes

The inputs to this step are the user response set to LQ^t, u^t and the class attribute vectors (Definition 3.3). Outputs are the probable classes and the class probability estimates for all classes computed based on u^t and vision's predictions (if pictures are available).

The class-probability estimate π_k^t for every class c_k $k = 1 \ldots m$ is computed as the conditional probability of c_k being the true class given image x and the responses U^t collected till t [3].

$$\pi^t(c_k) = p(\frac{c_k}{x, U^t}) = \frac{p(\frac{U^t, c_k}{x})}{p(\frac{U^t}{x})} = \frac{p(\frac{U^t}{c_k, x}) * p(\frac{c_k}{x})}{\sum_k p(\frac{U^t}{c_k, x}) * p(\frac{c_k}{x})} \tag{4}$$

Note that $p(\frac{U^0}{c_k, x}) = 1 \quad \because U^0 = \{\}$

When pictures are unavailable, we consider $p(\frac{c_k}{x})$ as uniform prior probabilities p(c_k). It is assumed that the questions are answered independently given the class [3]. Equation 4 can then be written as

$$\pi^t(c_k) = p(\frac{c_k}{x, U^t}) = \frac{p(\frac{U^{t-1}}{c_k, x}) * p(\frac{u^t}{c_k, x}) * p(\frac{c_k}{x})}{\sum_k p(\frac{U^{t-1}}{c_k, x}) * p(\frac{u^t}{c_k, x}) * p(\frac{c_k}{x})} \tag{5}$$

$$where, \quad p(\frac{u^t}{c_k, x}) = \prod_{u_i^t} p(\frac{u_i^t}{c_k, x}), \quad u_i^t \in u^t$$

$$p(\frac{u_i^t}{c_k, x}) = p(\frac{a_i, r_i}{c_k, x}) = \alpha_{r_i} * p(\frac{a_i}{c_k, x}) = \alpha_{r_i} * p(\frac{a_i}{c_k})$$

$$p(\frac{a_i}{c_k}) = \left\{ \begin{array}{ll} CAV^0_{c_k}[i], & \text{if } a_i = True \\ (1 - CAV^0_{c_k}[i]), & \text{if } a_i = False \end{array} \right\}$$

Here, we make the assumption that p($\frac{a_i}{c_k, x}$) =p($\frac{a_i}{c_k}$). It means that the types of noise or randomness that we see in user responses is class-dependent and not image-dependent. We use α_{r_i} to scale the responses according to the response's confidence as given by the user to ensure effective contribution of user's responses. i.e,

$$\alpha_{guess} < \alpha_{probably} < \alpha_{def}$$

Once $\pi^t(c_1), \ldots, \pi^t(c_m)$ are calculated, we determine probable classes C^t at $t+1$. A class is considered *probable* at $t+1$ (after t) if its class-probability estimate $\pi^t(c_k)$ is not too less than the mean estimate $p(\frac{C}{U^t, x})_{Avg}$. i.e,

$$C^t = [c_k \mid \quad (p(\frac{C}{U^t, x})_{Avg} - \pi^t(c_k)) \geq \gamma \; \forall \, k \in 1, \ldots, m]. \qquad (6)$$

Here, γ is a threshold for pruning.

The probable classes and the class-probability estimates are then passed as input to the question mining stage for processing in the next time-step. Algorithm 2 shows the pseudocode.

