Enhanced Max Margin Learning on Multimodal Data Mining in a Multimedia Database

Zhen Guo, Zhongfei (Mark) Zhang
Computer Science Department
SUNY Binghamton
Binghamton, NY, 13902
{zguo,zhongfei}@cs.binghamton.edu

Eric P. Xing, Christos Faloutsos
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213
{epxing,christos}@cs.cmu.edu

ABSTRACT

The problem of multimodal data mining in a multimedia database can be addressed as a structured prediction problem where we learn the mapping from an input to the structured and interdependent output variables. In this paper, built upon the existing literature on the max margin based learning, we develop a new max margin learning approach called Enhanced Max Margin Learning (EMML) framework. In addition, we apply EMML framework to developing an effective and efficient solution to the multimodal data mining problem in a multimedia database. The main contributions include: (1) we have developed a new max margin learning approach — the enhanced max margin learning framework that is much more efficient in learning with a much faster convergence rate, which is verified in empirical evaluations; (2) we have applied this EMML approach to developing an effective and efficient solution to the multimodal data mining problem that is highly scalable in the sense that the query response time is independent of the database scale, allowing facilitating a multimodal data mining querying to a very large scale multimedia database, and excelling many existing multimodal data mining methods in the literature that do not scale up at all; this advantage is also supported through the complexity analysis as well as empirical evaluations against a state-of-the-art multimodal data mining method from the literature. While EMML is a general framework, for the evaluation purpose, we apply it to the Berkeley Drosophila embryo image database, and report the performance comparison with a state-of-the-art multimodal data mining method.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining,Image databases*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*; I.5.1 [**Pattern Recognition**]: Models—*Structural*; J.3 [**Computer Applications**]: Life and Medical Sciences—*Biology and genetics*

KDD'07, August 12–15, 2007, San Jose, California, USA.

General Terms

Algorithms, experimentation

Keywords

Multimodal data mining, image annotation, image retrieval, max margin

1. INTRODUCTION

Multimodal data mining in a multimedia database is a challenging topic in data mining research. Multimedia data may consist of data in different modalities, such as digital images, audio, video, and text data. In this context, a multimedia database refers to a data collection in which there are multiple modalities of data such as text and imagery. In this database system, the data in different modalities are related to each other. For example, the text data are related to images as their annotation data. By multimodal data mining in a multimedia database it is meant that the knowledge discovery to the multimedia database is initiated by a query that may also consist of multiple modalities of data such as text and imagery. In this paper, we focus on a multimedia database as an image database in which each image has a few textual words given as annotation. We then address the problem of multimodal data mining in such an image database as the problem of retrieving similar data and/or inferencing new patterns to a multimodal query from the database.

Specifically, in the context of this paper, multimodal data mining refers to two aspects of activities. The first is the multimodal retrieval. This is the scenario where a multimodal query consisting of either textual words alone, or imagery alone, or in any combination is entered and an expected retrieved data modality is specified that can also be text alone, or imagery alone, or in any combination; the retrieved data based on a pre-defined similarity criterion are returned back to the user. The second is the multimodal inferencing. While the retrieval based multimodal data mining has its standard definition in terms of the semantic similarity between the query and the retrieved data from the database, the inferencing based mining depends on the specific applications. In this paper, we focus on the application of the fruit fly image database mining. Consequently, the inferencing based multimodal data mining may include many different scenarios. A typical scenario is the across-stage multimodal inferencing. There are many interesting questions a biologist may want to ask in the fruit fly research given such a multimodal mining capability. For example, given an embryo

image in stage 5, what is the corresponding image in stage 7 for an image-to-image three-stage inferencing? What is the corresponding annotation for this image in stage 7 for an image-to-word three-stage inferencing? The multimodal mining technique we have developed in this paper also addresses this type of across-stage inferencing capability, in addition to the multimodal retrieval capability.

In the image retrieval research area, one of the notorious bottlenecks is the semantic gap [18]. Recently, it is reported that this bottleneck may be reduced by the multimodal data mining approaches [3, 11] which take advantage of the fact that in many applications image data typically co-exist with other modalities of information such as text. The synergy between different modalities may be exploited to capture the high level conceptual relationships.

To exploit the synergy among the multimodal data, the relationships among these different modalities need to be learned. For an image database, we need to learn the relationship between images and text. The learned relationship between images and text can then be further used in multimodal data mining. Without loss of generality, we start with a special case of the multimodal data mining problem — image annotation, where the input is an image query and the expected output is the annotation words. We show later that this approach is also valid to the general multimodal data mining problem. The image annotation problem can be formulated as a structured prediction problem where the input (image) $\mathbf{x}$ and the output (annotation) $\mathbf{y}$ are structures. An image can be partitioned into blocks which form a structure. The word space can be denoted by a vector where each entry represents a word. Under this setting, the learning task is therefore formulated as finding a function $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ such that

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \tag{1}$$

is the desired output for any input $\mathbf{x}$.

In this paper, built upon the existing literature on the max margin learning, we propose a new max margin learning approach on the structured output space to learn the above function. Like the existing max margin learning methods, the image annotation problem may be formulated as a quadratic programming (QP) problem. The relationship between images and text is discovered once this QP problem is solved. Unlike the existing max margin learning methods, the new max margin learning method is much more efficient with a much faster convergence rate. Consequently, we call this new max margin learning approach as Enhanced Max Margin Learning (EMML). We further apply EMML to solving the multimodal data mining problem effectively and efficiently.

Note that the proposed approach is general that can be applied to any structured prediction problems. For the evaluation purpose, we apply this approach to the Berkeley Drosophila embryo image database. Extensive empirical evaluations against a state-of-the-art method on this database are reported.

2. RELATED WORK

Multimodal approaches have recently received the substantial attention since Barnard and Duygulu et al. started their pioneering work on image annotation [3, 10]. Recently there have been many studies [4, 17, 11, 7, 9, 23] on the multimodal approaches.

The learning with structured output variables covers many natural learning tasks including named entity recognition, natural language parsing, and label sequence learning. There have been many studies on the structured model which include conditional random fields [14], maximum entropy model [15], graph model [8], semi-supervised learning [6] and max margin approaches [13, 21, 20, 2]. The challenge of learning with structured output variables is that the number of the structures is exponential in terms of the size of the structure output space. Thus, the problem is intractable if we treat each structure as a separate class. Consequently, the multiclass approach is not well fitted into the learning with structured output variables.

As an effective approach to this problem, the max margin principle has received substantial attention since it was used in the support vector machine (SVM) [22]. In addition, the perceptron algorithm is also used to explore the max margin classification [12]. Taskar et al. [19] reduce the number of the constraints by considering the dual of the loss-augmented problem. However, the number of the constraints in their approach is still large for a large structured output space and a large training set.

