Relevance Topic Model for Unstructured Social Group Activity Recognition

Authored by:

Fang Zhao
Yongzhen Huang
Liang Wang
Tieniu Tan

Abstract

Unstructured social group activity recognition in web videos is a challenging task due to 1) the semantic gap between class labels and low-level visual features and 2) the lack of labeled training data. To tackle this problem, we propose a relevance topic model" for jointly learning meaningful mid-level representations upon bag-of-words (BoW) video representations and a classifier with sparse weights. In our approach, sparse Bayesian learning is incorporated into an undirected topic model (i.e., Replicated Softmax) to discover topics which are relevant to video classes and suitable for prediction. Rectified linear units are utilized to increase the expressive power of topics so as to explain better video data containing complex contents and make variational inference tractable for the proposed model. An efficient variational EM algorithm is presented for model parameter estimation and inference. Experimental results on the Unstructured Social Activity Attribute dataset show that our model achieves state of the art performance and outperforms other supervised topic model in terms of classification accuracy, particularly in the case of a very small number of labeled training videos."

1 Paper Body

The explosive growth of web videos makes automatic video classification important for online video search and indexing. Classifying short video clips which contain simple motions and actions has been solved well in standard datasets (such as KTH [1], UCF-Sports [2] and UCF50 [3]). However, detecting complex activities, specially social group activities [4], in web videos is a more difficult task because of unstructured activity context and complex multi-object interaction. In this paper, we focus on the task of automatic classification of unstructured social group activities (e.g., wedding dance, birthday party and graduation ceremony in Figure 1), where the low-level features have innate limitations in

semantic description of the underlying video data and only a few labeled training videos are available. Thus, a common method is to learn human-defined (or semi-human-defined) semantic concepts as mid-level representations to help video classification [4]. However, those human defined concepts are hardly generalized to a larger or new dataset. To discover more powerful representations for classification, we propose a novel supervised topic model called ?relevance topic model? to automatically extract latent ?relevance? topics from bag-of-words (BoW) video representations and simultaneously learn a classifier with sparse weights. Our model is built on Replicated Softmax [5], an undirected topic model which can be viewed as a family of different-sized Restricted Boltzmann Machines that share parameters. Sparse Bayesian learning [6] is incorporated to guide the topic model towards learning more predictive topics which are associated with sparse classifier weights. We refer to those topics corresponding to non-zero weights as ?relevance? topics. Meanwhile, binary stochastic units in Replicated Softmax are replaced by rectified linear units [7], which allows each unit to express more information for better 1

(a) Wedding Dance

(b) Birthday Party

(c) Graduation Ceremony

Figure 1: Example videos of the ?Wedding Dance?, ?Birthday Party? and ?Graduation Ceremony? classes taken from the USAA dataset [4]. explaining video data containing complex content and also makes variational inference tractable for the proposed model. Furthermore, by using a simple quadratic bound on the log-sum-exp function [8], an efficient variational EM algorithm is developed for parameter estimation and inference. Our model is able to be naturally extended to deal with multi-modal data without changing the learning and inference procedures, which is beneficial for video classification tasks.

2 Related work

The problems of activity analysis and recognition have been widely studied. However, most of the existing works [9, 10] were done on constrained videos with limited contents (e.g., clean background and little camera motions). Complex activity recognition in web videos, such as social group activity, is not much explored. Most relevant to our work is a recent work that learns video attributes to analyze social group activity [4]. In [4], a semi-latent attribute space is introduced, which consists of user-defined attributes, class-conditional and background latent attributes, and an extended Latent Dirichlet Allocation (LDA) [11] is used to model those attributes as topics. Different from that, our work discovers a set of discriminative latent topics without extra human annotations on videos. From the view of graphical models, most similar to our model are the maximum entropy discrimination LDA (MedLDA) [12] and the generative Classification Restricted Boltzmann Machines (gClassRBM) [13], both of which have been successfully applied to document semantic analysis. MedLDA integrates the max-margin learning and hierarchical directed topic models by optimizing a single objective function with a set of expected margin constraints. MedLDA tries to estimate parameters and find latent topics in a max-margin

sense, which is different from our model relying on the principle of automatic relevance determination [14]. The gClassRBM used to model word count data is actually a supervised Replicated Softmax. Different from the gClassRBM, instead of point estimation of classifier parameters, our proposed model learns a sparse posterior distribution over parameters within a Bayesian paradigm.

