Active Learning in Multiple Modalities for Semantic Feature Extraction from Video

Ming-yu Chen and Alexander Hauptmann

School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, Pennsylvania
{mychen, alex}@cs.cmu.edu

Abstract

Active learning has been demonstrated to be a useful tool to reduce human labeling effort for many multimedia applications. However, most of the previous work on multimedia active learning has gloss the multi-modality problem very much. From several experimental results, multi-modality fusion plays an important role to boost performance of multimedia classification. In this paper, we present a multi-modality active learning approach which enhances the process of active learning approach from single-modality to multi-modality. The experimental results on the TRECVID 2004 semantic feature extraction task show that the proposed active learning approach works more effectively than single-modality approach and also demonstrate a significantly reduced amount of labeled data.

Introduction

As the amount of available multimedia information increases, it is driving the demand for content-based access to video data. Machine learning approaches show potential for multimedia classification and retrieval, but a large amount of annotated data is necessary for the typical training process. Unfortunately, manually annotating training data is not only labor and time consuming, but also subject to human errors. A fast way to reduce human effort is to annotate training data randomly, and propagate the labels to the whole training set though supervised learning algorithms. People only have to annotate an initial set and then correct the results from each iteration of supervised learning. However, random sampling can't always provide all the necessary data to annotate and it still requires a lot of time for people to label. Active learning provides an approach to non-random selection of data for annotation. The basic idea of the active learning approach is to select that data for annotation which will provide the most information for the learning algorithm in the next round. The training data to be selected for annotation should be that part which cannot be clearly explained by the current model.

The effectiveness of active learning to reduce labeling cost has been demonstrated by previous work in many contexts. An active learner may begin with a pool of unlabeled data, select a set of unlabeled examples to be manually labeled as positive or negative and learn from the newly obtained knowledge repetitively. This type or style of

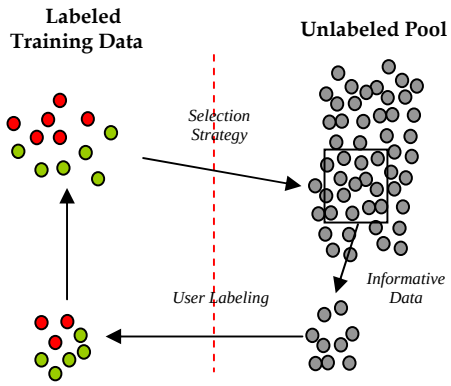

Fig. 1. Illustration of active learning

problem can also be called "query learning" or "selective sampling" [1]. Typically, the unlabeled examples can be selected by either minimization of the learner's expected error or maximization of the information gain or version space reduction [2, 3]. Figure 1 illustrates the basic process of the active learning approach.

Semantic feature extraction has become an interesting and challenging research issue in recent years. Researchers are trying to label video segments with semantic concepts and then utilize this information to achieve video understanding and multimedia retrieval. Multimedia classification is the main method for semantic feature extraction. As demonstrated by previous experimental results [4, 5], fusion of multiple modality features cam play an important role regarding performance. Multi-modality fusion is defined as combining classification results from multiple sources of multimedia content, like color, texture, edge, audio signal and many other low level features sets extracted from a multimedia source. Following this general approach, we want to extend multi-modality fusion into an active learning process and present this as multi-modality active learning approach. In

previous multimedia active learning work, researchers have tried to concatenate all low level features into a single feature vector (at times using PCA to reduce the dimensionality) and the perform single-modality active learning. Experiments on the TRECVID 2004 [6] semantic feature extraction task indicate that our multi-modality active learning approach works more effectively than a single-modality active learning approach.

Multi-modality Active Learning

In this section, we present a multi-modality active learning framework. In this framework, we employ Support Vector Machine (SVM) [7] as our base learning algorithm. The scheme for choosing informative data is to simply select examples close to the SVM margin boundary. After selecting informative data, a linear combination is adopted to fuse the multiple modalities while the newly annotated data provides a held-out set for learning the proper combination weights and for constructing a global multi-modality classifier. In the following paragraphs we will further describe the details of this approach.