For learning with structured output variables, Tsochantaridis et al. [21] propose a cutting plane algorithm which finds a small set of active constraints. One issue of this algorithm is that it needs to compute the most violated constraint which would involve another optimization problem in the output space. In EMML, instead of selecting the most violated constraint, we arbitrarily select a constraint which violates the optimality condition of the optimization problem. Thus, the selection of the constraints does not involve any optimization problem. Osuna et al. [16] propose the decomposition algorithm for the support vector machine. In EMML, we generalize their idea to the scenario of learning with structured output variables.

3. HIGHLIGHTS OF THIS WORK

This work is based on the existing literature on max margin learning, and aims at solving for the problem of multimodal data mining in a multimedia database defined in this paper. In comparison with the existing literature, the main contributions of this work include: (1) we have developed a new max margin learning approach — the enhanced max margin learning framework that is much more efficient in learning with a much faster convergence rate, which is verified in empirical evaluations; (2) we have applied this EMML approach to developing an effective and efficient solution to the multimodal data mining problem that is highly scalable in the sense that the query response time is independent of the database scale, allowing facilitating a multimodal data mining querying to a very large scale multimedia database, and excelling many existing multimodal data mining methods in the literature that do not scale up at all; this advantage is also supported through the complexity analysis as well as empirical evaluations against a state-of-the-art multimodal data mining method from the literature.

4. LEARNING IN THE STRUCTURED OUTPUT SPACE

Assume that the image database consists of a set of instances $S = \{(I_i, W_i)\}_{i=1}^{L}$, where each instance consists of

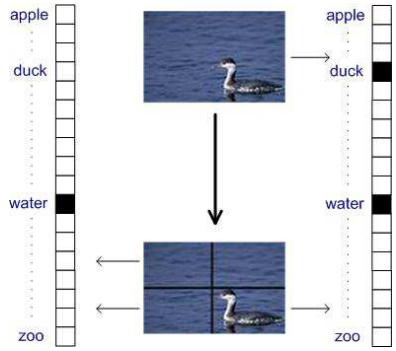


Figure 1: An illustration of the image partitioning and the structured output word space

an image object I_i and the corresponding annotation word set W_i. First we partition an image into a set of blocks. Thus, an image can be represented by a set of sub-images. The feature vector in the feature space for each block can be computed from the selected feature representation. Consequently, an image is represented as a set of feature vectors in the feature space. A clustering algorithm is then applied to the whole feature space to group similar feature vectors together. The centroid of a cluster represents a visual representative (we refer it to VRep in this paper) in the image space. In Figure 1, there are two VReps, *water* and *duck in the water*. The corresponding annotation word set can be easily obtained for each VRep. Consequently, the image database becomes the VRep-word pairs $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, where n is the number of the clusters, $\mathbf{x}_i$ is a VRep object and $\mathbf{y}_i$ is the word annotation set corresponding to this VRep object. Another simple method to obtain the VRep-word pairs is that we randomly select some images from the image database and each image is viewed as a VRep.

Suppose that there are W distinct annotation words. An arbitrary subset of annotation words is represented by the binary vector $\bar{\mathbf{y}}$ whose length is W; the j-th component $\bar{\mathbf{y}}_j = 1$ if the j-th word occurs in this subset, and 0 otherwise. All possible binary vectors form the word space $\mathcal{Y}$. We use $\mathbf{w}_j$ to denote the j-th word in the whole word set. We use $\mathbf{x}$ to denote an arbitrary vector in the feature space. Figure 1 shows an illustrative example in which the original image is annotated by *duck* and *water* which are represented by a binary vector. There are two VReps after the clustering and each has a different annotation. In the word space, a word may be related to other words. For example, *duck* and *water* are related to each other because *water* is more likely to occur when *duck* is one of the annotation words. Consequently, the annotation word space is a structured output space where the elements are interdependent.

The relationship between the input example VRep $\mathbf{x}$ and an arbitrary output $\bar{\mathbf{y}}$ is represented as the joint feature mapping $\Phi(\mathbf{x}, \bar{\mathbf{y}})$, $\Phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$ where d is the dimension of the joint feature space. It can be expressed as a linear combination of the joint feature mapping between $\mathbf{x}$ and all the unit vectors. That is

$$\Phi(\mathbf{x}, \bar{\mathbf{y}}) = \sum_{j=1}^{W} \bar{\mathbf{y}}_j \Phi(\mathbf{x}, \mathbf{e}_j)$$

where $\mathbf{e}_j$ is the j-th unit vector. The score between $\mathbf{x}$ and $\bar{\mathbf{y}}$ can be expressed as a linear combination of each component in the joint feature representation: $f(\mathbf{x}, \bar{\mathbf{y}}) = \langle \boldsymbol{\alpha}, \Phi(\mathbf{x}, \bar{\mathbf{y}}) \rangle$. Then the learning task is to find the optimal weight vector $\boldsymbol{\alpha}$ such that the prediction error is minimized for all the training instances. That is

$$\arg \max_{\bar{\mathbf{y}} \in \mathcal{Y}^{(i)}} f(\mathbf{x}_i, \bar{\mathbf{y}}) \approx \mathbf{y}_i, \quad i = 1, \cdots, n$$

where $\mathcal{Y}_i = \{\bar{\mathbf{y}} | \sum_{j=1}^W \bar{\mathbf{y}}_j = \sum_{j=1}^W \mathbf{y}_{ij}\}$. We use $\Phi_i(\bar{\mathbf{y}})$ to denote $\Phi(\mathbf{x}_i, \bar{\mathbf{y}})$. To make the prediction to be the true output $\mathbf{y}_i$, we must follow

$$\boldsymbol{\alpha}^\top \Phi_i(\mathbf{y}_i) \geq \boldsymbol{\alpha}^\top \Phi_i(\bar{\mathbf{y}}), \quad \forall \bar{\mathbf{y}} \in \mathcal{Y}_i \backslash \{\mathbf{y}_i\}$$

where $\mathcal{Y}_i \backslash \{\mathbf{y}_i\}$ denotes the removal of the element $\mathbf{y}_i$ from the set $\mathcal{Y}_i$. In order to accommodate the prediction error on the training examples, we introduce the slack variable ξ_i. The above constraint then becomes

$$\boldsymbol{\alpha}^\top \Phi_i(\mathbf{y}_i) \geq \boldsymbol{\alpha}^\top \Phi_i(\bar{\mathbf{y}}) - \xi_i, \quad \xi_i \geq 0 \quad \forall \bar{\mathbf{y}} \in \mathcal{Y}_i \backslash \{\mathbf{y}_i\}$$

