Formulating Queries for Collecting Training Examples in Visual Concept Classification

Rami Albatal, Kevin McGuinness, Feiyan Hu, Alan F. Smeaton
Insight Centre for Data Analytics
Dublin City University
Glasnevin, Dublin 9, Ireland.
{rami.albatal}@insight-centre.org

Abstract

Video content can be automatically analysed and indexed using trained classifiers which map low-level features to semantic concepts. Such classifiers need training data consisting of sets of images which contain such concepts and recently it has been discovered that such training data can be located using text-based search to image databases on the internet. Formulating the text queries which locate these training images is the challenge we address here. In this paper we present preliminary results on TRECVid data of concept classification using automatically crawled images as training data and we compare the results with those obtained from manually annotated training sets.

1 Introduction

Content-based access to video archives is based on learning the presence of semantic concepts in video content by mapping low-level features like colour and texture, to high-level concepts. Concept classification is typically based on training classifiers on a set of annotated ground truth images (called a training set) containing positive and negative example images of a given semantic concept. The manual creation of training sets for each concept is a time-consuming and costly task. An alternative is to automatically gather training examples using available resources on the Internet. Several recent papers have demonstrated the effectiveness of such an approach. (Griffin, 2006) used search engine results to gather material for learning the appearance of categories, (Chatfield and Zisserman, 2013) shows that effective classifiers can be trained on-the-fly at query time using examples collected from Google Image search. The AXES research search engine (McGuinness et al., 2014) uses a combination of pre-trained classifiers and on-the-fly classifiers trained using examples from Google Image search. (Kordumova et al., 2014) investigate four practices for collecting training negative and positive examples from socially tagged videos and images.

The above work exploits the visual content of the collected example images while the question of how to formulate a textual query for collecting the data is not yet considered. It is important to note here that current search engines are not use content-based image classifiers, they are based the text from the embedding pages, and that is not always accurate or scalable. This represents a unique relationship between vision (the images used to train a concept classifier) and language (the text used to find those training images). In this work, we initiate a first step to addressing the problem of formulating text queries that collect positive example images for classifier training. This first step is based on querying web resources with single-term queries and comparing the classification results with those from manually annotated training sets. The results show the potential of automatic crawling and open the way for enhancing query formulation by adding external lexical or semantic resources.

2 Automatic gathering of training examples

Our goal is to create a framework for automatic training of classifiers by gathering training examples from available resources on the Internet. The steps for formulating a query are straightforward and widely-used in information retrieval, especially query expansion. First, an initial query is pre-processed by removing stop words and applying stemming techniques; then external lexical and/or semantic bases are explored in order to enrich the query and add useful terms which help in retrieving relevant training examples and excluding false positives. The resulting query is then posted to a search engine or image databases and the retrieved images are used as positive examples for a classification algorithm. Our plan is to use the Natural Language Toolkit (Bird et al., 2009) for stemming and stop words removal, and WordNet as external lexical base (Fellbaum, 1998).

3 Experiments and results

Experiments were conducted on the TRECVid (Smeaton et al., 2006) 2014 semantic indexing development data set. Single-term queries were posted to two data sources: Google Images, and ImageNet (an image database organized according to the nouns of the WordNet hierarchy where each node is depicted by an average of +500 images). Unlike results from Google Images, examples gathered from ImageNet are classified by human annotators, and are therefore a "purer" source of training images. To ensure a high-quality training set, we first search for the concept in ImageNet; if the concept does not exist in as an ImageNet visual category, we use images retrieved using a search for the term on Google Images.

We carried out two experiments to evaluate the performance of classifiers trained on manually annotated (internal) data provided by TRECVid versus data gathered from external sources. These external sources are search engines that retrieve images using textual queries, as explained in section2. The first experiment used data from the 2013x subset of the TRECVid 2014 development data and the second used external training data gathered as discussed above in the first paragraph of this section. Accuracy in both cases was evaluated using inferred average precision (infAP) on the 2013y subset of the development data. One-vs-all linear SVM classifiers were used for both experiments, trained on visual features extracted using pre-trained deep convolutional neural networks (CNN) using the Caffe software (Jia, 2013).

Classifiers for 34 of the 60 concepts were trained using data from ImageNet and the remaining using examples from Google Images. All classifiers trained using images from Google Images demonstrated poorer infAP than those trained on internal data. Of the 34 classifiers trained on ImageNet, 7 demonstrated improved infAP (*airplane, beach, bicycling, classroom, computers, dancing, flowers, highway*). In all cases it was possible to find more positive examples on ImageNet than in the internal set. Internal out-performed ImageNet in the remaining 27 cases. There were several possible reasons for this. In many cases there were fewer examples from ImageNet than in the internal set (12/27 cases) and in some cases the ImageNet examples were incorrect. For example, in the case of the concept *"hand"*, several synsets matching the term consisted entirely of wristwatches. Finally, in other cases, the concept text (the query) was either ambiguous or insufficiently semantically rich, for example *"greeting"* (greeting cards were retrieved) and *"government leader"* (smaller subset of such leaders in internal training data).

4 Hypothesis, challenges, and future work

The experiments indicate that automatically-gathered external training data can, in some cases, outperform annotated internal training data when it is sufficiently plentiful and of high quality. Using both high-quality external data and internal examples during training has the potential to improve results overall. A more sophisticated method of gathering external training examples that takes into account the semantics of the concept and of related concepts could provide even higher-quality external data. A significant challenge is in combining such semantic query expansion with visual analysis to ensure that the additional examples collected are relevant. This could potentially be achieved by bootstrapping a classifier on internal examples and then using this to classify external examples gathered by iterative semantic query expansion, updating the classifier model with each batch of accepted training examples.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Ken Chatfield and Andrew Zisserman. 2013. Visor: Towards on-the-fly large-scale object category retrieval. In *Computer Vision ACCV 2012*, volume 7725 of *Lecture Notes in Computer Science*, pages 432–446.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Lewis D Griffin. 2006. Optimality of the basic colour categories for classification. *Multimedia Tools and Applications*, 3.6:71–85.

Yangqing Jia. 2013. Caffe: An open source convolutional architecture for fast feature embedding. `http://caffe.berkeleyvision.org/`.

S. Kordumova, X. Li, and C. G. M. Snoek. 2014. Best practices for learning video concept detectors from social media examples. *Multimedia Tools and Applications*, pages 1–25, May.

Kevin McGuinness, Robin Aly, Ken Chatfield, Omkar Parkhi, Relja Arandjelovic, Matthijs Douze, Max Kemman, Martijn Kleppe, Peggy Van Der Kreeft, Kay Macquarrie, et al. 2014. The AXES research video search system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Alan F Smeaton, Paul Over, and Wessel Kraaij. 2006. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330. ACM.