Large Scale Visual Semantic Extraction

Samy Bengio, Google

Frontiers of Engineering Workshop, 2011

Image Annotation: What is it?

Goal: Label a new image using a predefined set of possible annotations.

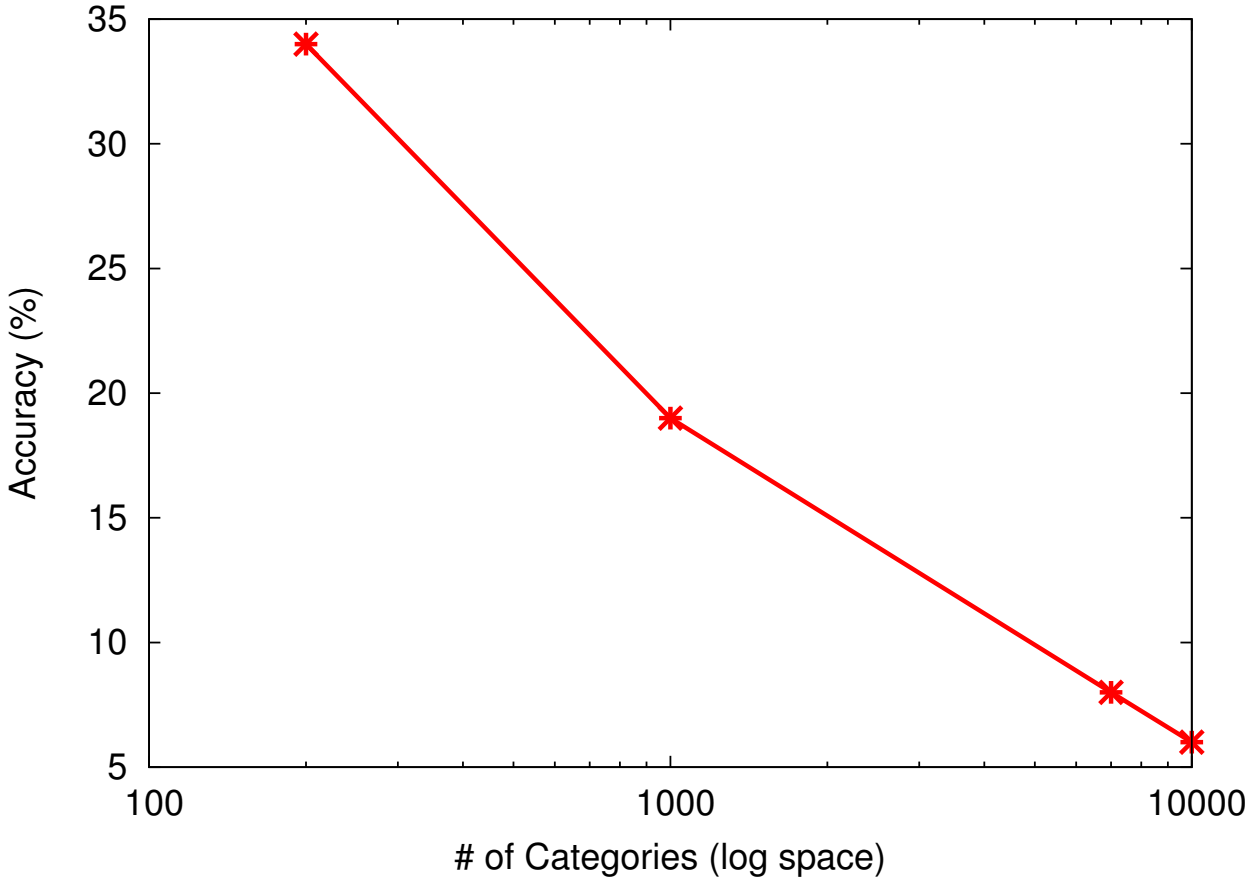
→ **obama**

→ **eiffel tower**

- Computer vision literature has mostly focused on getting better features to represent images.
- The number of possible annotations (dictionary) is usually small (from 20 to 1000 or even 10,000 very recently).
- In this work, we consider dictionaries of size 100,000 and more.

