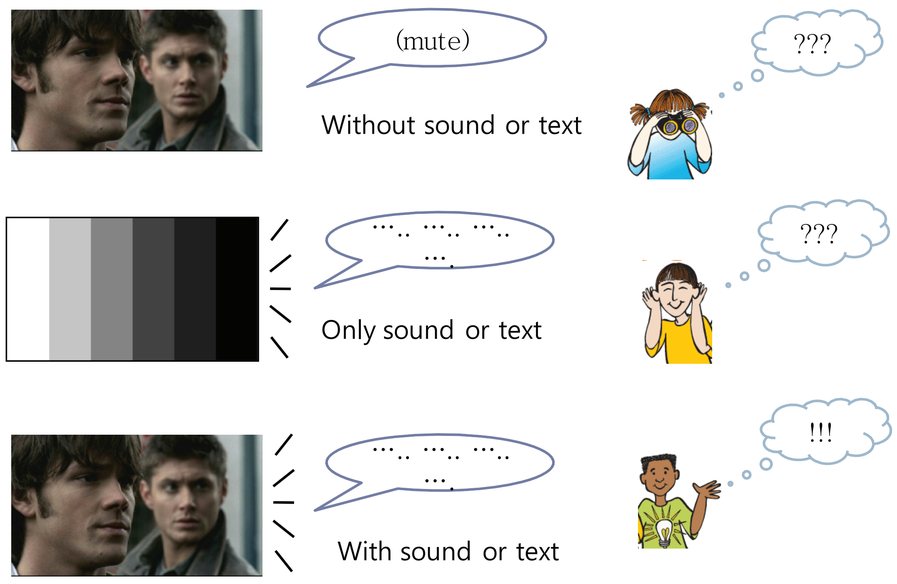
Watch, Listen & Learn: Co-training on Captioned Images and Videos

Sonal Gupta, Joohyun Kim, Kristen Grauman, Raymond Mooney
The University of Texas at Austin
{sonaluta, scimitar, grauman, mooney}@cs.utexas.edu

Introduction



Motivation

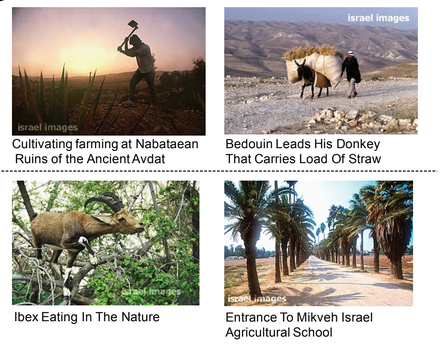
- **Image Recognition & Human Activity Recognition in Videos**
  - Hard to classify, visual cues are ambiguous
  - Expensive to manually label instances
- **Often images and videos have text captions**
  - Leverage multi-modal data
  - Use readily available unlabeled data to improve accuracy

Goals

- Classify images and videos with the help of associated text captions
- Use Co-training to achieve better classification accuracy for image and video classification task

Datasets

- Image : 362 instances with 2 classes



- Video : 221 instances with 4 classes

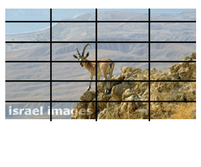
Approach

- **Combining two views (Text and Visual) of images and videos using Co-training (Blum and Mitchell '98) learning algorithm**
- **Text View**
  - Caption of image or video
  - Readily available
- **Visual View**
  - Color, texture, temporal information in image/video

Feature Extraction

Image Feature

Divide images into 4X6 grid



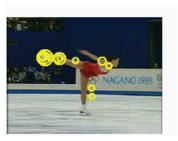
Capture texture and color distributions of each cell into 30-dim vector

Cluster the vectors using k-Means to quantize the features into a dictionary of *visual words*

Represent each image in terms of the dictionary

Video Feature

Detect Interest Points
Harris-Forstener Corner Detector for both spatial and temporal space



Describe Interest Points
Histogram of Oriented Gradients (HoG)

Create Spatio-Temporal Vocabulary
Quantize interest points to create 200 visual words dictionary

Represent each video in terms of the dictionary

Text Feature

Raw Text Commentary

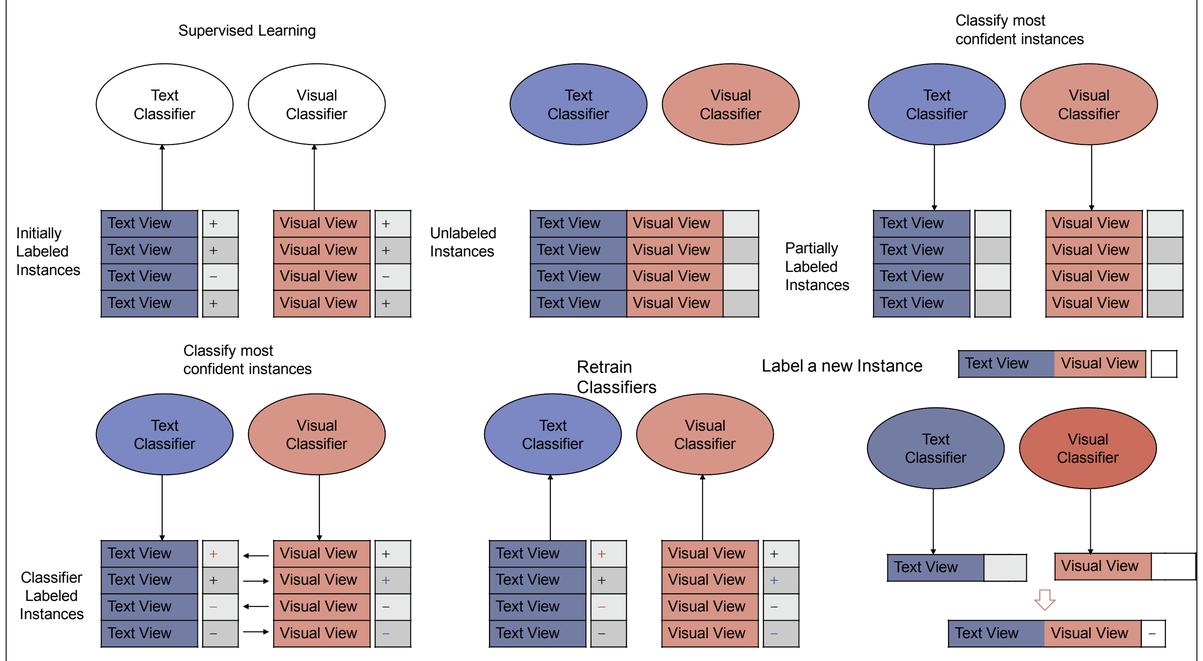
- That was a very nice forward camel.
- Well I remember her performance last time.
- He has some delicate hand movement.
- She gave a small jump while gliding
- He runs in to chip the ball with his right foot.
- He runs in to take the instep drive and executes it well.
- The small kid pushes the ball ahead with his tiny kicks.

