Learning from Unstructured Multimedia Data

Janani Kalyanam
University of California, San Diego
9500 Gilman Drive, La Jolla
CA, USA
jkalyana@ucsd.edu

Gert Lanckriet
University of California, San Diego
9500 Gilman Drive, La Jolla
CA, USA
gert@ece.ucsd.edu

ABSTRACT

Information in today's world is highly heterogeneous and unstructured. Learning and inferring from such data is challenging and is an active research topic. In this paper, we present and investigate an approach to learning from heterogeneous and unstructured multimedia data. Inspired by approaches in many fields including computer vision, we investigate a histogram based approach to represent multimodal unstructured data. While existing works have predominantly focused on histogram based approaches for unimodal data, we present a methodology to represent unstructured multimodal data. We explain how to discover the prototypical features or codewords over which these histograms are built. We present experimental results on classification and retrieval tasks performed on the histogram based representation.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

unstructured data, bag-of-words model

1. INTRODUCTION

An increasingly important aspect of information in today's world is that it is highly unstructured and heterogeneous. For example, a search for "government shutdown" yields fundamentally different modes of information about the topic via news articles, blogs, videos, pictures e.t.c. While there are some commonalities in the information provided across these modalities (i.e., they are about the government shutdown) the nature of information differs widely amongst the modalities. Moreover, such information is often available in a variety of unstructured formats: webpages may contain one or more pictures, or none, embedded videos, and/or audio tracks e.t.c. Ideally, an algorithm to make inferences from the data (e.g., whether or not the webpage supposes a government shutdown) would be given access to all available information, since each source may bring some unique information to the table. Machine learning with such unstructured heterogeneous data is challenging and an active topic of research. Existing approaches assume a rigidly structured input format. They cannot accommodate missing modalities (incomplete information) or modalities with multiple instances (over-complete information) [2].

We present and investigate an approach to learn from heterogeneous, unstructured multimedia data. Inspired by its success in fields like computer vision, we investigate a histogram based approach to represent multimodal unstructured data. Broadly speaking, such techniques have been predominantly developed for unimodal data. At the core of such techniques exist the mission of identifying recurring patterns (a.k.a. codewords, or prototypical features) across a large collection of data. Once such patterns have been identified, every datapoint is quantized and represented in terms of these patterns.

However, working with heterogeneous, unstructured and irregular data poses several new challenges that are not addressed by a direct extension of the said unimodal technique. In the unimodal case, a clustering algorithm is usually employed to identify the prototypical features. In our case, the data points may reside in a variety of feature spaces. Moreover, the lack of structure in the data precludes the usage of any kernel based clustering as well.

The prototypical features discovered from unstructured heterogeneous data themselves happen to be a collection of unimodal codewords. We discover these collections by clustering the unimodal codewords that often co-occur (or are highly correlated). In other words, after finding the codewords in each modality separately using existing techniques, we cluster them to form what are called heterogeneous codewords or heterogeneous prototypes. After discovering the prototypes, the data is quantized and represented as histograms over the prototypes.

2. DATA REPRESENTATION

The data that we are interested in is multimodal and unstructured data like the contents of a webpage (Figure 1). Here, each object-set is a *group* of datapoints from different modalities. Each datapoint in an object-set has a unimodal codeword-based representation. Most codeword based techniques attempt to uncover recurring interesting patterns across datapoints by some means of clustering over

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04
http://dx.doi.org/10.1145/2567948.2577318.

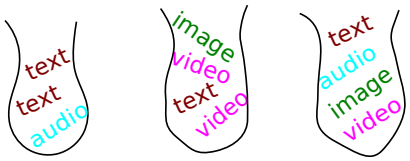

Figure 1: Each 'bag' (or an *object-set*) in this figure can be thought of as a webpage. Each webpage contains varying amounts of multimedia objects.

