Automated User-Generated Content Curation with Deep Learning Techniques

PhD Thesis Research Plan

Stalin Leonel Cruz

Supervisor: Dr. Ruben Tous
Dra. Beatriz Otero

Department of Computer Architecture
Universitat Politècnica de Catalunya

This Research Plan is submitted for the degree of
Doctor of Philosophy

Universitat Politècnica de Catalunya May 2017

Table of contents

Chapter 1

Introduction

1.1 Motivation

Nowadays, there is a growing interest in exploiting the photos that users share on social networks such as Instagram or Twitter [**?**], a part of the so-called user-generated content (UGC). On the one hand, users' photos can be analyzed to obtain knowledge about users behavior and opinions in general, or with respect to a certain products or brands. On the other hand, some users' photos can be of value themselves, as original and authentic content that can be used, upon users' permission, in different communication channels. However, discovering valuable images on social media streams is challenging. The amount of images to process is huge and, while they help, user defined tags are scarce and noisy. The most part of current solutions rely on costly manual curation tasks over random samples. This way many contents are not even processed, and many valuable photos go unnoticed.

Some works, such as [2], [3] or [1], propose scene-based and object-based image recognition techniques to enrich the metadata originally present in social media images in order to facilitate their processing. Automatically tagging the incoming images with tags that describe their semantics (e.g. "beach", "car", etc.) enables more expressive filters and search conditions, minimizing manual curation. This way the time between the image publication and its detection and usage is minimized, the quality of the resulting photos increases (as more photos are analyzed and only the best ones go through manual curation) and the cost is reduced.

Adoption of image recognition techniques in commercial UGC systems is currently very limited. In the best case they provide generic classifiers whose categories and original training data were not specific to UGC. Often these classifiers limit to the categories of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), but the vast majority of Instagram/Twitter photos are people-centric (selfies, food, clothes, etc.) while ILSVRC is

more generic (fauna, flora, etc.). An important particularity of UGC is the huge amount of *spam images*, i.e. images that, in the most usage scenarios, has no value neither as a knowledge carrier nor as a exploitable content. The incapacity of detecting the multiple types of spam images limits the usability and efficiency of existing solutions. Another difficulty of adopting image recognition techniques into UGC systems is the high computational cost of CNN-based image classifiers and object detectors. These systems need to process incoming streams of hundreds of images per second and a very volatile traffic. Any additional processing component need to be extremelly efficient and scalable.

For all these reasons, the development of UGC systems that overcome the constraints imposed by manual curation (cost, qualty and delay time) requires the development of efficient and scalable image recogition pipelines, tailored for the particularities of UGC.

1.2 Goals

In this thesis, our general goal is to improve some techniques currently applied for automatically annotating photos posted by social media users. The generated annotations facilitate searching, filtering and analyzing user-generated content (UGC). We will apply scene-based and object-based image recognition techniques to extend the metadata originally present in the images with tags that describe their visual content. In order to reach state-of-the-art precission values we will employ CNNs. Performing these tasks in the UGC scenario implies some problems that are currently constraining their applicability to commercial systems. The specific goals of the project, that aim solving or mitigating some of these problems, are the following:

- Development of an efficient and scalable solution to enable complex image recogition pipelines for automatic UGC curation, where multiple CNN-based classifiers and object detectors can be chained in a graph fashion. The resulting solution must provide the necessary throughtput under reasonable cost assumptions, and must be able to scale.

- Development of novel UGC-specific scene-based and object-based image recognition models, including but not limited to, a complete set of Instagram spam image detection models.

- Development of a solution for the scalable training of on-demand custom classifiers and object detectors (e.g. company logos). Because of the particularities of these models, the resulting solution need to minimize training times and to be able to deal with small datasets.

Chapter 2

State of the art

2.1 Image recognition

Image recognition is still an active research area. It can be addressed through different approaches. Among them, one of the most well-known techniques is bag of visual words. Two additional approaches can be applicable : Convolutional Neural Networks (CNN) and Part-based models.

