PinterNet: A Thematic Label Curation Tool for Large Image Datasets

Open Science in Big Data (OSBD) Workshop at BigData 2016

December 5, 2016

Ruoqian Liu[1], Diana Palsetia[2], Arindam Paul[1], Reda Al-Bahrani[1], **Dipendra Jha**[1], Wei-keng Liao[1], Ankit Agrawal[1] and Alok Choudhary

Northwestern University

Overview:

Without crowd-sourced labeling, we labeled a 110K image data set and made it ***open-access***.

Larger and Better Labeled Image Set

- Booming of deep learning and related techniques calls for larger, cleaner, better labeled image data set
- Example of existing large-scale image datasets:
  - ImageNet (1.4 million images, 1000 classes)
  - Microsoft COCO (300K images, 80 classes of object, fine-grained segmentation)
  - Visual Genome (100K images, much richer sources of descriptions per image)
- How are those labels become alive? Human labeler.

Limitations of Crowd Labeling

- biased (by individual labelers)
- expensive
- lacks view of entire dataset

Label Curation Tool

Automatic Label Curation

Annotations:

father's day, *12*
mother day ideas, *25*
mother's day gifts, *16*
toddler crafts, *3*
father's day craft, *4*
father's day ideas, *5*
diy father's day gift, *5*
mother's day, *30*
father day crafts, *11*
mother's day crafts for kids, *5*
mother's day crafts for toddlers, *2*
mother's day gifts ideas, *6*
diy mother's day gift, *2*