We measure the prediction error on the training instances by the loss function which is the distance between the true output $\mathbf{y}_i$ and the prediction $\bar{\mathbf{y}}$. The loss function measures the goodness of the learning model. The standard zero-one classification loss is not suitable for the structured output space. We define the loss function $l(\bar{\mathbf{y}}, \mathbf{y}_i)$ as the number of the different entries in these two vectors. We include the loss function in the constraints as is proposed by Taskar et al. [19]

$$\boldsymbol{\alpha}^\top \Phi_i(\mathbf{y}_i) \geq \boldsymbol{\alpha}^\top \Phi_i(\bar{\mathbf{y}}) + l(\bar{\mathbf{y}}, \mathbf{y}_i) - \xi_i$$

We interpret $\frac{1}{\|\boldsymbol{\alpha}\|} \boldsymbol{\alpha}^\top [\Phi_i(\mathbf{y}_i) - \Phi_i(\bar{\mathbf{y}})]$ as the margin of $\mathbf{y}_i$ over another $\bar{\mathbf{y}} \in \mathcal{Y}^{(i)}$. We then rewrite the above constraint as $\frac{1}{\|\boldsymbol{\alpha}\|} \boldsymbol{\alpha}^\top [\Phi_i(\mathbf{y}_i) - \Phi_i(\bar{\mathbf{y}})] \geq \frac{1}{\|\boldsymbol{\alpha}\|} [l(\bar{\mathbf{y}}, \mathbf{y}_i) - \xi_i]$. Thus, minimizing $\|\boldsymbol{\alpha}\|$ maximizes such margin.

The goal now is to solve the optimization problem

$$\begin{aligned} \min \quad & \frac{1}{2} \|\boldsymbol{\alpha}\|^2 + C \sum_{i=1}^{n} \xi_i^r \\ s.t. \quad & \boldsymbol{\alpha}^\top \Phi_i(\mathbf{y}_i) \geq \boldsymbol{\alpha}^\top \Phi_i(\bar{\mathbf{y}}) + l(\bar{\mathbf{y}}, \mathbf{y}_i) - \xi_i \\ & \forall \bar{\mathbf{y}} \in \mathcal{Y}_i \backslash \{\mathbf{y}_i\}, \quad \xi_i \geq 0, \quad i = 1, \cdots, n \end{aligned} \tag{2}$$

where $r = 1, 2$ corresponds to the linear or quadratic slack variable penalty. In this paper, we use the linear slack variable penalty. For $r = 2$, we obtain similar results. $C > 0$ is a constant that controls the tradeoff between the training error minimization and the margin maximization.

Note that in the above formulation, we do not introduce the relationships between different words in the word space. However, the relationships between different words are implicitly included in the VRep-word pairs because the related words is more likely to occur together. Thus, Eq. (2) is in fact a structured optimization problem.

4.1 EMML Framework

One can solve the optimization problem Eq. (2) in the primal space — the space of the parameters $\boldsymbol{\alpha}$. In fact this problem is intractable when the structured output space is large because the number of the constraints is exponential in terms of the size of the output space. As in the traditional support vector machine, the solution can be obtained by solving this quadratic optimization problem in the dual space — the space of the Lagrange multipliers. Vapnik [22]

and Boyd et al. [5] have an excellent review for the related optimization problem.

The dual problem formulation has an important advantage over the primal problem: it only depends on the inner products in the joint feature representation defined by Φ, allowing the use of a kernel function. We introduce the Lagrange multiplier $\mu_{i,\bar{\mathbf{y}}}$ for each constraint to form the Lagrangian. We define $\Phi_{i,\mathbf{y}_i,\bar{\mathbf{y}}} = \Phi_i(\mathbf{y}_i) - \Phi_i(\bar{\mathbf{y}})$ and the kernel function $K((\mathbf{x}_i,\bar{\mathbf{y}}),(\mathbf{x}_j,\tilde{\mathbf{y}})) = \langle \Phi_{i,\mathbf{y}_i,\bar{\mathbf{y}}}, \Phi_{j,\mathbf{y}_j,\tilde{\mathbf{y}}} \rangle$. The derivatives of the Lagrangian over $\boldsymbol{\alpha}$ and ξ_i should be equal to zero. Substituting these conditions into the Lagrangian, we obtain the following Lagrange dual problem

$$\begin{aligned} \min \ & \frac{1}{2}\sum_{\substack{i,j \\ \bar{\mathbf{y}}\neq\mathbf{y}_i \\ \tilde{\mathbf{y}}\neq\mathbf{y}_j}} \mu_{i,\bar{\mathbf{y}}}\mu_{j,\tilde{\mathbf{y}}} K((\mathbf{x}_i,\bar{\mathbf{y}}),(\mathbf{x}_j,\tilde{\mathbf{y}})) - \sum_{\substack{i \\ \bar{\mathbf{y}}\neq\mathbf{y}_i}} \mu_{i,\bar{\mathbf{y}}} l(\bar{\mathbf{y}},\mathbf{y}_i) \quad (3) \\ s.t. \ & \sum_{\bar{\mathbf{y}}\neq\mathbf{y}_i} \mu_{i,\bar{\mathbf{y}}} \leq C \quad \mu_{i,\bar{\mathbf{y}}} \geq 0 \quad i = 1,\cdots,n \end{aligned}$$

After this dual problem is solved, we have $\boldsymbol{\alpha} = \sum_{i,\bar{\mathbf{y}}} \mu_{i,\bar{\mathbf{y}}} \Phi_{i,\mathbf{y}_i,\bar{\mathbf{y}}}$.

For each training example, there are a number of constraints related to it. We use the subscript i to represent the part related to the i-th example in the matrix. For example, let $\boldsymbol{\mu_i}$ be the vector with entries $\mu_{i,\bar{\mathbf{y}}}$. We stack the $\boldsymbol{\mu_i}$ together to form the vector $\boldsymbol{\mu}$. That is $\boldsymbol{\mu} = [\boldsymbol{\mu_1}^\top \ \cdots \boldsymbol{\mu_n}^\top]^\top$. Similarly, let $\mathbf{S}_i$ be the vector with entries $l(\bar{\mathbf{y}},\mathbf{y}_i)$. We stack $\mathbf{S}_i$ together to form the vector $\mathbf{S}$. That is $\mathbf{S} = [\mathbf{S}_1^\top \cdots \mathbf{S}_n^\top]^\top$. The lengths of $\boldsymbol{\mu}$ and $\mathbf{S}$ are the same. We define $\mathbf{A}_i$ as the vector which has the same length as that of $\boldsymbol{\mu}$, where $\mathbf{A}_{i,\bar{\mathbf{y}}} = 1$ and $\mathbf{A}_{j,\bar{\mathbf{y}}} = 0$ for $j \neq i$. Let $\mathbf{A} = [\mathbf{A}_1 \cdots \mathbf{A}_n]^\top$. Let matrix $\mathbf{D}$ represent the kernel matrix where each entry is $K((\mathbf{x}_i,\bar{\mathbf{y}}),(\mathbf{x}_j,\tilde{\mathbf{y}}))$. Let $\mathbf{C}$ be the vector where each entry is constant C.

