MULTI-CONCEPT LEARNING WITH LARGE-SCALE MULTIMEDIA LEXICONS

Lexing Xie[†], *Rong Yan*[†], *Jun Yang*[‡] *

[†]IBM T. J. Watson Research Center, [‡]Carnegie Mellon University

ABSTRACT

Multi-concept learning is an important problem in multimedia content analysis and retrieval. It connects two key components in the multimedia semantic ecosystem: multimedia lexicon and semantic concept detection. This paper aims to answer two questions related to multi-concept learning: does a large-scale lexicon help concept detection? how many concepts are enough? Our study on a large-scale lexicon shows that more concepts indeed help improve detection performance. The gain is statistically significant with more than 40 concepts and saturates at over 200. We also compared a few different modeling choices for multi-concept detection: generative models such as Naive Bayes performs robustly across lexicon choices and sizes, discriminative models such as logistic regression and SVM performs comparably on specially selected concept sets, yet tend to over-fit on large lexicons.

Index Terms— Multimedia computing, Pattern recognition, Multimedia databases

1. INTRODUCTION

Multimedia lexicon, semantic concept detection, and multimodal search are three key components in the multimedia retrieval ecosystem. Detecting objects and scenes from visual input has long been one of the central problems in computer vision. Knowing what should be detected leads to the problem of designing and using a multimedia lexicon. All these components help build better search engines that has become the *de facto* user access model for large online repositories. The afore-mentioned multimedia retrieval ecosystem is visualized in Fig. 1. Many prior work has focused on the links that connect these three components. It has been shown that using concept detectors can significantly improve search results [7], and vice versa [6]. Studies [5, 9] have shown that lexicons of different sizes bring about different levels of improvement in retrieval performance.

In this paper, we investigate multi-concept learning, i.e., how to leverage large-scale multimedia lexicons for semantic concept detection. This problem is of much practical and theoretical interest, on which there has been a number of recent studies. The basic structure of such learning algorithms is to use baseline classifiers such as Support Vector Machine (SVM) to model the non-linearity in the features, and then train another layer of meta-classifier to model the relationships among concepts. Such relationships can be learned with a discriminative classifier such as SVM [1] or logistic regression [4], probabilistic graphical models [12], or discriminative models regularized with concept dependency information [8].

Three questions remain in order to adequately establish the link from multimedia lexicon to concept detectors: Does large-scale lexicons benefit multi-concept learning? How many concepts are enough? What computational models can learn multi-concept detectors robustly and consistently with a large lexicon? Most prior work studied no more than 40 concepts, and did not explore the potentials when the number of concepts grow to a much larger scale. But as a larger number of publicly annotated data become available, e.g., LSCOM [13], it is now possible to further explore these questions. In this paper, we investigate multi-concept learning with the largest multimedia lexicon to-date using both generative and discriminative models. We found that more concepts indeed help improve concept detection performance. The performance gain is statistically significant with more than 40 concepts, and it saturates at around 240. When measuring performance among a collection of concepts, it was somewhat surprising that generative models such as Naive Bayes perform robustly with consistent performance gain across lexicon choices and sizes, while discriminative models such as logistic regression and SVM perform comparably on specially selected concept sets, yet tend to over-fit on a large number of dependent dimensions. Performance gain varies for different target concepts and the nature of baseline input detectors, and these two areas warrant further investigation.

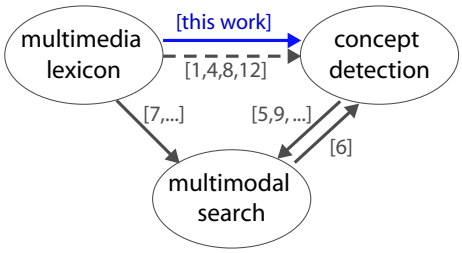

Fig. 1. The multimedia semantic ecosystem.

In the rest of this paper, Section 2 discusses different multi-concept models being used, Section 3 gives the background for obtaining baseline detectors, Section 4 presents our experiments and results, Section 5 concludes the paper.

2. MODELS FOR MULTI-CONCEPT LEARNING

Multi-concept learning is to estimate the label $y_c \in \{1, 0\}$ for concept c from a collection of relevance scores for concepts that are related to the concept c, denoted as $\mathbf{x} = [x_1, \ldots, x_M]$. Multi-concept models need to account for two factors of correlations and uncertainty. The first factor is ontological relationship, e.g., a *bus* is likely to be *outdoors* and unlikely to be in an *office*. The second factor is detector performance, e.g., knowing both the presence/absence of *people* and *desert* may help us decide if it is *military* setup, but *people* detectors are generally more reliable and hence shall be given more weight. Generative and discriminative models are two general classes of models commonly used to account for such uncertainties,

*This material is based upon work funded in part by the U. S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Government.

on which we shall experiment and compare.

