

Bees? Bees!

--Image Classification Using Apache Spark

Proposal submitted by

Chang, Xiaowen Dai, Xingchi Hwang, Jeewon

xiaowenchang@g.harvard.edu dai01@g.harvard.edu jeewonhwang@g.harvard.edu

Harvard University

Cambridge, Massachusetts

2015/11/5

**Problem:** Given five thousand images, find out how to detect the area of the image containing an object(bee) and then distinguish the species of them (honey bees or bumble bees).

**Objectives - Functionality and Performance**

* **Functionality:** detecting and distinguishing objects(bees) in images
* **Features**
  + **Mandatory:** detection of the area containing the object (bee) in one image.
  + **Optional:** differentiation the species of objects in one image
* **Performance:** 1 ~ 2 hours

**Background and Motivation**

* **Motivation:** Image is an important data, which contains useful, yet hard to retrieve, information. Many technology companies are trying to develop their own algorithms on processing images accurately to get powerful features, such as face recognition and self-driving cars. This project will implement knowledge taught in CS205, such as parallel computing, spark etc., and other machine learning concepts, such as PCA and KNN classifications to do a fast image processing procedure.
* **Background Researches:** 
  + Parallel computing algorithm in the image processing
  + Classification algorithms: Principal-Component-Analysis (PCA), K-Nearest-Neighbors (KNN) and Linear-Support-Vector-Machine (Linear SVM)
  + Metrics to estimate performances: Confusion Matrix and ROC curve

**Design Overview**

* **In-class Knowledge:** parallel computing in Spark, MLlib module, image processing

**Verification**

* **Test data, serial implementations, expected results:**

1. Data: 5,000 images, each one contains a honey bee or a bumble bee; a correct label has to be tagged for them.
2. Train/Test data: data will be divided into the train dataset and the test dataset. The train dataset would be used to develop the algorithm and the test dataset would be used to evaluate the performance of the algorithm.

* **Accuracy**

1. Cross-Validation: when estimating the parameters of the model, k-fold cross-validation approach would be applied to avoid over-fitting scenario.
2. ROC Curve: AUROC (area under ROC curve) will be used to evaluate the accuracy of results and the performance of algorithms.

**Schedules and Major milestones**

* **Nov.19th ~ Nov.22nd:** data manipulation and cleaning - paint on the image to label where the bee is, then divide up those images and the originals into parts. Any part that is more than 50% bee is a "bee part".
* **Nov.23rd ~ Nov.30th:** implementation of the mandatory functionality. The function could detect of the area containing bee(s) in one image.
* **Dec.1st ~ Dec.6th:** implementation of the optional functionality. The function could differentiate species of objects and name them; preparation of final presentations and reports
* **Dec.7th ~ Dec.10th:** website development

**Task Assignment**

The group will meet on the daily basis. Each member will report his/her progress and discuss the status of the project during the meeting. Each member is required to read and share any good resource he/she has found. During the project, each member takes the **same amount of responsibility** on designing algorithm and writing implementation. Besides that, the group also has the following structure.

* **Project Manager (Xiaowen Chang):** Xiaowen will be responsible for coordinating inside the group. She will lead the discussions in every group meeting and write down the meeting minutes.
* **Web (Jeewon Hwang):** Jeewon will lead the design of the project web, and assign any tasks related to web designing to other members.
* **Reports, Presentations (Xingchi Dai):** Xingchi will lead the writing of reports and presentations. He will be responsible for assigning any tasks related to reports and presentations to other members.

**Sample Image**