Size Matters!

(Figure from F. F. Li)



Despite several research advances, performance of best systems degrades significantly as the number of possible categories grows.

Datasets Used In This Work (to grasp the scale)

Statistics	ImageNet	Web
Number of Training Images	2,518,604	9,861,293
Number of Test Images	839,310	3,286,450
Number of Validation Images	837,612	3,287,280
Number of Labels	15,952	109,444

Classical Approach To Image Annotation

Feature Extraction

1. Interest point detection: which points in the image should we analyze.
2. Feature extraction: how do we represent each point. Examples: color histograms, edges (SIFT, HoG).
3. Aggregation of features: from a dictionary of commonly seen features, count how many of each common feature was in the image.

Model Training

1. Gather a large training set of labeled images.
2. Extract features for each training image.
3. Train a classifier for each label (so-called one-vs-rest).
4. Example of an often-used classifier: Support Vector Machine.
5. Does not scale well...

Our Proposed Solution In One Slide: Wsabie

Learn $\Phi_I(\cdot)$ and $\Phi_W(\cdot)$ to optimize precision@k.

To Label an Image is Equivalent to a Ranking Problem

- Label an image means selecting a few relevant labels from a large set of potential labels.
- That amounts to ranking (ordering) labels given the image.
- Learning-To-Rank is a known setting in machine learning.
- Classical approach to learning-to-rank:
  - for each image x,
  - for each proper label for that image y,
  - and for each wrong label for that image $\bar{y}$:
  - make sure the distance between x and y is smaller (by a margin) than the distance between x and $\bar{y}$.
- This can be done by sampling triplets $(x, y, \bar{y})$, compute loss, and change parameters accordingly (stochastic gradient descent) if necessary.

A Better Ranking Loss

Problem: All pairwise errors are considered the same.

Example:

Function 1: true annotations ranked 1st and 101st.

Function 2: true annotations ranked 50st and 52st.

Ranking Loss prefers these *equally* as both have 100 "violations".

We want to optimize the top of the ranked list!

Idea: weigh pairs according to the rank of the positive label

- Put more emphasis on the highly ranked positive labels
- Problem: how to estimate the rank efficiently?
  - Computing the scores of all labels is too slow (100,000 of them).
  - Instead sample negative labels until you find one that violates the loss.
  - This can be used to estimate the rank of the positive label.

