Webly Supervised Learning of Convolutional Network

Author: Xinlei Chen, Abhinav Gupta
Presented by : Vicky Xuening Wang

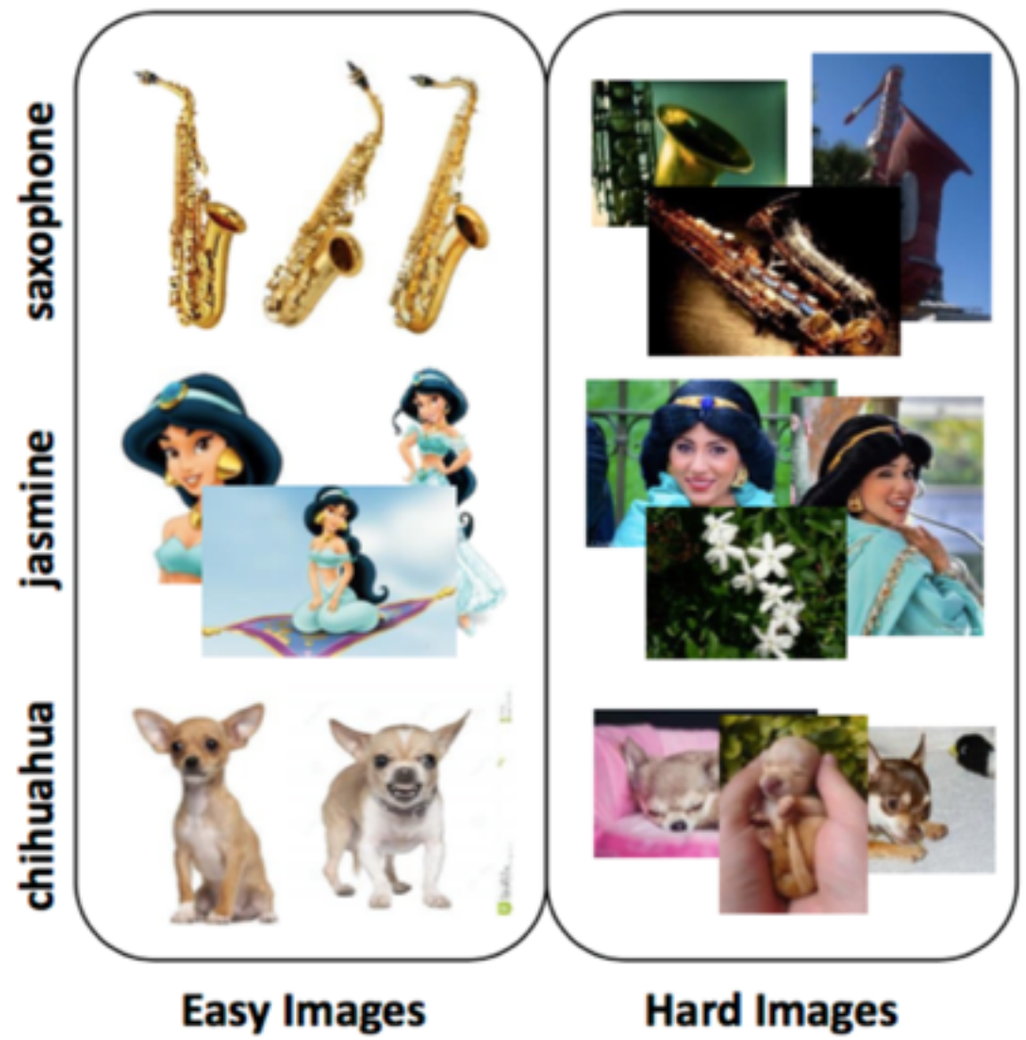
ECS 289G, Nov 2015, UC Davis

Motivation

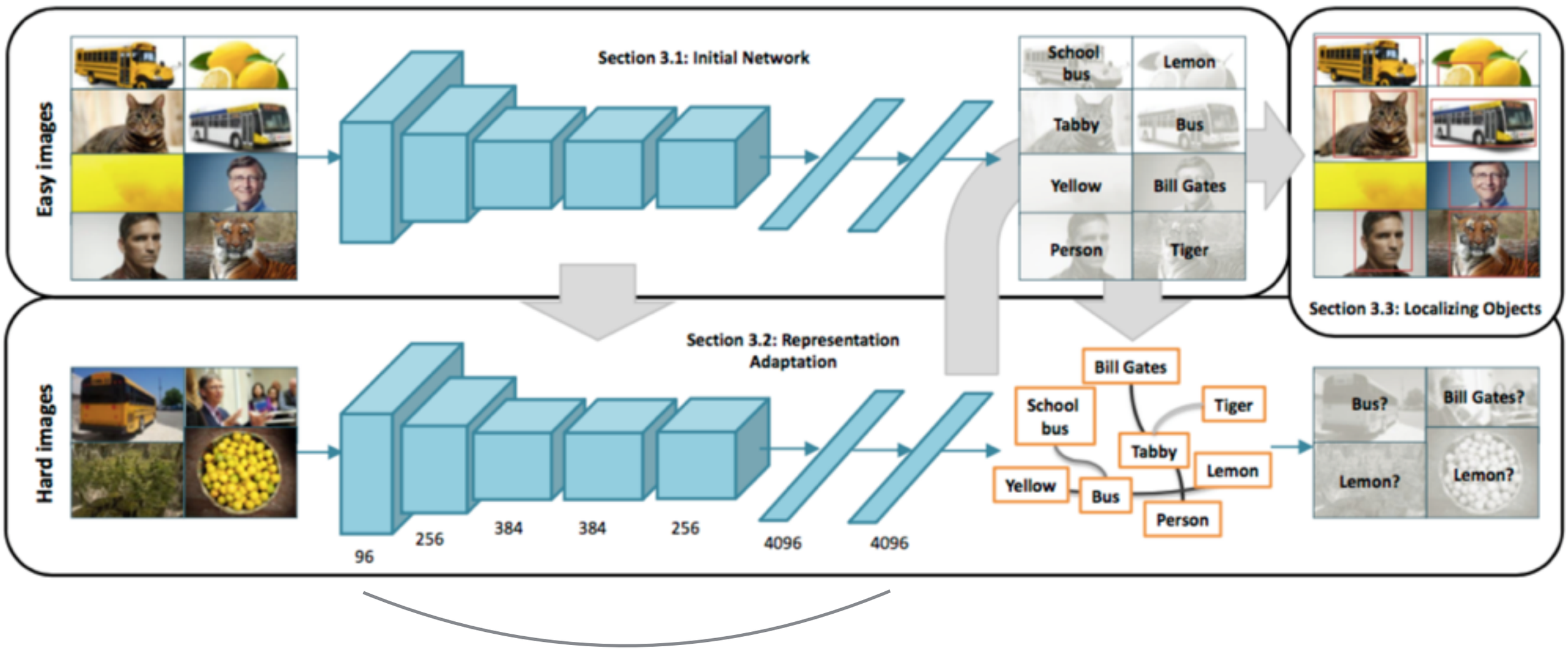
- Deeper Network + More Training Data => Better CNN
- Inspired by curriculum learning
- Two-step approach for learning CNNs:
    1. Train an initial visual representation with **easy images**
    2. Fine-tuning the CNN with **harder images**