2.1. Generative Models: Naive Bayes

Generative models estimate the class-conditional probability in order to find the most likely class based on Bayesian rules. They have efficient learning algorithms and can handle new classes or missing data fairly effectively. Naive Bayes is the simplest in this family, which models the pair-wise correlations between the target concept c and each input concept $i, i = 1, \ldots, M$. The maximum-likelihood estimate of the class-conditional probability $P(x_i|y_c)$ is obtained by averaging the confidence over all the training data on a given target concept y_c. Because naive Bayes models assume that input x_i are conditionally independent given label y_c, we can factorize the joint class-probabilities and use these estimates to obtain the log-likelihood ratio L_c as shown in Equation (2).

$$P(y_c|x_{1:M}) \propto P(y_c) \prod_{i=1:M} P(x_i|y_c) \quad (1)$$

$$\begin{aligned} L_c &= \log \frac{P(y_c=1|x_{1:M})}{P(y_c=0|x_{1:M})} \quad (2) \\ &= \log \frac{P(y_c=1)}{P(y_c=0)} + \sum_{i=1:M} \log \frac{P(x_i|y_c=1)}{P(x_i|y_c=0)} \end{aligned}$$

In our implementation, each input observation x_i is sigmoid-normalized and uniformly quantized into 15 bins, and the maximum likelihood estimates of $P(x_i|y_c)$ is smoothed by the prior of x_i as $\hat{P}(x_i|y_c) = (1-\alpha)P(x_i|y_c) + \alpha P(x_i)$, with $\alpha = 0.06$. The resulting likelihood scores L_c are then used to re-rank the original prediction score with simple averaging.

2.2. Discriminative Models: Logistic Regression and SVM

In contrast to generative models, discriminative models attempt to directly model the probability of a class given the data. Among various discriminative models, logistic regression is one of the most popular choices, which directly models the log-odd function as a linear function of the input observations [3]. If logistic regression is used to model concept relationship, the posterior probability of the target concept y_c can be written as,

$$P(y_c|x_{1:M}) = \frac{1}{Z} \exp\left[\sum_i (w_{0c} + \sum_j w_{ic}x_i)y_c\right], \quad (3)$$

where Z is a normalization factor for the conditional probability. The parameters w_{ic} can be estimated by using any gradient descent methods such as iterative reweighted least squares(IRLS) algorithm. To avoid optimization singularity, a small regularization factor is added to the log-likelihood function. Note that, if $P(x_i|y_c)$ in Naive Bayes chosen to be in exponential family, then it has been shown that logistic regression and Naive Bayes are representing the same family of conditional probability functions.

Support vector machines (SVM) are another type of discriminative models. Built on the structural risk minimization principle, SVMs seek a decision surface that can separate the training data into two classes with the maximal margin between them. In the case of multi-concept learning, the decision function of SVMs is as follows,

$$y_c = sign\left(\sum_{d=1}^{D} y_d \alpha_i K(\mathbf{x}, \mathbf{x}_d) + b\right), \quad (4)$$

where d is the index of training data, $K(\cdot)$ is the kernel function, the weights $\alpha = \{\alpha_1, ..., \alpha_M\}$ and offset b are the model parameters. Linear and RBF kernels are popular choices for modeling.

3. LEARNING INDIVIDUAL CONCEPT DETECTORS

The input to multi-concept learning is related concept scores provided by individual concept detectors. Each concept detector is learned from low-level visual features using SVMs with radial-basis (RBF) kernels. These features include a set of visual descriptors at various granularities for each representative keyframe of the video shots, such as color histogram, color correlogram, color moments, co-occurrence texture, wavelet texture, and edge histogram (see details in [2]). The performance of RBF-SVMs can vary significantly with respect to model parameters, hence the choice of parameters is crucial. To optimize the performance, we choose model parameters using a grid-search strategy. In our experiments, we build models for different values of the RBF kernel parameters, the relative cost factors of positive vs. negative examples, feature normalization schemes, and the weights between training error and margin. The optimal parameters are selected based on average precision using 2-fold cross validation.

