Utilizing Text Captions to Improve Image Classification

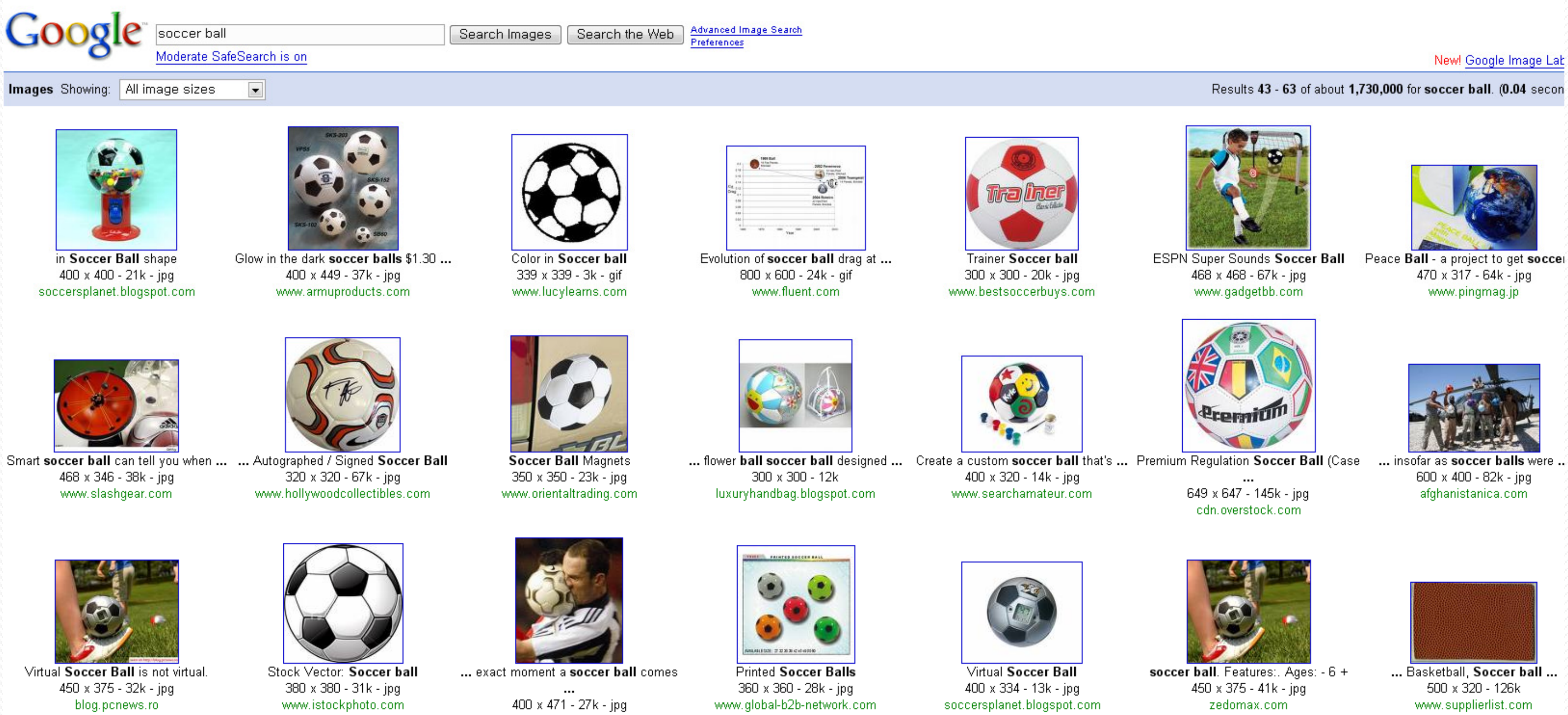
Joo Hyun Kim

Visual Recognition and Search
March 7, 2008

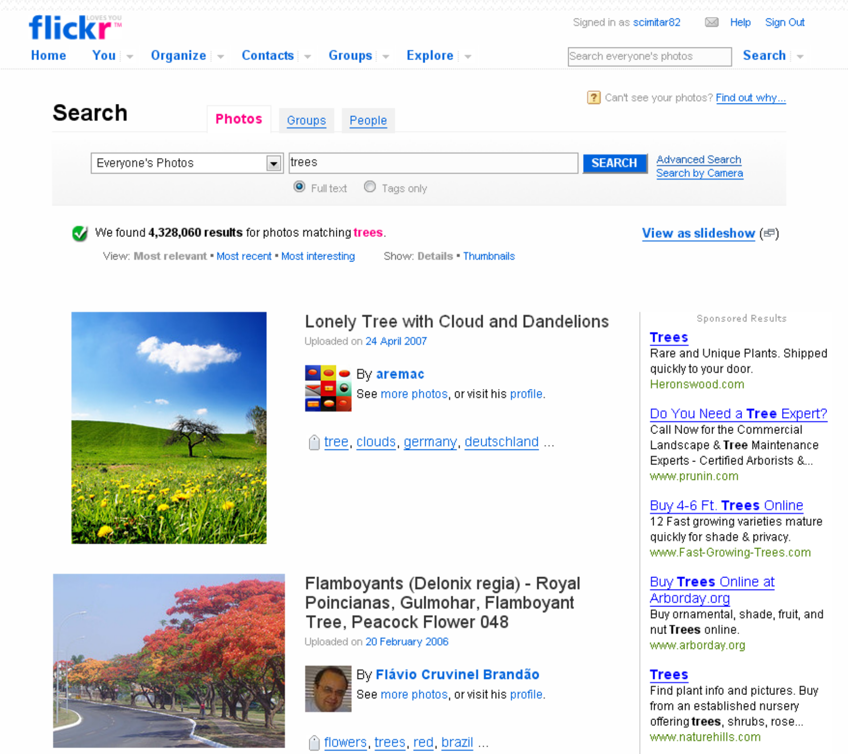
Outline

- Introduction
- Basics of co-training and how it works
- Datasets
- Experiment results
- Conclusion

Images with text captions



Introduction



- Images often come up with text captions
- Lots and lots of unlabeled data are available on the internet

Introduction

- Motivation
  - How can we use text captions for visual object recognition?
  - Use both text captions and image contents as two separate, redundant views
  - Use lots of unlabeled training examples with text captions to improve classification accuracy

Introduction

- Goal
  - Exploit multi-modal representation (text captions and image contents) and unlabeled data (usually easily available): **Co-training**
  - Learn more accurate image classifiers than standard supervised learning with abundant unlabeled data

Outline

- Introduction
- Basics of co-training and how it works
- Datasets
- Experiment results
- Conclusion

Co-training

- First proposed by Blum and Mitchell (1998)
- Semi-supervised learning paradigm that exploits two distinct, redundant views
- Features of dataset can be divided into two sets:
  - The instance space: $X = X_1 \times X_2$
  - Each example: $x = (x_1, x_2)$
- Proven to be effective at several domains
  - Web page classification (content and hyperlink)
  - E-mail classification (header and body)