With the above notations we rewrite the Lagrange dual problem as follows

$$\begin{aligned} \min \quad & \frac{1}{2}\boldsymbol{\mu}^\top \mathbf{D}\boldsymbol{\mu} - \boldsymbol{\mu}^\top \mathbf{S} \quad (4) \\ s.t. \quad & \mathbf{A}\boldsymbol{\mu} \preceq \mathbf{C} \\ & \boldsymbol{\mu} \succeq 0 \end{aligned}$$

where $\preceq$ and $\succeq$ represent the vector comparison defined as entry-wise less than or equal to and greater than or equal to, respectively.

Eq. (4) has the same number of the constraints as Eq. (2). However, in Eq. (4) most of the constraints are lower bound constraints ($\boldsymbol{\mu} \succeq 0$) which define the feasible region. Other than these lower bound constraints, the rest constraints determine the complexity of the optimization problem. Therefore, the number of constraints is considered to be reduced in Eq. (4). However, the challenge still exists to solve it efficiently since the number of the dual variables is still huge. Osuna et al. [16] propose a decomposition algorithm for the support vector machine learning over large data sets. We generalize this idea to learning with the structured output space. We decompose the constraints of the optimization problem Eq. (2) into two sets: the working set B and the nonactive set N. The Lagrange multipliers are also correspondingly partitioned into two parts $\boldsymbol{\mu}_B$ and $\boldsymbol{\mu}_N$. We are interested in the subproblem defined only for the dual variable set $\boldsymbol{\mu}_B$ when keeping $\boldsymbol{\mu}_N$=0 as follows

$$\begin{aligned} \min \quad & \frac{1}{2}\boldsymbol{\mu}^\top \mathbf{D}\boldsymbol{\mu} - \boldsymbol{\mu}^\top \mathbf{S} \quad (5) \\ s.t. \quad & \mathbf{A}\boldsymbol{\mu} \preceq \mathbf{C} \\ & \boldsymbol{\mu_B} \succeq 0, \quad \boldsymbol{\mu_N} = 0 \end{aligned}$$

It is clearly true that we can move those $\mu_{i,\bar{\mathbf{y}}} = 0, \mu_{i,\bar{\mathbf{y}}} \in \boldsymbol{\mu}_B$ to set $\boldsymbol{\mu_N}$ without changing the objective function. Furthermore, we can move those $\mu_{i,\bar{\mathbf{y}}} \in \boldsymbol{\mu}_N$ satisfying certain conditions to set $\boldsymbol{\mu_B}$ to form a new optimization subproblem which yields a strict decrease in the objective function in Eq. (4) when the new subproblem is optimized. This property is guaranteed by the following theorem.

THEOREM 1. *Given an optimal solution of the subproblem defined on $\boldsymbol{\mu_B}$ in Eq. (5), if the following conditions hold true:*

$$\begin{aligned} & \exists i, \quad \textstyle\sum_{\bar{\mathbf{y}}} \mu_{i,\bar{\mathbf{y}}} < C \\ & \exists \mu_{i,\bar{\mathbf{y}}} \in \boldsymbol{\mu_N}, \quad \boldsymbol{\alpha}^\top \Phi_{i,\mathbf{y}_i,\bar{\mathbf{y}}} - l(\bar{\mathbf{y}},\mathbf{y}_i) < 0 \quad (6) \end{aligned}$$

the operation of moving the Lagrange multiplier $\mu_{i,\bar{\mathbf{y}}}$ satisfying Eq. (6) from set $\boldsymbol{\mu_N}$ to set $\boldsymbol{\mu_B}$ generates a new optimization subproblem that yields a strict decrease in the objective function in Eq. (4) when the new subproblem in Eq.(5) is optimized.

PROOF. Suppose that the current optimal solution is $\boldsymbol{\mu}$. Let δ be a small positive number. Let $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu} + \delta e_r$, where e_r is the r-th unit vector and $r = (i,\bar{\mathbf{y}})$ denotes the Lagrange multiplier satisfying condition Eq. (6). Thus, the objective function becomes

$$\begin{aligned} \mathbf{W}(\bar{\boldsymbol{\mu}}) &= \frac{1}{2}(\boldsymbol{\mu}+\delta e_r)^\top \mathbf{D}(\boldsymbol{\mu}+\delta e_r) - (\boldsymbol{\mu}+\delta e_r)^\top \mathbf{S} \\ &= \frac{1}{2}(\boldsymbol{\mu}^\top \mathbf{D}\boldsymbol{\mu} + \delta e_r^\top \mathbf{D}\boldsymbol{\mu} + \delta \boldsymbol{\mu}^\top \mathbf{D} e_r + \delta^2 e_r^\top \mathbf{D} e_r) \\ &\quad -\boldsymbol{\mu}^\top \mathbf{S} - \delta e_r^\top \mathbf{S} \\ &= \mathbf{W}(\boldsymbol{\mu}) + \frac{1}{2}(\delta e_r^\top \mathbf{D}\boldsymbol{\mu} + \delta \boldsymbol{\mu}^\top \mathbf{D} e_r + \delta^2 e_r^\top \mathbf{D} e_r) \\ &\quad -\delta e_r^\top \mathbf{S} \\ &= \mathbf{W}(\boldsymbol{\mu}) + \delta e_r^\top \mathbf{D}\boldsymbol{\mu} - \delta e_r^\top \mathbf{S} + \frac{1}{2}\delta^2 e_r^\top \mathbf{D} e_r \\ &= \mathbf{W}(\boldsymbol{\mu}) + \delta(\boldsymbol{\alpha}^\top \Phi_{i,\mathbf{y}_i,\bar{\mathbf{y}}} - l(\bar{\mathbf{y}},\mathbf{y}_i)) + \frac{1}{2}\delta^2 \|\Phi_{i,\mathbf{y}_i,\bar{\mathbf{y}}}\|^2 \end{aligned}$$

Since $\boldsymbol{\alpha}^\top \Phi_{i,\mathbf{y}_i,\bar{\mathbf{y}}} - l(\bar{\mathbf{y}},\mathbf{y}_i) < 0$, for small enough δ, we have $\mathbf{W}(\bar{\boldsymbol{\mu}}) < \mathbf{W}(\boldsymbol{\mu})$. For small enough δ, the constraints $\mathbf{A}\bar{\boldsymbol{\mu}} \preceq \mathbf{C}$ is also valid. Therefore, when the new optimization subproblem in Eq. (5) is optimized, there must be an optimal solution no worse than $\bar{\boldsymbol{\mu}}$. □

In fact, the optimal solution is obtained when there is no Lagrange multiplier satisfying the condition Eq. (6). This is guaranteed by the following theorem.