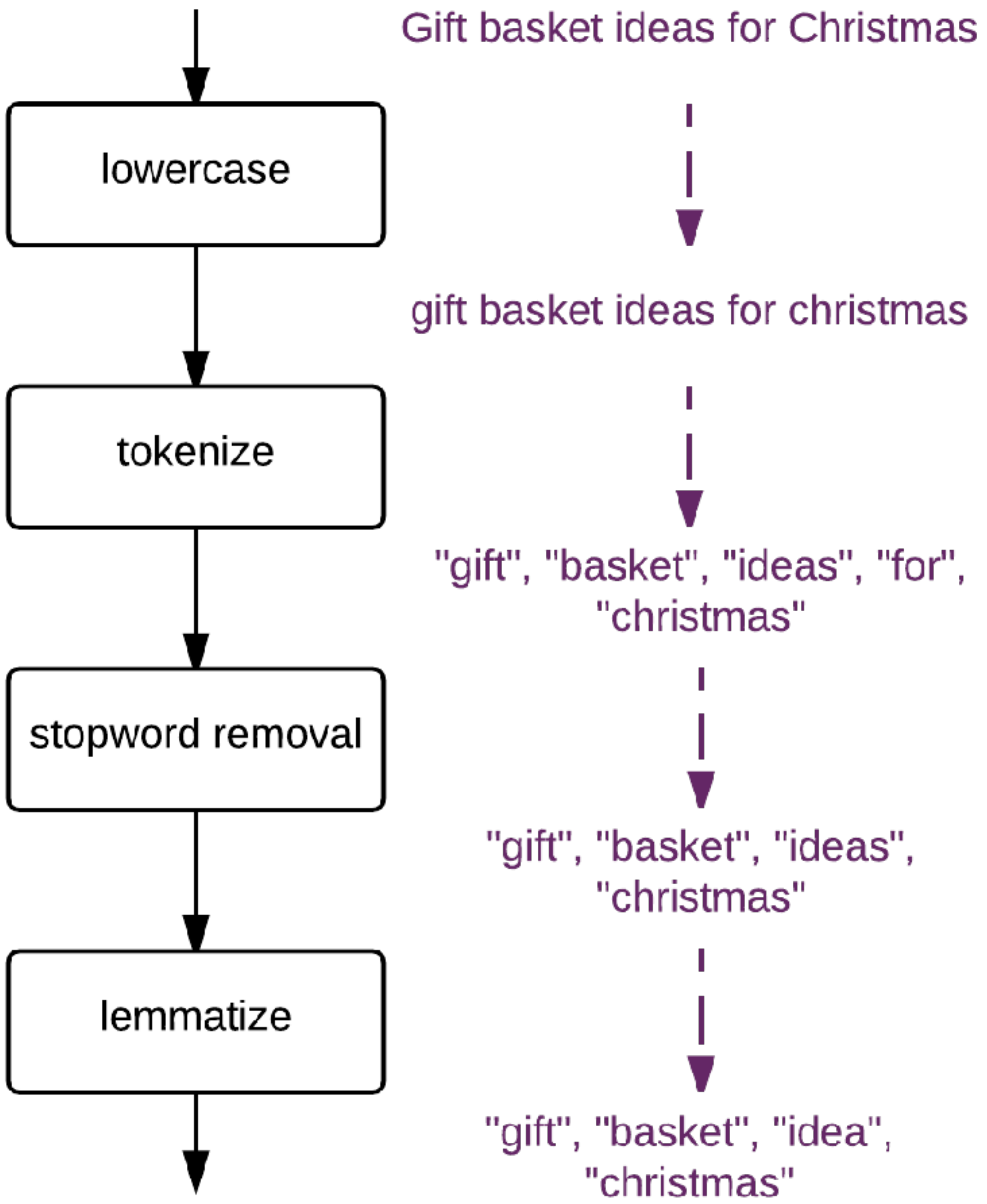
⇩ Refine

Labels:

mother's day
father's day
gift ideas
diy

- Automated
- Data-driven
- Modular
- Frequency based
- Multi-word label generated by automatic merging of frequent single words

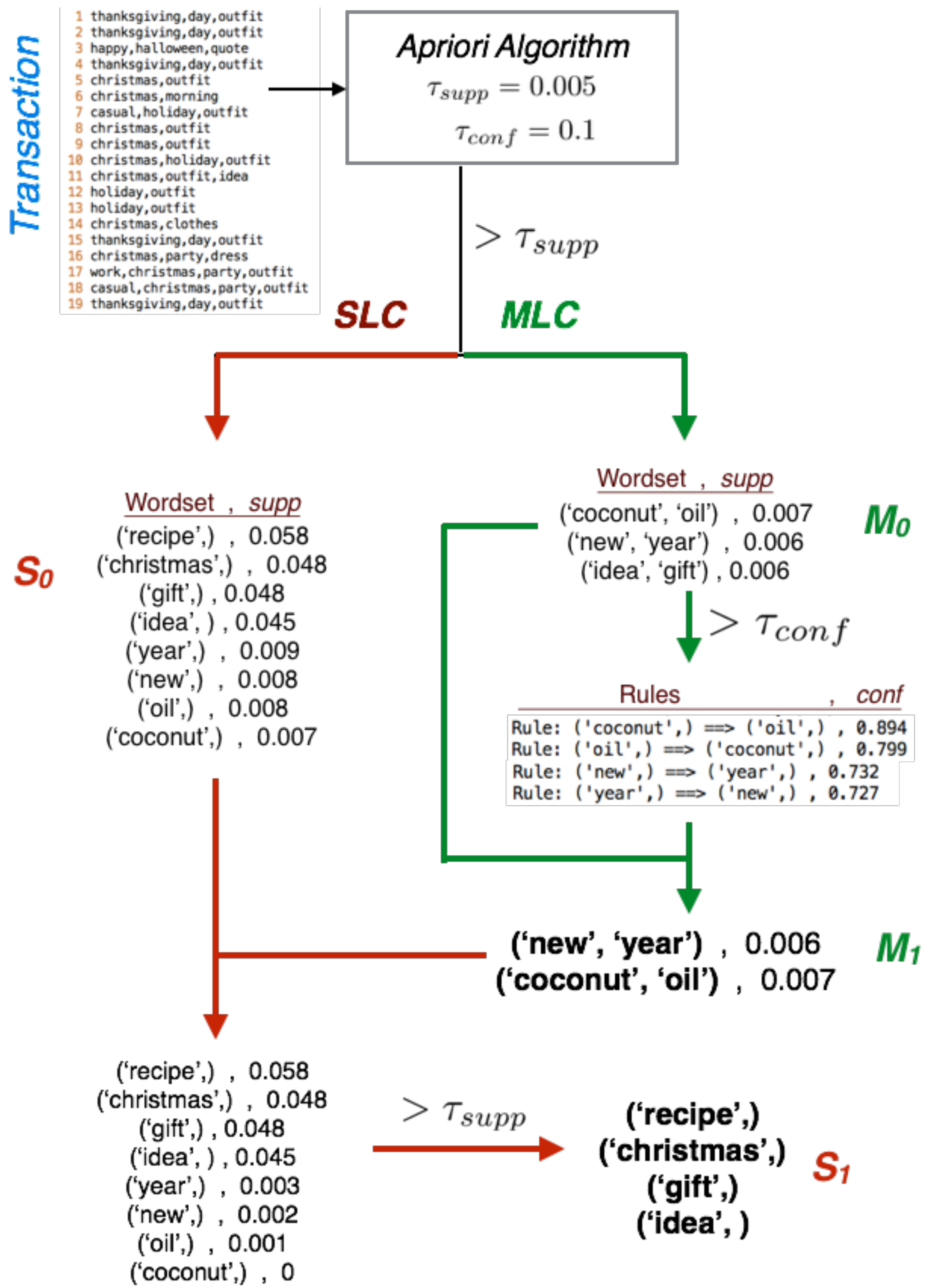
Frequent Itemset Mining



Idea is based on Association Rule Mining.