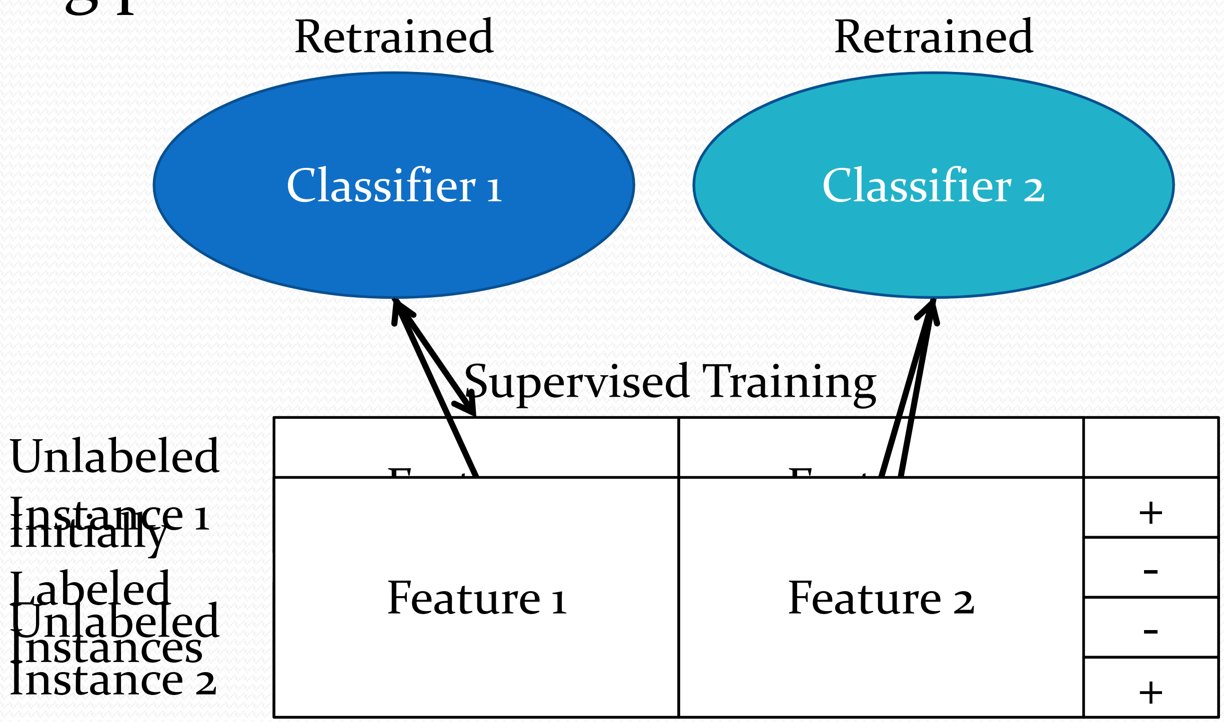
Two Assumptions of Co-training

- The instance distribution D is *compatible* with the target function $f=(f_1, f_2)$
  - Each set of features are *sufficient* to classify examples.

- The features in one set are *conditionally independent* of the features in the second set given a class
  - Informative as a random document

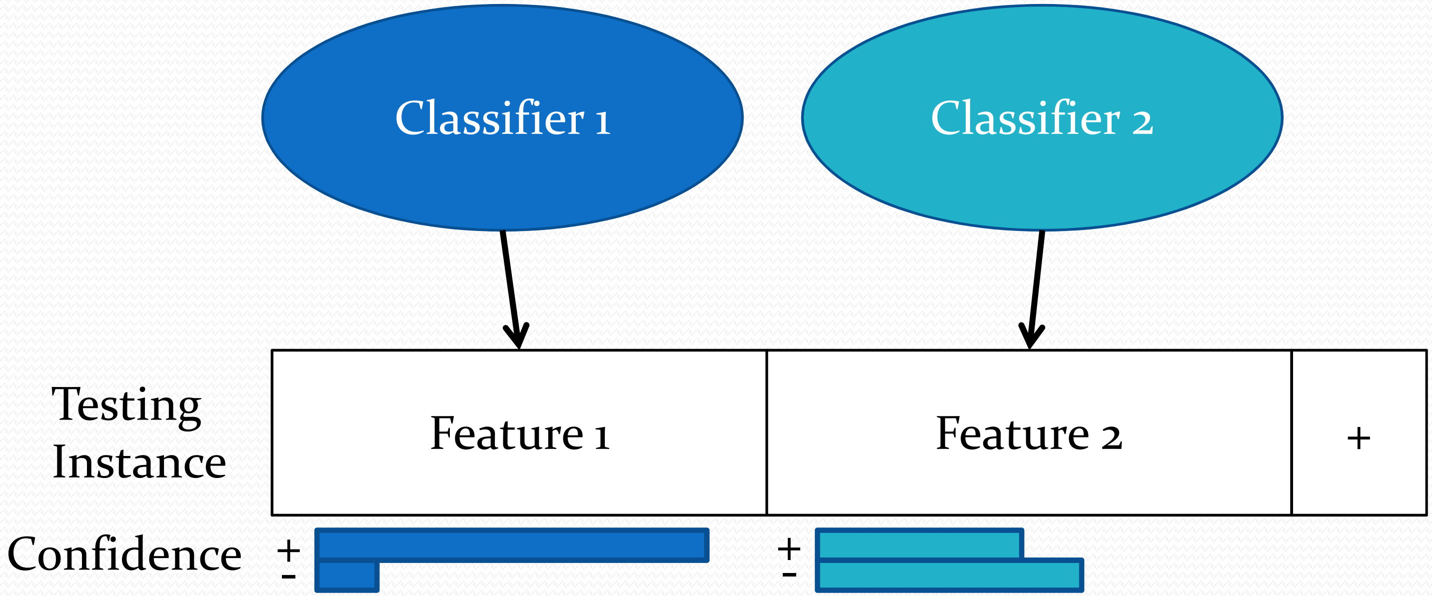
How Co-training Works

- Training process



How Co-training Works

- Testing process



Why Co-training Works?