Algorithm 2 Batch-based response modelling at t

Input: $u^t = u_1^t, \ldots, u_N^t$: response set to LQ^t.
Output: C^t: probable classes after t,
$\pi^t(c_1), \pi^t(c_2), \ldots, \pi^t(c_m)$: class-probability estimates of all classes after t.

```
for $c_k$, $k \in 1..m$ do
    for $u_i^t \in u^t$ do
        Compute $\pi^t(c_k)$ using Equation 5
Determine $C^t$ using Equation 6.
```

5 Experimental results

All the experiments are conducted on an Intel i5 processor with 8GB RAM running Ubuntu Linux operating system. To evaluate the performance of the proposed method, the experiments are conducted on three datasets: CUB-200-2011 (CUB) [14], Animals with attributes (AwA2) [15], and aPascal dataset [18]. Table 2 provides the details about datasets. In Table 2, the notations |Size|, |A| and |C| denote the number of labelled objects, binary attributes per object, and classes in the dataset respectively. The attributes for all the datasets are of binary type. However, the approach in [2] requires multiple-choice questions. So, we manually created the same for all the datasets.

We have evaluated the performance of the following approaches.

- **B-IQR**: Image and Binary question answering based Recognition [3]
- **M-IQR**: Image and Multiple-choice question answering based Recognition [2]
- **B-RAN**: Batch-based random question answering algorithm implemented by us. In this approach, we select the questions in the batch randomly. The objective for considering this approach is to show the effect of cluster-based local question mining in RIBQ.
- **RIBQ**: Recognition via Image and Batch-based local Question answering, which is the proposed approach.

We simulated the human-in-the-loop paradigm for all the preceding approaches by scaling the available instance-level attribute values in the datasets to integrals in the range [-3,3] (excluding 0), where a positive value {1, 2, 3} signifies the response 'yes' and a negative value {-1, -2, -3} signifies 'no' with the magnitude indicating the confidence of the user response (guessing, probably, definitely). For example, if the user responded with -2 for a question, then it indicates the response as 'no' with confidence as 'probably'.

The parameters of our simulation are selected to closely reflect real-world application requirements. Table 3 summarizes the parameters of our performance evaluation. The parameter *Resp* is employed to vary the total number of questions to be posed to the user (default=30). To vary the number of time-steps, we employ the parameter *Level* (default=15). To vary the number of questions in a time-step, we employ the parameter N (default=2). We employ the parameter γ (default=10^{-5}) to vary the number of improbable classes which are pruned at the end of each time-step. The parameters $\alpha_{probably}$ and α_{guess} are employed to vary the weights of the user's response: probably (default=0.75) and guessing (default=0.25) respectively. The performance was observed to be the best at the chosen default values for γ, $\alpha_{probably}$ and α_{guess}. We have not presented the actual results due to space limitation.

The following performance metrics are employed.

- Acc1: % of correct predictions in the test dataset without considering pictures.
- Acc2: % of correct predictions in the test dataset by considering pictures.
- CT: Average time (in seconds) taken for a target object's recognition. (It does not include the time consumed by the user a.k.a. human-in-the-loop to give the response.)

We use a test-train split of 50:50 on all the datasets. We use Pedro Morgado's publicly available source code [8] as the vision algorithm on the datasets CUB and AwA2.

Table 2: Dataset details

Data	|Size|	|A|	|C|
CUB	11,788	312	200
AwA2	37,322	85	50
aPascal	12,695	64	20

Table 3: Parameter details

Param	Def	Variations
|Resp|	30	10,20,40,50,60,70
N	2	1,4,8
Level	15	5,10,20,25,30
γ	10^{-5}	10^{-1},10^{-3},10^{-7}
α_{prob}	0.75	0.5,0.8,0.9
α_{guess}	0.5	0.25,0.6,0.8

Table 4: Performance results at default parameter values

	CUB			AwA2			aPascal	
	*Acc*1	*Acc*2	*CT*	*Acc*1	*Acc*2	*CT*	*Acc*1	*CT*
RIBQ	34.6	83.8	7	67.2	85.8	4	82.5	3
B-IQR	29.6	75.6	167	66.2	83.3	76	81.1	68
M-IQR	21.8	67.3	64	51.4	74.5	28	73.9	23
B-RAN	12.6	56.2	2	42.7	69.4	2	68.5	2