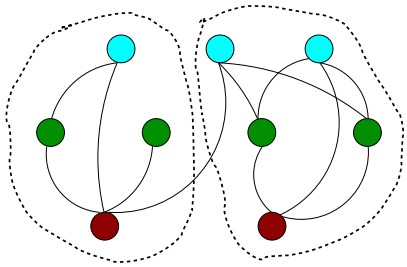

Figure 2: Each node in the graph is a unimodal or homogeneous codeword. There are 3 modalities. Nodes of the same color belong to the same modality. There are two clusters in this graph. Each cluster is a *heterogeneous codeword*.

the features. But when the datapoints reside in different spaces as in our problem setting, simultaneously clustering them becomes a challenge. Nonetheless, note that the features in the individual modalities have already been clustered in order to produce the homogeneous codeword-based representation. We can therefore exploit this fact, and travel to a higher level of abstraction and focus on clustering the *codewords* across modalities as opposed to clustering the *features* across modalities.

To cluster codewords across modalities, we can build a graph with codewords from the different modalities as the nodes, and some measure of affinity between the codewords as the weighted edges (Figure 2). This measure of affinity would, in some sense, be representative of how correlated the codewords within each object-set are. The important takeaway here is that working at the *codeword-level* makes it easier to model similarity between codewords across modalities. This is because the unimodal data themselves have been represented as a *histogram* over codewords and hence, modeling similarity becomes fairly straight forward (e.g. correlation). Once we have a good similarity measure at hand, clustering falls into place [1].

After discovering the codewords, all data is represented as a histogram over the codewords (generally through hard or soft quantization). Intuitively, the purpose of the histogram is to measure the ratio of the object-set that belongs to a particular codeword. Concatenating these ratios into a single vector is the histogram based representation of the object. In order to achieve this, we simply average the contributions from all the datapoints from an object-set towards a particular heterogeneous codeword, yielding the histogram based representation for the data.

[1]If the data is complete (without any missing or overcomplete information), feature concatenation or equivalently kernel addition can be used to cluster datapoints that reside in different feature spaces. However, the unstructured quality in the data precludes the usage of such approaches.

3. EXPERIMENTS AND CONCLUSION

Category	Our approach	Baseline
Architecture	**0.6281**	0.3061
Biology	**0.9359**	0.9151
Geography & Places	**0.7472**	0.7285
History	0.8219	**0.8431**
Literature & Theater	**0.4530**	0.4237
Media	**0.7858**	0.4878
Music	**0.8704**	0.8641
Royalty & nobility	0.8369	**0.8526**
Sports & recreation	**0.8894**	0.8647
Warfare	0.7798	**0.8642**
Average	**0.7742**	0.7156

Table 1: Category-level MAP scores for multimodal retrieval tasks.

While the central idea of histogram based approaches presented in this paper remains intact for all kinds of data, we consider predominantly textual and image data for experiments. We worked with Wikipedia Featured Articles and performed classification and retrieval. Each article is an object-set. We begin by representing each image as a histogram over SIFT descriptors, and each text as a histogram over LDA topics. After that, we build a graph as in Figure 2, and use [1] for clustering the unimodal codewords. After discovering the heterogeneous codewords, we represent each article as a histogram over the heterogeneous codewords. Each article was classified into one of 10 classes using logistic regression. As a baseline, we classified each image and text individually using logistic regression. We took a majority vote of all the labels in the object-set in order to obtain the label of an object-set. Our approach has a higher accuracy of 86.07% than the baseline accuracy of 82.86%.

For the multimodal retrieval task, after obtaining a heterogeneous codeword based representation for each article, we obtain a 10-class semantic multinomial by applying logistic regression. On this semantic multinomial, the cosine similarities between the query and the rest of the data points were calculated, and a sorted ranking was obtained. Mean Average Precision (MAP) scores were calculated on each of the rankings, and MAP scores were aggregated across all queries. In the baseline, for each example in each article, we obtain a 10-class semantic multinomial by subjecting it through a logistic regression. The 10-class semantic multinomial of an article is the average of all the examples contained in the article. Results are summarized in Table 1.

We conclude by stating that being able to learn from unstructured multimodal data is an important problem, and an active research topic. In this paper, we have presented a simple, and efficient approach to represent such data.

References

[1] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD*, pages 269–274, 2001.

[2] G. Ye, I.-H. Jhuo, D. Liu, Y.-G. Jiang, D. T. Lee, and S.-F. Chang. Joint audio-visual bi-modal codewords for video event detection. In *ICMR*, page 39, 2012.