- Intuitive explanation
  - One classifier finds an *easily classified* example (an example classified with high confidence) which maybe difficult for the other classifier
  - Provide useful information each other to improve overall accuracy

Simple Example on Image Classification

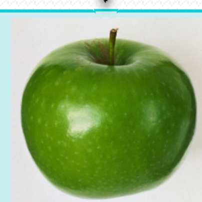
Initially labeled instances

Image	Text	Class
	red apple	Apple
	Korean pear	Pear

Confidence

$<$

New unlabeled instance

	green apple	**Apple**

Co-training Algorithm

- Given
  - labeled data L
  - unlabeled data U

- Create a pool U' of examples at random from U
- Loop for *k iterations:*
  - Train C_1 using L
  - Train C_2 using L
  - Allow C_1 to label *p positive, n negative examples from U'*
  - Allow C_2 to label *p positive, n negative examples from U'*
  - Add these self-labeled examples to L
  - Randomly choose 2p+2n examples from U to replenish U'

Modified Algorithm in the Experiment

- Inputs
  - Labeled examples set L and unlabeled examples set U represented by two sets of features, f_1 for image and f_2 for text
- Train image classifier C1 with f1 portion of L and text classifier C2 with f2 portion of L
- Loop until |U| = 0:
  - 1. Compute predictions and confidences of both classifiers for all instances of U
  - 2. For each f1 and f2, choose the *m* unlabeled instances for which its classifier has the highest confidence.
    For each such instance, if the confidence value is less than the threshold for this view, then ignore the instance and stop labeling instances with this view,
    else label the instance and add it to L
  - 3. Retrain the classifiers for both views using the augmented L
- Outputs
  - Two classifiers f_1 and f_2 whose predictions are combined to classify new test instances.
  - A test instance is labeled with the class predicted by the classifier with the higher confidence.

Outline

- Introduction
- Basics of co-training and how it works
- Datasets
- Experiment results
- Conclusion

Datasets Used

- IsraelImage dataset
  - Classes: Desert and Trees
  - Total 362 images
  - 25 image features and 363 text features
  - Image contents are more ambiguous and general
  - Text captions are natural and do not contain any particular words to directly represent class
  - www.israelimages.com

- Flickr dataset
  - Images are crawled from web with text captions & tags
  - Classes: Cars and Motorbike & Calculator and Motorbike
  - Total 907 images (Cars and Motorbike), 953 images (Calculator and Motorbike)
  - Image contents are more distinguishing between classes
  - Texts usually contain particular tags to represent class
  - www.flickr.com