THEOREM 2. *The optimal solution of the optimization problem in Eq. (4) is achieved if and only if the condition Eq. (6) does not hold true.*

PROOF. If the optimal solution $\hat{\boldsymbol{\mu}}$ is achieved, the condition Eq. (6) must not hold true. Otherwise, $\hat{\boldsymbol{\mu}}$ is not optimal according to the Theorem 1. To prove in the reverse

direction, we consider the Karush-Kuhn-Tucker (KKT) conditions [5] of the optimization problem Eq. (4).

$$\mathbf{D}\boldsymbol{\mu} - \mathbf{S} + \mathbf{A}^\top \boldsymbol{\gamma} - \boldsymbol{\pi} = 0$$
$$\boldsymbol{\gamma}^\top (\mathbf{C} - \mathbf{A}\boldsymbol{\mu}) = 0$$
$$\boldsymbol{\pi}^\top \boldsymbol{\mu} = 0$$
$$\boldsymbol{\gamma} \succeq 0$$
$$\boldsymbol{\pi} \succeq 0$$

For the optimization problem Eq. (4), the KKT conditions provide necessary and sufficient conditions for optimality. One can check that the condition Eq. (6) violates the KKT conditions. On the other hand, one can check that the KKT conditions are satisfied when the condition Eq. (6) does not hold true. Therefore, the optimal solution is achieved when the condition Eq. (6) does not hold true. □

The above theorems suggest the Enhanced Max Margin Learning (EMML) algorithm listed in Algorithm 1. The correctness (convergence) of EMML algorithm is provided by Theorem 3.

Algorithm 1 EMML Algorithm

Input: n labeled examples, dual variable set $\boldsymbol{\mu}$.
Output: Optimized $\boldsymbol{\mu}$

1: **procedure**
2: Arbitrarily decompose $\boldsymbol{\mu}$ into two sets: $\boldsymbol{\mu_B}$ and $\boldsymbol{\mu_N}$.
3: Solve the subproblem in Eq. (5) defined by the variables in $\boldsymbol{\mu_B}$.
4: While there exists $\mu_{i,\bar{\mathbf{y}}} \in \boldsymbol{\mu_B}$ such that $\mu_{i,\bar{\mathbf{y}}} = 0$, move it to set $\boldsymbol{\mu_N}$.
5: While there exists $\mu_{i,\bar{\mathbf{y}}} \in \boldsymbol{\mu_N}$ satisfying condition Eq. (6), move it to set $\boldsymbol{\mu_B}$. If no such $\mu_{i,\bar{\mathbf{y}}} \in \boldsymbol{\mu_N}$ exists, the iteration exits.
6: Goto Step 3.
7: **end procedure**

THEOREM 3. *EMML algorithm converges to the global optimal solution in a finite number of iterations.*

PROOF. This is the direct result from Theorems 1 and 2. Step 3 in Algorithm 1 strictly decreases the objective function of Eq. (4) at each iteration and thus the algorithm does not cycle. Since the objective function of Eq. (4) is convex and quadratic, and the feasible solution region is bounded, the objective function is bounded. Therefore, the algorithm must converge to the global optimal solution in a finite number of iterations. □

Note that in Step 5, we only need find one dual variable satisfying Eq. (6). We need examine all the dual variables in the set $\boldsymbol{\mu}_N$ only when no dual variable satisfies Eq. (6). It is fast to examine the dual variables in the set $\boldsymbol{\mu}_N$ even if the number of the dual variables is large.

4.2 Comparison with other methods

In the max margin optimization problem Eq. (2), only some of the constraints determine the optimal solution. We call these constraints active constraints. Other constraints are automatically met as long as these active constraints are valid. EMML algorithm uses this fact to solve the optimization problem by substantially reducing the number of the dual variables in Eq. (3).

In the recent literature, there are also other methods attempting to reduce the number of the constraints. Taskar et al. [19] reduce the number of the constraints by considering the dual of the loss-augmented problem. However, the number of the constraints in their approach is still large for a large structured output space and a large training set. They do not use the fact that only some of the constraints are active in the optimization problem. Tsochantaridis et al. [21] also propose a cutting plane algorithm which finds a small set of active constraints. One issue of this algorithm is that it needs to compute the most violated constraint which would involve another optimization problem in the output space. In EMML, instead of selecting the most violated constraint, we arbitrarily select a constraint which violates the optimality condition of the optimization problem. Thus, the selection of the constraint does not involve any optimization problem. Therefore, EMML is much more efficient in learning with a much faster convergence rate.

5. MULTIMODAL DATA MINING

The solution to the Lagrange dual problem makes it possible to capture the semantic relationships among different data modalities. We show that the developed EMML framework can be used to solve for the general multimodal data mining problem in all the scenarios. Specifically, given a training data set, we immediately obtain the direct relationship between the VRep space and the word space using the EMML framework in Algorithm 1. Given this obtained direct relationship, we show below that all the multimodal data mining scenarios concerned in this paper can be facilitated.

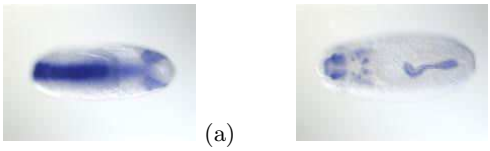
5.1 Image Annotation

Image annotation refers to generating annotation words for a given image. First we partition the test image into blocks and compute the feature vector in the feature space for each block. We then compute the similarity between feature vectors and the VReps in terms of the distance. We return the top n most-relevant VReps. For each VRep, we compute the score between this VRep and each word as the function f in Eq. (1). Thus, for each of the top n most relevant VReps, we have the ranking-list of words in terms of the score. We then merge these n ranking-lists and sort them to obtain the overall ranking-list of the whole word space. Finally, we return the top m words as the annotation result.

In this approach, the score between the VReps and the words can be computed in advance. Thus, the computation complexity of image annotation is only related to the number of the VReps. Under the assumption that all the images in the image database follow the same distribution, the number of the VReps is independent of the database scale. Therefore, the computation complexity in this approach is $O(1)$ which is independent of the database scale.