Web Data

- Easy Images:
    - e.g. results of Google search engine
    - Biased
- Harder Images
    - e.g. Flickr style images
    - Noisy



Main Idea



BLVC in Caffe

Main idea

- First, bootstrap CNN training with easy examples
- Then, conduct a more extensive and comprehensive learning procedure with **similarity constraints** to learn visual representation
- Constraints: **Relationship graph** learned from easy examples

Network Set-up

- Initialization:
  - Category - a combination of categorical lists from **ImageNet** Challenge, **SUN** database and **NEIL** knowledge base
  - Query from Google Image search engine
    - ~600 per category -> exclude unreadable
  - Query from Flickr with tag search
    - ~1500 categories close to uniform distribution

Similarity constraint

- A relationship graph represented by a matrix
- Learned from the first stage, and applied to the second stage
- Remove noisy example mistake

- confusion matrix

$$R_{ij} = P(i|j) = \frac{\sum_{k \in C_i} CNN(j|I_k)}{|C_i|}, \qquad (1)$$

- Rij is used to characterize the label-flip noise

- Softmax loss function:

$$L = \sum_k \sum_i R_{il_k} \log(CNN(i|I_k)), \qquad (2)$$

- context of a category is obtained

Noise

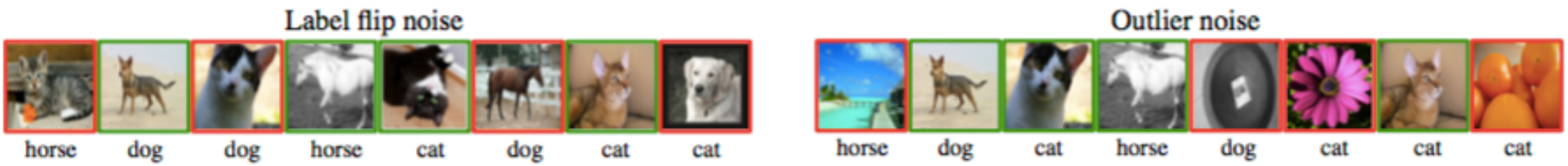
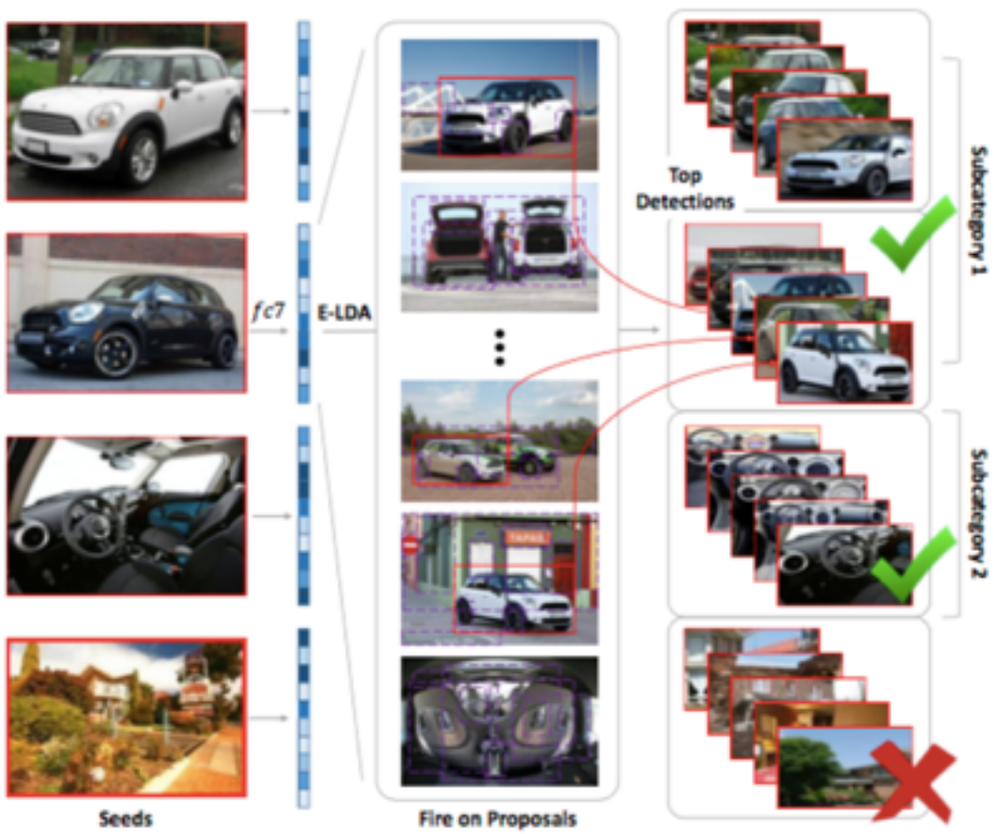


Figure 1: A toy classification example with 3 classes, illustrating the two types of label noise encountered on real datasets. In the label flip case, the images all belong to the 3 classes, but sometimes the labels are confused between them. In the outlier case, some images are unrelated to the classification task but possess one of the 3 labels.

* Image source: S. Sukhbaatar and R. Fergus. Learning from noisy labels with deep neural networks. arXiv:1406.2080, 2014.

Localizing Objects



- **Seed**: full images returned from Google as seed bounding box
- **Nearest Neighbor propagation:**
  - train Exemplar-LDA detector using fc7 features
  - fire on remaining images to find top k nearest neighbors
  - use EdgeBox instead of sliding window for efficiency
- **Clustering into subdirectories**
  - based on E-LDA similarity scores and density estimation
  - iteratively merge neighbors from bottom up

Localizing Objects con'd

Moreover

- train R-CNN detector for each categories
- negative examples: random patches from YFCC
- positive examples: hundreds of instances per category from original training dataset
- increase positive bounding boxes
    - **EdgeBox Augmentation(EA):** add IoU>0.5
    - **Category Expansion(CE):** verify with semantics

Experiment

2,240 objects, 89 attributes, 874 scenes

Two networks are trained:

- The object-attribute network (GoogleO), output dimension: 2,329
    - ~1.5 million images downloaded from Google Image Search
    - ~1.2 million Flickr images for fine-tuning
- All included network (GoogleA), output dimension: 3,203
    - plus scene images = ~2.1 million images

Experiment- 2 stages

Learning:

- batch size: 256
- learning rate: 0.01 /10^i, per 150K iterations
- termination: 450K iterations

Fine-tuning:

- step size: 30K
- termination: 100K iterations

Experiment

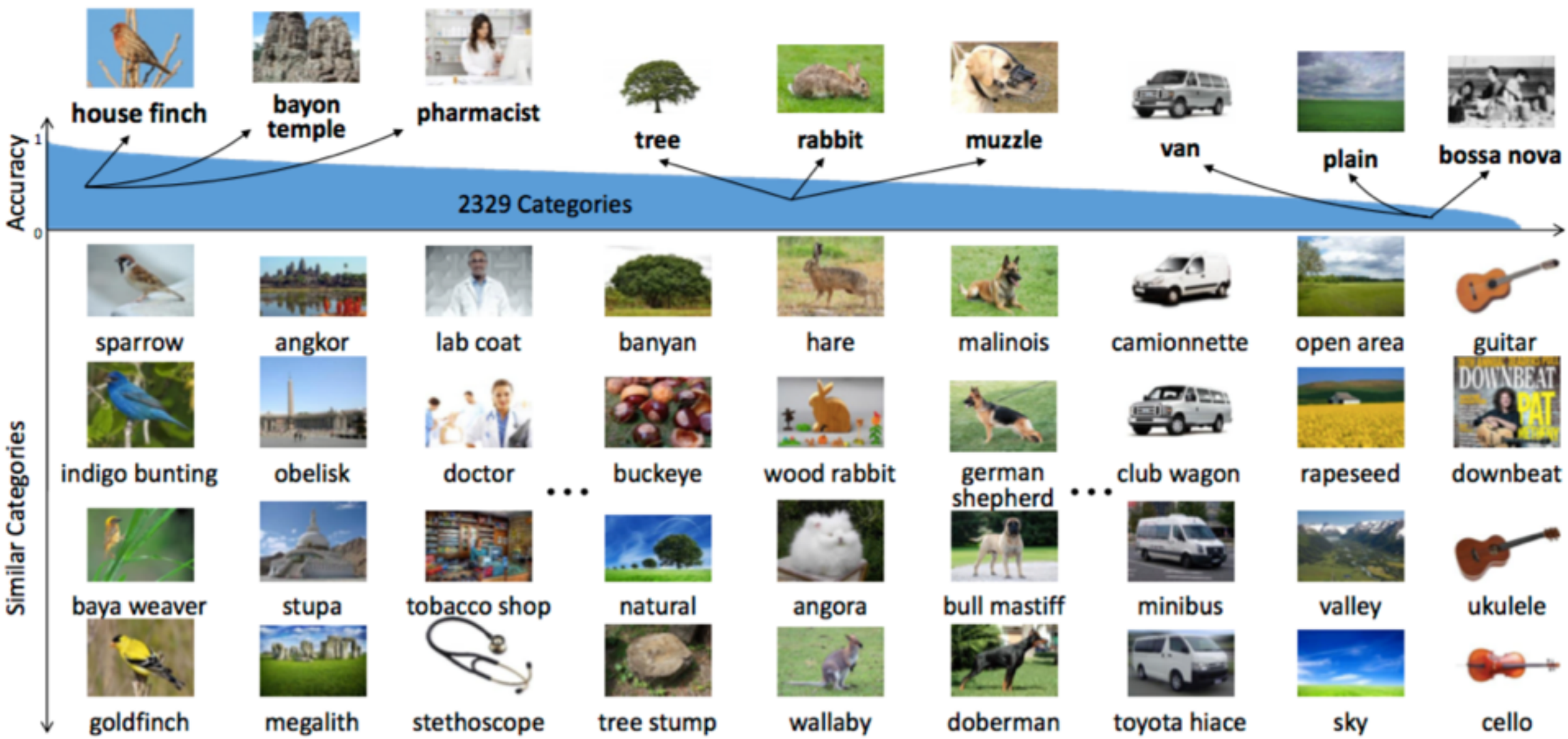
#1. PASCAL VOC Object Detection

#2. Clean up web images

#3. Failure Modes for Webly Trained Detectors

#4. Scene Classification

Result - confusion matrix



#1. PASCAL VOC Object Detection

- VOC 2007
    - fine-tuning with trainval images
    - step size: 20K
    - total: 100K iterations
- VOC 2012
    - step size: 50K
    - total: 200K iterations
- Did not tune R-CNN parameters, same SVM training settings with ImageNet
- GoogleO-FT, GoogleA-FT, Flickr-FT

Result #1

VOC 2007 and 2012 dataset

	VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	pers	plant	sheep	sofa	train	tv	mAP
	ImageNet [20]	**57.6**	57.9	38.5	**31.8**	23.7	51.2	58.9	**51.4**	20.0	**50.5**	**40.9**	**46.0**	**51.6**	55.9	**43.3**	23.3	**48.1**	35.3	**51.0**	**57.4**	**44.7**
w/o VOC FT	GoogleO [Obj.]	57.1	59.9	35.4	30.5	21.9	53.9	59.5	40.7	18.6	43.3	37.5	41.9	49.6	57.7	38.4	22.8	45.2	37.1	48.0	54.5	42.7
w/o VOC FT	GoogleA [Obj. + Sce.]	54.9	58.2	35.7	30.7	22.0	54.5	59.9	44.7	19.9	41.0	34.5	40.1	46.8	56.2	40.0	22.2	45.8	36.3	47.5	54.2	42.3
w/o VOC FT	FlickrS [Flickr Obj.]	50.0	55.9	29.6	26.8	18.7	47.6	56.3	34.4	14.5	35.9	33.3	34.2	43.2	52.2	36.7	21.5	43.3	31.6	48.5	48.4	38.1
w/o VOC FT	GFAll [All Obj., 1-stage]	52.1	57.8	38.1	25.6	21.2	47.6	56.4	43.8	19.6	42.6	30.3	37.6	45.1	50.8	39.3	22.9	43.5	34.2	48.3	52.2	40.5
w/o VOC FT	FlickrF [2-stage]	53.9	60.7	37.0	31.6	23.8	**57.7**	60.8	44.1	20.3	46.5	31.5	39.8	49.7	59.0	41.6	23.0	44.4	36.2	49.9	56.2	43.4
w/o VOC FT	FlickrG [2-stage, Graph]	55.3	**61.9**	**39.1**	29.5	**24.8**	55.1	**62.7**	43.5	**22.7**	49.3	36.6	42.7	48.9	**59.7**	41.2	**25.4**	47.7	**41.9**	48.8	56.8	**44.7**
w/ VOC FT	VOC-Scratch [2]	49.9	60.6	24.7	23.7	20.3	52.5	64.8	32.9	20.4	43.5	34.2	29.9	49.0	60.4	47.5	28.0	42.3	28.6	51.2	50.0	40.7
w/ VOC FT	ImageNet [20]	64.2	**69.7**	**50.0**	**41.9**	**32.0**	62.6	71.0	**60.7**	**32.7**	58.5	46.5	**56.1**	60.6	66.8	54.2	31.5	52.8	48.9	57.9	**64.7**	**54.2**
w/ VOC FT	GoogleO	**65.0**	68.1	45.2	37.0	29.6	65.4	73.8	54.0	30.4	57.8	48.7	51.9	**64.1**	64.7	54.0	**32.0**	**54.9**	44.5	57.0	64.0	53.1
w/ VOC FT	GoogleA	64.2	68.3	42.7	38.7	26.5	65.1	72.4	50.7	28.5	**60.9**	**48.8**	51.2	60.2	65.5	**54.5**	31.1	50.5	**48.5**	56.3	60.3	52.3
w/ VOC FT	FlickrG	63.7	68.5	46.2	36.4	30.2	**68.4**	**73.9**	56.9	31.4	59.1	46.7	52.4	61.5	**69.2**	53.6	31.6	53.8	44.5	**58.1**	59.6	53.3