Image Features – IsraelImage

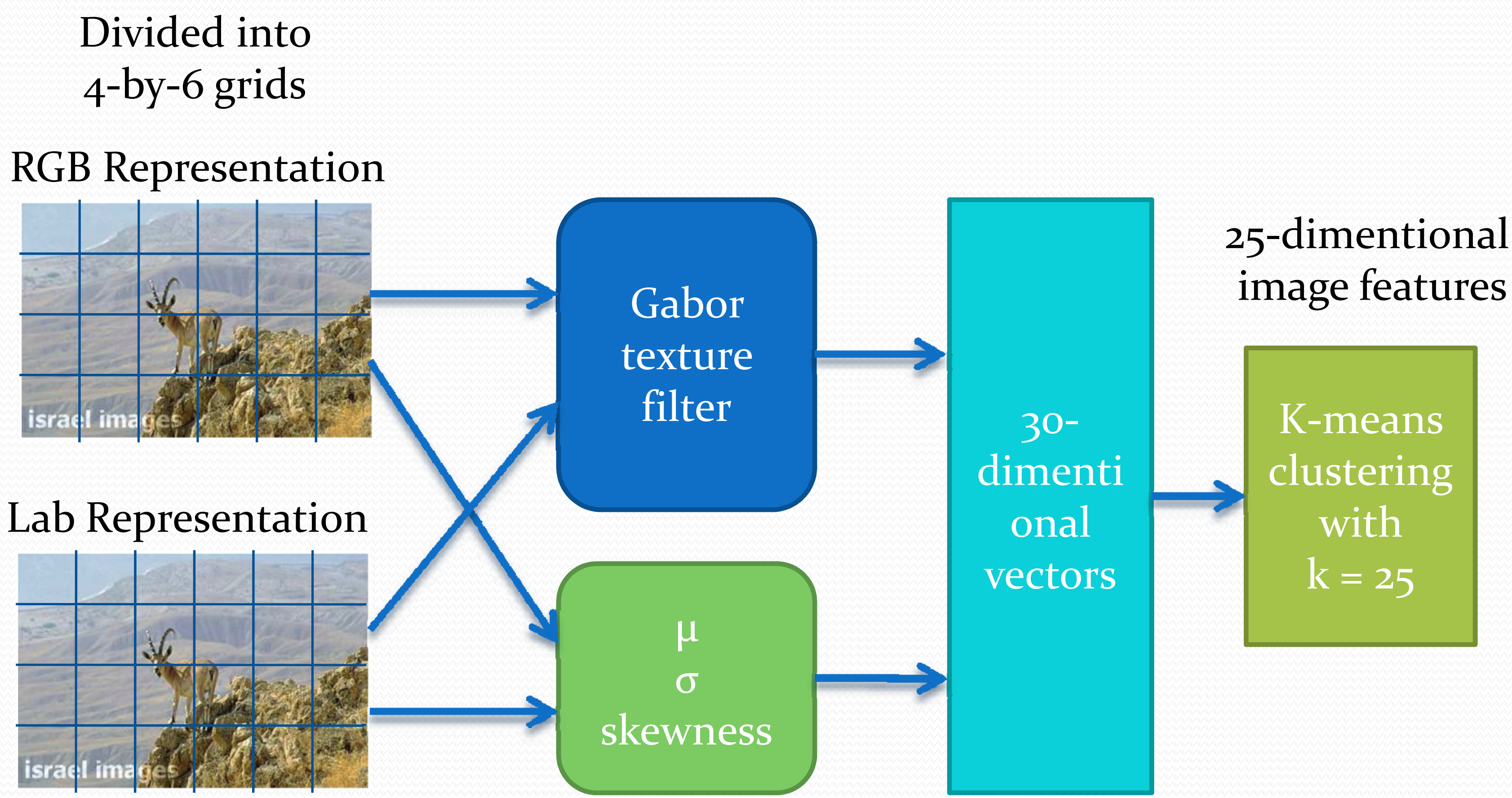


Image Features – Flickr

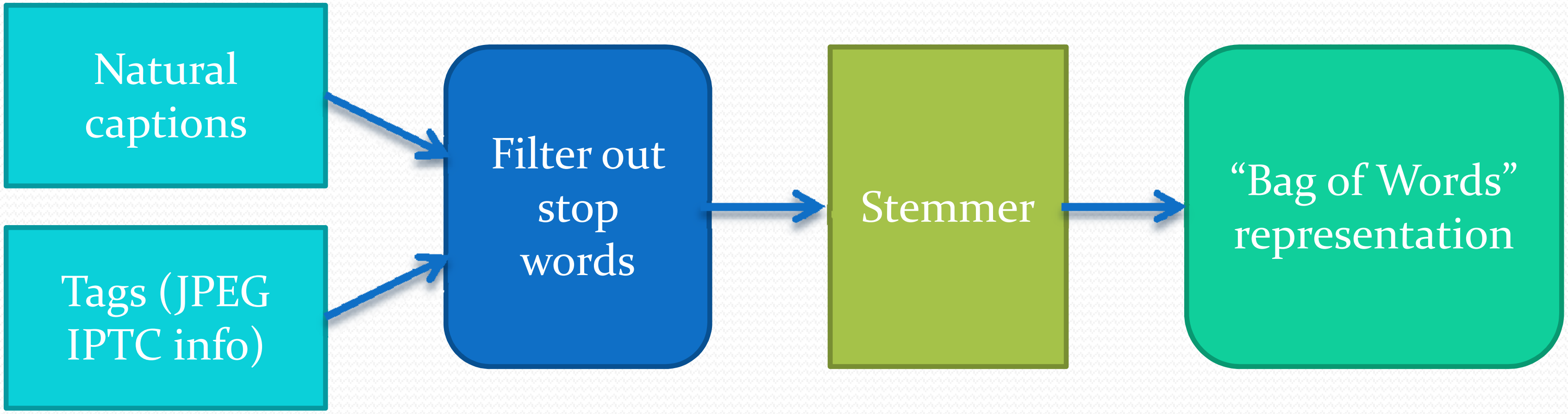
Images downloaded from flickr.com

SIFT extractor

K-means clustering with k = 75

75-dimentional image features

Text Features



An Example

- IsraelImages dataset

(a) Caption: Ibex in Judean Desert

(b) Caption: Ibex eating in the Nature

Class: Desert

Class: Trees

An Example

- Flickr dataset

• Caption: Arguably one of the most energy efficient pocket calculators ever made
• Tag: pocket, calculator, casio, macro, 2902, ddmm, daily, …

Class: Calculator

• Caption: 2008 Paeroa Battle of the Streets
• Tag: elm-pbos4, Paeroa battle of the streets, paeroa, motocycle, motorbike, race, racing, speed, …

