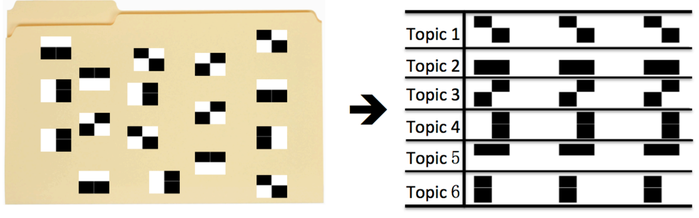
Finding Topics of Egocentric Data through Convolutional Neural Network Based Representations

Kai Zhen, David Crandall
Indiana University, Bloomington IN

Motivation

- Life-logging cameras create huge collections of photos, even for a single person on a single day, which makes it difficult for users to browse or organize their photos effectively.
- Here we use a Convolutional Neural Network (CNN) to extract visual words and apply topic modeling on top. We hope that the large collection of egocentric images can be organized under a certain set of topics.



Methods

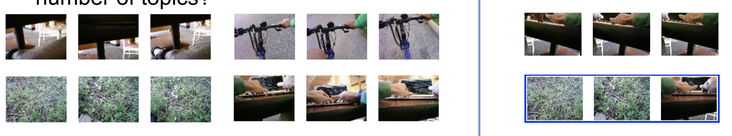
- As our intermediate representation, we extract deep image features using AlexNet [2]:
  - We extract (1,000 dimensional) features at the fc7 and fc8 layers.
  - We use top 50 responses from each layer as a bag of words.
- Collapsed Gibbs sampling for LDA [1]:
  - Here each image is analogous to a single document. Like conventional LDA, we assume that each image is a multinomial distribution over topics and each topic is a multinomial distribution over words.
  - To visualize results, we look into the θ matrix in which $\theta_{i,z}$ means the probability of image i being generated by topic z.

$$\theta_{i,z} = \frac{n(i,z)+\alpha}{\sum_Z (n(i,z)+\alpha)}$$

, where $n(i,z)$ is the count of visual words in image i being assigned to topic z; α is a hyper prior.

Experiment

- Data: Using a Narrative Clip life-logging camera, we collected 7,927 images over 12 days covering a wide variety daily activities including commuting to work, having meetings, preparing and eating meals, interacting with friends and family, etc.
- Environment: Experiments were run on a Dell PowerEdge T630 server with a NVidia Tesla K40 GPU for feature extraction via Caffe.
- Results:
  - Can we get higher semantic levels of visual concepts by decreasing the number of topics?



Left: 4 topics Right: 2 topics

> In the case of 4 topics, some topics are highly correlated, however higher level of semantic visual concepts cannot be acquired by simply decreasing the number of topics.

  - Are 1,000 labels enough to categorize these egocentric images?

> They may not accurately describe the corresponding categories, but will suffice to measure the dissimilarity among images.

[Flagpole, window]

[Laptop, bannister]

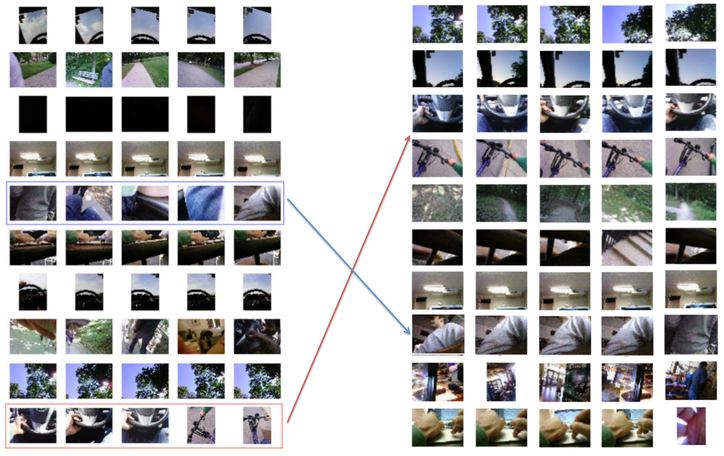
[Park, chainlink, picket]

[Parking, car, minivan]

[Jean, seat, wool]

  - What is the difference between the features extracted from fc7 and fc8?

> 10 topics are picked out manually from 30 topics due to limited space.



fc7 fc8

Conclusion

- We find a way to organize egocentric images by grouping them with recurring topics via a combination of CNN features and LDA.
- Topics discovered by the method are related more to **homogeneity and frequency** of images in each topics, whereas a human would likely use **higher levels of semantics** to make grouping decisions.

References:

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

[2] 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