Fast Supervised LDA for Discovering Micro-Events in Large-Scale Video Datasets

Aristotle University of Thessaloniki

Angelos Katharopoulos, Despoina Paschalidou, Christos Diou, Anastasios Delopoulos

Multimedia Understanding Group, Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, Greece

ACM multimedia

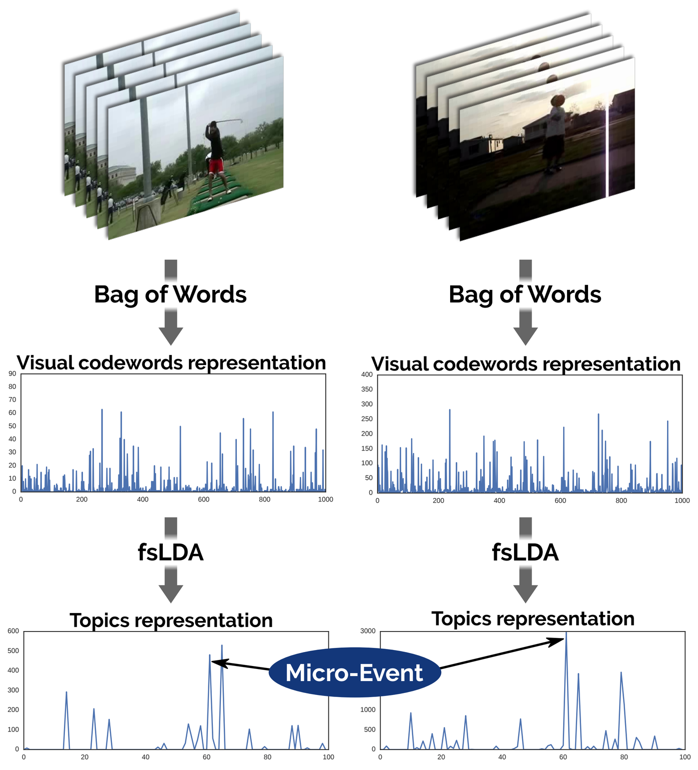
Approach

Can topic modelling be used to infer video structure for video event detection?

- Issues with existing topic modelling algorithms:
  - **Latent Dirichlet Allocation (LDA)** can result in class irrelevant topics
  - **Supervised LDA (sLDA)** is intractable for large-scale datasets
  - **LDA** and **sLDA** have similar performance

We propose a new variational inference method, **Fast Supervised Latent Dirichlet Allocation (fsLDA)**, able to:

- Identify meaningful discriminative components in videos, which we call **micro-events**
- **Retain class relevant information** so that the topics are relevant to the performed actions



Fast Supervised LDA

Fast Supervised LDA (fsLDA) reduces the computational compexity of sLDA and increases the influence of class relevant information on the infered topics to improve classifcation performance.

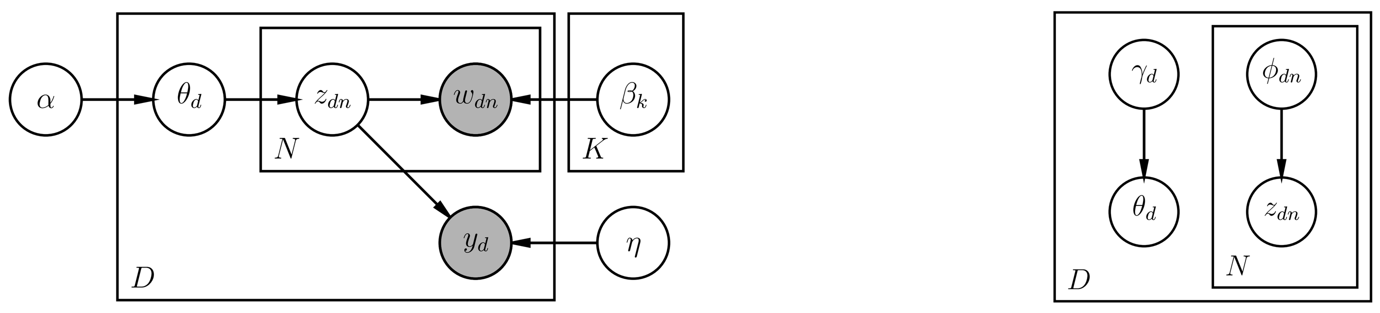


Figure 1: (Left) The graphical model representation of fsLDA. (Right) The graphical model representation of the variational distribution used to approximate the posterior of fsLDA

Given a document and the corresponding class label y_d, the posterior distribution of the latent variables $p(\theta, z \mid w, y, \alpha, \beta, \eta)$ is intractable. Therefore, we use variational methods to approximate this posterior.

