Finding Egocentric Image Topics through Convolutional Neural Network Based Representations

Kai Zhen, David Crandall
School of Informatics and Computing, Indiana University.

Life-logging cameras create huge collections of photos, even for a single person on a single day [1, 6], which makes it difficult for users to browse or organize their photos effectively. Unlike text corpora in which words create intermediate representations that carry semantic meaning for higher-level concepts such as topics, images have no such obvious intermediate representation to connect raw pixels and semantics. Egocentric photos are particularly challenging because they were taken opportunistically, so they are often blurry and poorly-composed compared to consumer-style images.

This paper applies topic modeling on deep features to extract visual "concept clusters" from egocentric datasets. We discretize features to form a better analogy to the word-document model, which we find yields faster convergence during inference. We also find that removing frequent, less informative features helps to prevent outliers and improve the semantic meaning of extracted topics, analogous to removing stop words in the text mining domain. In a generative process similar to that proposed in LDA [2], we model an image as being generated by first choosing topics, and then sampling features (visual words) from selected topics,

$$p(w,z,\theta,\phi|\alpha,\beta) = p(\theta|\alpha)p(z|\theta)p(\phi|\beta)p(w|\phi,z), \qquad (1)$$

where α and β are hyperpriors, θ is the distribution over topics for each document (image), ϕ is the distribution over words for each topic, and z represents the topic allocations. Parameters z, θ, and ϕ are inferred from the posterior distribution.

For inference, we use collapsed Gibbs sampling, which is based on the observation that ϕ and θ can be represented in closed form as functions of z [3],

$$\begin{aligned}\theta_{i,z} &= \frac{n(i,z)+\alpha}{\sum_Z (n(i,z)+\alpha)},\\ \phi_{z,w} &= \frac{n(z,w)+\beta}{\sum_W (n(z,w)+\beta)},\end{aligned} \qquad (2)$$

where $n(i,z)$ is the number of words in image i being assigned to topic z, $n(z,w)$ is the total number of times the word w is assigned to the topic z, and Z and W are the total number of topics and number of distinct words, respectively. In this work, we are particularly interested in Θ, which is a matrix which maps topics and images, and lets us sample representative images from each topic for visualization purposes.

The collapsed Gibbs sampler needs to calculate the probability of the n^{th} word in the m^{th} image being assigned to topic k, given all other topic assignments of the remaining words in all images. We calculate this probability by integrating out the multinomial parameters,

$$\begin{aligned}&p(z_{(m,n)} = k|z_{-(m,n)}, w, \alpha, \beta)\\ &\quad\approx \frac{\alpha + C^{-(m,n)}_{k,m,*}}{Z*\alpha + C^{-(m,n)}_{*,m,*}} * \frac{\beta + C^{-(m,n)}_{k,*,w_{m,n}}}{W*\beta + C^{-(m,n)}_{k,*,*}}\\ &\quad\approx \frac{(\alpha + C^{-(m,n)}_{k,m,*})(\beta + C^{-(m,n)}_{k,*,w_{m,n}})}{W*\beta + C^{-(m,n)}_{k,*,*}}\end{aligned} \qquad (3)$$

where $-(m,n)$ refers to all words but the n^{th} word in the m^{th} image, and k refers to a specifc topic (of the K possible topics). For example, $C^{-(m,n)}_{k,m,*}$ means the count of all the words in image m that have been assigned to topic k, except for the n^{th} word; $C^{-(m,n)}_{k,*,*}$ means the count for all words from all images with the topic assignment of k, except for the n^{th} word in the m^{th} image. In practice, the term $C^{-(m,n)}_{k,*,*}$ is dropped out as it is a constant in each image.

Preprocessing. The extracted features have relatively low frequency, and some appear only once in the whole dataset. We address this problem by discretizing the features into discrete bins. To choose a precision, we conducted sensitivity analysis of the Gibbs sampling convergence rate with respect to the degree of quantization and the number of topics, and results showed that 2 decimal digits of precision led to a higher convergence rate given a specific number of topics.