Support Vector Machine (SVM)

The basic idea of SVM is to separate samples with a hyperplane that has a maximal margin between two classes. To formulate the problem of classifying synthesized feature vectors, the training data are represented as $\{x_i, y_i\}$, i = 1,2, … , n, y_i is either -1 (negative examples) or 1 (positive examples), n is the number of training samples. Suppose all training data satisfy the following constraints:

$$\begin{aligned} x_i \bullet w + b \geq +1 \text{ when } y_i = 1 \\ x_i \bullet w + b \leq -1 \text{ when } y_i = -1 \end{aligned} \quad (1)$$

The distance between the hyperplane $x_i \bullet w + b \geq +1$ and the hyperplane $x_i \bullet w + b \geq -1$ is $2/\|w\|$, where $\|w\|^2$ is the Euclidean norm of w. Therefore, by minimizing $\|w\|$ we get the two hyperlanes with maximal margins. Quadratic programming provides well-studied optimizations to maximize the quadratic functions subject to the linear constraints in equation 1, which guarantees finding the global maximum.

More generally, SVM can project the original training data in space $\boldsymbol{X}$ to a higher dimensional feature space $\boldsymbol{F}$ via a Mercer kernel operator $\boldsymbol{K}$.

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) \quad (2)$$

When $\boldsymbol{K}$ satisfies Mercer's condition [7] we can write: $\boldsymbol{K}(u,v) = \Phi(u) \cdot \Phi(u)$ where $\Phi : \boldsymbol{X} \rightarrow \boldsymbol{F}$ and "$\cdot$" denotes an inner product. We can then rewrite $\boldsymbol{f}$ as:

$$f(x) = w \bullet \Phi(x), \text{ where } w = \sum_{i=1}^{n} \alpha_i \Phi(x_i) \quad (3)$$

With the $\boldsymbol{K}$ function, we are implicitly projecting the training examples into a different feature space $\boldsymbol{F}$ and employ the same optimization problem as equation 1 to maximize margin of hyperplane in $\boldsymbol{F}$. By choosing different kernel functions we can project the training data to different spaces to make more complex decision boundaries than in the original space. A commonly used kernel is the radial basis function kernel $\boldsymbol{K}(u,v) = (e^{-r(u-v)*(u-v)})$ which induces boundaries by placing weighted Gaussians. Our base learning algorithm is this radiant based SVM.

In active learning, we want to choose the most informative data to annotate. Following the procedure of [3], we learn a SVM on the existing labeled data and choose as the next examples those which come closest to the hyperplane in $\boldsymbol{F}$. This scheme for choosing new examples will reduce the corresponding version space of the SVM.

Multi-modality Fusion

For any multimedia source, there are many different variants of features (various texture computations, alternate color spaces, different audio feature types, etc.) to represent its content. Assume we have r different feature sets, our training data x_i is composed of $\{x_{ij}\}$, j=1,2, …, r. Most of the time, the easiest way to deal with this kind of data, is to concatenate it as a larger feature vector x_i and employ a machine learning algorithm, such as SVM. This creates two main problems, first and foremost, the curse of dimensionality [8]. One ends up needing much more labeled data for the learning algorithm due to the increase in dimensionality of the feature vector. Second, it becomes more difficult for a human to understand and analyze the relative importance and the performance corresponding to a particular feature set. Furthermore, we effectively eliminate the variations of individual feature sets and only maintain one, undifferentiated global model to explain all the data. From our TRECVID experiments, concatenating feature vectors always perform worse in evaluation than intelligently selected feature sets.

Therefore, multi-modality fusion can lead us to a better approach than the concatenation method. Assume we have r different feature sets; we can construct r individual sub-models for each feature set. Each model represents its own information according to the its feature space.

We fuse the sub-models by linear combination via a held-out set to obtain a global model for the multi-modality data. This approach is motivated by an attempt to keep the locality of different feature spaces but still have a global model to represent the classification concept. The α in Equation 3 is the weight parameter for each sub-model.

$$f(x) = \sum_{j=1}^{r} \alpha_j g_j(x_j) \quad (3)$$

The fusion approach requires a held-out set to learn the combinational parameter. Usually, a split of training data is required and this reduces the number of examples we can use in training the classifier. However, with the active learning algorithm, we will choose some informative data

from the unlabeled data pool iteratively, and this data has not already been used in training process. This provides us with the held-out set we need for multi-modality fusion.

Multi-modality active learning approach works as follows:

1. Randomly select examples from unlabeled data pool. This is the initial training set for active learning.
2. Build r individual sub-models for the training set according to the different feature sets and apply their leaned models to the unlabeled data.
3. For each sub-model, choose k examples which are closest to its hyperplane. In total, k*r unlabeled examples will be chosen for annotation in each iteration.
4. Annotate these examples. The multi-modality fusion weights are then trained using these new annotated examples. A global model can then be constructed and evaluated.
5. Add the newly annotated examples into training set.
6. Repeat step 2 to step 4.