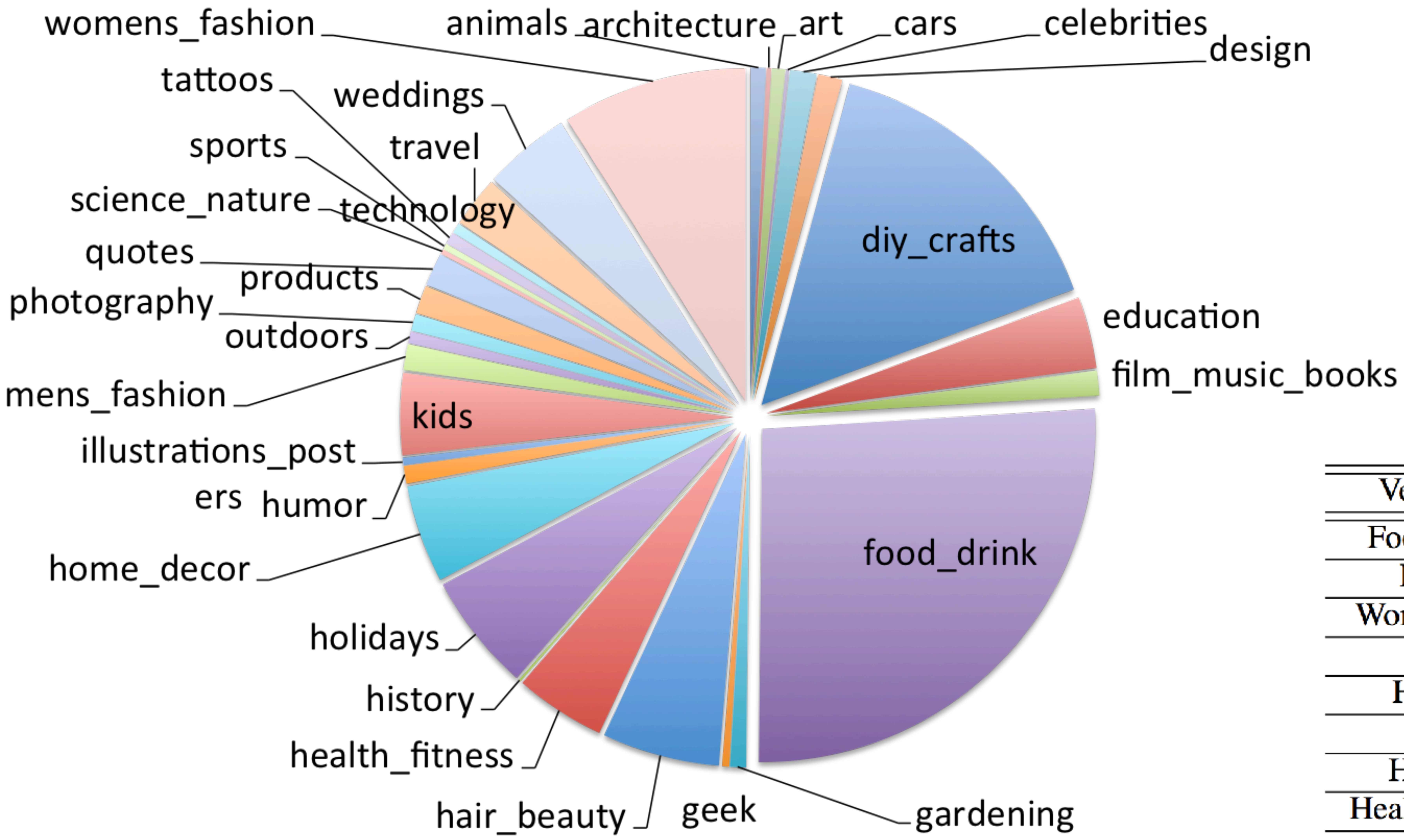
1. Build a word transaction
2. Word preprocessing with natural language processing (NLP) steps
3. frequent word set mining
4. Generate both single word and multi-word labels (next page)

