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

Sonal Gupta, Joohyun Kim, Kristen Grauman, Raymond Mooney
The University of Texas at Austin, U.S.A.
Outline

- Introduction
- Motivation
- Approach
- How does Co-training work?
- Experimental Evaluation
- Conclusions
Outline

- Introduction
- Motivation
- Approach
- How does Co-training work?
- Experimental Evaluation
- Conclusions
Introduction

(mute)

Without sound or text
Introduction

Without sound or text
Introduction

Without sound or text

(mute)

Only sound or text

????
Introduction

(mute)
Without sound or text

...... ...... ....
Only sound or text

???

???
Introduction

- **(mute)**
  - Without sound or text

- **...... ...... ....**
  - Only sound or text

- **...... ...... ....**
  - With sound or text

The University of Texas at Austin
Introduction

Without sound or text

Only sound or text

With sound or text
Motivation
Motivation

- Image Recognition & Human Activity Recognition in Videos
Motivation

- Image Recognition & Human Activity Recognition in Videos
  - Hard to classify, ambiguous visual cues
Motivation

- Image Recognition & Human Activity Recognition in Videos
  - Hard to classify, ambiguous visual cues
  - Expensive to manually label instances
Motivation

- Image Recognition & Human Activity Recognition in Videos
  - Hard to classify, ambiguous visual cues
  - Expensive to manually label instances

- Often images and videos have text captions
Motivation

- Image Recognition & Human Activity Recognition in Videos
  - Hard to classify, ambiguous visual cues
  - Expensive to manually label instances

- Often images and videos have text captions
  - Leverage multi-modal data
Motivation

- Image Recognition & Human Activity Recognition in Videos
  - Hard to classify, ambiguous visual cues
  - 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 visual information and associated text captions
Goals

- Classify images and videos with the help of visual information and associated text captions
- Use unlabeled image and video examples
Image Examples

Desert
- Cultivating farming at Nabataean Ruins of the Ancient Avdat
- Bedouin Leads His Donkey That Carries Load Of Straw

Trees
- Ibex Eating In The Nature
- Entrance To Mikveh Israel Agricultural School
Video Examples

Kicking

He runs in and hits ball with the inside of his shoes to reach the target

Dribbling

Using the sole to tap the ball she keeps it in check.

Dancing

Her last spin is going to make her win

Spinning

God, that jump was very tricky
**Video Examples**

**Kicking**

He runs in and hits ball with the inside of his shoes to reach the target.

**Dribbling**

Using the sole to tap the ball she keeps it in check.

**Dancing**

Her last spin is going to make her win.

**Spinning**

God, that jump was very tricky.
Related Work
Related Work

- Images + Text
  - Barnard et al. (JMLR 03) and Duygulu et al. (ECCV 02) generated models to annotate image regions with words.
  - Bekkerman and Jeon (CVPR 07) exploited multi-modal information to cluster images with captions.
  - Quattoni et al. (CVPR 07) used unlabeled images with captions to improve learning in future image classification problems with no associated captions.
Related Work

- **Images + Text**
  - Barnard et al. (JLMR 03) and Duygulu et al. (ECCV 02) generated models to annotate image regions with words.
  - Bekkerman and Jeon (CVPR 07) exploited multi-modal information to cluster images with captions.
  - Quattoni et al. (CVPR 07) used unlabeled images with captions to improve learning in future image classification problems with no associated captions.

- **Videos + Text**
  - Wang et al. (MIR 07) used co-training to combine visual and textual ‘concepts’ to categorize TV ads., retrieved text using OCR and used external sources to expand the textual features.
  - Everingham et al. (BMVC 06) used visual information, closed-captioned text, and movie scripts to annotate faces.
  - Fleischman and Roy (NAACL 07) used text commentary and motion description in baseball games to retrieve relevant video clips given text query.
Outline

- Introduction
- Motivation
- Approach
- How does Co-training work?
- Experimental Evaluation
- Conclusions
Approach

- Combining two views of images and videos using Co-training (Blum and Mitchell '98) learning algorithm

- Views: Text and Visual
  - Text View
    - Caption of image or video
    - Readily available
  - Visual View
    - Color, texture, temporal information in image/video
Outline

- Introduction
- Motivation
- Approach
- How does Co-training work?
- Experimental Evaluation
- Conclusions
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)
Co-training