We also found that filtering out visual "stop words" was critical to applying LDA in this domain: when we kept the common features that occur in most images, the very generic "uninteresting" topics containing those terms overwhelmed other topics. We filter visual words that occur in more than half of the images, which removed roughly 30% of the visual vocabulary in our dataset.

Experiments. To investigate the ability of LDA to summarize lifelogging data, we conducted preliminary experiments on a dataset of first-person images captured by one of the authors. We wore a Narrative Clip lifelogging camera, which takes pictures about every 30 seconds, for a week during Summer 2015. The camera captured 7,927 (12 days) images of a wide variety daily activities including commuting to work, having meetings, interacting with friends and family, etc. The lifelogging user removed about 20 images that he felt too private to share. We implemented collapsed Gibbs sampling for LDA in C++, with the hyperprior α being 0.5 and β being 0.1. We experimented with using both the output of the second to the last fully connected layer (fc7) and the output of the last fully connected layer (fc8) of AlexNet [5]. The results were similar, so we used fc8 (so that each feature corresponds with the confidence of the image belonging to a certain ImageNet object category) since this yielded a lower-dimensional representation. The experiments were conducted on a Dell PowerEdge T630 server with a NVidia Tesla K40 GPU for feature extraction via Caffe [4]. Feature extraction took approximately 200 minutes for all images in the dataset, and LDA inference took about 20 minutes on the dataset of a single day, and about 150 minutes on the dataset of a week.

We first learned topics on a small subset of 130 egocentric images from a single day. The images included a visit to the library, studying in front of a laptop, and riding a bicycle home. We used this small dataset with roughly three obvious topics in order to test the discretization and stopping components of our application of LDA to this domain. We extracted different numbers of topics from the dataset and selected the top five images from each topic for visualization purposes, as shown in Fig. 1. In the case of two topics, the first topic (1*st* row) seems to on "T"-shaped objects, while the second (2*nd* row) seems related to a prominent dark object. For the case of three topics, the first seems to correspond with "T"-shaped objects that are thinner, while the third is relatively thicker. While this pattern makes sense, a human user might have hoped for a topic grouping at a higher, semantic level of abstraction, such as differentiating the table scenarios from the bicycle ones. One possible explanation for this is that these topics are being selected primarily based on very common but uninformative features that occur in almost every image. For example, as Tab. 1 shows, the second topic is highly influential even for those images that seem to be generated by other topics.

Removing "visual stop words," which we define to be features occurring in more than half of images, improves the situation, as shown in the lower part of Tab. 1. The major topic plays the leading role in generating the corresponding images, even though the predominant topic is still quite influential compared to others. Fig. 1(c) and (d) shows a visualization of the discovered topics once stop words have been removed. We notice that the "dark pattern" topic has disappeared, as it occurs in over half of the images and should be regarded as noise. In the case of two topics (Fig. 1(c)), the T-shaped bicycle is now merged with T-shaped tables, and the second topic is

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

Table 1: Distribution over topics for the top image in each topic, with and without stop words removed. Without removing stop words, the predominant topic is highly influential even for those images from other topics. By removing common terms that appear in more than half of the images, we dampen the influence of the predominant topic.

Without stop word filtering			
images	topic 1	topic 2*	topic 3
1st image of topic 1	0.200600	0.489745	0.309655
1st image of topic 2	0.001498	0.969046	0.029456
1st image of topic 3	0.079460	0.556222	0.364318
With stop word filtering			
images	topic 1	topic 2	topic 3*
1st image of topic 1	0.382138	0.232796	0.385066
1st image of topic 2	0.103614	0.534940	0.361446
1st image of topic 3	0.006993	0.006993	0.986014

mainly for the laptop scenario. In the case of three topics (Fig. 1(d)), LDA appears to break the second topic from the two topic case into one related to computer screens and another topic for desks. The first topic is cleaner than before, with all top images being bicycles.

We next tested our approach on a larger dataset of 3,075 images from a whole week of lifelogging. The results for $K = 4$ topics are shown in Fig. 1(e). The first topic seems related to leafy trees and blue sky, while the third topic is about the dark pattern in the left corner and the sky. For the second image in this topic, it has a tree, which is darker, at the left corner and the sky, and has also been categorized under this topic. The fourth topic seems to correspond with humans. Notice that images under the second topic are not pure black (some of them were taken when the camera wearer was in a tunnel), and that those images are associated with distinct feature representations.