Table 1. Results on VOC 2007 (PASCAL data used). Please see Section 4.1 for more details.

	VOC 2012 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
w/ VOC FT	ImageNet [20]	68.1	63.8	46.1	29.4	27.9	56.6	57.0	65.9	26.5	48.7	39.5	**66.2**	57.3	65.4	53.2	26.2	54.5	38.1	50.6	51.6	49.6
w/ VOC FT	ImageNet-TV	**73.3**	67.1	46.3	31.7	30.6	59.4	61.0	**67.9**	27.3	**53.1**	39.1	64.1	**60.5**	70.9	57.2	26.1	**59.0**	40.1	56.2	**54.9**	52.3
w/ VOC FT	GoogleO	72.2	67.3	46.0	**32.3**	**31.6**	**62.6**	62.5	66.5	27.3	52.1	38.9	64.0	59.1	71.6	58.0	27.2	57.6	41.3	56.3	53.7	52.4
w/ VOC FT	FlickrG	72.7	**68.2**	**47.3**	32.2	30.6	62.3	**62.6**	65.9	**28.1**	52.2	**39.5**	65.1	60.0	**71.7**	**58.2**	**27.3**	58.0	**41.5**	**57.2**	53.8	**52.7**

Table 2. Results on VOC 2012. Since [20] only fine-tuned on the train set, we also report results on trainval (ImageNet-TV) for fairness.