Test Set Performance Results

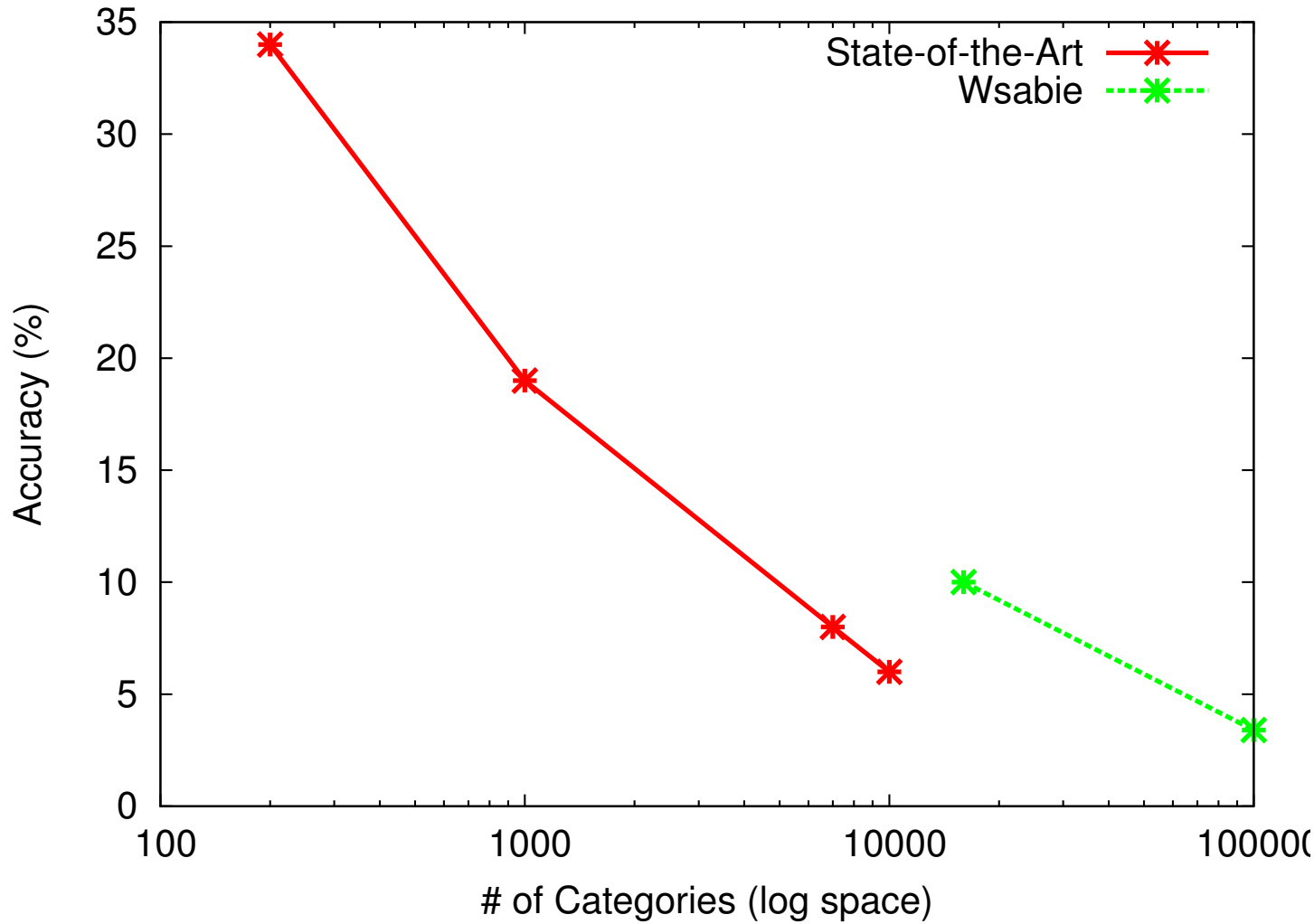
On ImageNet

Algorithm	precision@1	precision@10
Approx. k-NN	1.55%	0.41%
One-vs-Rest	2.27%	1.02%
Multiclass	3.14%	1.26%
Wsabie	4.03%	1.48%
3 Wsabie models	6.14%	2.09%
3 Wsabie w/ better features	10.03%	3.02%

On Web Images

Algorithm	precision@1	precision@10
Approx. k-NN	0.30%	0.34%
One-vs-Rest	0.52%	0.29%
Multiclass	0.32%	0.16%
Wsabie	1.03%	0.44%
4 Wsabie w/ better features	3.43%	1.27%

Size Matters - Revisited



Learned Annotation Embedding (on Web Data)

Annotation	Neighboring Annotations
barack obama	barak obama, obama, barack, barrack obama, bow wow
david beckham	beckham, david beckam, alessandro del piero, del piero
santa	santa claus, papa noel, pere noel, santa clause, joyeux noel
dolphin	delphin, dauphin, whale, delfin, delfini, baleine, blue whale
cows	cattle, shire, dairy cows, kuh, horse, cow, shire horse, kone
rose	rosen, hibiscus, rose flower, rosa, roze, pink rose, red rose
pine tree	abies alba, abies, araucaria, pine, neem tree, oak tree
mount fuji	mt fuji, fuji, fujisan, fujiyama, mountain, zugspitze
eiffel tower	eiffel, tour eiffel, la tour eiffel, big ben, paris, blue mosque
ipod	i pod, ipod nano, apple ipod, ipod apple, new ipod
f18	f 18, eurofighter, f14, fighter jet, tomcat, mig 21, f 16