Class: Motorbike

Outline

- Introduction
- Basics of co-training and how it works
- Datasets
- Experiment results
- Conclusion

Experiments

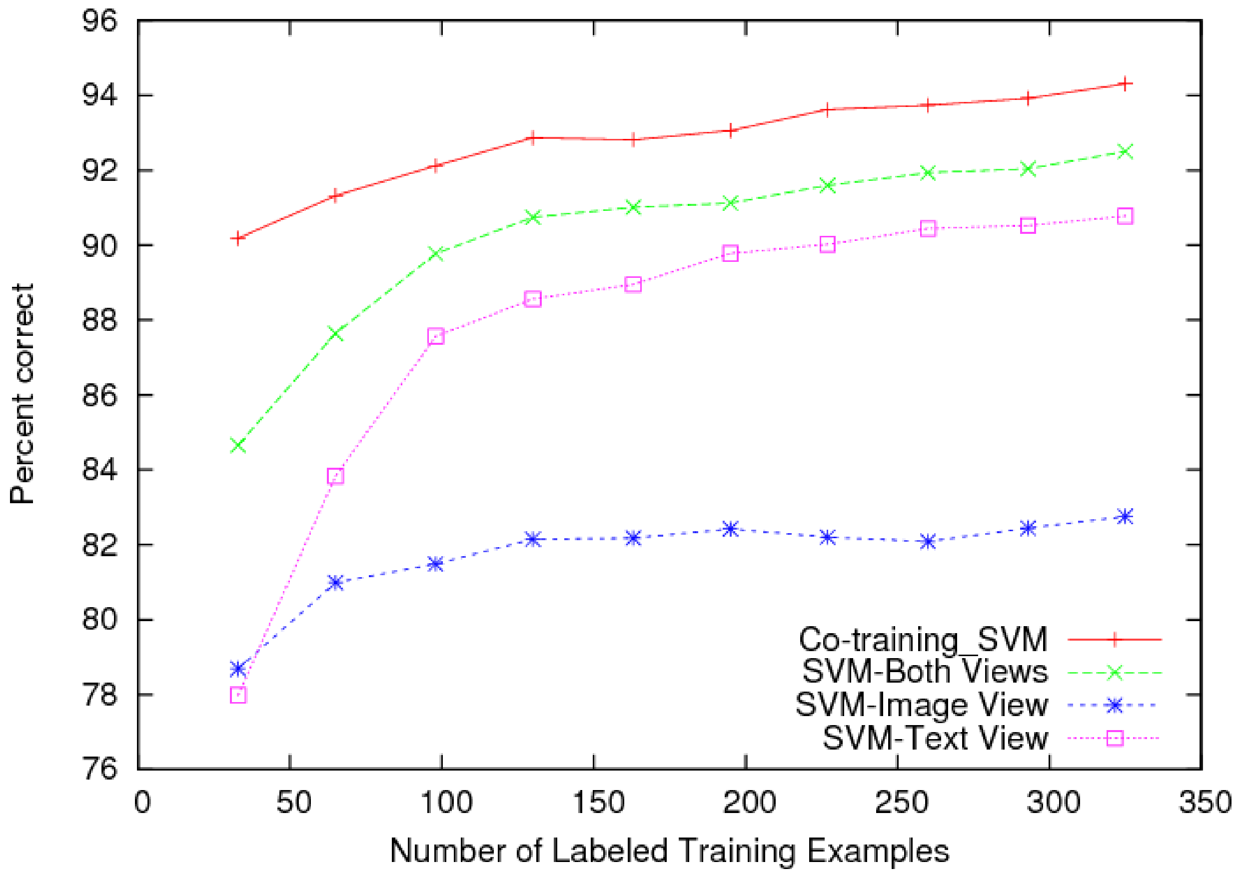
- Using WEKA (Witten, 2000), experiments are conducted with 10-fold cross validation, 1 run
- In the Co-training experiment, use SVM as base classifiers for both image and text classifiers
- Comparing Co-training with supervised SVM classifiers on concatenated features, only image, and only text features

Experiments

- Datasets are manually labeled
- Plot graphs based on the number of labeled examples and the classification accuracy
- Pick labeled examples from the training set, the other examples are used as unlabeled examples