5.2 Word Query

Word query refers to generating corresponding images in response to a query word. For a given word input, we compute the score between each VRep and the word as the function f in Eq. (1). Thus, we return the top n most relevant VReps. Since for each VRep, we compute the similarity between this VRep and each image in the image database in terms of the distance, for each of those top n most rele-



(a) (b)

Figure 2: An example of 3-stage image-to-image inferencing.

vant VReps, we have the ranking-list of images in terms of the distance. Then we merge these n ranking-lists and sort them to obtain the overall ranking-list in the image space. Finally, we return the top m images as the query result.

For each VRep, the similarity between this VRep and each image in the image database can be computed in advance. Similar to the analysis in Sec. 5.1, the computation complexity is only related to the number of the VReps, which is $O(1)$.

5.3 Image Retrieval

Image retrieval refers to generating semantically similar images to a query image. Given a query image, we annotate it using the procedure in Sec. 5.1. In the image database, for each annotation word j there are a subset of images S_j in which this annotation word appears. We then have the union set $S = \cup_j S_j$ for all the annotation words of the query image.

On the other hand, for each annotation word j of the query image, the word query procedure in Sec. 5.2 is used to obtain the related sorted image subset T_j from the image database. We then merge these subsets T_j to form the sorted image set T in terms of their scores. The final image retrieval result is $R = S \cap T$.

In this approach, the synergy between the image space and the word space is exploited to reduce the semantic gap based on the developed learning approach. Since the complexity of the retrieval methods in Secs. 5.1 and 5.2 are both $O(1)$, and since these retrievals are only returned for the top few items, respectively, finding the intersection or the union is $O(1)$. Consequently, the overall complexity is also $O(1)$.

5.4 Multimodal Image Retrieval

The general scenario of multimodal image retrieval is a query as a combination of a series of images and a series of words. Clearly, this retrieval is simply a linear combination of the retrievals in Secs. 5.2 and 5.3 by merging the retrievals together based on their corresponding scores. Since each individual retrieval is $O(1)$, the overall retrieval is also $O(1)$.

5.5 Across-Stage Inferencing

For a fruit fly embryo image database such as the Berkeley Drosophila embryo image database which is used for our experimental evaluations, we have embryo images classified in advance into different stages of the embryo development with separate sets of textual words as annotation to those images in each of these stages. In general, images in different stages may or may not have the direct semantic correspondence (e.g., they all correspond to the same gene), not even speaking that images in different stages may necessarily exhibit any visual similarity. Figure 2 shows an example of a pair of identified embryo images at stages 9-10 (Figure 2(a)) and stages 13-16 (Figure 2(b)), respectively, in which they

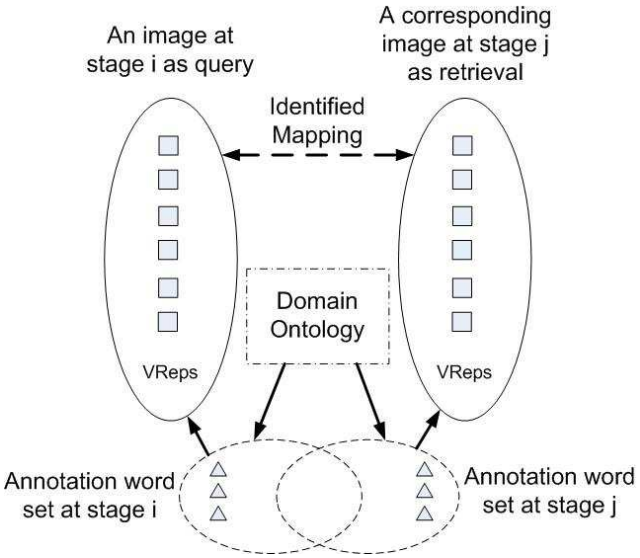


Figure 3: An illustrative diagram for image-to-image across two stages inferencing

both correspond to the same gene as a result of the image-to-image inferencing between the two stages[1]. However, it is clear that they exhibit a very large visual dissimilarity.

Consequently, it is not appropriate to use any pure visual feature based similarity retrieval method to identify such image-to-image correspondence across stages. Furthermore, we also expect to have the word-to-image and image-to-word inferencing capabilities across different stages, in addition to the image-to-image inferencing.

Given this consideration, this is exactly where the proposed approach for multimodal data mining can be applied to complement the existing pure retrieval based methods to identify such correspondence. Typically in such a fruit fly embryo image database, there are textual words for annotation to the images in each stage. These annotation words in one stage may or may not have the direct semantic correspondence to the images in another stage. However, since the data in all the stages are from the same fruit fly embryo image database, the textual annotation words between two different stages share a semantic relationship which can be obtained by a domain ontology.

In order to apply our approach to this across-stage inferencing problem, we treat each stage as a separate multimedia database, and map the across-stage inferencing problem to a retrieval based multimodal data mining problem by applying the approach to the two stages such that we take the multimodal query as the data from one stage and pose the query to the data in the other stage for the retrieval based multimodal data mining. Figure 3 illustrates the diagram of the two stages (state i and state j where $i \neq j$) image-to-image inferencing.

Clearly, in comparison with the retrieval based multimodal data mining analyzed in the previous sections, the only additional complexity here in across-stage inferencing

[1]The Berkeley Drosophila embryo image database is given in such a way that images from several real stages are mixed together to be considered as one "stage". Thus, stages 9-10 are considered as one stage, and so are stages 13-16.

is the inferencing part using the domain ontology in the word space. Typically this ontology is small in scale. In fact, in our evaluations for the Berkeley Drosophila embryo image database, this ontology is handcrafted and is implemented as a look-up table for word matching through an efficient hashing function. Thus, this part of the computation may be ignored. Consequently, the complexity of the across-stage inferencing based multimodal data mining is the same as that of the retrieval based multimodal data mining which is independent of database scale.

6. EMPIRICAL EVALUATIONS

While EMML is a general learning framework, and it can also be applied to solve for a general multimodal data mining problem in any application domains, for the evaluation purpose, we apply it to the Berkeley Drosophila embryo image database [1] for the multimodal data mining task defined in this paper. We evaluate this approach's performance using this database for both the retrieval based and the across-stage inferencing based multimodal data mining scenarios. We compare this approach with a state-of-the-art multimodal data mining method MBRM [11] for the mining performance.

In this image database, there are in total 16 stages of the embryo images archived in six different folders with each folder containing two to four real stages of the images; there are in total 36,628 images and 227 words in all the six folders; not all the images have annotation words. For the retrieval based multimodal data mining evaluations, we use the fifth folder as the multimedia database, which corresponds to stages 11 and 12. There are about 5,500 images that have annotation words and there are 64 annotation words in this folder. We split the whole folder's images into two parts (one third and two thirds), with the two thirds used in the training and the one third used in the evaluation testing. For the across-stage inferencing based multimodal data mining evaluations, we use the fourth and the fifth folders for the two stages inferencing evaluations, and use the third, the fourth and the fifth folders for the three stages inferencing evaluations. Consequently, each folder here is considered as a "stage" in the across-stage inferencing based multimodal data mining evaluations. In each of the inferencing scenarios, we use the same split as we do in the retrieval based multimodal data mining evaluations for training and testing.