Initially Labeled Instances

<table>
<thead>
<tr>
<th>Text View</th>
<th>Visual View</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>
Co-training

Initially Labeled Instances

<table>
<thead>
<tr>
<th>Text View</th>
<th>+</th>
<th>Visual View</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text View</td>
<td>+</td>
<td>Visual View</td>
<td>+</td>
</tr>
<tr>
<td>Text View</td>
<td>-</td>
<td>Visual View</td>
<td>-</td>
</tr>
<tr>
<td>Text View</td>
<td>+</td>
<td>Visual View</td>
<td>+</td>
</tr>
</tbody>
</table>
Co-training

Supervised Learning

Text Classifier

Visual Classifier
Co-training

Unlabeled Instances

<table>
<thead>
<tr>
<th>Text View</th>
<th>Visual View</th>
<th>Text View</th>
<th>Visual View</th>
<th>Text View</th>
<th>Visual View</th>
<th>Text View</th>
<th>Visual View</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Co-training

Classify most confident instances

Text Classifier

Text View
Text View
Text View
Text View

Visual Classifier

Visual View
Visual View
Visual View
Visual View
Co-training

Classify most confident instances

Text Classifier

Visual Classifier

Partially Labeled Instances

<table>
<thead>
<tr>
<th>Text View</th>
<th>Text View</th>
<th>Text View</th>
<th>Text View</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Visual View</th>
<th>Visual View</th>
<th>Visual View</th>
<th>Visual View</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Co-training

Classify most confident instances

Partially Labeled Instances

Text View
Text View
Text View
Text View

Visual View
Visual View
Visual View
Visual View

Text Classifier

Visual Classifier

+ 

- 

The University of Texas at Austin
Co-training

Label all views in instances

Classifier Labeled Instances

Text View
Text View
Text View
Text View

+ + - -

Visual View
Visual View
Visual View
Visual View

+ + - -

Text Classifier

Visual Classifier
Co-training

Text Classifier

Visual Classifier

<table>
<thead>
<tr>
<th>Text View</th>
<th>Visual View</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Co-training

Text Classifier

Visual Classifier

<table>
<thead>
<tr>
<th>Text View</th>
<th>Visual View</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Co-training

Retrain Classifiers

Text Classifier

Visual Classifier
Co-training

Label a new Instance

Text View  Visual View
Co-training

Label a new Instance

Text View  Visual View

Text Classifier

Text View  Visual View
Co-training

Label a new Instance

Text View  Visual View

Text Classifier

Text View  + -

Visual Classifier

Visual View  + -
Co-training

Label a new Instance

Text View  Visual View

Text Classifier

Text View  +  -

Visual Classifier

Visual View  +  -

Text View  Visual View  -
Features

- Visual Features
  - Image Features
  - Video features
- Textual features
Image Features

Divide images into 4x6 grid

[Fei-Fei et al. ‘05, Bekkerman & Jeon ‘07]

The University of Texas at Austin
Image Features

Divide images into 4X6 grid

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

[Fei-Fei et al. ‘05, Bekkerman & Jeon ‘07]
Image Features

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*

[Fei-Fei et al. ‘05, Bekkerman & Jeon ‘07]
Image Features

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 as histogram of visual words

[Fei-Fei et al. '05, Bekkerman & Jeon '07]
Video Features

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

[Laptev, IJCV ‘05]
Video Features

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

Describe Interest Points
Histogram of Oriented Gradients (HoG)

[Laptev, IJCV ‘05]
Video Features

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

[Laptev, IJCV ‘05]
Video Features

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 as histogram of visual words

[Laptev, IJCV ‘05]
Textual Features

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
Outline

- Introduction
- Motivation
- Approach
- How does Co-training work?
- Experimental Evaluation
- Conclusions
Experimental Methodology

- Test set is disjoint from both labeled and unlabeled training set

- For plotting learning curves, vary the percentage of training examples labeled

- SVM is used as base classifier for both visual and text classifiers
  - SMO implementation in WEKA (Witten & Frank ‘05)
  - RBF Kernel ($\gamma = 0.01$)

- All experiments are evaluated with 10 iterations of 10-fold cross-validation
Baselines - Overview

- Uni-modal
  - Visual View
  - Textual View

- Multi-modal (Snoek et al. ICMI '05)
  - Early Fusion
  - Late Fusion

- Supervised SVM
  - Uni-modal, Multi-modal

- Other Semi-Supervised methods
  - Semi-Supervised EM - Uni-modal, Multi-modal
  - Transductive SVM - Uni-modal, multi-modal
Baseline - Individual Views