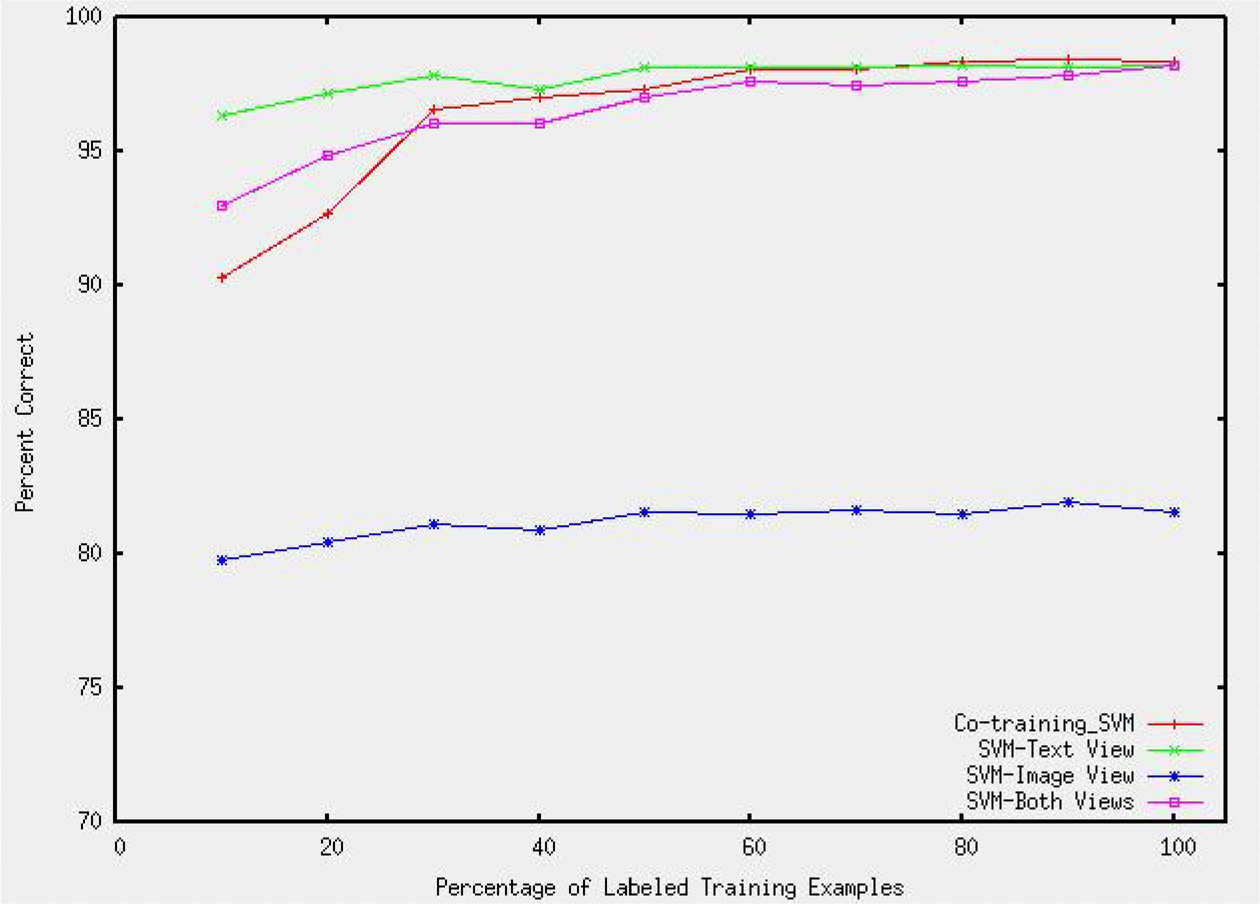
Results

- IsraelImage dataset
- Co-training vs. Supervised SVM



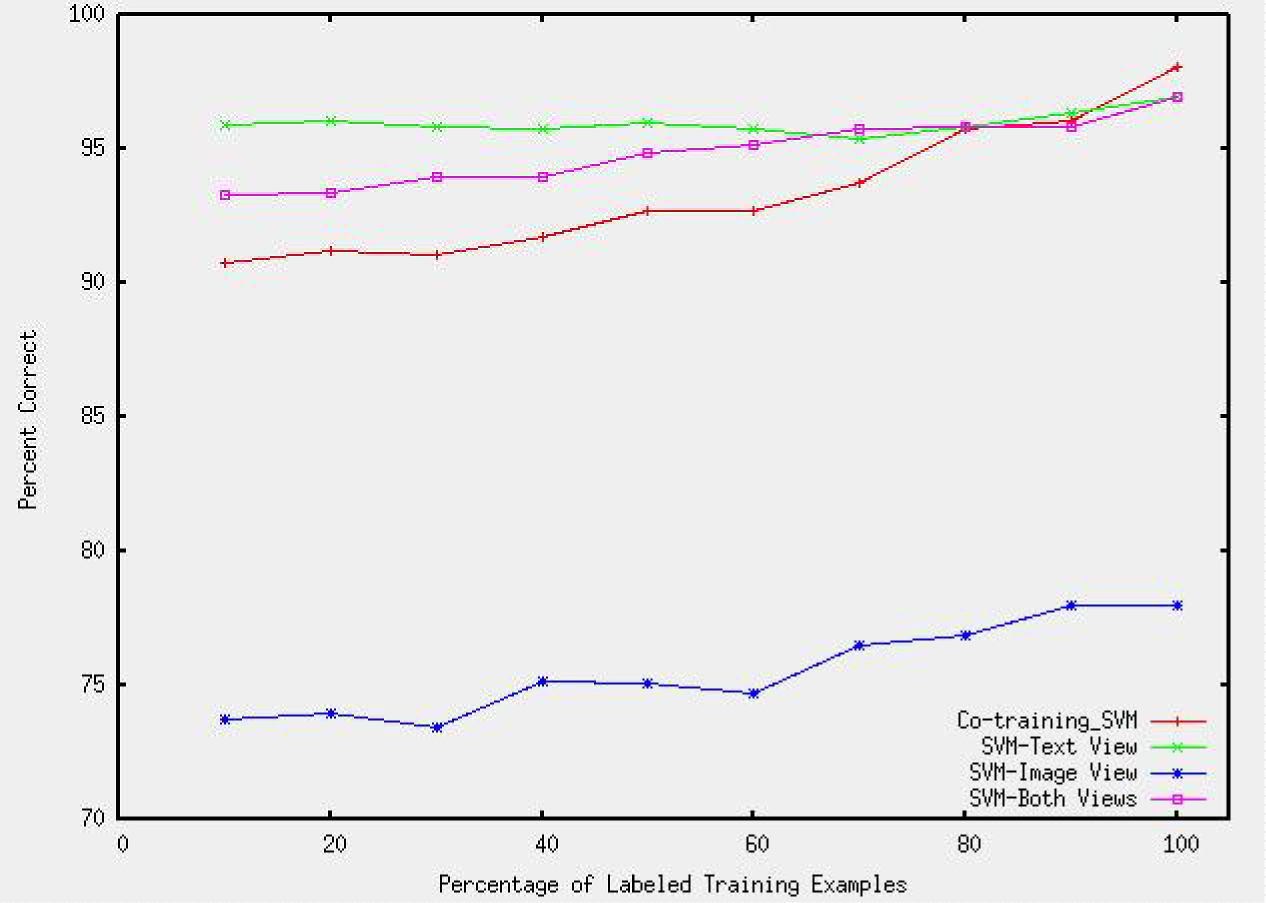
Results

- Flickr dataset
- Cars & Motorbike
- Co-training vs. Supervised SVM



Results

- Flickr dataset
- Calculator & Motorbike
- Co-training vs. Supervised SVM



Discussion

- Why IsraelImage set only shows improvement with co-training?
  - Image and text classifiers are both sufficient to classify
  - Both classifiers are helping each other well

- Why Flickr set shows worse performance?
  - Text classifier was too good (tag information is nearly as good as actual labels)
  - Image classifier actually harms the whole classification

Outline

- Introduction
- Basics of co-training and how it works
- Datasets
- Experiment results
- Conclusion

Conclusion

- Using both image contents and textual data helps classification of images
- Exploiting redundant separate views improves classification accuracy on visual object recognition
- Using unlabeled data improves supervised learning

- To use co-training effectively, the two assumptions should be met (compatibility and conditional independence)

References

- Papers
  - Co-training with Images and Text Captions – Gupta, Kim, and Mooney (2008), Under Review
  - Combining labeled and unlabeled data with co-training – Blum and Mitchell (1998), Proceedings of the 11th Annual Conference on Computational Learning Theory
  - Analyzing the effectiveness and applicability of co-training (2000) – Nigam and Ghani, Proceedings of the Ninth International Conference on Information and Knowledge Mangement
- Tools
  - WEKA system (http://www.cs.waikato.ac.nz/ml/weka/)
  - Matlab Central (http://www.mathworks.com/matlabcentral/)
  - Oxford Visual Geometry Group (http://www.robots.ox.ac.uk:5000/~vgg/index.html)

Thank You!