In order to facilitate the across-stage inferencing capabilities, we handcraft the ontology of the words involved in the evaluations. This is simply implemented as a simple look-up table indexed by an efficient hashing function. For example, *cardiac mesoderm primordium* in the fourth folder is considered as the same as *circulatory system* in the fifth folder. With this simple ontology and word matching, the proposed approach may be well applied to this across-stage inferencing problem for the multimodal data mining.

The EMML algorithm is applied to obtain the model parameters. In the figures below, the horizonal axis denotes the number of the top retrieval results. We investigate the performance from top 2 to top 50 retrieval results. Figure 4 reports the precisions and recalls averaged over 1648 queries for image annotation in comparison with MBRM model where the solid lines are for precisions and the dashed lines are for recalls. Similarly, Figure 5 reports the precisions and recalls averaged over 64 queries for word query in comparison with MBRM model. Figure 6 reports the precisions and recalls averaged over 1648 queries for image retrieval in comparison with MBRM model.

Table 1: Comparison of scalability

Database Size	50	100	150
EMML	1	1	1
MBRM	1	2.2	3.3

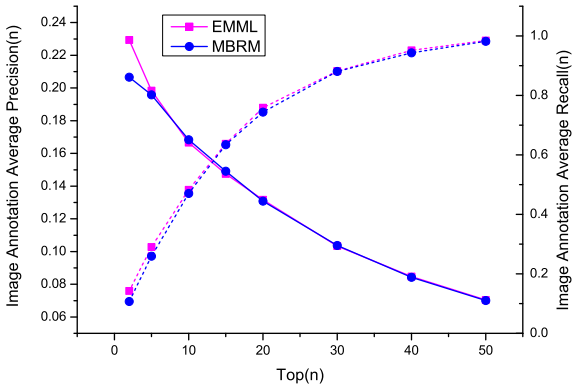


Figure 4: Precisions and Recalls of image annotation between EMML and MBRM (the solid lines are for precisions and the dashed lines are for recalls)

For the 2-stage inferencing, Figure 7 reports the precisions and recalls averaged over 1648 queries for image-to-word inferencing in comparison with MBRM model, and Figure 8 reports the precisions and recalls averaged over 64 queries for word-to-image inferencing in comparison with MBRM model. Figure 9 reports the precisions and recalls averaged over 1648 queries for image-to-image inferencing in comparison with MBRM model. Finally, for the 3-stage inferencing, Figure 10 reports precisions and recalls averaged over 1100 queries for image-to-image inferencing in comparison with MBRM model.

In summary, there is no single winner for all the cases. Overall, EMML outperforms MBRM substantially in the scenarios of word query and image retrieval, and slightly in the scenario of 2-stage word-to-image inferencing and 3-stage image-to-image inferencing. On the other hand, MBRM has a slight better performance than EMML in the scenario of 2-stage image-to-word inferencing. For all other scenarios the two methods have a comparable performance.

In order to demonstrate the strong scalability of EMML approach to multimodal data mining, we take image annotation as a case study and compare the scalability between EMML and MBRM. We randomly select three subsets of the embryo image database in different scales (50, 100, 150 images, respectively), and apply both methods to the subsets to measure the query response time. The query response time is obtained by taking the average response time over 1648 queries. Since EMML is implemented in MATLAB environment and MBRM is implemented in C in Linux environment, to ensure a fair comparison, we report the scalability as the relative ratio of a response time to the baseline response time for the respective methods. Here the baseline

response time is the response time to the smallest scale subset (i.e., 50 images). Table 1 documents the scalability comparison. Clearly, MBRM exhibits a linear scalability w.r.t the database size while that of EMML is constant. This is consistent with the scalability analysis in Sec. 5.

In order to verify the fast learning advantage of EMML in comparison with the existing max margin based learning literature, we have implemented one of the most recently proposed max margin learning methods by Taskar et al. [19]. For the reference purpose, in this paper we call this method as TCKG. We have applied both EMML and TCKG to a small data set randomly selected from the whole Berkeley embryo database, consisting of 110 images along with their annotation words. The reason we use this small data set for the comparison is that we have found that in MATLAB platform TCKG immediately runs out of memory when the data set is larger, due to the large number of the constraints, which is typical for the existing max margin learning methods. Under the environment of 2.2GHz CPU and 1GB memory, TCKG takes about 14 hours to complete the learning for such a small data set while EMML only takes about 10 minutes. We have examined the number of the constraints reduced in both methods during their executions for this data set. EMML has reduced the number of the constraints in a factor of 70 times more than that reduced by TCKG. This explains why EMML is about 70 times faster than TCKG in learning for this data set.

7. CONCLUSION

We have developed a new max margin learning framework — the enhanced max margin learning (EMML), and applied it to developing an effective and efficient multimodal data mining solution. EMML attempts to find a small set of active constraints, and thus is more efficient in learning than the existing max margin learning literature. Consequently, it has a much faster convergence rate which is verified in empirical evaluations. The multimodal data mining solution based on EMML is highly scalable in the sense that the query response time is independent of the database scale. This advantage is also supported through the complexity analysis as well as empirical evaluations. While EMML is a general learning framework and can be used for general multimodal data mining, for the evaluation purpose, we have applied it to the Berkeley Drosophila embryo image database and have reported the evaluations against a state-of-the-art multimodal data mining method.

8. ACKNOWLEDGEMENT

This work is supported in part by NSF (IIS-0535162, EF-0331657), AFRL (FA8750-05-2-0284), AFOSR (FA9550-06-1-0327), and a CAREER Award to EPX by the National Science Foundation under Grant No. DBI-054659.

9. REFERENCES

[1] `http://www.fruitfly.org/`.

[2] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden markov support vector machines. In *Proc. ICML*, Washington DC, 2003.

[3] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Maching Learning Research*, 3:1107–1135, 2003.