3

Models and algorithms

We start with the description of Replicated Softmax, and then by integrating it with sparse Bayesian learning, propose the relevance topic model for videos. Finally, we develop an efficient variational algorithm for inference and parameter estimation. 3.1

Replicated Softmax

The Replicated Softmax model is a two-layer undirected graphical model, which can be used to model sparse word count data and extract latent semantic topics from document collections. Replicated Softmax allows for very efficient inference and learning, and outperforms LDA in terms of both the generalization performance and the retrieval accuracy on text datasets. As shown in Figure 2 (left), this model is a generalization of the restricted Boltzmann machine (RBM). The bottom layer represents a multinomial visible unit sampled K times (K is the total number of words in a document) and the top layer represents binary stochastic hidden units. 2

tr

h W1

?c

...

W

W2

?c

v

y

v

C

Figure 2: Left: Replicated Softmax model: an undirected graphical model. Right: Relevance topic model: a mixed graphical model. The undirected part models the marginal distribution of video BoW vectors v and the directed part models the conditional distribution of video classes y given latent topics tr by using a hierarchical prior on weights ? . Let a word count vector v ? NN be the visible unit (N is the size of the vocabulary), and a binary topic vector h ? {0, 1}F be the hidden units. Then the energy function of the state {v, h} is defined as follows: F N F N X X X X Wij vi hj ? ai vi ? K bj hj , (1) E(v, h; ?) = ? i=1 j=1

i=1

j=1

where ? = {W, a, b}, Wij is the weight connected with vi and hj , ai and bj are the bias terms of visible and hidden units respectively. The joint distribution

over the visible and hidden units is defined by: XX 1 exp(?E(v, h; ?)), Z(?) = exp(?E(v, h; ?)), (2) P (v, h; ?) = Z(?) v h

where Z(?) is the partition function. Since exact maximum likelihood learning is intractable, the contrastive divergence [15] approximation is often used to estimate model parameters in practice. 3.2

Relevance topic model

The relevance topic model (RTM) is an integration of sparse Bayesian learning and Replicated Softmax, the main idea of which is to jointly learn discriminative topics as mid-level video representations and sparse discriminant function as a video classifier. We represent the video dataset with class labels y ? {1, ..., C} as D = {(vm , ym)}M m=1 , where each video is represented as a BoW vector v ? NN . Consider modeling video BoW vectors using the Replicated Softmax. Let tr = [tr1 , ..., trF] denotes a F-dimensional topic vector of one video. According to Equation 2, the marginal distribution over the BoW vector v is given by: 1 X P (v; ?) = exp(?E(v, tr ; ?)), (3) Z(?) tr Since videos contain more complex and diverse contents than documents, binary topics are far from ideal to explain video data. We replace binary hidden units in the original Replicated Softmax with rectified linear units which are given by: trj

= max(0, tj), P (tj —v; ?) = N (tj —Kbj +

N X

Wij vi , 1),

(4)

i=1

where N (?—?, ?) denotes a Gaussian distribution with mean ? and variance ? . The rectified linear units taking nonnegative real values can preserve information about relative importance of topics. Meanwhile, the rectified Gaussian distribution is semi-conjugate to the Gaussian likelihood. This facilitates the development of variational algorithms for posterior inference and parameter estimation, which we describe in Section 3.3. ?y }C Let ? = {? y=1 denote a set of class-specific weight vectors. We define the discriminant function as a linear combination of topics: F (y, tr , ?) = ? Ty tr . The conditional distribution of classes is 3

defined as follows: P (y—tr , ?) = PC

exp(F (y, tr , ?))

y 0 =1

exp(F (y 0 , tr , ?))

,

(5)

and the classifier is given by: y? = arg max E[F (y, tr , ?)—v].

(6)

y?C

The weights ? are given a zero-mean Gaussian prior: ?—? ?) = P (?