Result #1

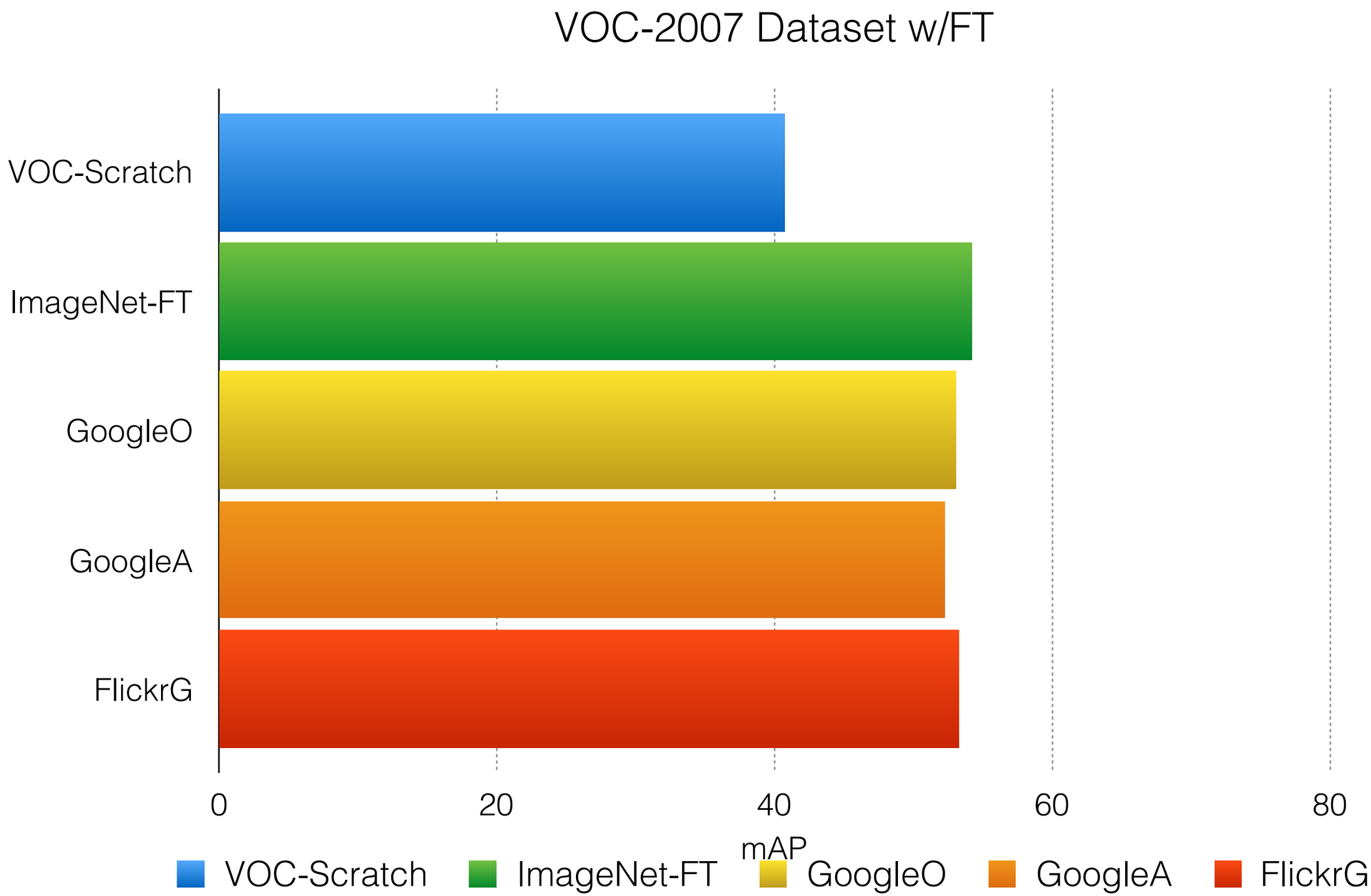
VOC 2007 and 2012 dataset

	VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	pers	plant	sheep	sofa	train	tv	mAP
w/o VOC FT	ImageNet [20]	**57.6**	57.9	38.5	**31.8**	23.7	51.2	58.9	**51.4**	20.0	**50.5**	**40.9**	**46.0**	**51.6**	55.9	**43.3**	23.3	**48.1**	35.3	**51.0**	**57.4**	**44.7**
	GoogleO [Obj.]	57.1	59.9	35.4	30.5	21.9	53.9	59.5	40.7	18.6	43.3	37.5	41.9	49.6	57.7	38.4	22.8	45.2	37.1	48.0	54.5	42.7
	GoogleA [Obj. + Sce.]	54.9	58.2	35.7	30.7	22.0	54.5	59.9	44.7	19.9	41.0	34.5	40.1	46.8	56.2	40.0	22.2	45.8	36.3	47.5	54.2	42.3
	FlickrS [Flickr Obj.]	50.0	55.9	29.6	26.8	18.7	47.6	56.3	34.4	14.5	35.9	33.3	34.2	43.2	52.2	36.7	21.5	43.3	31.6	48.5	48.4	38.1
	GFAll [All Obj., 1-stage]	52.1	57.8	38.1	25.6	21.2	47.6	56.4	43.8	19.6	42.6	30.3	37.6	45.1	50.8	39.3	22.9	43.5	34.2	48.3	52.2	40.5
	FlickrF [2-stage]	53.9	60.7	37.0	31.6	23.8	**57.7**	60.8	44.1	20.3	46.5	31.5	39.8	49.7	59.0	41.6	23.0	44.4	36.2	49.9	56.2	43.4
	FlickrG [2-stage, Graph]	55.3	**61.9**	**39.1**	29.5	**24.8**	55.1	**62.7**	43.5	**22.7**	49.3	36.6	42.7	48.9	**59.7**	41.2	**25.4**	47.7	**41.9**	48.8	56.8	**44.7**
w/ VOC FT	VOC-Scratch [2]	49.9	60.6	24.7	23.7	20.3	52.5	64.8	32.9	20.4	43.5	34.2	29.9	49.0	60.4	47.5	28.0	42.3	28.6	51.2	50.0	40.7
	ImageNet [20]	64.2	**69.7**	**50.0**	**41.9**	**32.0**	62.6	71.0	**60.7**	**32.7**	58.5	46.5	**56.1**	60.6	66.8	54.2	31.5	52.8	48.9	57.9	**64.7**	**54.2**
	GoogleO	**65.0**	68.1	45.2	37.0	29.6	65.4	73.8	54.0	30.4	57.8	48.7	51.9	**64.1**	64.7	54.0	**32.0**	**54.9**	44.5	57.0	64.0	53.1
	GoogleA	64.2	68.3	42.7	38.7	26.5	65.1	72.4	50.7	28.5	**60.9**	**48.8**	51.2	60.2	65.5	**54.5**	31.1	50.5	**48.5**	56.3	60.3	52.3
	FlickrG	63.7	68.5	46.2	36.4	30.2	**68.4**	**73.9**	56.9	31.4	59.1	46.7	52.4	61.5	**69.2**	53.6	31.6	53.8	44.5	**58.1**	59.6	53.3

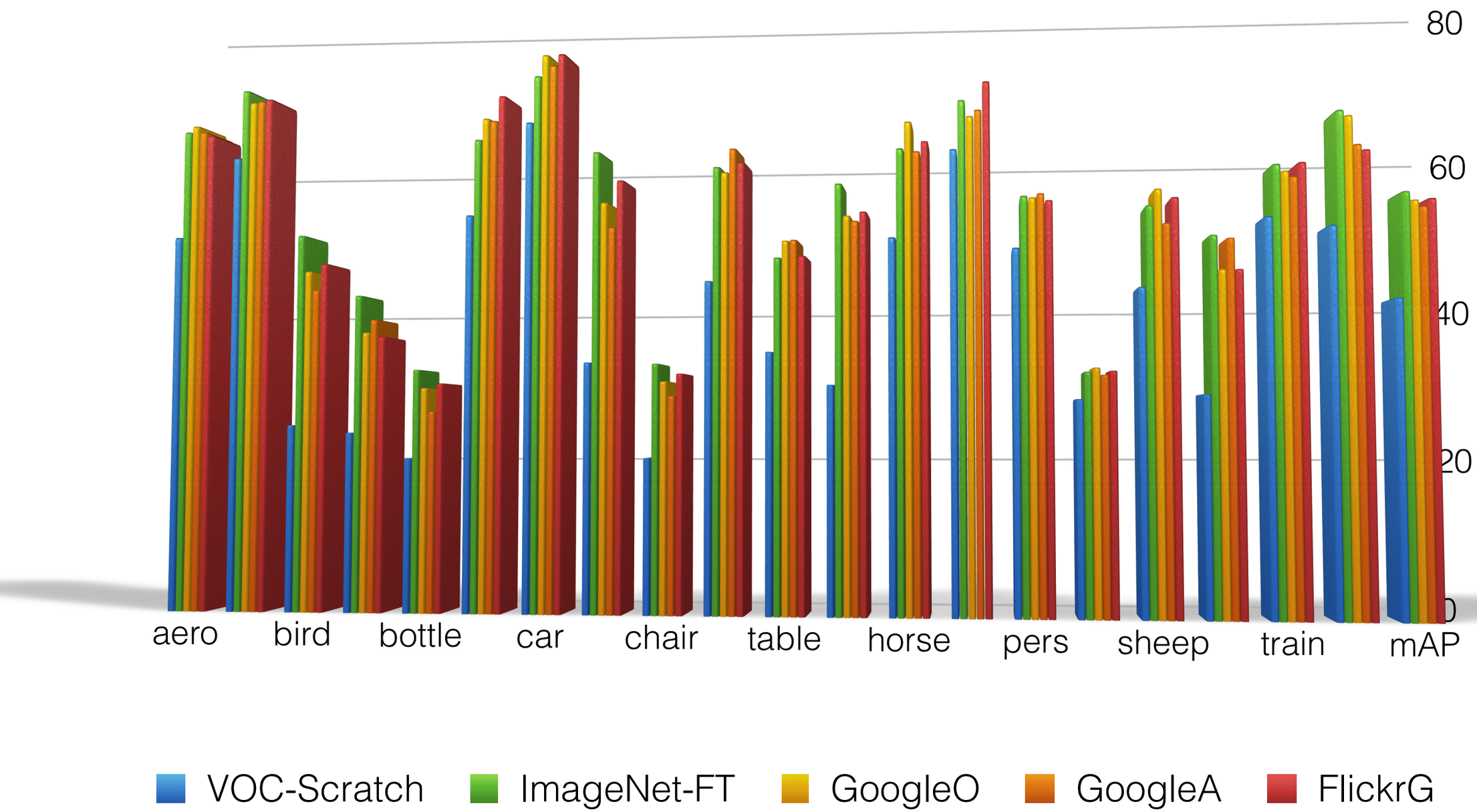
Table 1. Results on VOC 2007 (PASCAL data used). Please see Section 4.1 for more details.