5.1 Performance comparison

Table 4 shows the performance results of the approaches for *Acc*1, *Acc*2, and *CT* on the datasets at default parameter values (Table 3) except *Level* as *Level* and *N* are inter-dependent at fixed *Resp* ($Resp = Level * N$) and N varies with each approach. Between B-IQR and M-IQR, the results show that M-IQR does well in terms of *CT* due to the multiple-choice questions employed in M-IQR. However, the performance in *Acc*1, *Acc*2 for M-IQR is lesser over B-IQR due to the lesser diversity among the questions in M-IQR compared to B-IQR. The B-RAN approach shows the best *CT* performance and lowest *Acc*1 and *Acc*2 as this approach selects a batch of questions at random. Between the proposed RIBQ and B-RAN, as both are batch-based approaches, *CT* performance of RIBQ is comparable to the best performing B-RAN. However, for RIBQ, *Acc*1 and *Acc*2 are high over B-RAN due to the careful selection of questions in RIBQ. The *CT* in both B-IQR and M-IQR is significantly higher than that in RIBQ due to the costly information gain metric used in B-IQR and M-IQR. The *Acc*1 and *Acc*2 are high for RIBQ than B-IQR and M-IQR because of the cluster-centric local question mining method employed in RIBQ. Overall, the results show that our proposed approach, RIBQ significantly improves accuracy over all the approaches and *CT* over B-IQR and M-IQR due to batch-based cluster-centric local question mining.

As an alternative to RIBQ, we considered an experimental approach by extending B-IQR to multiple binary questions. In the experimental approach, at each time-step, *N* attributes with the highest information gain are posed as questions to the user. However, the overlap between the discriminative information gained by the chosen *N* attributes is very high. As a result, the performance of the experimental approach is even lesser than B-IQR, so we have not presented the comparison results due to space limitation.

5.2 Effect of variation in *Resp*

The parameter *Resp* is employed to vary the total number of questions to be posed to the user. Fig. 2 shows the results of the approaches on the datasets w.r.t. variations in *Resp*. From Fig. 2(a), we can see that with an increase in *Resp*, RIBQ improves *Acc*1 significantly over other approaches on CUB dataset. The results show that RIBQ at N=2 improves the *Acc*1 performance over B-IQR, and M-IQR, due to the multiple discriminative questions posed by RIBQ. Between B-IQR and M-IQR, B-IQR performs better for all *Resp* values, since the diversity of the responses received to questions in B-IQR is better than that in M-IQR as the net response for a single multiple-choice question in M-IQR is the total number of choices of the question (say a user selects one of the choices, it implies that other choices are not possible). B-RAN fails to improve *Acc*1 due to the random nature of posed questions. The *Acc*1 performance on AwA2 dataset in Fig. 2(b) and aPascal dataset in Fig. 2(c) exhibit a similar trends as that of Fig. 2(a). Notably, as the number of attributes and classes are few, the performance improvement of RIBQ on AwA2 and aPascal over other approaches is less as compared to the performance improvement on CUB dataset. The results demonstrate that RIBQ exhibits superior performance over other methods with the increase in the number of attributes and classes.

Similar to the results of Fig. 2(a), the results of Fig. 2(d) and Fig. 2(e) show that, with the increase in *Resp*, the proposed RIBQ improves *Acc*2 over other approaches on CUB dataset and AwA2 datasets. The justification for the performance improvement is similar to the case of performance improvement of Acc1. The results of Fig. 2(f) show *CT* results on CUB dataset with the increase in *Resp*. As expected, B-RAN at all responses gives the lowest *CT* over other approaches due to its batch-basele to gain more information by posing more number of questions compared to other appd random question answering. *CT* for RIBQ is slightly higher than B-RAN as it has to compute cluster-centric local questions. *CT* in B-IQR is significantly high due to the high computational overhead in gain-based question mining. M-IQR improves performance in comparison to B-IQR by posing a single multiple-choice question every time-step leading to reduced computation. The results of Fig. 2(g) and Fig. 2(h) show *CT* results on AwA2 and aPascal datasets respectively with the increase in *Resp*. The results trend are similar to the performance results shown in Figure 2(f).