C Y F Y

P (?yj —?yj) =

y=1 j=1

C Y F Y

$?1$ N $(?yj —0, ?yj)$,

(7)

y=1 j=1

$?$ y $\}C$ where $? = \{?$ y=1 is a set of hyperparameter vectors, and each hyperparameter $?yj$ is assigned independently to each weight $?yj$. The hyperpriors over $?$ are given by Gamma distributions: $?) = P (?$

C Y F Y

C Y F Y

$P (?yj) =$

y=1 j=1

$?1$ c $c?1 ?d?$ d $?yj$ e ,

$?(c)$

(8)

y=1 j=1

where $?(c)$ is the Gamma function. To obtain broad hyperpriors, we set c and d to small values, e.g., c = d = 10?4 . This hierarchical prior, which is a type of automatic relevance determination prior [14], enables the posterior probability of the weights $?$ to be concentrated at zero and thus effectively switch off the corresponding topics that are considered to be irrelevant to classification. Finally, given the parameters $?$, RTM defines the joint distribution:

Y Y F C Y F P $(?yj —?yj)P (?yj)$. P $(tj —v; ?)$ P $(v, y, tr , ? , ? ; ?) =$ P $(v; ?)P (y—tr , ?)$ j=1

(9)

y=1 j=1

Figure 2 (right) illustrates RTM as a mixed graphical model with undirected and directed edges. The undirected part models the marginal distribution of video data and the directed part models the conditional distribution of classes given latent topics. We can naturally extend RTM to Multimodal RTM by using the undirected part to model the multimodal data v = {vmodl }L l=1 . Accordingly, QL modl modl P $(v; ?)$ in Equation 9 is replaced with l=1 P (v ;?). In Section 3.3, we can see that it will not change learning and inference rules. 3.3

Parameter estimation and inference

For RTM, we wish to find parameters $? = \{W, a, b\}$ that maximize the log likelihood on D: Z M ?d? ?, log P $(D; ?)$ = log P ({vm , ym , trm }M (10) m=1 , ? , ? ; ?)d{tm }m=1 d? ?, ? —D; ?) = P (? ?, ? , D; ?)/P (D; ?). Since exactly computand learn the posterior distribution P (? ing P (D; ?) is intractable, we employ variational methods to optimize a lower bound L on the log likelihood by introducing a variational distribution to approximate P ({tm }M m=1 , ? , ? —D; ?): Y

M Y F ?)q(? ?). Q({tm }M q(tmj) q(? (11) m=1 , ? , ?) = m=1 j=1

Using Jensens inequality, we have: Z log P $(D; ?)$ ¿ Q({tm }M m=1 , ? , ?) Q

M r ?—? ?)P (? ?) P (v ; ?)P (y —t , ?)P (t —v ; ?) P (? m m m m m m=1 ?d? ?. (12) d{tm }M log m=1 d? Q({tm }M m=1 , ? , ?) Note that P (ym —trm , ?) is not conjugate to the Gaussian prior, which makes it

5

intractable to compute ?) and q(tmj). Here we use a quadratic bound on the log-sum-exp (LSE) the variational factors q(? function [8] to derive a further bound. We rewrite P (ym —trm , ?) as follows: T P (ym —trm , ?) = exp(ym Trm ? ? lse(Trm ?)),

4

(13)

where Trm ? = [(trm)T ? 1 , ..., (trm)T ? C?1], ym = I(ym = c) is the one-of-C encoding of class PC?1 label ym and lse(x) , log(1 + y0 =1 exp(xy0)) (we set ? C = 0 to ensure identifiability). In [8], the LSE function is expanded as a second order Taylor series around a point ? , and an upper bound ?) with a fixed matrix A = 12 [IC ? ? C ?1+1 1C ? 1TC ?] is found by replacing the Hessian matrix H(? ? ?), where C = C ? 1, IC ? is the identity matrix of size M ? M and 1C ? is a such that A H(? M -vector of ones. Thus, similar to [16], we have: 1 T log P (ym —trm , ?) ¿ J(ym , trm , ? , ? m) = ym Trm ? ? (Trm ?)T ATrm ? + sTm Trm ? ? ?i , 2 ?m ? exp(? ?m ? lse(? ?m)), sm = A? 1 ?m ? ?Tm exp(? ?m ? lse(? ?m)) + lse(? ?m), ?i = ?Tm A? 2