	VOC 2012 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
w/ VOC FT	ImageNet [20]	68.1	63.8	46.1	29.4	27.9	56.6	57.0	65.9	26.5	48.7	39.5	**66.2**	57.3	65.4	53.2	26.2	54.5	38.1	50.6	51.6	49.6
	ImageNet-TV	**73.3**	67.1	46.3	31.7	30.6	59.4	61.0	**67.9**	27.3	**53.1**	39.1	64.1	**60.5**	70.9	57.2	26.1	**59.0**	40.1	56.2	**54.9**	52.3
	GoogleO	72.2	67.3	46.0	**32.3**	**31.6**	**62.6**	62.5	66.5	27.3	52.1	38.9	64.0	59.1	71.6	58.0	27.2	57.6	41.3	56.3	53.7	52.4
	FlickrG	72.7	**68.2**	**47.3**	32.2	30.6	62.3	**62.6**	65.9	**28.1**	52.2	**39.5**	65.1	60.0	**71.7**	**58.2**	**27.3**	58.0	**41.5**	**57.2**	53.8	**52.7**

Table 2. Results on VOC 2012. Since [20] only fine-tuned on the train set, we also report results on trainval (ImageNet-TV) for fairness.


VOC-2007 Dataset w/FT
VOC-Scratch
ImageNet-FT
GoogleO
GoogleA
FlickrG
0
20
40
60
80
mAP
VOC-Scratch
ImageNet-FT
GoogleO
GoogleA
FlickrG

VOC-2007 Dataset w/FT



#2. Discover subdirectories and localize the object in the web images

- Train R-CNN based detector with VOC 2007
- compared with Google-Ngrams

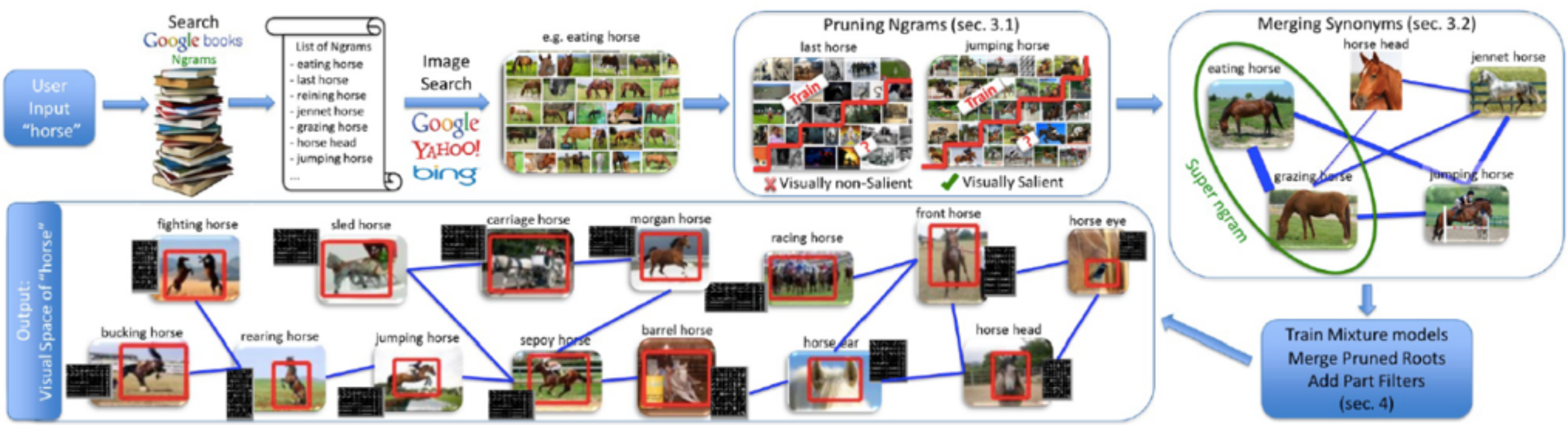


Image source: S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In CVPR, 2014.

#2. Discover subdirectories and localize the object in the web images

	feature	bounding box
GoogleO	GoogleO CNN	easy images
GoogleA	GoogleA CNN	easy images
Flickr	2-stage CNN	easy images
Flickr-M	2-stage CNN	easy+hard images
Flickr-C	2-stage CNN	original and related categories