Before the learning process, the distribution between positive and negative data are re-balanced by randomly down-sampling the negative data to a smaller amount. For each low-level feature, we select one optimal configuration to generate the concept model. Finally, four best-performed models are combined to be a composite classifier by averaging their results. In the detection stage, we apply the optimal model to evaluate the target images for the presence/absence of the concepts, and generate a confidence measure that can be used to rank the testing images.

4. EXPERIMENTS

We evaluate the multi-concept learning algorithms on a large broadcast video collection from the TRECVID 2007 video retrieval benchmark [10]. This collection contains news magazine, science news, news reports, documentaries, educational programming, and archival videos in MPEG-1 format from the Netherlands Institute for Sound and Vision. There are 209 programs in total with over 120 hours of content, in which 110 were designated as the development set for training and tuning the algorithms, and the rest 109 as the test set for evaluating learning performance. In addition, about 80 hours of annotated multi-lingual news videos from TRECVID-2005 were also used in training concept detectors from extended lexicons.

The TRECVID-2007 development set is partitioned into two halves, with the first to learn the SVM-fusion detectors for individual concepts and the second half to learn the multi-concept detectors with naive Bayes and logistic regression, as described in Sec. 2. Rotating the two halves gives us two multi-concept models per target concept. These models are also fused with the baseline detection score, by normalizing with the logistic function and then averaging.

The performance metric is inferred average precision (infAP) [11]. As an approximation of average precision (AP) which characterizes the area under the precision-recall curve of a binary detector, inferred average precision allows reliable assessments to be made given a smaller set of available ground-truth. Studies have shown that while the absolute performance number may change, infAP especially preserves the relative performance across different systems, and is hence a good metric for benchmark comparison.

4.1. Visual lexicons

Our experiments use a few different versions of visual concept lexicon for training and evaluating the multi-concept detectors, as enu-

merated below,

- LSlite36 [10]. This lexicon consists of 36 visual concepts covering seven essential semantic categories of broadcast content, which include *people*, *activities*, *scene*, *objects* and etc. A handful of representative concepts are selected in each category, such as *crowd*, *people-marching*, *court*, *car*. A collaborative annotation effort was completed by the TRECVID-2007 participants to annotate each of the 21,532 shots in the development set with this lexicon.
- LS364. 364 visual concepts selected from the LSCOM (Large Scale Concept Ontology for Multimedia) ontology. This ontology was designed to broadly cover visual semantics in broadcast news, and meet additional criteria relating to utility (usefulness), observability (by humans), and feasibility (by automatic detection). The detectors are built by averaging three SVMs learned on different visual descriptors. Their training data are from TRECVID-2005. Detection scores on the TRECVID-2007 collection were donated as a shared resource for TRECVID-2007 [13], and 10 concepts that either overlap with LSlite36 or has missing detection results are excluded from the original 374-concept set.
- LS157. A subset of 157 LSCOM concepts from LS364. It is selected based on relevance to the sound-and-vision data, presence in the collection and possibility for building reasonable detectors. The detectors were built with a process described in Sec. 3 and an earlier report [2].
- LS157b. This is the same set of concepts as in LS157, with detectors from LS364. The purpose of this subset is to validate the effect of different training setup on multi-concept detection performance.
- LSlite10. This is a subset of the LSlite36 lexicon. Partial ground truth were available for 20 out of the original 36 concepts in the benchmark. Then we selected 10 most-frequent ones since they have more reliable detection outputs to compare different experiments – few of the other ten concepts yielded statistically significant results in comparisons described in the rest of this section. These concepts are: *waterscape-waterfront*, *car*, *computer/TV screen*, *boat-ship*, *animal*, *office*, *meeting*, *truck*, *airplane*, *sports*.

4.2. Performance versus lexicon size

We first conduct an experiment to see whether or not multi-concept learning can improve detection performance, and see what new observations and conclusions can be made with a large-scale lexicon.

This experiment uses the union of LSlite36 and LS364 as input. The *baseline* performance is measured on the TRECVID-2007 test set from the concept detectors in LSlite36. We randomly sample K concepts from a total of 400 concepts (LSlite36+LS364), $K = 4, 8, \ldots, 380, 400$. We repeat the random sampling 5 times, and plot the mean-infAP over LSlite10 concepts in Fig. 2. Two sets of *t*-tests are conducted for naive Bayes, results are shown below the x-axis in Fig. 2. The top row tests if the mean-infAP is greater than the baseline, and the bottom row tests if the mean-infAP is less than the best mean-infAP with all 400 input concepts. The detector performance vs. lexicon size, along with their statistical properties, can help us answer the following two key questions.

