Topic Modeling for Unsupervised Image Clustering and Retrieval

Crystal Wang, Yaateh Richardson, Ryan Sander

Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

Project Motivation

- Annotation of large imagery datasets has historically been a significant obstacle for training supervised machine learning models.
- Automated techniques for clustering and retrieving images via latent object classes can accelerate this annotation pipeline through human-in-the-loop automated annotation.
- Topic modeling is a set of techniques that can be applied to data to cluster it based off of observed features and latent topics.

Research Question and Problem Statement

Research Question: Which input features, clustering algorithms, and topic modeling algorithms lead to the lowest distributional differences between images assigned to the same topic?

Problem Statement: Given a set of unlabeled RGB images, how can we cluster these images according to the objects contained within them to accelerate the image annotation process?

Dataset

Dataset: ADE20K dataset, which consists of **22210** RGB images and class-segmented ground truth labels over **758** classes.

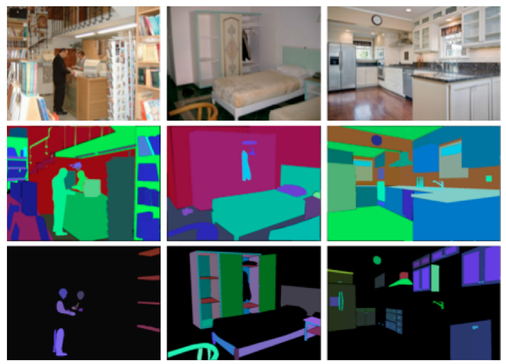


Figure 1: Samples from the ADE20K dataset, with RGB image input (top), scene segmentation by ground truth classes (middle), and part segmentation (bottom).

The main purpose of the dataset was scene segmentation, so the class labels had large sample imbalance as shown below. To compensate we limited training and testing to same-size subsets across classes. This effectively reduced our train set size to **2000** images over **25** classes with **80** samples-per-image.

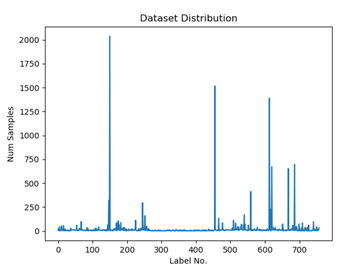


Figure 2: A graph of the distribution over labels. we handled this by restricting training and evaluation.

Technique: Feature Extraction, Clustering, and topic modeling Pipeline

The system we designed is composed of a pipeline with the following steps:

- **Feature Extraction**: Uses Computer Vision techniques such as SIFT and CNNs to extract important features for input to our clustering algorithm.
- **Clustering**: Uses k-means clustering to cluster features in a meaningful, discrete way for subsequent topic modeling.
- **Topic Modeling**: Uses clustered features and Latent Dirichlet Allocation (LDA) to assign distributions of topics to each image.

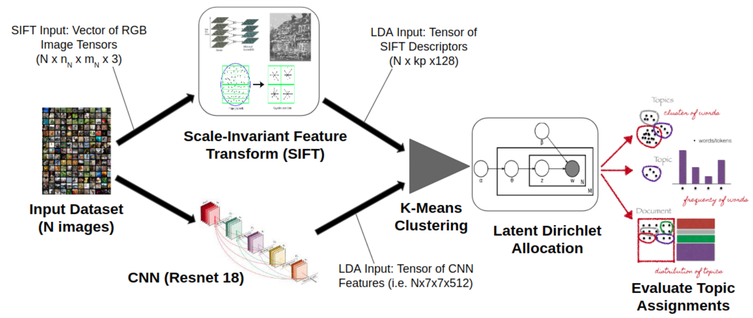


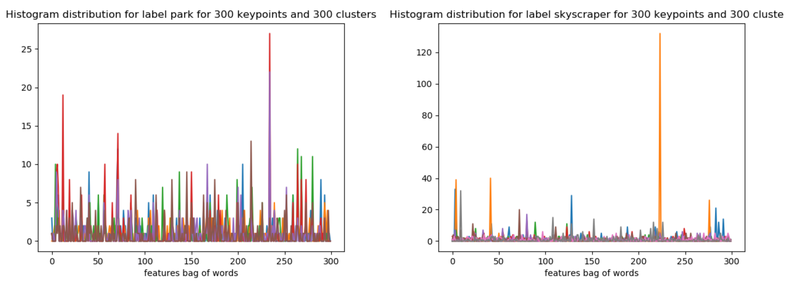
Figure 3: Block diagram detailing our unsupervised image clustering pipeline.

Results and Conclusions

[1] Intra-label Quantitative Validation: LDA creates latent topics that don't necessary map to any one of our given labels. Thus, to evaluate our performance we devised intra-label validation metrics using *KL-Divergence* and the MSE (L_2 Norm) between histograms:

$$D(k,c,t) = \frac{2}{N^2 - N} \sum_{i=1}^{N} \sum_{j=i+1}^{N} d(h(i;k,c,t), h(j;k,c,t)) \tag{1}$$

Where N denotes the number of images in a given label, k denotes the number of keypoints, c the number of clusters, t the number of LDA topics, $d(\cdot,\cdot)$ denotes our distance function, and $h(\cdot;k,c,t)$ denotes our histogram, implicitly parameterized by k, c, and t.



-			Class Label					
-			"living_room"		"skyscraper"		"park"	
Keypoints	Clusters	Topics	KL Div.	L_2 Dist.	KL Div.	L_2 Dist.	KL Div.	L_2 Dist.
150	150	100	6.624	0.3354	6.734	0.394	4.818	0.408
150	150	20	4.317	0.4001	4.295	0.4001	3.416	0.4001
200	200	20	4.291	0.378	4.364	0.446	3.202	0.502
300	300	20	4.195	0.339	4.165	0.438	3.017	0.506
350	300	20	4.254	0.322	4.403	0.322	3.141	0.322
400	300	20	4.099	0.286	4.309	0.415	3.142	0.447
500	400	20	4.172	0.312	4.320	0.423	3.320	0.550

Table 1: Average KL Divergence and L_2 Distances for the labels "living_room", "skyscraper", and "park".

Results and Conclusions (cont'd)

[2] Intra-label Qualitative Validation: To verify our algorithm qualitatively, we looked at images that were sorted into the same topic via LDA. Below are our results for three different latent topics.

Figure 4: Images from topic 4 of a model trained with 300 keypoints, 300 k-means clusters, and 20 topics.

Figure 5: Images from topic 3 of a model trained with 300 keypoints, 300 k-means clusters, and 20 topics.

Figure 6: Images from topic 5 of a model trained with 300 keypoints, 300 k-means clusters, and 20 topics.

Main Conclusions: Our findings lead us to the following conclusions:

- Latent Dirichlet Allocation, and potentially other topic modeling algorithms, is applicable not only for grouping/clustering language data, but for spatial, namely imagery, data as well.
- Feature extraction largely impacts the performance of this image clustering pipeline, as can be seen in Table 1: generally, larger ratios of **keypoints** and **clusters** to **topics** lead to better results.
- This framework has strong potential for image annotation automation and retrieval.

References

I. David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

II. David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60.2 (2004), pp. 91–110.

III. Xiaogang Wang and Eric Grimson. "Spatial latent dirichlet allocation". In: *Advances in neural information processing systems*. 2008, pp. 1577–1584.

IV. Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.

V. Bolei Zhou et al. "Scene Parsing through ADE20K Dataset". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.