2.1.1 Bag of visual words

The bag of visual words model, proposed by Csurka et. al. [**?**], has been widely used in computer vision. This model produces a description of each image feature using a visual vocabulary, and uses machine learning models to perform image recognition. By collecting information from labeled images having common characteristics, this method can classify and recognize a large number of images from different categories.

This technique typically involves two main phases. Training is an offline process for creating the visual dictionary and computing image histograms. First, we detect and describe image feature points as presented in Figure **??**, using SIFT (Scale Invariant Features Transform) [**?**] in our work. The vocabulary is generated by clustering the resulting descriptors with k-means algorithm [**?**].

The visual words are clusters centroids which represents representative or common features, as illustrated in **??**.

Each image feature is indexed by the cluster in which it belongs. Thus, each image is represented as a "bag of visual words", by generating an histogram (figure **??**) representing the importance of each entry of the visual vocabulary.

The second phase is predicting and represents the effective image classification against previously-labeled images. The process is executed on the given image to be recognized : extract SIFT feature points and calculate descriptor for each feature point. After that, we match image feature descriptors with the visual vocabulary and build the histogram which counts how many times each of the visual words occurs in the image. Images from a specific category will have more representative features from their category as compared to other features present in dictionary. Images must have appropriate labels describing the class that they represent. We will train an SVM model, a binary classifier that builds a linear decision boundary between two classes of inputs.The histograms obtained as a result will be given to the SVM to predict respective category labels.

2.1.2 Part-based models

Part-based models are also extensively used in image recognition. The idea is that even if the general object aspect may have significant variations within a class, appearance and relative locations between some parts can still be useful for learning object models, as illustrated in figure **??**.

These techniques are based on representing objects as an assembly of parts and modeling their spatial relationships. Among part-based models, the constellation model [**?**] is popularly used in image recognition. Constellation model attempts to represent an object class by a set of N parts under mutual geometric constraints. Further improvements in constellation models develop models for shape, appearance and relative scale of objects for better image recognition. Such methods perform excellently on objects that have a very defined structure, like faces.

2.1.3 Deep Convolutional Neural Networks

Recently, deep learning methods like convolutional neural networks (CNNs) are being used for image recognition [**?** **?**]. They consist of multiple layers of small neuron collections which look at small portions of the input image. The results of these collections are then tiled so that they overlap to obtain a better representation of the original image [**?**]. Some of the works are available as specialized collection of machine learning libraries, e.g., TensorFlow [**?**]. It will be interesting to see the efficiency of such solutions in the context of massive image flows.

2.2 Apache Spark

In this work, we will investigate how to leverage Apache Spark, as an example modern big data platform for image recognition tasks.

Apache Spark [**?**] is an open source framework designed to support fast iterative data analysis on very large datasets, as is common in large scale machine learning. It originated at UC Berkeley AMPLab in 2009 and was contributed to Apache Software Foundation in 2010. It becomes widely adopted in industry, and academic research. Compared to other solutions (such as Hadoop MapReduce [**?**]), Apache is fast because of his computations in-memory nature. Its Resilient Distributed Dataset (RDD) abstraction enables developers to materialize any point in a processing pipeline into memory across the cluster. These items can be manipulated in parallel using functional APIs. Spark executes a more general Directed Acyclic Graph (DAG) of operations. It can directly pass results to the next step in the pipeline without writing intermediate results to the distributed file system. This technique also enhances the solution fault-tolerance, by describing parallel tasks, tracking computational lineage and optimizing for data locality when scheduling work. By default, Spark uses Hadoop Distributed File System (HDFS) [**?**] for persistent storage. These features make Apache Spark an interesting candidate for our image processing distributed. Apache provides Scala, Java, and Python programming interfaces. We will focus on Java language for implementing our solution.