Porter Stemmer — Remove Stop Words

Standard Bag-of-Words Representation

Algorithm

- **Co-training**
  - Semi-supervised learning paradigm that exploits two mutually independent and sufficient 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 in several domains**
  - Web page classification (content and hyperlink)
  - E-mail classification (header and body)



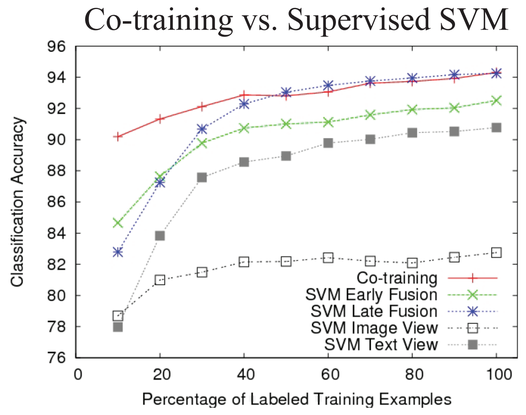
Experimental Results

Baselines

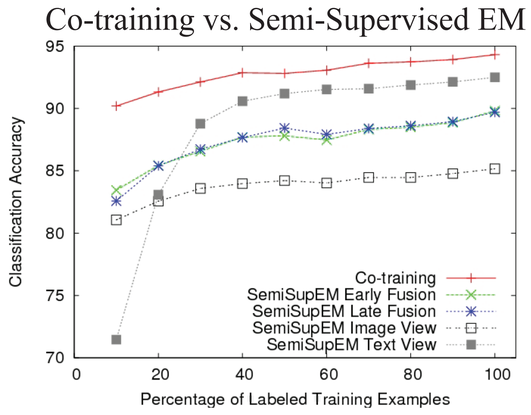
- **Uni-modal**
  - Image/Video View : Only image/video features are used
  - Text View : Only textual features are used
- **Multi-modal**
  - Early Fusion : Concatenate visual and textual features and train classifier
  - Late Fusion : Run separate classifiers on each view and concatenate the results

Image Dataset

Co-training vs. Supervised SVM

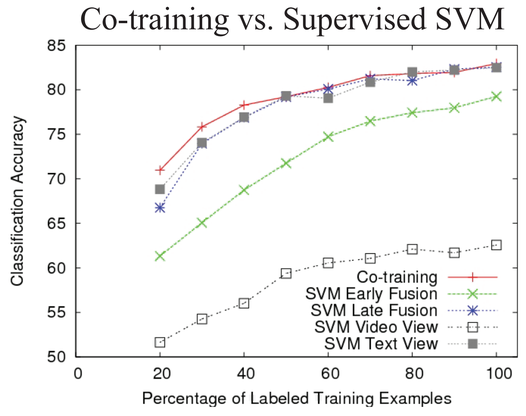


Co-training vs. Semi-Supervised EM

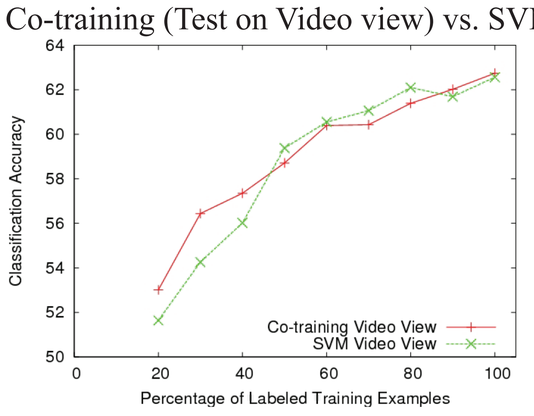


Video Dataset

Co-training vs. Supervised SVM



Co-training (Test on Video view) vs. SVM



Conclusion

- Combining textual and visual features can help improve accuracy
- Co-training can be useful to combine textual and visual features to classify images and videos
- Co-training helps in reducing labeling of images and videos

References

[1] Bekkerman and Jeon, Multi-modal Clustering for Multimedia Collections. CVPR 2007
[2] Blum and Mitchell, Combining labeled and unlabeled data with co-training, COLT 1998
[3] Laptev, On space-time interest points, IJCV 2005
[4] Weka Data Mining Tool (Witten and Frank)