Image Annotation Examples: Dolphin

delfini, orca, dolphin, mar, delfin, dauphin, whale, cancun, killer whale, sea world

blue whale, whale shark, great white shark, underwater, white shark, shark, manta ray, dolphin, requin, blue shark, diving

Image Annotation Examples: Obama & Eiffel Tower

barrack obama, barak obama, barack hussein obama, barack obama, james marsden, jay z, obama, nelly, falco, barack

eiffel, paris by night, la tour eiffel, tour eiffel, eiffel tower, las vegas strip, eifel, tokyo tower, eifel tower

Image Annotation Examples: Ipod

ipod, ipod nano, nokia, i pod, nintendo ds, nintendo, lg, pc, nokia 7610, vino

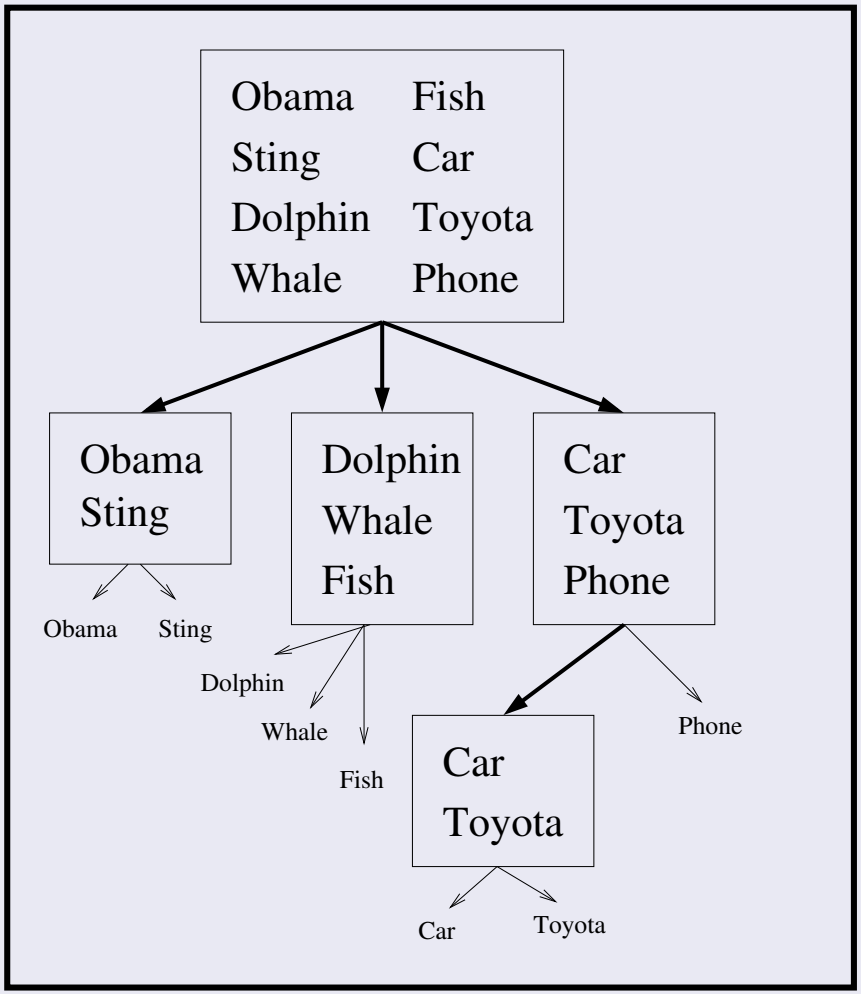
radioactive, ipod ad, post it, smiley, yellow, smiley face, smile, iowa hawkeyes, a style, caution, soda stereo, kill bill, idance