Chapter 3

Methodology

3.1 Project phases

We aim to address all the above mentioned goals by setting out following stages:

Phase 1: Implementation of a distributed version of the Bag of Visual Words algorithm Our scope is to understand how Spark operates and use it to implement a distributed version of bag of visual words image recognition technique. Spark's core and Machine Learning library contains most of the components of bag of visual words model. Fist of all, we need to construct input RDD of image paths and metadata to be processed. Next, we need to perform all the required processing steps of training and predicting described in the state of art chapter.

Phase 2: Evaluation of the computational performance and scalability of the obtained system over the MareNostrum III Supercomputer In the second part of the work, we will analyze the performance of Spark in a cluster. The main aspect of this phase is to run spark jobs on many nodes and study how this will affect the computation time. We will use Spark on MareNostrum (spark4mn), which is basically an Apache Spark cluster setup on MareNostrum supercomputer. In order to evaluate the scalability of our solution, and because the framework provides functionalities to evaluate different configurations, we will have to inspect the impact of :
- increasing the workload (size up)
- increasing the number of clusters
- equally increasing the workload along with the number of computing nodes (scale up)
- strong scaling (speed up)

Phase 3: Application of the obtained system to a specific object recognition task in continuous streams of Instagram photos One of the challenges of our work is handling continuous flows social pictures. In order to effectively run and evaluate the obtained system, we need to create a dataset to simulate scenarios in real world. We will have to choose a real dataset, taken in a constrained way : from smartphones and posted to Instagram social media. Images should represent a particular field such as food recognition and should be labeled with many categories in order to allow the classification step. This dataset can contain multiple modalities : images and text metadata. A possible use case is food recognition. We need to study how different dataset characteristics (quality, size, color, dictionary size, etc.) influences scalability and the overall performance recognition : in the training and prediction phases. These performance should be provided by different metrics (accuracy, precision).

Phase 4: Comparison of the results with other alternatives In order to validate our solution, we need to perform experiments to compare performance of different image recognition techniques. Comparison will take into account computing and recognition quality performance.

3.2 Associated work plan

I plan to finish my Ph.D in a total of three years. A tentative work plan of my future activities can be found in Table **??**.