- Individual views
  - Image/Video View: Only image/video features are used
  - Text View: Only textual features are used
Baseline - Early Fusion

- Concatenate visual and textual features

<table>
<thead>
<tr>
<th>Text View</th>
<th>Visual View</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text View</td>
<td>Visual View</td>
<td>-</td>
</tr>
</tbody>
</table>

Training

Classifier

Testing

| Text View | Visual View | - |
Baseline - Late Fusion

<table>
<thead>
<tr>
<th>Text View</th>
<th>+</th>
<th>Visual View</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text View</td>
<td>-</td>
<td>Visual View</td>
<td>-</td>
</tr>
</tbody>
</table>

Text Classifier

Visual Classifier
Baseline - Late Fusion

Text View  +  Visual View  +
Text View  -  Visual View  -

Text Classifier

Visual Classifier
Baseline - Late Fusion

Text View + Visual View +
Text View - Visual View -

Training

Text Classifier

Visual Classifier
Baseline - Late Fusion

Text View  +  Visual View  +
Text View    -

Training

Text Classifier

Label a new instance

Text View  +  -
Visual View  +  -
Baseline - Late Fusion

Text View + - Visual View + -

Text View - Visual View -

Training

Text Classifier

Visual Classifier

Label a new instance

Text View + - Visual View + -

Text View Visual View -

The University of Texas at Austin
Baseline - Other Semi-Supervised

- Semi-Supervised Expectation Maximization (SemiSup EM)
  - Introduced by Nigam et al. CIKM ‘00
  - Used Naïve bayes as the base classifier

- Transductive SVM in Semi-Supervised setting
  - Introduced by Joachims ICML ‘99, Bennett & Demiriz ANIPS ‘99
Image Dataset

- Our image data is taken from the Israel dataset
  (Bekkerman & Jeon CVPR ‘07, www.israelimages.com)
- Consists of images with short text captions
- Used two classes, Desert and Trees
- A total of 362 instances
Image Examples

Desert

Cultivating farming at Nabataean Ruins of the Ancient Avdat

Bedouin Leads His Donkey That Carries Load Of Straw

Trees

Ibex Eating In The Nature

Entrance To Mikveh Israel Agricultural School
Results

Co-training v. Supervised SVM

![Graph showing classification accuracy vs. percentage of labeled training examples for different methods including Co-training, SVM Early Fusion, SVM Late Fusion, SVM Image View, and SVM Text View.]
Results

Co-training v. Supervised SVM

Co-training

Classification Accuracy vs. Percentage of Labeled Training Examples

- Co-training
- SVM Early Fusion
- SVM Late Fusion
- SVM Image View
- SVM Text View
Results

Co-training v. Supervised SVM

The diagram shows a graph with the title "SVM Late Fusion". The x-axis represents the percentage of labeled training examples, ranging from 0 to 100, and the y-axis represents classification accuracy, ranging from 76 to 96. There are several lines indicating different methods, including Co-training, SVM Early Fusion, SVM Late Fusion, SVM Image View, and SVM Text View. The graph illustrate how each method's performance changes with the increase in labeled training examples.
Results
Co-training v. Supervised SVM

![Graph showing classification accuracy vs. percentage of labeled training examples for different methods: Co-training, SVM Early Fusion, SVM Late Fusion, SVM Image View, SVM Text View. SVM Early Fusion performs the best throughout the range of labeled data percentages.](image)
Results

Co-training v. Supervised SVM

![Graph showing comparison between Co-training and Supervised SVM]
Results

Co-training v. Supervised SVM

![Graph showing classification accuracy vs. percentage of labeled training examples for different methods: Co-training, SVM Early Fusion, SVM Late Fusion, SVM Image View, SVM Text View. The graph illustrates the performance of these methods across varying percentages of labeled data.]
Results
Co-training v. Supervised SVM
Results
Co-training v. Supervised SVM

~5%
Results
Co-training v. Supervised SVM

~7%
Results

Co-training v. Supervised SVM

![Graph showing classification accuracy vs. percentage of labeled training examples for different methods. The graph indicates a significant improvement with co-training, approximately 12%.](image-url)
Results
Co-training v. Supervised SVM

