Advances in Computer Vision and Pattern Recognition

Radu Tudor Ionescu
Marius Popescu

Knowledge Transfer between Computer Vision and Text Mining

Similarity-based Learning Approaches

Springer

Advances in Computer Vision and Pattern Recognition

More information about this series at http://www.springer.com/series/4205

Radu Tudor Ionescu · Marius Popescu

Knowledge Transfer between Computer Vision and Text Mining

Similarity-based Learning Approaches

Springer

Radu Tudor Ionescu
Department of Computer Science
University of Bucharest
Bucharest
Romania

Marius Popescu
Department of Computer Science
University of Bucharest
Bucharest
Romania

ISSN 2191-6586 ISSN 2191-6594 (electronic)
Advances in Computer Vision and Pattern Recognition
ISBN 978-3-319-30365-9 ISBN 978-3-319-30367-3 (eBook)
DOI 10.1007/978-3-319-30367-3

Library of Congress Control Number: 2016932522

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

To our dear families and friends

Foreword

The present book basically studies similarity-based learning approaches for two different fields: computer vision and string processing. However, the discussed text goes far beyond the goal of a general or even of a comprehensive presentation. From the very beginning, the reader is faced with a genuine scientific challenge: accepting the authors' view according to which image and text can and should be treated in a similar fashion.

Computer vision and string processing seem and are traditionally considered two unrelated fields of study. A question which naturally arises is whether this classical view of the two fields can or should be modified.

While learning from data is a central scientific issue nowadays, no one should claim to be a data analyst without having performed string processing. Information retrieval and extraction ultimately depend on string manipulation. From a different angle, computer vision is concerned with the theory behind artificial systems that extract information from images. One is finally concerned with a goal of the same nature: information acquisition. From this perspective, the approach proposed by the authors seems more natural and indeed scientifically justified.

The authors consider treating images as text and improving text processing techniques with knowledge coming from computer vision. Indeed, corresponding concepts like word and visual word do exist, while the existing literature regards, for instance, modeling object recognition as machine translation. The authors present improved methods that exploit such concepts as well as novel approaches, while broadening the meaning of classical concepts like string processing by taking into account tasks ranging from phylogenetic analysis and DNA sequence alignment to native language identification and text categorization by topic. All in all, the authors gradually build a strong case in favor of the theory they are promoting:

knowledge transfer from one of the studied domains to the other being extremely productive. The very topic of this book represents a scientific challenge, one that the authors master with a great precision and that offers interesting perspectives for future scientific research.

Florentina Hristea

Preface

Machine learning is currently a vast area of research with applications in a broad range of fields such as computer vision, bioinformatics, information retrieval, natural language processing, audio processing, data mining, and many others. Among the variety of state-of-the-art machine learning approaches for such applications are the similarity-based learning methods. Learning based on similarity refers to the process of learning based on pairwise similarities between the training samples. The similarity-based learning process can be both supervised and unsupervised, and the pairwise relationship can be either a similarity, a dissimilarity, or a distance function.

This book studies several similarity-based learning approaches, such as nearest neighbor models, local learning, kernel methods, and clustering algorithms. A nearest neighbor model based on a novel dissimilarity for images is presented in this book. It is used for handwritten digit recognition and achieves impressive results. Kernel methods are used in several tasks investigated in this book. First, a novel kernel for visual word histograms is presented. It achieves state-of-the-art performance for object recognition in images. Several kernels based on a pyramid representation are presented next. They are used for facial expression recognition from static images. The same pyramid representation is successfully used for text categorization by topic. Moreover, an approach based on string kernels for native language identification is also presented in this work. The approach achieves state-of-the-art performance levels, while being language independent and theory neutral. An interesting pattern can already be observed, namely that the machine learning tasks approached in this book can be divided into two different areas: computer vision and string processing.

