# Supervised Topic Models for Video Activity Recognition

**Michael C. Hughes**
Department of Computer Science
Brown University
Providence, RI 02906
mhughes@cs.brown.edu

## Abstract

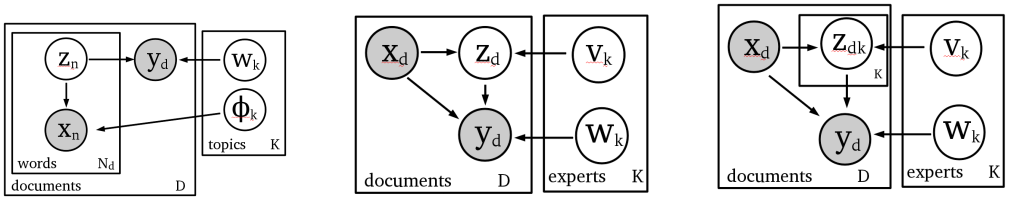
Topic models successfully capture latent structure useful for unsupervised analysis of bag-of-words data. Applying these models to domains such as video activity recognition requires two critical extensions: (1) incorporating supervised information (activity labels) to recover topic structure with greater discriminative power and (2) moving beyond the bag-of-words assumption to model temporal dynamics. We propose two parallel investigations to accomplish these tasks. First, we will study generic supervision techniques for topic models, exposing shortcomings in previously published generative approaches and exploring new discriminative models based on Mixtures of Experts. Second, we will apply these supervised models to video activity classification on the challenging Hollywood2 and Olympic Sports datasets, and explore extensions that capture chronological structure inherent in real-world activities.

## 1 Introduction

Organizing and understanding information captured in video is an exciting and useful application of modern machine learning research. As digital videos become more pervasive, automated tools are necessary to make this data accessible to human users. These tools should be capable of automatically discovering and labelling the relevant events in a video. For example, given footage of a baseball game, users might like the computer to identify whenever a home run occurs. This problem is called activity recognition and is the motivating application for this research project.

In recent years, the dominant framework for activity recognition has been to represent videos as "bags of features" and build classifiers that operate over these representations. Some recent efforts have examined applying *topic models* such as Latent Dirichlet Allocation (LDA) [1] to this task, in the hope of uncovering latent structure that is more expressive than the raw features alone. In this problem, each video clip in the collection is cast as a document, its low-level spatio-temporal features are vector quantized into "codewords", and the behaviors or events of interest are described by topics (groups of related codewords). For example, Niebles et al. [2] use LDA to classify video segments that portray solitary human actors engaging in simple repetitive behavior (e.g. jumping jacks or hand waving). They achieve state-of-the-art classifier accuracy ($\approx$80%) on benchmark datasets, but this approach suffers from two primary drawbacks.

First, this approach is unable to distinguish between complicated real-world activities (e.g. hitting a homerun or baking a cake). This shortcoming is a consequence of the bag-of-words assumption, which ignores the temporal structure necessary to describe more complicated activities. The second shortcoming in directly applying LDA to classification is that LDA is a fundamentally unsupervised method. This means the associated topics extracted from the corpus are not directly tied to the desired labelling task, and thus have questionable predictive value.




(a) Supervised LDA (b) Mixture of Experts (c) Admixture of Experts

Figure 1: Various Models for Supervised Learning on Bag-of-Words data

The goal of this research project is to overcome both of these shortcomings. We will pursue two parallel investigations: (1) a study developing and evaluating different techniques for applying topic models to supervised tasks, and (2) a study evaluating various supervised hierarchical models for activity recognition. The first investigation will make important contributions to understanding strengths and weaknesses of generative and discriminative models of latent structure in bag-of-words data. The second will explore applying these supervised topic models to activity recognition, studying different approaches to model temporal dynamics as well. The remainder of this document consists of two sections, outlining the two proposed investigations in greater detail.

## 2 Supervised Topic Models

Several models have been proposed for learning topics in a supervised fashion, encouraging recovery of topics that are some how discriminative to the predictive task at hand. For example, if we hope to accuractly estimate a reviewer's rating of a movie given a text review, we should expect that topics focused on emotive words such as "exciting" or "dull" would be more helpful than topics describing genre that might come out of classic LDA.