It can be noted that, even when the value of *Resp* is kept same for all four approaches, the proposed approach RIBQ is giving better *Acc*1, *Acc*2 and *CT* over other approaches for all datasets. It means that RIBQ is posing questions with better discriminating power as compared to other approaches.

5.3 Effect of variation in *Level*

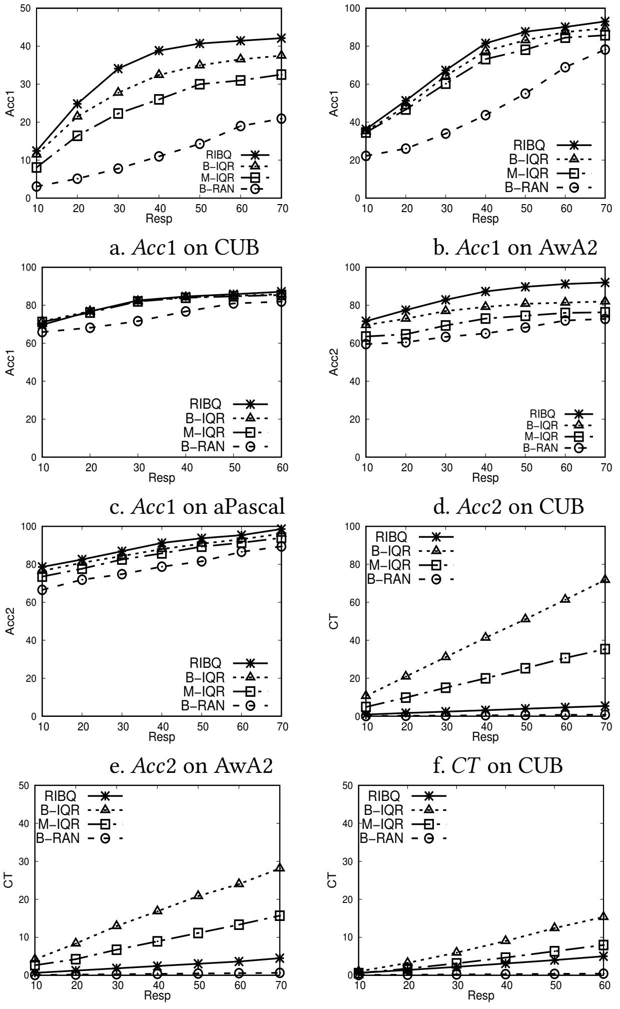
The parameter *Level* is employed to vary the total number of time-steps. Fig. 3 shows the results of the approaches on the datasets w.r.t. variations in the value of parameter *Level*. The results in Fig. 3(a) show that with the increase in *Level*, the proposed approach, RIBQ improves *Acc*1 significantly over other approaches on CUB dataset. This is because, with the increase in *Level*, the number of diverse questions asked by RIBQ is double (as default N=2) of that asked in B-IQR and M-IQR leading to faster knowledge gain. Also, M-IQR performs slightly better than B-IQR as *Level* increases. The reason is that for any *Level* value, M-IQR poses *Level* number of informative multiple-choice questions whereas B-IQR poses *Level* number of informative binary questions. As a result, the knowledge gain in M-IQR is slightly higher than in B-IQR leading to its better performance. B-RAN fails to improve performance due to the random nature of posed questions every level. The *Acc*1 performance on AwA2 dataset in Fig. 3(b) and aPascal dataset in Fig, 3(c) exhibit a similar trend to Fig. 3(a) in the performance improvement.

Similar to the results of Figure 3(a), the results of Figure 3(d) and 3(e) show that, with the increase in *Level*, RIBQ improves the *Acc*2 performance over other approaches on CUB dataset and AwA2 datasets. The reason for the performance improvement is the same as that mentioned in the case of Acc1. The results of Fig. 3(f), 3(g), 3(h) show *CT* results on CUB, AwA2 and aPascal datasets respectively with the increase in *Level*. The trend and justification for the trend is similar to that in 2(f).