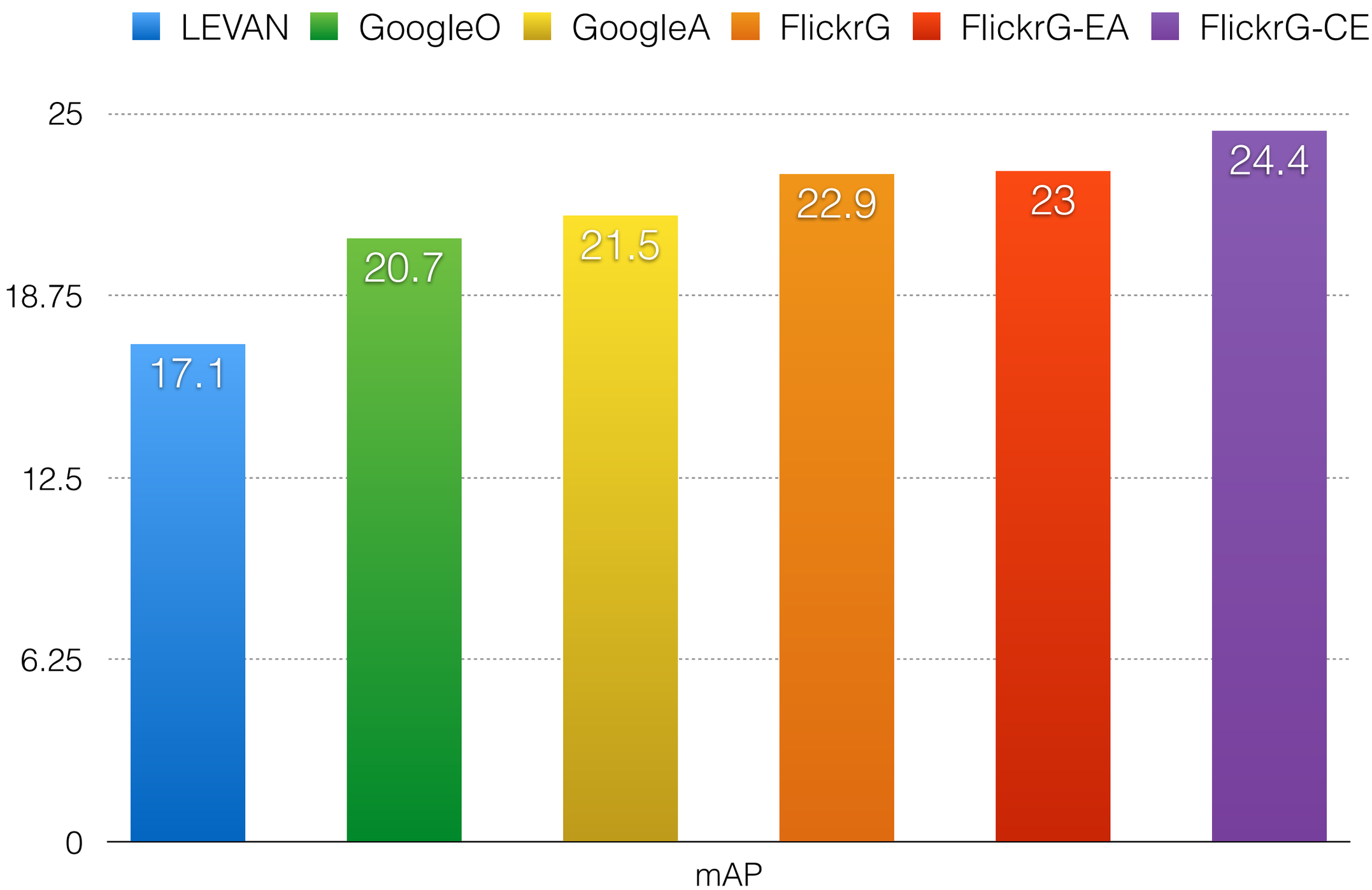
Result #2

VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
LEVAN [14]	14.0	36.2	12.5	10.3	9.2	35.0	**35.9**	8.4	**10.0**	17.5	6.5	12.9	**30.6**	27.5	6.0	1.5	18.8	10.3	23.5	16.4	17.1
GoogleO	30.2	34.3	16.7	13.3	6.1	43.6	27.4	22.6	6.9	16.4	10.0	21.3	25.0	35.9	7.6	9.3	21.8	17.3	31.0	18.1	20.7
GoogleA	29.5	38.3	15.1	14.0	9.1	44.3	29.3	24.9	6.9	15.8	9.7	22.6	23.5	34.3	9.7	12.7	21.4	15.8	33.4	19.4	21.5
FlickrG	32.6	42.8	19.3	13.9	9.2	46.6	29.6	20.6	6.8	17.8	10.2	22.4	26.7	40.8	11.7	**14.0**	19.0	19.0	34.0	21.9	22.9
FlickrG-EA	**32.7**	**44.3**	17.9	14.0	**9.3**	**47.1**	26.6	19.2	8.2	18.3	10.0	22.7	25.0	42.5	12.0	12.7	**22.2**	20.9	35.6	18.2	23.0
FlickrG-CE	30.2	41.3	**21.7**	**18.3**	9.2	44.3	32.2	**25.5**	9.8	**21.5**	**10.4**	**26.7**	27.3	**42.8**	**12.6**	13.3	20.4	**20.9**	**36.2**	**22.8**	**24.4**

Result #2

VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
LEVAN [14]	14.0	36.2	12.5	10.3	9.2	35.0	**35.9**	8.4	**10.0**	17.5	6.5	12.9	**30.6**	27.5	6.0	1.5	18.8	10.3	23.5	16.4	17.1
GoogleO	30.2	34.3	16.7	13.3	6.1	43.6	27.4	22.6	6.9	16.4	10.0	21.3	25.0	35.9	7.6	9.3	21.8	17.3	31.0	18.1	20.7
GoogleA	29.5	38.3	15.1	14.0	9.1	44.3	29.3	24.9	6.9	15.8	9.7	22.6	23.5	34.3	9.7	12.7	21.4	15.8	33.4	19.4	21.5
FlickrG	32.6	42.8	19.3	13.9	9.2	46.6	29.6	20.6	6.8	17.8	10.2	22.4	26.7	40.8	11.7	**14.0**	19.0	19.0	34.0	21.9	22.9
FlickrG-EA	**32.7**	**44.3**	17.9	14.0	**9.3**	**47.1**	26.6	19.2	8.2	18.3	10.0	22.7	25.0	42.5	12.0	12.7	**22.2**	20.9	35.6	18.2	23.0
FlickrG-CE	30.2	41.3	**21.7**	**18.3**	9.2	44.3	32.2	**25.5**	9.8	**21.5**	**10.4**	**26.7**	27.3	**42.8**	**12.6**	13.3	20.4	**20.9**	**36.2**	**22.8**	**24.4**