Variational distribution: $q(\theta, z_{1:N} \mid \gamma, \phi_{1:N}) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n)$

Kullback-Leibler (KL) divergence: $\mathrm{KL}(q \parallel p) = -(\mathbb{E}_q[\log p(\theta, z, w, y, \alpha, \beta, \eta)] - \mathbb{E}_q[\log q(\theta, z)]) + \log p(w, y, \alpha, \beta, \eta) = -\mathcal{L}(\gamma, \phi \mid \alpha, \beta, \eta) + \log p(w, y, \alpha, \beta, \eta)$

Evidence Lower Bound (ELBO): $\mathcal{L}(\gamma, \phi \mid \alpha, \beta, \eta) = \mathbb{E}_q[\log p(\theta \mid \alpha)] + \mathbb{E}_q[\log p(z \mid \theta)] + \mathbb{E}_q[\log p(w \mid \beta, z)] + H(q) + \underbrace{\mathbb{E}_q[\log p(y \mid z, \eta)]}_{\textbf{Problematic term}}$

Problematic term: $\mathbb{E}_q\left[\log p(y \mid z, \eta)\right] = \eta_y^T \, \mathbb{E}_q\left[\bar{z}\right] - \mathbb{E}_q\left[\log \sum_{\hat{y}=1}^{C} \exp(\eta_{\hat{y}}^T \bar{z})\right]$

1. We use **Jensen's inequality** for the problematic term

$$-\mathbb{E}_q\left[\log \sum_{\hat{y}=1}^{C} \exp(\eta_{\hat{y}}^T \bar{z})\right] \geq -\log \sum_{\hat{y}=1}^{C} \mathbb{E}_q\left[\exp(\eta_{\hat{y}}^T \bar{z})\right]$$

2. We approximate using **Second-order Taylor expansion**

$$-\log \sum_{\hat{y}=1}^{C} \mathbb{E}_q\left[\exp(\eta_{\hat{y}}^T \bar{z})\right] \approx -\log \sum_{\hat{y}=1}^{C} \exp(\eta_{\hat{y}}^T \, \mathbb{E}_q[\bar{z}]) \left(1 + \frac{1}{2} \eta_{\hat{y}}^T \, \mathbb{V}_q[\bar{z}] \eta_{\hat{y}}\right)$$

3. The **variance term** $\mathbb{V}_q[\bar{z}] = \frac{1}{N^2}\left(\sum_{n=1}^{N} \sum_{m \neq n} \phi_n \phi_m^T + \sum_{n=1}^{N} \mathrm{diag}(\phi_n)\right)$ is very small in the case of Mutlimedia due to N, the word counts, which exceeds 10,000 and thus it **can be omitted**

4. The derivative of $\mathcal{L}$ w.r.t. ϕ_n, having added the Lagrange Multipliers λ_n, is

$$\frac{d\mathcal{L}_{\phi_n}}{d\phi_n} = \left(\Psi(\gamma) - \Psi(\sum_{j=1}^{K} \gamma_j)\right) + \log \beta_n - \log \phi_n - 1 + \lambda_n + \frac{1}{N}\left(\eta_y - \underbrace{\sum_{\hat{y}=1}^{C} s_{\hat{y}} \eta_{\hat{y}}}_{s = \mathrm{softmax}(\mathbb{E}_q[\bar{z}], \eta)}\right)$$

5. s **changes very slowly** w.r.t ϕ_n, thus we derive a closed form update rule for ϕ_n

Closed form update rules	
$\phi_n \propto \beta_n \exp\left(\Psi(\gamma) + \frac{\mathcal{C}}{\max(\eta)}\left(\eta_y - \sum_{\hat{y}=1}^{C} s_{\hat{y}} \eta_{\hat{y}}\right)\right)$	$\beta_{ij} \propto \sum_{d,n} \phi_{dni} \mathbf{1}(j = w_n)$
$\gamma = \alpha + \sum_{n=1}^{N} \phi_n$	$\eta = \underset{\eta}{\mathrm{argmax}} \left(\sum_{d=1}^{D} \eta_{y_d}^T \, \mathbb{E}_q[\bar{z}_d] - \sum_{d=1}^{D} \log \sum_{\hat{y}=1}^{C} \exp(\eta_{\hat{y}}^T \, \mathbb{E}_q[\bar{z}_d])\right)$

Experimental Results

We conducted **qualitative** and **quantitative** experiments in **UCF-11** and **UCF-101** datasets using state-of-the-art local features such as **Improved Dense Trajectories** (IDT) and **Deep Convolutional Neural Networks** (DCNNS).

Qualitative analysis of a topic

Figure 2: Qualitative analysis shows that topics are semantic and transcend classes

fsLDA **outperforms** both sLDA and LDA in UCF-11 and UCF-101 in a variety of motion and visual content descriptors with respect to **classification accuracy** (*see Table*).

We observe that this superiority is accentuated when reducing the feature dimensions using either mRMR feature selection or training with a smaller number of topics.