The algorithm can be formulated as follows:

Multi - modality Active Learning
unlabel data $D = \{x_i\}\ i = 1,..., n$
$D_0 = \{x_l\}$ which randomly choose from D
$D_{0m} = D_0$
for $j = 1$. to t
{
  for $m = 1$ to r
  {
    g $_{jm}$ is the model constructe d from D_{j-1m}
    $d_{jm} = \{x_{jm}\}$ the set of examples cloest to hyperplane of g_{jm}
    $D_{jm} = D_{j-1m} \cup d_{jm}$
  }
  $F_j(x) = \sum_{m=1}^{r} \alpha_m g_{jm}$ combinatio n parameters are trained by d_{jm}
}

Fig. 2. Multi-modality Active Learning Algorithm

Some interesting issues are raised by this approach. The main idea we want to achieve is to train and select each feature set individually. Therefore, we split the training data for each feature set, let's call it D_{jm}, which is the training data of feature m in j iteration. After each iteration, we select new examples for each feature to annotate and obtain d_{jm} of them. This means, that for m different features, we select m sets of data according to each feature and build sub-models. The reason we keep every feature set separately is to maintain the specificity of that feature. We want to train locally for each feature set instead of a global model. Through experiments, we found the problem of active learning is that it makes a strong assumption about the correctness of the previous model and the selected data is to improve the boundary. However, this assumption leads the whole model to a more and more restricted area in the feature space with each iteration. Our hope is that with separate sub-models for each feature set, we can expend the selected data from different feature spaces and avoid this problem.

Experiments

In this section, we describe experiments on semantic feature extraction using the development set of TRECVID 2004 feature extraction task to demonstrate the performance of our multi-modality active learning approach. We selected 20 topics in TRECVID 2003 and 2004 semantic feature extraction tasks. The development set is the collection of news video from ABC and CNN. It contains 52943 shots and is totally around 60 hours.

Low-level features including color, edge, texture and face are generated to learn the semantic features. After dividing an image into 5 by 5 grids, the color feature in each grid is computed as the mean and variance of color histogram from HSV color space. A canny edge detector was applied to extract edges from the images. The edge histogram for 5 by 5 grids is quantized at 45 degree intervals. Six orientated Gabor filter is applied to extract texture feature. Schneiderman's face detection algorithm [9] was used to extract frontal and profile faces. The size and location of faces represent the face detection result.

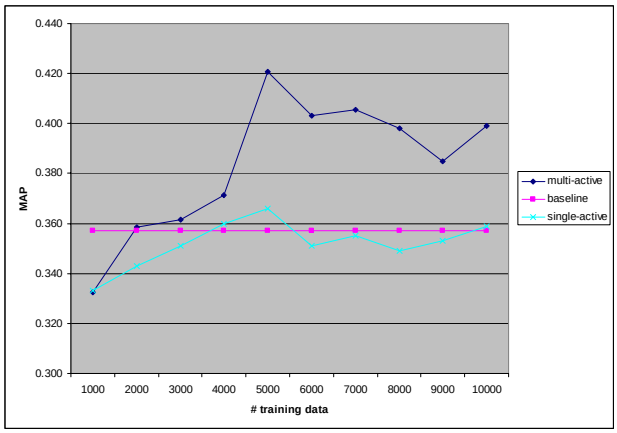

Fig. 3. Classification performance for multi-modality active learning and basic active learning

Figure 3 compares the performance between the multi-modality active learning approach and single-modality active learning. We start the initial data with 1000 examples and during each iteration we choose 250 new examples from the 4 individual feature sets (for a total of 1000 new examples). The curve labeled multi-active is depicts the results of the new approach and the curve labeled "single-active" is the approach which concatenates the 4 feature sets into one larger feature vector. The baseline uses the complete training data set without any active learning. Our evaluation experiments are performed on the TRECVID 2003 and 2004 ground truth provided by NIST in a separate test set. Our measurement is the macro-average mean average precision (MAP) of those 20 topics. From

figure 3, we note that active learning is very effective. Even the single-modality active learning approach can reach the same performance as using the whole data set with only 7% of the labeled data (4000 over 52943). Furthermore, the new approach works much more effectively than the single-modality approach. Its performance was comparable to the baseline with only 3% of the training data.