Label Curation Algorithm



Curated Image Set Statistics

Two Level Hierarchy: Verticals, then Label Classes



Top 10 Verticals

Vertical name	#Images	#Classes
Food and drink	24762	287
DIY crafts	14223	126
Women's fashion	8438	87
Holidays	5441	52
Hair beauty	5361	53
Kids	5201	77
Home decor	4450	28
Health and fitness	4141	30
Weddings	4004	43
Education	3292	44

Example of thematic labels

Thematic labels of images

Example of images under the same label, but from different verticals

Benchmark: Deep Net Training

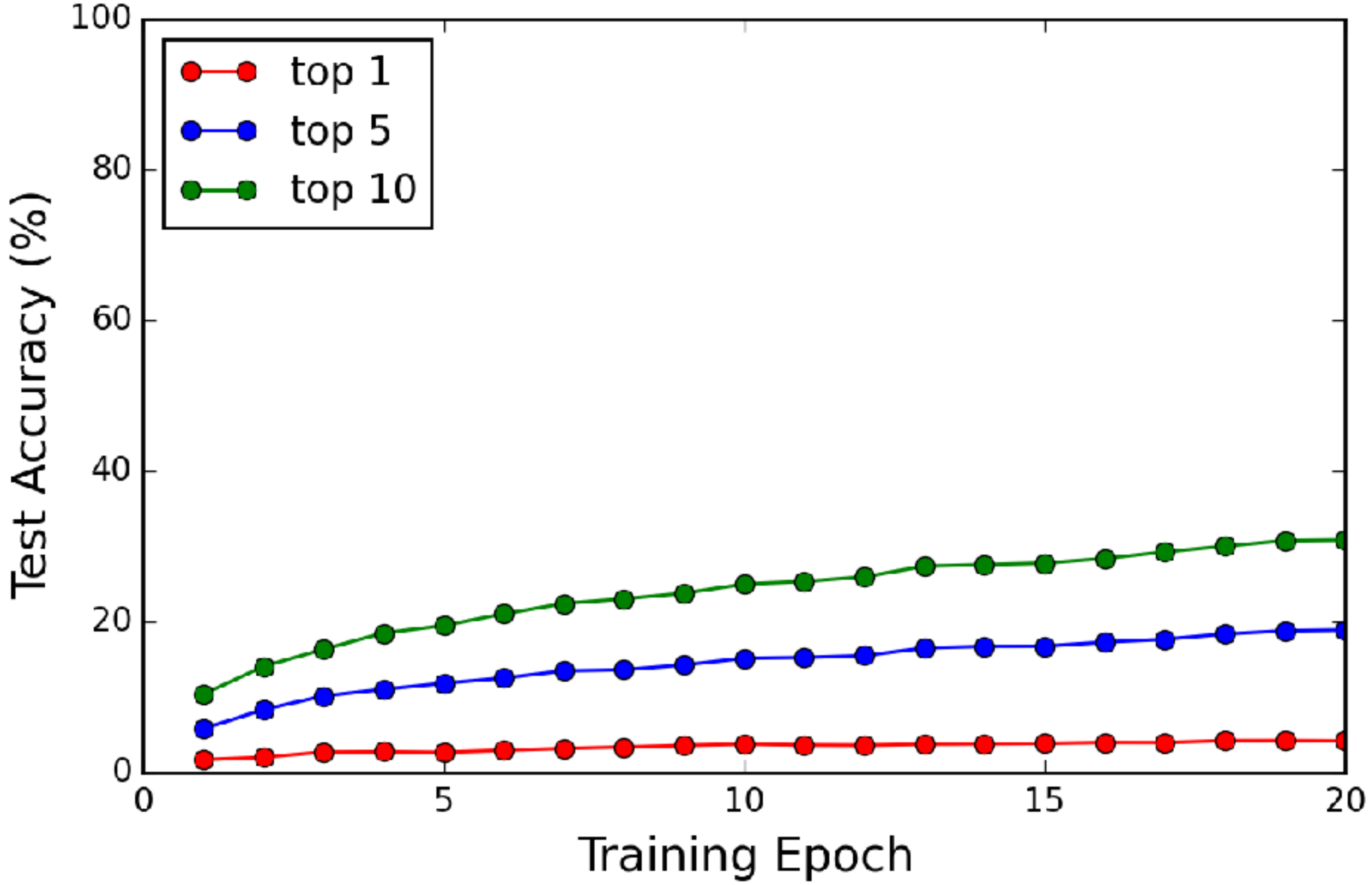
Experiments with AlexNet-OWT:

- (a) Take all images and treat all 536 classes as on one flat level. No pre-training.