Two candidate generative models of interest for supervised learning are "Supervised LDA" (sLDA) [3], and more recently "Discriminative LDA" (DiscLDA) [4]. These differ fundamentally in the generative story for the output or response variable $y_d$ for a document. sLDA provides a "down-stream" model in which a response is drawn conditioned on the topic distribution observed in the document, as shown in figure 1(a). DiscLDA provides an upstream model where the document's label determines the available topics for a document. sLDA supports inference for any type of response variable (real-valued, categorical, integer, etc.) while DiscLDA is restricted to categorical responses only. We focus on sLDA in this analysis because of its flexible support of different response types, tractable inference schemes, and wider adoption (thus far) in the research community.

### 2.1 Disadvantages of sLDA model

Previous work by the authors of sLDA has shown that the method improves upon unsupervised LDA in its ability to discover topics useful for prediction [3]. However, we have noticed a previously unpublished weakness in this model: the power of supervision decreases as the number of words per document increases.

We can see this just by inspecting the generative model for sLDA, shown in figure 1(a). Given a single document, sLDA encourages the discovery of topic assignments $z_d$ (one topic per word) that maximize the following generative log-likelihood

$$\log p(z_d|x_d, y_d) \propto \log p(y_d|z_d) + \sum_{n=1}^{N_d} \log \left[p(x_{d,n}|z_{d,n})p(z_{d,n})\right] \tag{1}$$

where $x_d$ denotes a vector of observed word-counts in the document and $y_d$ denotes the document's scalar response (movie rating, essay score, etc.). Notice that this log-likelihood has two essential terms: a supervised term and then an unsupervised term identical to LDA's log-likelihood.

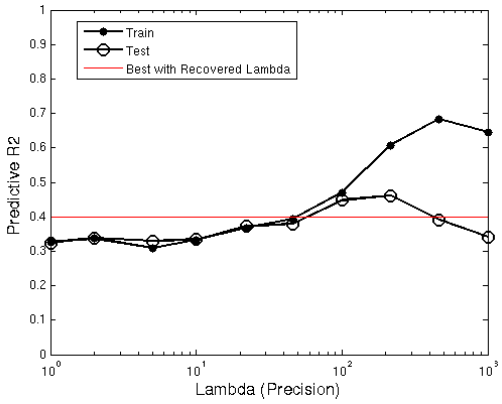



Figure 2: Tunable $\lambda$ sLDA performance on movie review rating prediction over a range of $\lambda$, compared to standard sLDA (red line). $K = 10$ topics.

Inspecting this equation reveals that when the number of words per document ($N_d$) is very large, the influence of the unsupervised term will dominate. In fact, as $N_d$ grows the model's optimization objective *asymptotically converges* to that of unsupervised LDA. For some small text datasets, this is not an issue. However, when applying sLDA to supervised tasks in image or video analysis, where each image "document" can have thousands or even tens of thousands of "words", we should not expect sLDA to deliver any appreciable benefit over unsupervised LDA. Indeed, in brief experiments (figure 3) we found that the two models give equivalent performance in classifying human actions in the Hollywood2 dataset, with LDA actually marginally better than its supervised counterpart.

Given this result, we observe that *generative likelihood may not be the best criterion when the goal is predictive power*. We can confirm this observation by considering the sLDA model for the moview reviews dataset in which the response $y_d$ for a document has the following Gaussian log-likelihood

$$\log p(y_d|z_d) \propto \frac{\lambda}{2}(y_d - w^T \bar{z}_d)^2 \tag{2}$$

Classic sLDA views the precision parameter $\lambda$ and weight vector $w$ as random variables to be infered based on training data. Instead, we can adopt an alternative approach in which we fix $\lambda$ to a specific value corresponding to how much emphasis we wish to the model to place on predictive power. Figure 2 shows the resulting predictive performance of this *tunable* $\lambda$ sLDA model on the moview reviews dataset[1] (as measured by Blei et al.'s predictive $R^2$ metric[2] ). We can see that there exists a certain interval for $\lambda$ (around 100-300) for which this model has noticeably better predictive power ($pR^2$ improves about 12% over that of sLDA). We can perhaps expect even more improvement for datasets with larger $N_d$.

### 2.2 Proposed Work

As a down-stream generative model, sLDA does not exhibit the best possible predictive performance, especially when documents contain many words. We propose in this investigation to consider discriminative models that would perform well on bag-of-words structured domains such as text or video. These models would not attempt to explain the origin of the words of a document, but instead concentrate on explaining the response conditioned on the observed words. However, we do wish to retain some features of the sLDA model, such as:

- Latent structure which provides an expressive probabilistic model
- Flexible support for different response types (real-values, integers, binary, etc.)