(14) (15) (16)

?

where ? m ? RC is a vector of variational parameters. Substituting J(ym , trm , ? , ? m) into Equation 11, we can obtain a further lower bound: log P (D; ?) ¿ L(?, ?) =

M X

log P (vm ; ?) + EQ

m=1

+

M X

X M

J(ym , trm , ? , ? m)

m=1

?—? ?) + log P (? ?) ? Q({tm }M log P (tm —vm ; ?) + log P (? , ? , ?) . (17) m=1

m=1

Now we convert the problem of model training into maximizing the lower bound L(?, ?) with ?), q(? ?) and q(t) = {q(tmj)} as well as the parameters respect to the variational posteriors q(? ?m }. We can give some insights into the objective function L(?, ?): the first term ? and ? = {? is exactly the marginal log likelihood of video data and the second term is a variational bound of the conditional log likelihood of classes, thus maximizing L(?, ?) is equivalent to finding a set of model parameters and latent topics which could fit video data well and simultaneously make good predictions for video classes. Due to the conjugacy properties of the chosen distributions, we can directly calculate free-form ?), q(? ?) and parameters ? : variational posteriors q(? ?) = N (? ?—E? , V?), q(? ?) = q(?

F C Y Y

(18)

Gamma(?yj —? c, d?yj),

6

(19)

y=1 j=1

? = hTrm iq(t) E? ,

(20)

where h?iq denotes an exception with respect to the distribution q and V? =

X M

r T (Tm) ATrm q(t) + diagh?yj iq(??)

?1 , E? = V?

m=1

M X

(Trm)T

q(t)

(ym + sm),

m=1

(21) 1 2 1 . c? = c + , d?yj = d + ?yj ?) q(? 2 2

(22)

For q(t), the calculation is not immediate because of the rectification. Inspired by [17], we have the following free-form solution: q(tmj) =

?pos ?neg 2 2 N (tmj —?pos , ?pos)u(tmj) + N (tmj —?neg , ?neg)u(?tmj), Z Z

(23)

where u(?) is the unit step function. See Appendix A for parameters of q(tmj). Given ?, through repeating the updates of Equations 18-20 and 23 to maximize L(?, ?), we can ?), q(? ?) and q(t). Then given q(? ?), q(? ?) and q(t), we estimate obtain the variational posteriors q(? ? by using stochastic gradient descent to maximize L(?, ?), and the derivatives of L(?, ?) with 5

respect to ? are given by:

M N X

r 1 X ?L(?, ?) r = vi tj data ? vi tj model + vmi htmj iq(t) ? Wij vmi ? Kbj , ?Wij M m=1 i=1 ?L(?, ?) = hvi idata ? hvi imodel , ?ai

M N X

r ?L(?, ?) r K X = tj data ? tj model + htmj iq(t) ? Wij vmi ? Kbj , ?bj M m=1 i=1 where the derivatives of

PM

m=1

(24) (25) (26)

log P (vm ; ?) are the same as those in [5].

This leads to the following variational EM algorithm: ?), q(? ?) and q(t). E-step: Calculate variational posteriors q(? M-step: Estimate parameters ? = {W, a, b} through maximizing L(?, ?). These two steps are repeated until L(?, ?) converges. For the Multimodal RTM learning, we just additionally calculate the gradients of ?modl for each modality l in the M-step while the updating rules are not changed. After the learning is completed, according to Equation 6 the prediction for new videos can be easily obtained:

y? = arg max ? Ty q(??) htr ip(t—v;?) . (27) y?C

4
Experiments

We test our models on the Unstructured Social Activity Attribute (USAA) dataset 1 for social group activity recognition. Firstly, we present quantitative evaluations of RTM in the case of different modalities and comparisons with other supervised topic models (namely MedLDA and gClassRBM). Secondly, we compare Multimodal RTM with some baselines in the case of plentiful and sparse training data respectively. In all experiments, the contrastive divergence is used to efficiently approximate the derivatives of the marginal log likelihood and the unsupervised training on Replicated Softmax is used to initialize ?. 4.1