Outdoors	News subject monologue
News subject face	Non-studio setting
People	Sporting event
Building	Weather news
Road	Boat/Ship
Vegetation	Bill Clinton
Animal	Beach
Female speech	Basket scored
Car/truck/bus	People walking/running
Aircraft	Road

Fig. 4. The 20 topics from TRECVID 2003 and 2004 semantic feature extraction task used in our experiments.

Conclusion and Future Work

We present a multi-modality active learning approach that considers the multiple modality problem during active learning. Our experiments on semantic feature extraction demonstrate that this approach can achieve good performance with much less labeled data and get significant improvements compared to a single-modality approach.

We ascribe the improvement to two factors. First, is the dimensionality issue. If we concatenate all feature sets together; we get a feature vector with 555 dimensions (color 50, edge 200, texture 300 and 5 for face information). According to an analysis of the curse of dimensionality, we would need an exponential increase in training data with increasing dimensions. By contrast, our approach builds more robust individual sub-models and fuses them with newly labeled and held-out data. The second source of the performance improvement is the variance. As we mention before, we keep an individual training set for each feature set. This maintains the individuality of each feature space. We only combine all the labeled data for a final global model to use in predicting the actual evaluation data. By preserving the feature individuality, we exploit the variance in the different aspects of the different feature spaces. Earlier experimental results and experience had informed us that active learning, especially the new data selection scheme we use, that is, to retrieve the examples to closest to the margin boundary, will lead the whole model to a limited and narrow space within the whole feature space. If we keep the individuality of each feature set, we at least have a chance to explore each feature space with a different view on the data and gain more variants of examples from the different sub-models. In our experiments there were only an average of 3% redundant examples in each iteration.

However, many challenges remain. First is the estimation of a stopping criterion. From the performance graphs, we can suspect that there is a point at which active learning performs best. This implies that with the current feature sets and algorithm; we can stop at this number of training examples and don't have to annotate more data. The estimation of this stopping point is an open issue. We have tried to estimate it from increases in positive examples and training error. Unfortunately, there did not seem to be any relation between them. Another open issue is the data selection scheme. At present, we don't know what would happen if we changed the selection scheme from closest margin examples to some other criteria which will enhance the variance [10], and whether our approach would still work effectively. These questions will be addressed in subsequent experiments and we expect to have more definitive conclusions in the near future.

Acknowledgements

This material is based on work supported by the Advanced Research and Development Activity (ARDA) under contract numbers H98230-04-C-0406 and NBCHC040037.

References

[1] C. Campbell, N. Chrisianini and A. Smola, "*Query learning with large margin classifiers*," in Proc. 17th International Conf. on Machine Learning (ICML00), 2000, pp. 111-118.

[2] G. Schohn and D. Cohn, "*Less is more: Active learning with support vector machines*," in Proc. 17th International Conf. on Machine learning (ICML00), 2000, pp. 839-846

[3] S. Tong and E. Chang, "*Support vector machine active learning for image retrieval*," in ACM Intl. Conf. on Multimedia, 2001, pp. 107-118

[4] A. Hauptmann, D. Ng, R. Baron, M-Y. Chen, M. Christel, Duygulu, C. Huang, W-H. Lin, H. Wactlar, N. Moraveji, N. Papernick, C. Snoek, G.Tzanetakis, J. Yang, R. Yan, and R. Jin, "*Informedia at TRECVID 2003: Analyzing and Searching Broadcast New Video,*" in Proc. of TRECVID 2003, 2003

[5] C.-Y. Lin, M. Naphade, A.P. Natsev, B. Tseng, Y. Wu, D.Zhang, G. Iyengar, C. Neti, H. Nock, A. Amir, M. Berg, S.-F. Chang and W. Hsu, "*IBM Research TRECVID-2003 Video Retrieval System,*" in Proc. of TRECVID 2003, 2003

[6] TRECVID, The NIST TREC Video Retrieval Evaluation, homepage http://www-nlpir.nist.org/project/t01v/

[7] C.J.C. Burges, "*A Tutorial on Support Vector Machines for Pattern Recognition,*" in Knowledge Discovery and Data Mining, 1998, Vol.2, No. 2

[8] T. Hastie, R. Tibshirani and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," published by Springer, Chap 2, p. 22-27, 2001

[9] H. Schneiderman and T. Kanade, "*Probailistic Modeling of Local Appearance and Spatial Rlatiionships of Object Recognition,*" IEEE CVPR, 1998

[10] K. Goh, E.Y. Chang, and W.-C. Lai, "Concep-dependent Multimodal Active Learning for Image Retrieval," in Proc. ACM International Conference on Multimedia, 2004, pp. 564-571