It can be noted that, even when the value of *Level* is kept same for all four approaches, the proposed approach RIBQ is giving better *Acc*1, *Acc*2 and *CT* over other approaches for all datasets. It means that RIBQ is able to gain more information by posing more number of questions compared to other approaches.

5.4 Effect of variation in *N*

The parameter *N* is employed to vary the number of questions to be posed to the user at each time-step. Table 5 reports the performance of the four methods in the comparison metrics at varying *N*.



a. *Acc*1 on CUB

b. *Acc*1 on AwA2

c. *Acc*1 on aPascal

d. *Acc*2 on CUB

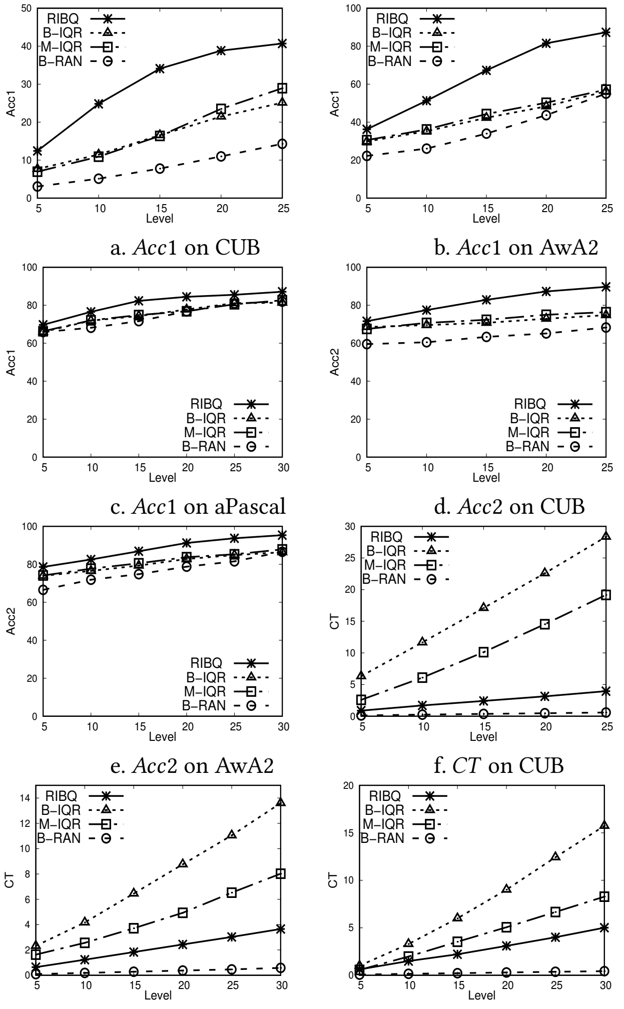
e. *Acc*2 on AwA2

f. *CT* on CUB

g. *CT* on AwA2

h. *CT* on aPascal

Fig 2: Effect of variations in Resp



a. *Acc*1 on CUB

b. *Acc*1 on AwA2

c. *Acc*1 on aPascal

d. *Acc*2 on CUB

e. *Acc*2 on AwA2

f. *CT* on CUB

g. *CT* on AwA2

h. *CT* on aPascal

Fig 3: Effect of variations in Level

(Note that in this experiment the value of *Level* is kept constant. Also, as the value of N increases the total number of questions *Resp* increases.) As expected, *Acc*1 and *Acc*2 of RIBQ increase with *N* for three datasets. The reason is that as *N* increases, more number of questions are processed at each time-step. Also, it can be observed that *CT* is increasing slightly with *N* for CUB. This is due to increased computation in identifying more number of discriminating attributes (questions) in a batch as *N* increases. Overall, the results show that there is scope to improve the performance using batch-based local question answering.