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data/

[2] Larger $pR^2$ values indicate better performance, perfect implies $pR^2 = 1$

This wishlist leads us to consider a classic family of models known as Mixtures of Experts (MoE) [5]. Briefly, Mixtures of Experts enable prediction by associating each input $x_d$ (e.g. bag-of-words vector for a document) with a single "expert" which indicates specific parameters to plug into a generalized linear model alongside $x_d$ for producing $y_d$, as diagramed in 1(b). We call this a "mixture" of experts because for a novel document, there is posterior uncertainty over which expert it should associate with, yielding a conditional likelihood of the form

$$p(y_d|x_d) = \sum_z p(y_d|x_d, z_d)p(z_d|x_d) \tag{3}$$

The first step in this investigation will be to evaluate the mixture of experts approach on bag-of-words data. We plan to investigate synthetic data for validation purposes, and use the movie reviews dataset as in [3] as a standard for evaluating a model's predictive power.

#### 2.2.1 Admixture of Experts Model

We anticipate that mixture of experts alone will likely be insufficiently flexible to truly maximize predictive performance. This is due to the fact that sLDA allows a more expressive model in which each *word* is assigned to a latent topic (aka expert), rather than the document as a whole.

Motivated by this, we propose a novel model called the Admixture of Experts (adMoE), in which each document is expressly associated with a soft assignment to experts indicated by a probability vector. That is, instead of $z_d$ taking one of $K$ discrete values as in MoE, we instead allow it to be a vector of length $K$ whose entries are positive and sum to unity.

One possible probabilistic formulation for adMoE might draw a document's admixture of experts $\bar{z_d}$ from a Dirichlet distribution, and consider the response normally distribution with mean determined by a weighted combination over $\bar{z_d}$, just as in sLDA.

$$z_d \sim \text{Dirichlet}(e^{v_1^T x_d}, e^{v_2^T x_d}, \dots e^{v_K^T x_d}) \tag{4}$$

$$y_d \sim \mathcal{N}(w^T z_d, \tau) \tag{5}$$

As part of this investigation, we hope to (1) identify tractable inference methods for this model, especially MCMC sampling techniques, and (2) explore different model families for obtaining the admixtures $z_d$ as well as the response $y_d$. Time permitting, we also hope to investigate non-parametric versions of the MoE and adMoE

## 3 Hierarchical Models for Human Action Recognition

### 3.1 Current state-of-the-art

Most existing research (e.g. [6] ) adopts a bag-of-words approach that ignores spatial and temporal dependencies among the observed video features. The typical pipeline represents each video as a bag of code words, and uses discriminative methods such as SVMs to classify these bags. Researchers have identified several low-level feature representations (such as Histograms of Gradients and Histograms of Optical Flow [6]) that allow near perfect classification performance on simple exercise datasets. Recent efforts have attempted to move on to videos from the wild, such as sports footage or YouTube videos. The bag-of-words classifiers do perform significantly better than chance in these settings, but leave much to be desired in order to be used in meaningful applications.

### 3.2 Disadvantages of current methods

As this field moves beyond hand-crafted datasets of simple repetitive actions to unconstrained videos showing complex activities, two important complexities make bag-of-words approaches undesierable. First, modeling activities that have complex chronological structure requires modeling temporal dynamics. Second, analyzing real-world video clips requires ability to localize activities to

| | LDA | sLDA | SIFT |
|---|---|---|---|
| AnswerPhone | 0.1375 | 0.1434 | 0.1596 |
| DriveCar | 0.6592 | 0.5927 | 0.6017 |
| Eat | 0.0557 | 0.0631 | 0.0711 |
| FightPerson | 0.2196 | 0.2184 | 0.3002 |
| GetOutCar | 0.3604 | 0.3583 | 0.3830 |
| HandShake | 0.1562 | 0.1238 | 0.1619 |
| HugPerson | 0.1365 | 0.1361 | 0.1329 |
| Kiss | 0.3508 | 0.3388 | 0.3843 |
| Run | 0.4909 | 0.4984 | 0.5194 |
| SitDown | 0.1756 | 0.1844 | 0.2017 |
| SitUp | 0.0590 | 0.0545 | 0.0730 |
| StandUp | 0.2599 | 0.2452 | 0.2556 |
| MEAN AP | 0.2551 | 0.2464 | 0.2704 |

Figure 3: Comparing SVM classifier performance via average precision on Hollywood2 activity recognition dataset. Data in the SIFT column based on SVM classifier using raw SIFT codewords, while other columns used the topic distribution of each document as features. All topic models used fixed $K = 50$ topics.

segments of that video. It is never guaranteed that every single frame in the video contains evidence of the action. Instead of treating each video as an indivisable whole that either contains an activity or not, models must be able to localize activities within segments of a video clip.