Conlusion. While the results we obtain here are reasonable, they are preliminary and a human observer would likely select different groupings than those selected by LDA. For example, LDA's top topics are related to homogeneity and frequency of images in each of those topics, whereas a human would likely ignore redundant images and instead group more based on novelty or "interestingness" of images in a topic. LDA has no notion of aesthetics or novelty, however, so it lacks this ability. Moreover, a human would likely use higher levels of semantics than those discovered by LDA to make grouping decisions, e.g. based on activities occurring in the photos instead of purely based on visual scene appearance alone.

[1] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1949–1957, 2015.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[4] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[6] Michael Ryoo and Larry Matthies. First-person activity recognition: What are they doing to me? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2730–2737, 2013.

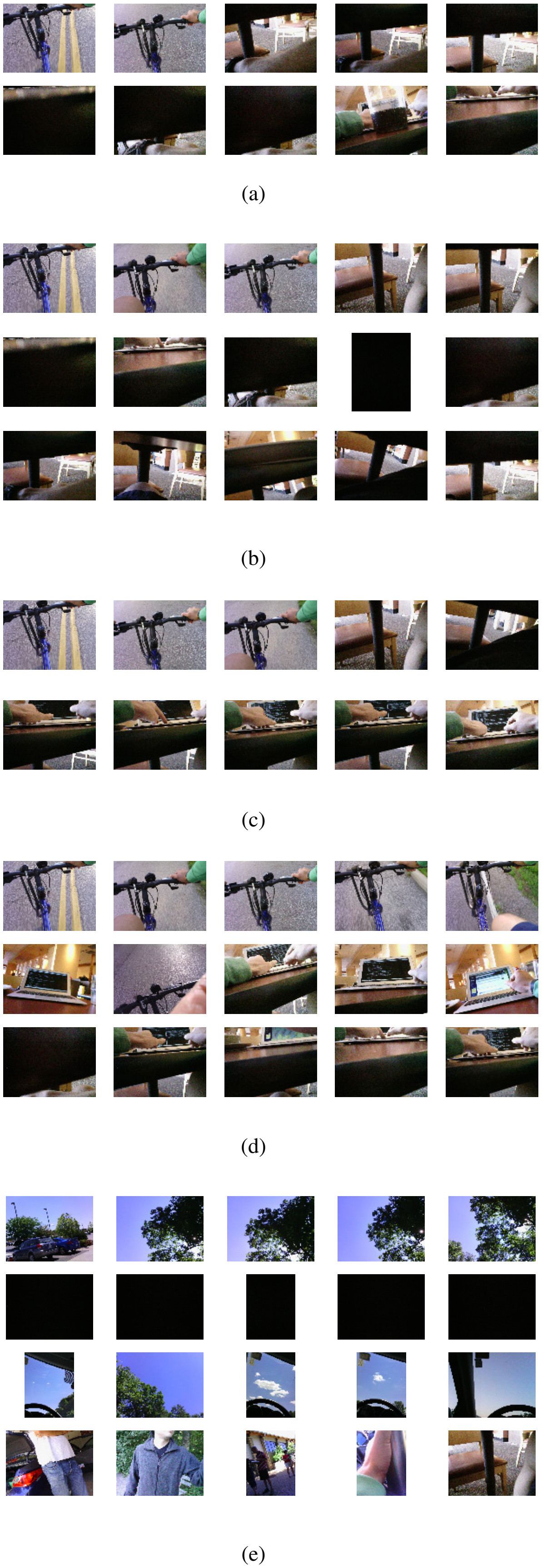

Figure 1: Top five images from each topic inferred by LDA, for (a) two topics from a single day with stop words; (b) three topics from a single day with stop words; (c) two topics on a single day without stop words; (d) three topics on a single day without stop words; and (e) four topics on weekly data without stop words (we picked four topics with consistent representatives out of ten topics).