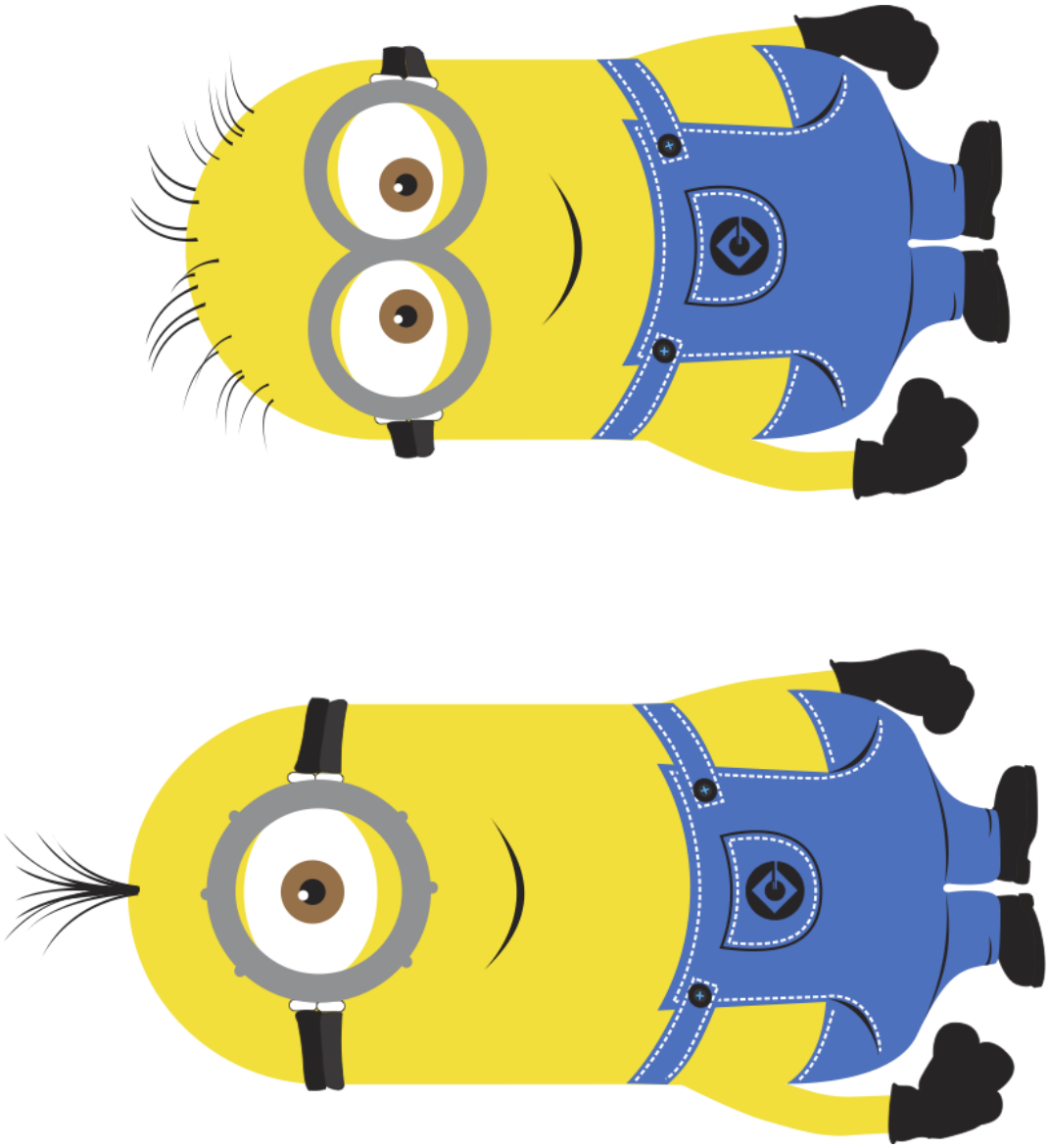

Fig. 3.1 Here you can place the GANTT diagram as an image

- C1: This publication will describe the performance of the Spark distributed implementation of bag of visual words algorithm with different parameters which can affect the image recognition process performance.
- C2: This publication will evaluate in more details the scalability performance of the proposed solution while running on the supercomputer.
- C3 : This publication will detail the application and the performance of our solution on real datasets of continuous Instagram image flows.
- J1: This publication will summarize previous contributions on real datasets of continuous Instagram image flows, taking into account associated metadata.
- J2: This publication will present a comparative performance study with other image recognition alternative

3.3 Completed activities

In this last year, the first six months have been principally spent studying the topic and having theoretical research. Then I implemented a standalone version of the bag of words using Opensource Computer Vision (OpenCV) library in Java. OpenCV library includes a wide variety of optimized machine learning algorithms. The C/C++ interface comes with a bag of words implementation. However, it is absent in Java interface and we created a bag of words that is fully compatible with OpenCV Java library. This step allowed me to fully understand the process and its variation, i.e different parameters which can be applied and which can have incidence on the recognition performance. Then, from September, I started studying distributed computing, Apache Spark architecture and implementing the distributed version of bag of visual words.

References

[1] Denton, E., Weston, J., Paluri, M., Bourdev, L., and Fergus, R. (2015). User conditional hashtag prediction for images. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1731–1740, New York, NY, USA. ACM.

[2] Park, M., Li, H., and Kim, J. (2016). HARRISON: A benchmark on hashtag recommendation for real-world images in social networks. *CoRR*, abs/1605.05054.

[3] Tous, R., Torres, J., and Ayguadé, E. (2015). Multimedia big data computing for in-depth event analysis. In *BigMM*, pages 144–147. IEEE.