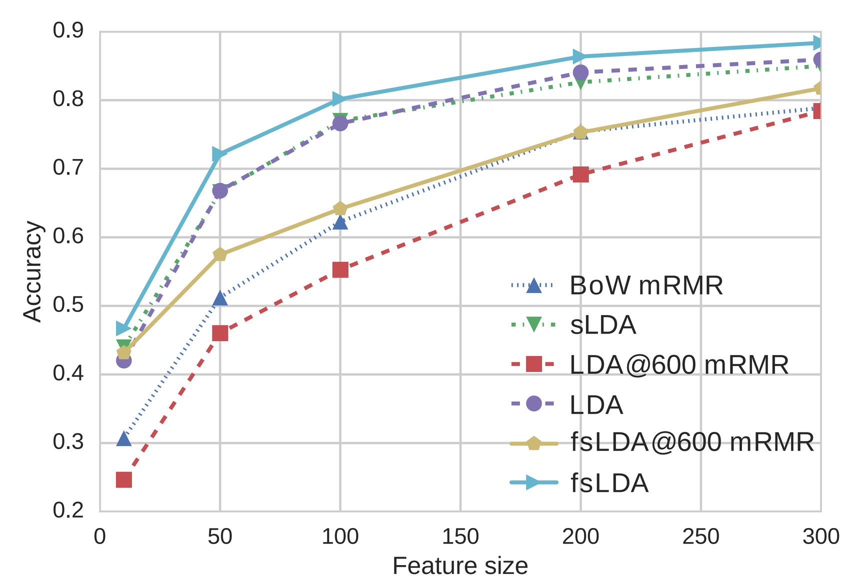


Figure 3: Comparison of fsLDA, sLDA, LDA and BOW using few dimensions to represent videos (UCF-11 idt-hog)

Dataset	*Feature*	*fsLDA*	*sLDA*	*LDA*
UCF-11	idt-hog	**0.9299**	0.9018	0.9118
UCF-11	idt-hof	0.8530	**0.8592**	0.8374
UCF-11	idt-mbhx	**0.8449**	0.8323	0.8336
UCF-11	idt-mbhy	**0.8580**	0.8455	0.8480
UCF-11	idt-traj	**0.7904**	0.7748	0.7754
UCF-11	dsift	**0.9280**	0.9143	**0.9280**
UCF-101	VGG 2014 conv5_2	**0.6237**	Intractable	0.5603
UCF-101	idt-hof	**0.5607**	Intractable	0.5272

Table 1: Comparison of fsLDA, sLDA and LDA with respect to classification accuracy

We observe that fsLDA is **comparably fast** with LDA while being **30-200 times faster** than sLDA.

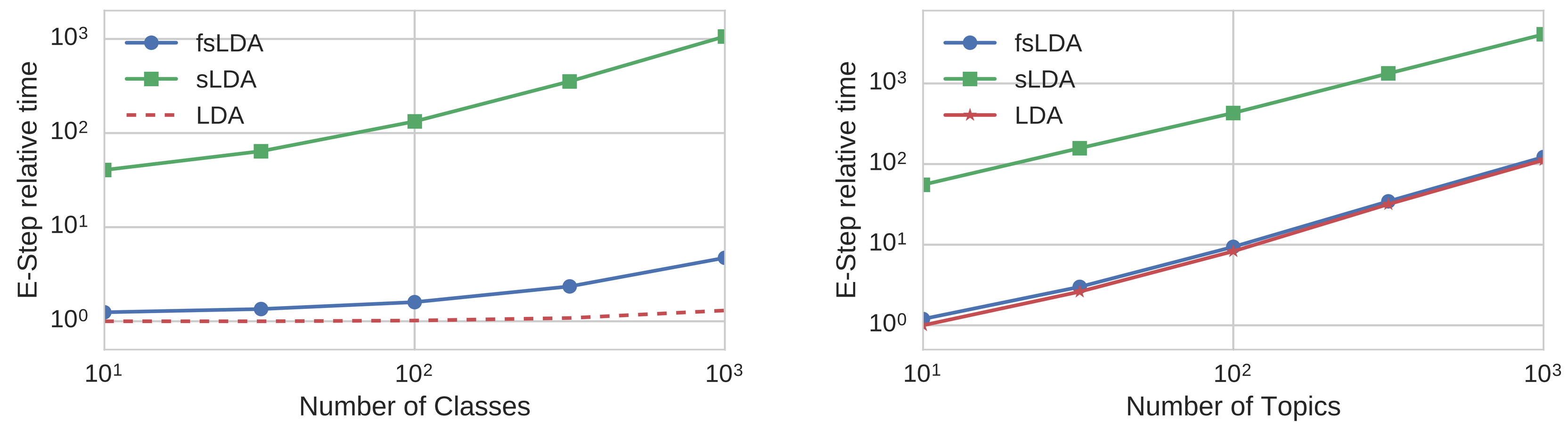


Figure 4: Speed comparison between fsLDA, sLDA and LDA on artificial data

Conclusions

We developed a **new variational inference method**, fsLDA, which

- is able to infer topics in a supervised manner
- in contrast to sLDA, is **faster**, **more discriminative** and **tractable** for large-scale datasets
- is able to decompose videos into **semantic components**, called micro-events
- outperforms both LDA and sLDA with respect to classification accuracy

Code & Data

Efficient C++ implementations for fsLDA, sLDA and LDA as well as all the data used in this paper are available at `http://ldaplusplus.com/r/research`