Does multi-concept models help? Yes. As can be seen from Fig. 2, mean-infAP is improved from 0.103 to 0.110 with a statistically significant margin. Moreover, the naive Bayes model exhibits

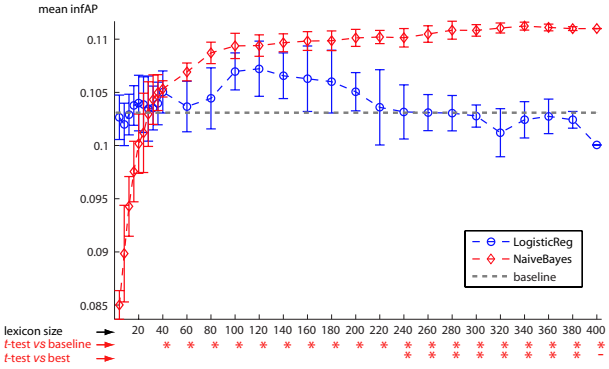

Fig. 2. Mean infAP vs lexicon size. The error bars center at the mean and stretch twice the standard deviation. *t*-tests with a confidence level more than 0.05 is marked with an asterisk (*). See Sec. 4.2 for detailed discussions.

robust performance gain and consistent improvement with the increase of the lexicon size, while the performance of logistic regression drops and significantly fluctuates as the lexicon becomes larger. This performance contrast lend itself to two reasons: 1) the concept dimensions are highly dependent on any large corpus; 2) naive Bayes estimates the conditional probability given each input dimension independently, while logistic regression optimizes training error by simultaneously adjusting weights on all dimensions, thus it is prone to be over-fitting on a lot of dependent inputs. We also observed that SVMs suffer from similar degradations as logistic regression. Their results are not included because the model training did not terminate in time. SVM-based models takes several hours with cross-validated hyper-parameter selection, while naive Bayes and logistic regression finish within minutes for all target concepts, taking only a fraction of learning time as SVM.

How many concepts do we need? 40 or 240, depending on required performance. Looking at the performance vs. lexicon size by naive Bayes models in Fig. 2, we can see that the mean-infAP increases as the lexicon size increase, and the benefits seems to diminish beyond 200 concepts. The statistical test results (shown below the x-axis) tell us that concept detection performance is significantly better than baseline for lexicon size greater than 40, and the performance difference larger than 240 is statistically insignificant.

4.3. Per-concept performance with different lexicon

We now discuss another experiment that examines the effects of using different learning models, or different choices of lexicons and input detectors.

This experiment involves a few different lexicon and their corresponding baseline detectors. LSlite36 is a manually selected set of only a few dozens of concepts, on which most prior studies were based [12, 8]. LS193 is a combination of LSlite36 and LS157, expanding the former to a much richer selection of concepts, trained using one version of SVM-fusion model. CU193 contains the same set of concepts as LS193, except with detectors trained with the LS374 set. CU400 is the same as the one used in Sec. 4.2, which is the largest lexicon available. We evaluate naive Bayes, logistic regression and SVM with two different kernels on this data. The results are visualized in Fig. 3.

From the performance of naive Bayes (gray bars) in Fig. 3 we can probe into the following two questions: *Is a hand-selected lexicon better?* No, both logistic regression and naive Bayes on