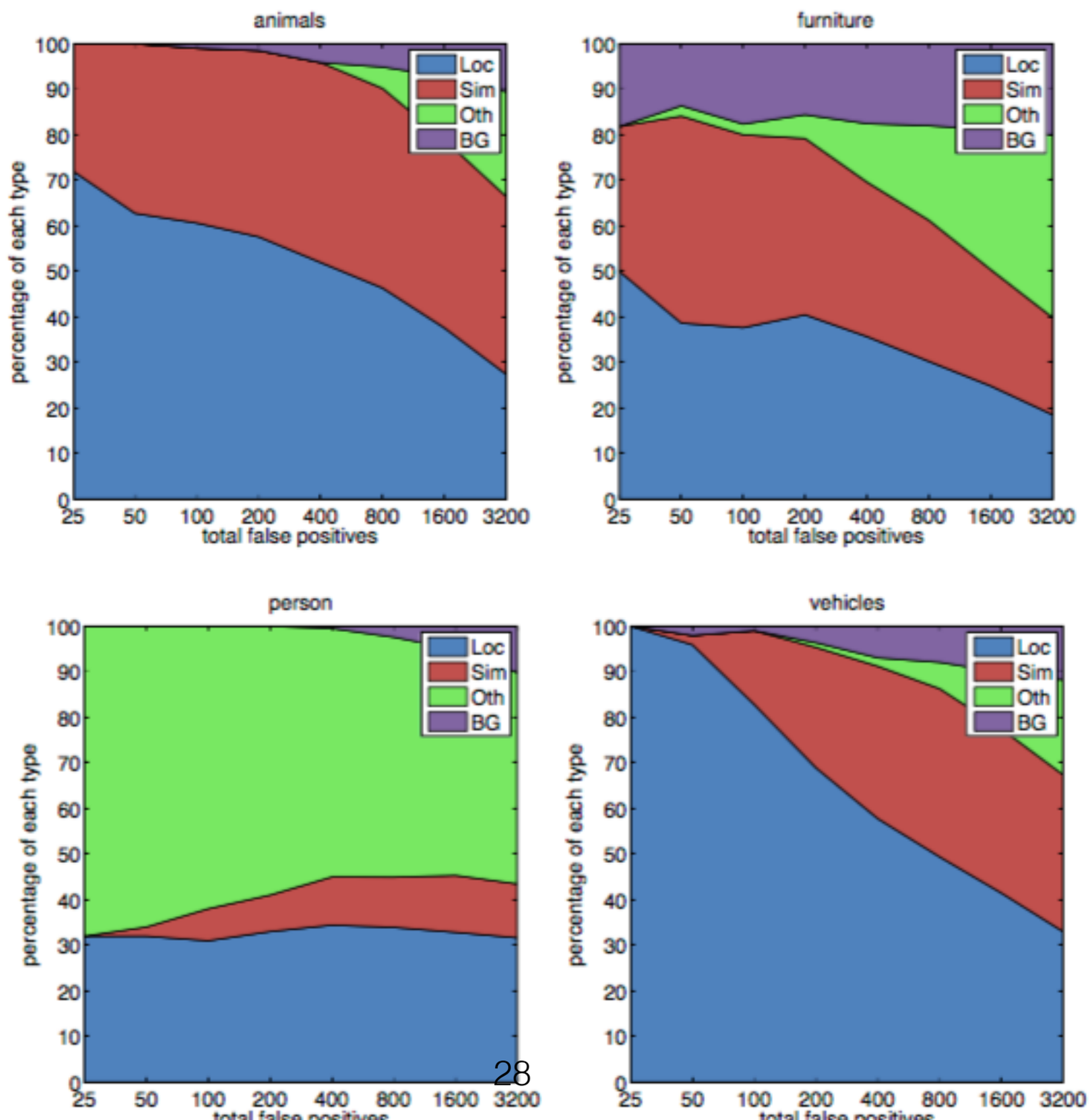
#2 Object localization results



#3. Failure Modes for Webly Trained Detectors

- Localization error - false positive
    - background
    - multiple instances
    - spatial invariant
- Semantic drift between Google categories and PASCAL categories
    - sense disambiguation

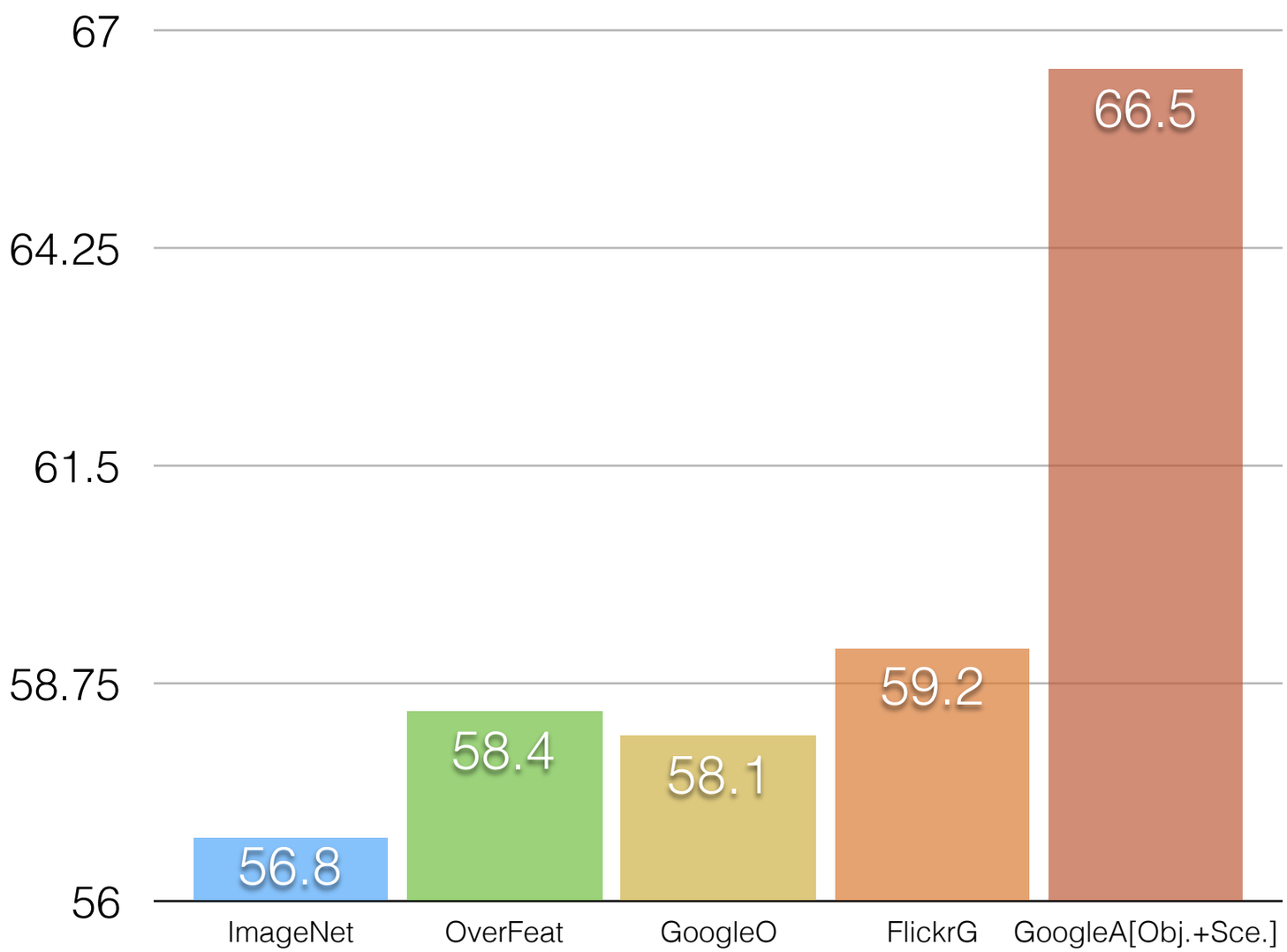
Result #3


animals
furniture
person
vehicles
Loc
Sim
Oth
BG
percentage of each type
total false positives
0
10
20
30
40
50
60
70
80
90
100
25
50
100
200
400
800
1600
3200

#4. Scene Classification

- MIT Indoor-67 dataset
- compute the fc7 feature vector: 4096 dimensions
- normalize to unit l2 length
- fixed SVM parameter (C=1)

Result #4



Scene classification results on MIT Indoor-67

Conclusion

Conclusion

- The two-stage CNN comes close to ImageNet pertained architecture on VOC 2007, out performs in VOC 2012
- only cost is providing a category to query from

Discussion

- What are good sources of data?
- What types of data can be exploited?
- How do we exploit the data?
- What should we lear from Web data?

Discussion

- **What are good sources of data?**
    - search engine + everyday images
- **What types of data can be exploited?**
    - images-only is enough
- **How do we exploit the data?**
    - learn from the queried domain from search engine
    - fine-tuning with everyday images
- **What should we learn from Web data?**
    - intra-category similarity relationship based on easy images (remove noise)
    - low-level features supplemented by Flickr style images (remove bias)