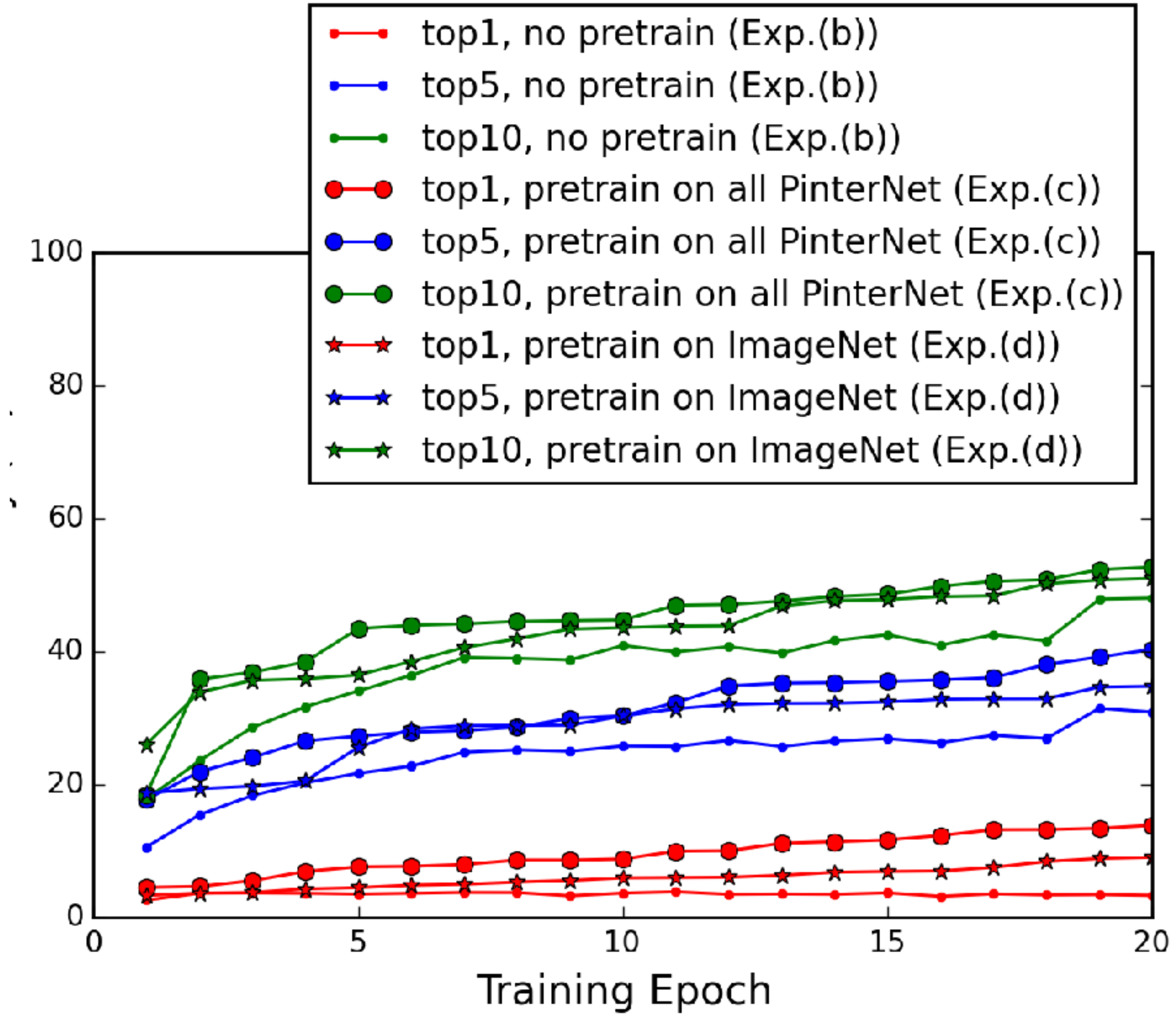
- (b) Take all images, first train a binary support vector machine (SVM) classifier that classifies an image as whether food or not food. Train CNN with images in the food vertical with 287 classes. No pre-training.

- (c) Same as (b), but with pre-training from all PinterNet image classes.

- (d) Same as (b), but with pre-training from ImageNet classes.



Experiment (a): AlexNet, no pretraining, flat 536 classes



Experiment (b), (c) and (d): AlexNet, hierarchical classification (first decide on vertical, then label class), with/without pretraining

Summary

- Deep learning on computer vision calls for more labelled images.
- We provide a tool, *PinterNet*, that labels image faster without human supervision.
- Labeled by this tool, the 110k image data set is analyzed, released, and benchmarked.