Dataset and video representation

The USAA dataset consists of 8 semantic classes of social activity videos collected from the Internet. The eight classes are: birthday party, graduation party, music performance, non-music performance, parade, wedding ceremony, wedding dance and wedding reception. The dataset contains a total of 1466 videos and approximate 100 videos per-class for training and testing respectively. These videos range from 20 seconds to 8 minutes averaging 3 minutes and contain very complex and diverse contents, which brings significant challenges for content analysis. Each video is represented using three modalities, i.e., static appearance, motion, and auditory. Specifically, three visual and audio local keypoint features are extracted for each video: scaleinvariant feature transform (SIFT) [18], spatial-temporal interest points (STIP) [19] and melfrequency cepstral coefficients (MFCC) [20]. Then the three features are collected into a BoW vector (5000 dimensions for SIFT and STIP, and 4000 dimensions for MFCC) using a soft-weighting clustering algorithm, respectively. 4.2

Model comparisons

To evaluate the discriminative power of video topics learned by RTM, we present quantitative classification results compared with other supervised topic models (MedLDA and gClassRBM) in the case of different modalities. We have tried our best to tune these compared models and report the best results. 1

Available at http://www.eecs.qmul.ac.uk/?yf300/USAA/download/.

6

Table 1: Classification accuracy of different supervised topic models for single-modal features.

Feature	Model	Accuracy (%)				
		20 topics	30 topics	40 topics	50 topics	60 topics
SIFT	Med LDA	44.72	44.17	43.07	42.80	40.74
	gClass RBM	45.40	46.11	47.08	46.81	49.72
	RTM	51.99	53.09	55.83	54.17	54.03
STIP	Med LDA	37.28	38.93	40.85	39.75	41.54
	gClass RBM	42.39	42.25	42.39	41.70	43.35
	RTM	48.29	49.11	50.62	51.71	51.17
MFCC	Med LDA	34.71	38.55	41.15	41.98	38.27
	gClass RBM	41.70	43.62	45.00	44.31	43.48

RTM 45.35 46.67 48.15 47.46 47.33

Table 2: Classification accuracy of different methods for multimodal features.
Method

100 Inst Accuracy (%) 10 Inst

Multimodal RTM 60 topics 90 topics 120 topics 150 topics 180 topics 60 topics 90 topics 120 topics 150 topics 180 topics

60.22 62.69 63.79 64.06 64.72 38.68 41.29 43.48 43.72 44.99

RS+SVM 54.60 56.10 57.34 59.26 60.63 23.73 28.53 30.59 33.47 35.94

Direct

SVM-UD+LR

SLAS+LR

66.0

65.0

65.0

29.0

37.0

40.0

Table 1 shows the classification accuracy of different models for three single-modal features: SIFT, STIP and MFCC. We can see that RTM achieves higher classification accuracy than MedLDA and gClassRBM in all cases, which demonstrates that through leveraging sparse Bayesian learning to incorporate class label information into topic modeling, RTM can find more discriminative topical representations for complex video data. 4.3

Baseline comparisons

We compare Multimodal RTM with the baselines in [4] which are the best results on the USAA dataset: Direct Direct SVM or KNN classification on raw video BoW vectors (14000 dimensions), where SVM is used for experiments with more than 10 instances and KNN otherwise. SVM-UD+LR SVM attribute classifiers learn 69 user-defined attributes, and then a logistic regression (LR) classifier is performed according to the attribute classifier outputs. SLAS+LR Semi-latent attribute space is learned, and then a LR classifier is performed based on the 69 user-defined, 8 class-conditional and 8 latent topics. Besides, we also perform a comparison with another baseline where different modal topics extracted by Replicated Softmax are connected together as video representations, and then a multi-class SVM classifier [21] is learned from the representations. This baseline is denoted by RS+SVM. The results are illustrated in Table 2. Here the number of topics of each modality is assumed to be the same. When the labeled training data is plentiful (100 instances per class), the classification performance of Multimodal RTM is similar to the baselines in [4]. However, We argue that our model learns a lower dimensional latent semantic space which provides efficient video representations and is able to be better generalized to a larger or new dataset because extra human defined concepts are not required in our model. When considering the classification scenario where only a very small number of training data are available (10 instances per class), Multimodal RTM can achieve better performance with an appropriate number (e.g., ¿ 90) of topics because the sparsity of relevance topics learned by RTM

can effectively prevent overfitting to specific training instances. In addition, our model outperforms RS+SVM in both cases, which demonstrates the advantage of jointly learning topics and classifier weights through sparse Bayesian learning. It is also interesting to examine the sparsity of relevance topics. Figure 3 illustrates the degree of correlation between topics and two different classes. We can see that the learned relevance topics are very sparse, which leads to good generalisation for new instances and robustness for small datasets. 7

Figure 3: Relevance topics discovered by RTM for two different classes. Vertical axis indicates the degree of positive and negative correlation.