Despite the fact that computer vision and string processing seem to be unrelated fields of study, image analysis and string processing are in some ways similar. As will be shown by the end of this book, the concept of treating image and text in a similar fashion has proven to be very fertile for specific applications in computer vision. In fact, one of the state-of-the-art methods for image categorization is inspired by the *bag of words* representation, which is very popular in information

retrieval and natural language processing. Indeed, the *bag of visual words* model, which builds a vocabulary of visual words by clustering local image descriptors extracted from images, has demonstrated impressive levels of performance for image categorization and image retrieval. By adapting string processing techniques for image analysis or the other way around, knowledge from one domain can be transferred to the other. In fact, many breakthrough discoveries have been made by transferring knowledge between different domains. This book follows this line of research and presents novel approaches or improved methods that rely on this concept. First of all, a dissimilarity measure for images is presented. The dissimilarity measure is inspired by the rank distance measure for strings. The main concern is to extend rank distance from one-dimensional input (strings) to two-dimensional input (digital images). While rank distance is a highly accurate measure for strings, the empirical results presented in this book suggest that the proposed extension of rank distance to images is very accurate for handwritten digit recognition and texture analysis. Second of all, a kernel that stems from the same idea is also presented in this book. The kernel is designed to encode the spatial information in an efficient way and it shows performance improvements in object class recognition and text categorization by topic. Third of all, some improvements to the popular bag of visual words model are proposed in the present book. As mentioned before, this model is inspired by the bag of words model from natural language processing and information retrieval. A new distance measure for strings is introduced in this work. It is inspired by the image dissimilarity measure based on patches that is also described in the present book. Designed to conform to more general principles and adapted to DNA strings, it comes to improve several state-of-the-art methods for DNA sequence analysis. Furthermore, another application of this novel distance measure for strings is discussed. More precisely, a kernel based on this distance measure is used for native language identification. To summarize, all the contributions presented in this book come to support the concept of treating image and text in a similar manner.

It is worth mentioning that the studied methods exhibit state-of-the-art performance levels in the approached tasks. A few arguments come to support this claim. First of all, an improved bag of visual words model described in this work obtained the fourth place at the Facial Expression Recognition (FER) Challenge of the ICML 2013 Workshop in Challenges in Representation Learning (WREPL). Second of all, the system based on string kernels presented in this book ranked on third place in the closed Native Language Identification Shared Task of the BEA-8 Workshop of NAACL 2013. Third of all, the PQ kernel for visual word histograms described in this work received the Caianiello Best Young Paper Award at ICIAP 2013. Together, these achievements reflect the significance of the methods described in the present book.

Contents

List of Figures

List of Tables

Chapter 1
Motivation and Overview

1.1 Introduction

Machine learning is a branch of artificial intelligence that studies computer systems that can learn from data. In this context, learning is about recognizing complex patterns and making intelligent decisions based on data. In the early years of artificial intelligence, the idea that human thinking could be rendered logically in a numerical computing machine emerged, but it was unclear if such a machine could model the complex human brain, until Alan Turing proposed a test to measure its performance in 1950. The Turing test states that a machine exhibits human-level intelligence if a human judge engages in a natural language conversation with the machine and cannot distinguish it from another human. Despite the fact that intelligent machines that can pass the Turing test have not been developed yet, many interesting and useful systems that can learn from data have been proposed since then.

Several learning paradigms have been proposed in the context of machine learning. The two most popular ones are supervised and unsupervised learning. *Supervised learning* refers to the task of building a classifier using labeled training data. The most studied approaches in machine learning are supervised and they include Support Vector Machines (Cortes and Vapnik 1995), Naïve Bayes classifiers (Manning et al. 2008), neural networks (Bishop 1995; Krizhevsky et al. 2012; LeCun et al. 2015), Random Forests (Breiman 2001), and many others (Caruana and Niculescu-Mizil 2006). *Unsupervised learning* refers to the task of finding hidden structure in unlabeled data. The best known form of unsupervised learning is *cluster analysis*, which aims at clustering objects into groups based on their similarity. Among the other learning paradigms are *semi-supervised learning*, which combines both labeled and unlabeled data, and *reinforcement learning*, which learns to take actions in an environment in order to maximize a long-term reward. Depending on the desired outcome of the machine learning algorithm or on the type of training input available for an application, a particular learning paradigm may be more suitable than the others.