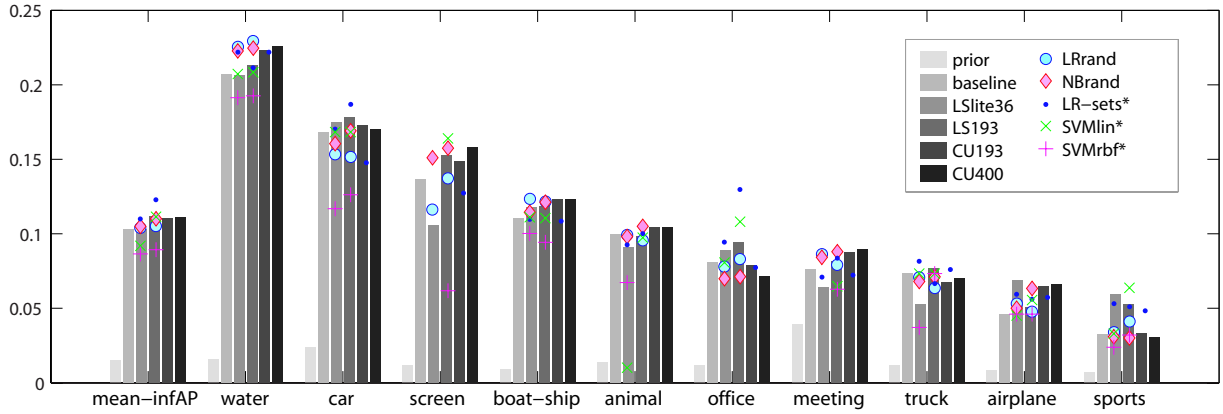

Fig. 3. Per-concept performance for different visual vocabularies or computational models. The gray-scale bars are concept prior, SVM-fusion baseline, as well as Naive Bayes model with four different versions of lexicon/input detector combination described in Sec. 4.3. The LRrand and NBrand markers are respective averages of infAP of randomly selected vocabularies of size 36 and 200 from CU400 (Fig. 2), using logistic regression and Naive Bayes models. LR-sets*, SVMlin-sets* and SVMrbf-sets* are infAP obtained by logistic regression, SVM with linear and RBF kernels on the LSlite36 and LS193 sets, respectively. These markers are only shown when statistically different from LRrand, aligned with bars from their respective lexicons.

the LS193/CU193 are comparable to their randomly-selected counterparts with respect to mean-infAP. On selected concepts such as *office*, both LS193 and CU193 models outperformed the full CU400 set, indicating that more concepts than necessary can hurt performance. An extreme case was *truck*, where none of the additional input concepts seemed to help. *Does different features/detectors make a difference?* Again the answer is concept-dependent – *sports*, for example, clearly benefited from the sharing of LSlite36 and LS157, and hurt by the system choices in LS374. On the contrary, *waterfront-waterscape*, *meeting* and *animal* performed better with the LS374 detectors.

From the color markers in Fig. 3, we can observe the behaviors of different modeling algorithms. On average (in terms of mean-infAP) the models perform similarly. Naive Bayes is the most stable across all lexicon choices and sizes. Logistic regression does a bit better on the manually-pruned lexicons, but over-fits in high dimensions. It is worth noting that each concept has significant improvement over its baseline for a few best-performing models, although relative performance on each concept tend to vary. This thus calling for concept-specific modeling that takes into account higher-order concept dependency and possibly a blend of models and visual lexicon.

5. CONCLUSION

This paper investigates multi-concept learning, i.e. leveraging large-scale multimedia lexicons for visual concept detection. This is an important problem for linking multimedia lexicon and semantic concept detection, two key components in the multimedia semantic ecosystem. We confirmed that a large-scale lexicon indeed help improve detection performance with statistical significance, and about two hundred concepts can generate maximum benefit. Areas of future interest include: concept-specific modeling, automatic selection of lexicon and combination strategies of different models.

6. REFERENCES

[1] A. Amir et al. IBM Research TRECVID-2003 video retrieval system. In *Worshop of TRECVID 2003*, 2003.

[2] M. Campbell et al. IBM research TRECVID-2007 video retrieval system. In *NIST TRECVID Workshop*, Gaithersburg, MD, November 2007.

[3] F. C. Gey. Inferring probability of relevance using the method of logistic regression. In *Proc. ACM SIGIR'94*, pages 222–231, Dublin, Ireland, 1994. Springer-Verlag New York, Inc.

[4] A. Hauptmann et al. Confounded Expectations: Informedia at TRECVID 2004. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD, 2004.

[5] A. Hauptmann et al. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *Multimedia, IEEE Transactions on*, 9(5):958–966, 2007.

[6] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev. To search or to label?: predicting the performance of search-based automatic image classifiers. In *ACM MIR Workshop 2006*, pages 249–258, 2006. ACM.

[7] A. P. Natsev et al. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 991–1000, New York, NY, USA, 2007. ACM.

[8] G.-J. Qi et al. Correlative multi-label video annotation. In *Proc. of the 15th ACM Int'l Conf. on Multimedia*, pages 17–26, 2007.

[9] C. G. M. Snoek et al. Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia*, 9(5):975986, August 2007.

[10] The National Institute of Standards and Technology (NIST). TREC video retrieval evaluation, 2001–2007. http://www-nlpir.nist.gov/projects/trecvid/.

[11] NIST. Inferred average precision and TRECVID-2006, 2006. http://www-nlpir.nist.gov/projects/tv2006/infAP.html.

[12] R. Yan, M.-Y. Chen, and A. G. Hauptmann. Mining relationship between video concepts using probabilistic graphical model. In *IEEE Int'l Conf. on Multimedia and Expo*, 2006.

[13] A. Yanagawa et al. Columbia University's baseline detectors for 374 LSCOM semantic visual concepts. Technical report, Columbia University, March 2007.