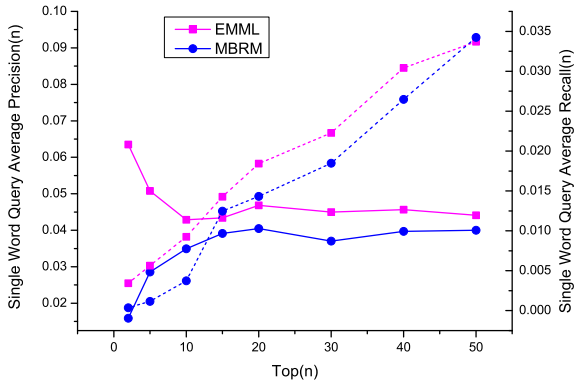


Figure 5: Precisions and Recalls of word query between EMML and MBRM

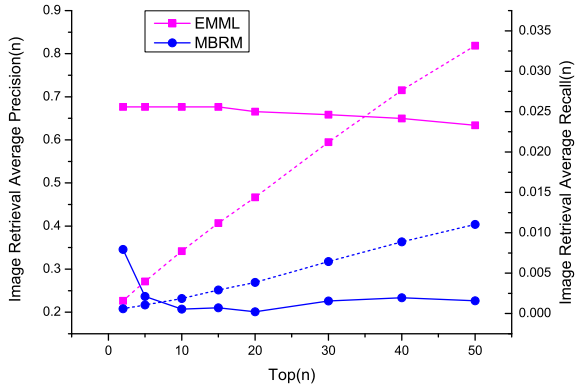


Figure 6: Precisions and Recalls of image retrieval between EMML and MBRM

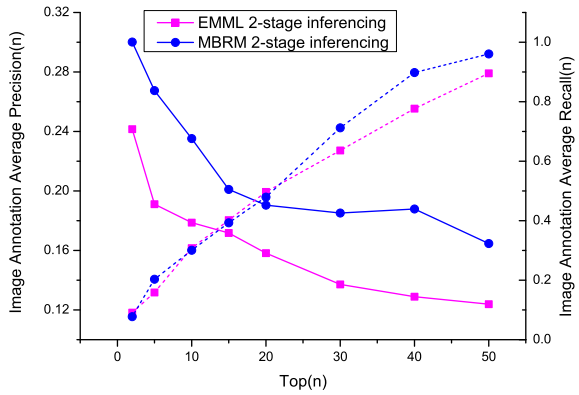


Figure 7: Precisions and Recalls of 2-stage image to word inferencing between EMML and MBRM

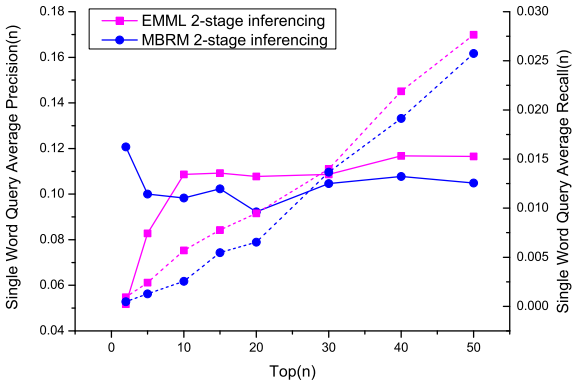


Figure 8: Precisions and Recalls of 2-stage word to image inferencing between EMML and MBRM

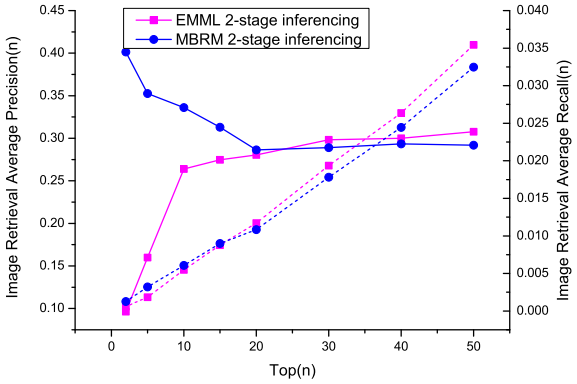


Figure 9: Precisions and Recalls of 2-stage image to image inferencing between EMML and MBRM

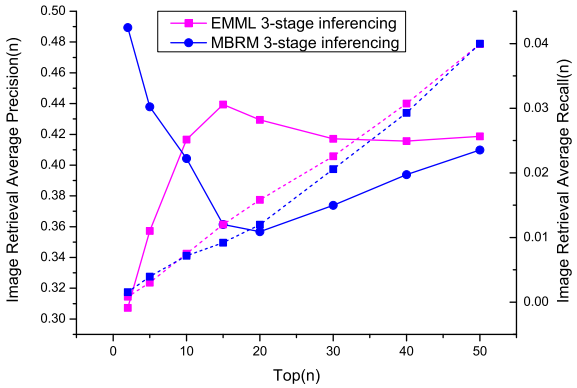


Figure 10: Precisions and Recalls of 3-stage image to image inferencing between EMML and MBRM

[4] D. Blei and M. Jordan. Modeling annotated data. In *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134, 2003.

[5] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[6] U. Brefeld and T. Scheffer. Semi-supervised learning for structured output variables. In *Proc. ICML*, Pittsburgh, PA, 2006.

[7] E. Chang, K. Goh, G. Sychay, and G. Wu. Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Trans. on Circuits and Systems for Video Technology*, 13:26–38, Jan 2003.

[8] W. Chu, Z. Ghahramani, and D. L. Wild. A graphical model for protein secondary structure prediction. In *Proc. ICML*, Banff, Canada, 2004.

[9] R. Datta, W. Ge, J. Li, and J. Z. Wang. Toward bridging the annotation-retrieval gap in image search by a generative modeling approach. In *Proc. ACM Multimedia*, Santa Barbara, CA, 2006.

[10] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Seventh European Conference on Computer Vision*, volume IV, pages 97–112, 2002.

[11] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *International Conference on Computer Vision and Pattern Recognition*, Washington DC, 2004.

[12] Y. Freund and R. E. Schapire. Large margin classification using the perceptron algorithm. In *Maching Learning*, volume 37, 1999.

[13] H. D. III and D. Marcu. Learning as search optimization: Approximate large margin methods for structured prediction. In *Proc. ICML*, Bonn, Germany, 2005.

[14] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, 2001.

[15] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy markov models for information extraction and segmentation. In *Proc. ICML*, 2000.

[16] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *Proc. of IEEE NNSP'97*, Amelia Island, FL, Sept. 1997.

[17] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *Proceedings of the 10th ACM SIGKDD Conference*, Seattle, WA, 2004.

[18] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22:1349–1380, 2000.

[19] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin. Learning structured prediction models: A large margin approach. In *Proc. ICML*, Bonn, Germany, 2005.

[20] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *Neural Information Processing Systems Conference*, Vancouver, Canada, 2003.

[21] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proc. ICML*, Banff, Canada, 2004.

[22] V. N. Vapnik. *The nature of statistical learning theory.* Springer, 1995.

[23] Y. Wu, E. Y. Chang, and B. L. Tseng. Multimodal metadata fusion using causal strength. In *Proc. ACM Multimedia*, pages 872–881, Hilton, Singapore, 2005.