R.T. Ionescu and M. Popescu, *Knowledge Transfer between Computer Vision and Text Mining*, Advances in Computer Vision and Pattern Recognition,
DOI 10.1007/978-3-319-30367-3_1

Machine learning is currently a vast area of research with applications in a broad range of fields, such as computer vision (Fei-Fei and Perona 2005; Forsyth and Ponce 2002; Sebastiani 2002; Zhang et al. 2007), bioinformatics (Dinu and Ionescu 2013; Inza et al. 2010; Leslie et al. 2002) information retrieval (Chifu and Ionescu 2012; Ionescu et al. 2015b; Manning et al. 2008), natural language processing (Lodhi et al. 2002; Popescu and Grozea 2012; Sebastiani 2002), and many others (Ionescu et al. 2015a). Among the variety of state-of-the-art machine learning approaches for such applications are the similarity-based learning methods (Chen et al. 2009).

This book studies similarity-based learning approaches such as nearest neighbor models, kernel methods (Shawe-Taylor and Cristianini 2004), and clustering algorithms. The studied approaches have interesting applications and exhibit state-of-the-art performance levels in two different areas: computer vision and string processing. It is important to note that, in this book, *string processing* refers to any task that needs to process string data such as text documents, DNA sequences, and so on. This work investigates string processing tasks ranging from phylogenetic analysis (Ionescu 2013) and sequence alignment (Dinu et al. 2014) to native language identification (Ionescu et al. 2014b; Popescu and Ionescu 2013) and text categorization by topic, from a machine learning perspective. These tasks belong to one of two separate fields, namely text mining or computational biology, but they are gathered under one umbrella called string processing. On the other hand, a broad variety of computer vision tasks are also investigated in this book, including object recognition (Ionescu and Popescu 2013b, 2015a, b), facial expression recognition (Ionescu et al. 2013), optical character recognition (Dinu et al. 2012; Ionescu and Popescu 2013a) and texture classification (Ionescu et al. 2014a). While all the topics enumerated so far seem to be unrelated, each and every one of them includes at least a concept that is borrowed from the other fields of study covered by this book. Further details about this transfer of knowledge between domains are given in the following section. Before going into the next section, it is worth mentioning that the core part of this book is mostly based on recently published works by the authors, yet, it also includes (previously) unpublished work and results.

1.2 Knowledge Transfer between Image and Text

In recent years, computer science specialists are faced with the challenge of processing massive amounts of data. The largest part of this data is actually unstructured and semi-structured data, available in the form of text documents, images, audio files, video files, and so on. Researchers have developed methods and tools that extract relevant information and support efficient access to unstructured and semi-structured content. Such methods that aim at providing access to information are mainly studied by researchers in machine learning and related fields. In fact, a tremendous amount of effort has been dedicated to this line of research (Agarwal and Roth 2002; Lazebnik et al. 2005, 2006; Leung and Malik 2001; Manning et al. 2008). In the context of machine learning, the aim is to obtain a good representation of the data that can later

be used to build an efficient classifier. In computer vision, image representations are obtained by *feature detection* and *feature extraction*. Most of the feature extraction methods are handcrafted by researchers that have a good understanding of the application and a vast experience. This is the case of the bag of visual words model (Leung and Malik 2001; Sivic et al. 2005) in computer vision. A different approach is *representation learning*, which aims at discovering a better representation of the data provided during training. This is the case of deep learning algorithms (Bengio 2009; LeCun et al. 2015; Montavon et al. 2012) that aim at discovering multiple levels of representation, or a hierarchy of features. Deep algorithms learn to transform one representation into another, by better disentangling the factors of variation that explain the observed data.

Whether the representation of the data is obtained through a handcrafted method or learned by a fully automatic process, common concepts of treating different kinds of unstructured and semi-structured data, such as image and text, naturally arise. Despite the fact that computer vision and string processing seem to be unrelated fields of study, the concept of treating image and text in a similar fashion has proven to be very fertile for several applications. Furthermore, by adapting string processing techniques to image analysis or the other way around, knowledge from one domain can be transferred to the other.