### 3.3 Proposed Work

We propose several lines of inquiry in the realm of video understanding and activity recognition. First, we are interested in directly applying supervised topic models to the task. Additionally, we are interested in extending these models to overcome the shortcomings listed above.

The primary dataset of interest is Hollywood2 [6]. This dataset contains clips from Hollywood movies such as Pirates of the Carribbean, and presents a valuable challenge due to multiple actors, diverse viewpoints, occlusions, etc. This will be the primary benchmark for evaluating our proposed models. However, when considering temporal dynamics the activities in the Hollywood2 dataset may not be too interesting, as most consist of repetitive actions like running or driving. To this end, we will consider the recent Olympic Sports dataset released by Fei-Fei Li's group at Stanford [7]. This dataset contains visually similar, chronologically-complicated activities such as high-jump and long-jump that would be ideal for testing methods attempting to model temporal dynamics.

#### 3.3.1 Direct Application of Supervised Topic Models

Preliminary experiments (see table 3 ) have shown that topics discovered by classic unsupervised LDA are roughly as discriminative as raw dense SIFT features for predicting the action categories in Hollywood2. We hope that the results of our earlier investigation into supervised topic models will provide methods that promote topics even more discriminative on this task, thus pushing the state-of-the-art for bag-of-words approaches.

We plan to compare and contrast several approaches to the task of labeling a novel test video with a activity label such as running or swimming. First, we can evaluate the direct labelling output of a supervised model. Second, we can consider instead describing each video by its latent representation (e.g. the distribution topics or experts assigned to it), and use this latent representation as input to a standard classifier such as an SVM (as was done in figure 3).

#### 3.3.2 Determining relevance of individual frames in a video clip

We plan to consider models of varying complexity to tackle the localization problem. As a first step, we are interested in developing techniques that can identify relevant and irrelevant frames in a video labeled as containing a certain action. For example, if only the last few seconds of a video show a person driving a car, we would like the algorithm to only use these as positive examples

of that behavior. This is a classic Multiple Instance Learning problem, where we have weak labels describing an entire collection (video clip), and the inference problem is identifying which members (frames) within the collection are actually positive examples of a particular category.

Possible inference approaches we intend to explore include

- mi-SVM [8] – a well-studied baseline MIL approach based on support vector machines
- MILES [9] – a more recent idea showing good results in object recognition tasks

If these appear promising on bag-of-words data, we can later consider incorporating an MIL approach into supervised topic models.

#### 3.3.3 Modeling chronological structure for activities in a video clip

Incorporating temporal dynamics into hierarchical model is a difficult problem. We expect to borrow ideas from classic time-dynamics models such as Hidden Markov Models here. A preliminary model might represent observed features in each frame as generated by latent topics. The available topics for a particular frame might be drawn conditioned on those available in the previous frame. A major challenge will be identifying efficient inference procedures for these complex models, and we hope to make important contributions here.

#### 3.3.4 Exploring video clustering in unsupervised settings

If the models discussed above prove sufficiently powerful on real-world videos, a number of applications beyond simple activity classification are worth investigating. One application we hope to consider is the task of automatically grouping videos that have related content (ala the related videos panel on YouTube). We suggest that models capturing temporal dynamics and capable of localizing activity segments within video clips would be naturally powerful candidates to help discover similar videos given a query. We could also incorporate text tags as supervised side information in order to create even more meaningful clusters.

## References


[1] David Blei, Andrew Ng, and M. Jordan. Latent dirichlet allocation. In *Journal of Machine Learning Research*, volume 3, pages 993–1022, 2003.

[2] Juan Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79:299–318, 2008. 10.1007/s11263-007-0122-4.

[3] David M. Blei and Jon D. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS) 21*, 2007.

[4] Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems (NIPS) 20*, 2008.

[5] Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the em algorithm. In *Neural Computation*, volume 6, pages 181–214, 1994.

[6] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2008.

[7] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *Proceedings of the 11th European Conference on Computer vision: Part II*, ECCV'10, pages 392–405, Berlin, Heidelberg, 2010. Springer-Verlag.

[8] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems (NIPS) 15*, volume 15, pages 577–584. MIT Press, 2003.

[9] Yixin Chen, Jinbo Bi, and James Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1931–1947, 2006.