Table 5: Effect of variation in N

	CUB			AwA2			aPascal	
RIBQ	*Acc*1	*Acc*2	*CT*	*Acc*1	*Acc*2	*CT*	*Acc*1	*CT*
N=2	34.6	83.8	2.4	51.3	82.6	1.3	69.6	0.6
N=4	38.3	90.1	2.6	78.4	88	1.3	79	0.6
N=6	40.5	91.6	2.7	87.5	91.3	1.4	81.9	0.6
N=8	42.2	91.9	2.8	92.2	95.2	1.4	82.7	0.7

6 Conclusions

Fine-grained object recognition has received much attention with the advent of neural nets due to its potential applications. However, many of the existing works are either excessively picture-dependent making them cost-sensitive or are too slow and weak. In this paper, we propose a batch-based local feature extraction method leveraging a human-in-the-loop's input for making the recognition process cost-sensitive, robust and fast. When pictures are available our model facilitates the plug-in of vision algorithms into the framework for better performance. Experiments on three real datasets show significant improvement in performance with respect to both the accuracy and computation time.

As a part of the future work, we are planning to investigate the applicability of the proposed approach in building human-in-the-loop based decision support systems in agriculture and medial domains for crop problem identification and disease diagnosis respectively.

References

[1] David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 1027–1035.

[2] Steve Branson, Grant Van Horn, Catherine Wah, Pietro Perona, and Serge Belongie. 2014. The ignorant led by the blind: A hybrid human–machine vision system for fine-grained categorization. *International Journal of Computer Vision* 108, 1-2 (2014), 3–29.

[3] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. 2010. Visual recognition with humans in the loop. In *European Conference on Computer Vision*. Springer, 438–451.

[4] Xi Stephen Chen, He He, and Larry S Davis. 2016. Object detection in 20 questions. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–9.

[5] Jia Deng, Jonathan Krause, Michael Stark, and Li Fei-Fei. 2016. Leveraging the wisdom of the crowd for fine-grained recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 4 (2016), 666–676.

[6] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 3 (2014), 453–465.

[7] Xiao Liu, Jiang Wang, Shilei Wen, Errui Ding, and Yuanqing Lin. 2017. Localizing by describing: Attribute-guided attention localization for fine-grained recognition. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[8] Pedro Morgado and Nuno Vasconcelos. 2017. Semantically consistent regularization for zero-shot recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6060–6069.

[9] Devi Parikh, Adriana Kovashka, Amar Parkash, and Kristen Grauman. 2012. Relative attributes for enhanced human-machine communication. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[10] Orod Razeghi, Guoping Qiu, Hywel Williams, Kim Thomas, and I VIPLAB. 2012. Skin lesion image recognition with computer vision and human in the loop. *Medical Image Understanding and Analysis (MIUA), Swansea, UK* (2012), 167–172.

[11] Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. 2015. Best of both worlds: human-machine collaboration for object annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2121–2131.

[12] Yaser Souri, Erfan Noury, and Ehsan Adeli. 2016. Deep relative attributes. In *Asian Conference on Computer Vision*. Springer, 118–133.

[13] Catherine Wah, Steve Branson, Pietro Perona, and Serge Belongie. 2011. Multi-class recognition and part localization with humans in the loop. In *2011 International Conference on Computer Vision*. IEEE, 2524–2531.

[14] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset, California Institute of Technology. (2011).

[15] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[16] Yichao Yan, Bingbing Ni, and Xiaokang Yang. 2017. Fine-grained recognition via attribute-guided attentive feature aggregation. In *Proceedings of the 25th ACM International Conference on Multimedia*. ACM, 1032–1040.

[17] Aron Yu and Kristen Grauman. 2014. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 192–199.

[18] Y. Yu. 2009. aPascal-aYahoo Image Data Collection, Department of Computer Science, University of Illinois at Urbana-Champaign. (2009).