An example of similarity between text and image is discussed next. It refers to word sense disambiguation and object recognition in images. *Word sense disambiguation* (WSD) is a core research problem in computational linguistics and natural language processing, which was recognized since the beginning of the scientific interest in machine translation, and in artificial intelligence, in general. WSD is about determining the meaning of a word in a specific context. Actually all the WSD methods use the context to determine the meaning of an ambiguous word, because the entire information about the word sense is contained in the context (Agirre and Edmonds 2006). The basic concept is to extract features from the context that could help the WSD process. In a similar fashion, an object in an image can be recognized using the entire image as a context. For example, a method that could detect the presence of a kitchen glove in the image would have to look for distinctive features such as the texture of the material, the shape, and perhaps even the color. However, there could be other objects that have similar shape or color, and in more difficult situations, such as illustrated in Fig. 1.1a, it may be almost impossible to distinguish the glove. Thus, a better approach could be to look for other distinctive features in the image provided by the context. For instance, a human can easily figure out that a glove is hanging by a kitchen cabinet knob in the scene illustrated in Fig. 1.1b. It is easier to understand the entire scene as a whole than taking the glove out of context. In conclusion, the idea of using the context can help to avoid any confusion. Not surprisingly, this intuitive idea has already been studied in the computer vision literature (Galleguillos and Belongie 2010; Rabinovich et al. 2007). In (Rabinovich et al. 2007), the semantic context is incorporated into object categorization to reduce ambiguity in objects' visual appearance and to improve accuracy. The paper of (Galleguillos and Belongie 2010) goes even further and makes a distinction between three types of context, namely semantic context, spatial context, and scale context.

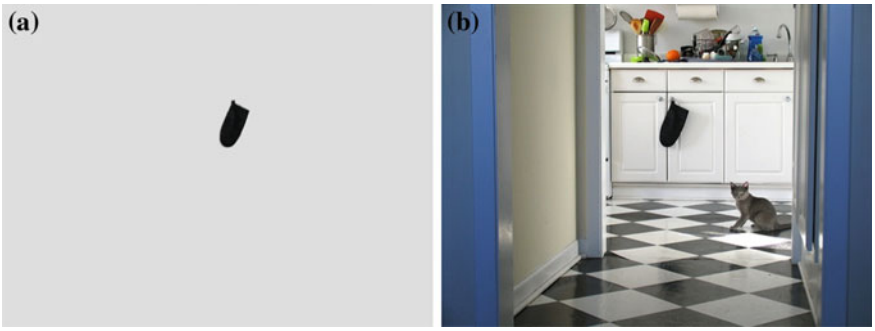

Fig. 1.1 An example in which the context helps to disambiguate an object (kitchen glove), which can easily be mistaken for something else if the rest of the image is not seen. The image belongs to the Pascal VOC 2007 data set. **a** A picture of a kitchen glove. **b** A picture of the same glove with context

Another example of treating image and text in a similar manner is a state-of-the-art method for image categorization and image retrieval inspired by the *bag of words* representation, which is very popular in information retrieval and natural language processing. The bag of words model represents a text as an unordered collection of words, completely disregarding grammar, word order, and syntactic groups. The bag of words model has many applications from information retrieval (Manning et al. 2008) to natural language processing (Manning and Schütze 1999) and word sense disambiguation (Agirre and Edmonds 2006; Chifu and Ionescu 2012). In the context of image analysis, the concept of *word* needs to be somehow defined. Computer vision researchers have introduced the concept of *visual word*. Local image descriptors, such as SIFT (Lowe 1999), are vector quantized to obtain a vocabulary of visual words. The vector quantization process can be done, for example, by k-means clustering (Leung and Malik 2001) or by probabilistic Latent Semantic Analysis (Sivic et al. 2005). The frequency of each visual word is then recorded in a histogram which represents the final feature vector for the image. This histogram is the equivalent of the bag of words representation for text. The idea of representing images as *bag of visual words* has demonstrated very good performance for image categorization (Zhang et al. 2007) and image retrieval (Philbin et al. 2007).

One of the most important problems in computer vision is object recognition. Machine learning methods represent the state-of-the-art approach for the object recognition problem. A common approach is to make some assumptions in order to treat object recognition as a classification problem. First, object categories are considered to be fixed and known. Second, each instance belongs to a single category. However, some researchers argue that these assumptions do not adequately describe the reality. The following example shows that these assumptions are indeed wrong. The object presented in Fig. 1.2 can be described either as a toy, a bear, or both. It is clear that the object does not belong to a single category. Furthermore, the category of the object might be irrelevant for particular applications. Another