Learning Label Trees

Learning a Label Tree

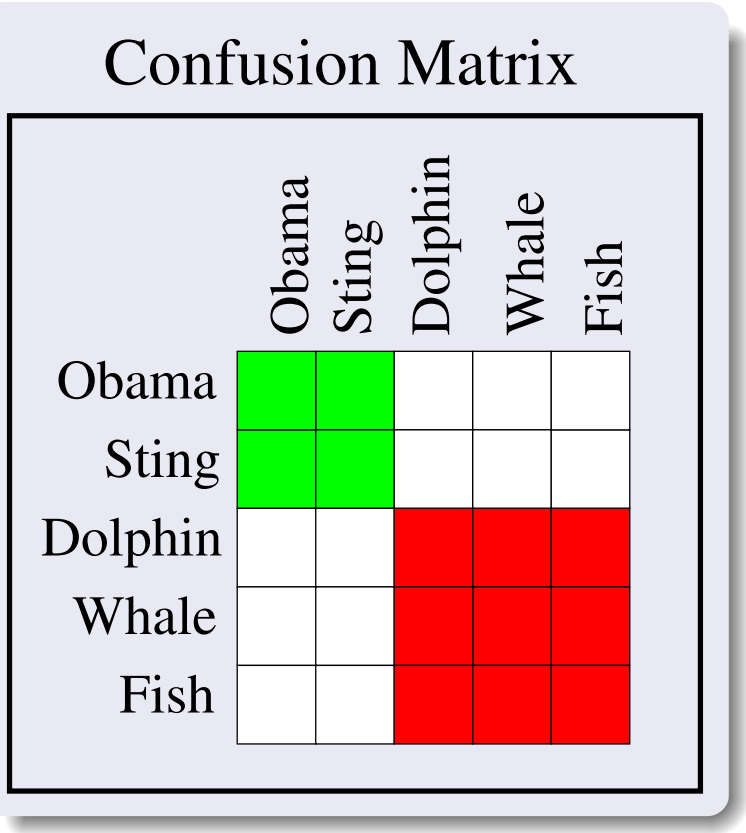
- How can we organize 100,000 labels?
- A tree would be useful:
  - Faster to take a decision: $\mathcal{O}(logN)$.
  - Can be used to stop labeling with too specific labels.
  - Provide a semantic organization of labels.
- Such a tree does not exist in general.
- Can we learn it from data?

Label Tree



Label Tree Building Block: Confusion Matrix

- Good tree structures have learnable label sets.
- Confused classes make learning hard.
- Idea: recursively partition the labels into label sets between which there is little confusion (measured with a surrogate learned model).
- Several algorithms exist to do so (such as spectral clustering).



Using a Learned Label Tree

Examples of Clustered Labels

- great white sharks, imagenes de delfines, liopleurodon meduse, mermaid tail, monstre du loch ness, monstruo del lago ness, oarfish, oceans, sea otter, shark attacks, sperm whale, tauchen, whales
- apple iphone 3gs, apple ipod, apple tablet, bumper, iphone 4, htc diamond, htc hd, htc magic, htc touch pro 2, iphone 2g, iphone 3, iphone 5g, iphone app, iphone apple, iphone apps, iphone nano
- chevy colorado, custom trucks, dodge ram, f 250, ford excursion, ford f 150, mini truck, nissan frontier, offroad, pickup, toyota tundra

Performance

- Precision performance remains the same
- Faster to label images ($\mathcal{O}(\log(n))$ instead of $\mathcal{O}(n)$).

Conclusion and Challenges

- Image annotation can be used in many applications such as automatically annotate your photos.
- When the number of labels is very high, need to structure them (tree).
- Learning jointly to label images and to semantically structure the labels is a challenge.
- Adding other media in the soup (videos, music, text documents).
- Ultimately learning a rich semantic space.
- Large datasets are rarely well labeled: need to be robust to label noise.
- How to use parallelism efficiently to learn such models?