5

Conclusion

This paper has proposed a supervised topic model, the relevance topic model (RTM), to jointly learn discriminative latent topical representations and a sparse classifer for recognizing unstructured social group activity. In RTM, sparse Bayisian learning is integrated with an undirected topic model to discover sparse relevance topics. Rectified linear units are employed to better fit complex video data and facilitate the learning of the model. Efficient variational methods are developed for parameter estimation and inference. To further improve video classification performance, RTM is also extended to deal with multimodal data. Experimental results demonstrate that RTM can find more predictive video topics than other supervised topic models and achieve state of the art classification performance, particularly in the scenario of lacking labeled training videos.

Appendix A. Parameters of free-form variational posterior q(tmj) The expressions of parameters in q(tmj) (Equation 23) are listed as follows: 2 2 ?pos = N (?—?, ? + 1), ?pos = (? ?1 + 1)?1 , ?pos = ?pos (

? + ?), ?

2 ?neg = N (?—0, ?), ?neg = 1, ?neg = ?,

1 ??pos 1 ?neg Z = ?pos erfc q + ?neg erfc q , 2 2 2 2 2?pos 2?neg

where erfc(?) is the complementary error function and

+ P * ? ?j ym + sm ? j 0 6=j ? ?j 0 Atrmj 0 ?= , ??j A? ?T?j ?)q(t) q(? N D E?1 X ?T?j ? = ? ?j A? , ?= Wij vmi + Kbj . ?) q(?

(28) (29) (30)

(31)

(32)

i=1

We can see that q(tmj) depends on expectations over ? and {tmj 0 }j 0 6=j , which is consistent with the graphical model representation of RTM in Figure 2.

2 References

[1] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In ICPR, 2004. 8

[2] M. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In CVPR, 2008. [3] Ucf50 action dataset. ?http://vision.eecs.ucf.edu/data/ucf50.rar?. [4] Y.W. Fu, T.M. Hospedales, T. Xiang, and S.G. Gong. Attribute learning for understanding unstructured social activity. In ECCV, 2012. [5] R. Salakhutdinov and G.E. Hinton. Replicated softmax: an undirected topic model. In NIPS, 2009. [6] M.E. Tipping. Sparse bayesian learning and the relevance vector machine. JMLR, 2001. [7] V. Nair and G.E. Hinton. Rectified linear units improve restricted boltzmann machines. In ICML, 2010. [8] D. Bohning. Multinomial logistic regression algorithm. AISM, 1992. [9] P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea. Machine recognition of human activities: a survey. TCSVT, 2008. [10] J. Varadarajan, R. Emonet, and J.-M. Odobez. A sequential topic model for mining recurrent activities from long term video logs. IJCV, 2013. [11] D. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. JMLR, 2003. [12] J. Zhu, A. Ahmed, and E.P. Xing. Medlda: Maximum margin supervised topic models. JMLR, 2012. [13] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio. Learning algorithms for the classification restricted boltzmann machine. JMLR, 2012. [14] R.M. Neal. Bayesian learning for neural networks. PhD thesis, University of Toronto, 1995. [15] G.E. Hinton. Training products of experts by minimizing contrastive divergence. Neural Computation, 2002. [16] K.P. Murphy. Machine learning: a probabilistic perspective. MIT Press, 2012. [17] M. Harva and A. Kaban. Variational learning for rectified factor analysis. Signal Processing, 2007. [18] D.G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 2004. [19] I. Laptev. On space-time interest points. IJCV, 2005. [20] B. Logan. Mel frequency cepstral coefficients for music modeling. ISMIR, 2000. [21] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector learning for interdependent and structured output spaces. In ICML, 2004.

9