![Graph showing classification accuracy vs. percentage of labeled training examples for Co-training, SVM Early Fusion, SVM Late Fusion, SVM Image View, and SVM Text View.]
Results
Co-training v. Semi-Supervised EM

![Graph showing classification accuracy vs. percentage of labeled training examples for different methods including Co-training, SemiSupEM Early Fusion, SemiSupEM Late Fusion, SemiSupEM Image View, and SemiSupEM Text View.]
Results

Co-training v. Semi-Supervised EM

Co-training

Classification Accuracy

Co-training
SemiSupEM Early Fusion
SemiSupEM Late Fusion
SemiSupEM Image View
SemiSupEM Text View

Percentage of Labeled Training Examples
Results
Co-training v. Semi-Supervised EM

![Graph showing classification accuracy vs. percentage of labeled training examples. The graph includes lines for Co-training, SemiSupEM Early Fusion, SemiSupEM Late Fusion, SemiSupEM Image View, and SemiSupEM Text View. The graph highlights SemiSup-EM Late Fusion.]
Results

Co-training v. Semi-Supervised EM

![Graph showing classification accuracy vs. percentage of labeled training examples for different methods. The graph highlights SemiSup-EM Early Fusion.](image-url)
Results
Co-training v. Semi-Supervised EM

SemiSup-EM Text View

Classification Accuracy

Percentage of Labeled Training Examples

Co-training
SemiSupEM Early Fusion
SemiSupEM Late Fusion
SemiSupEM Image View
SemiSupEM Text View
Results

Co-training v. Semi-Supervised EM

[Graph showing comparison between Co-training and different variants of SemiSupEM: Early Fusion, Late Fusion, Image View, Text View. The x-axis represents the percentage of labeled training examples, and the y-axis represents classification accuracy. The graph shows that Co-training generally outperforms the SemiSupEM variants, especially at lower percentages of labeled data.]

SemiSup-EM Image View

The University of Texas at Austin
Results

Co-training v. Semi-Supervised EM
Results
Co-training v. Semi-Supervised EM
Results

Co-training v. Transductive SVM

![Graph showing classification accuracy versus percentage of labeled training examples for different methods: Co-training, TSVM Early Fusion, TSVM Late Fusion, TSVM Image View, TSVM Text View. The Co-training method consistently shows the highest accuracy.]
Results
Co-training v. Transductive SVM

- Classification Accuracy vs. Percentage of Labeled Training Examples
  - Co-training
  - TSVM Early Fusion
  - TSVM Late Fusion
  - TSVM Image View
  - TSVM Text View

Approximately 4% improvement
Video Dataset
Video Dataset

- Manually collected video clips of
  - kicking and dribbling from soccer game DVDs
  - dancing and spinning from figure skating DVDs
Video Dataset

- Manually collected video clips of
  - kicking and dribbling from soccer game DVDs
  - dancing and spinning from figure skating DVDs
- Manually commented the clips
Video Dataset

- Manually collected video clips of
  - kicking and dribbling from soccer game DVDs
  - dancing and spinning from figure skating DVDs
- Manually commented the clips
- Significant variation in the size of the person across the clips
Video Dataset

- Manually collected video clips of
  - kicking and dribbling from soccer game DVDs
  - dancing and spinning from figure skating DVDs
- Manually commented the clips
- Significant variation in the size of the person across the clips
- Number of clips
  - dancing: 59, spinning: 47, dribbling: 55 and kicking: 60
Video Dataset

- Manually collected video clips of
  - kicking and dribbling from soccer game DVDs
  - dancing and spinning from figure skating DVDs
- Manually commented the clips
- Significant variation in the size of the person across the clips
- Number of clips
  - dancing: 59, spinning: 47, dribbling: 55 and kicking: 60
- The video clips
  - resized to 240x360 resolution
  - length varies from 20 to 120 frames
## Video Examples

<table>
<thead>
<tr>
<th>Kicking</th>
<th>Dribbling</th>
</tr>
</thead>
<tbody>
<tr>
<td>He runs in and hits ball with the inside of his shoes to reach the target</td>
<td>Using the sole to tap the ball she keeps it in check.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dancing</th>
<th>Spinning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Her last spin is going to make her win</td>
<td>God, that jump was very tricky</td>
</tr>
</tbody>
</table>
Results
Co-training v. Supervised SVM
Results
Co-training v. Supervised SVM

![Graph showing comparison between Co-training and supervised SVM models. The x-axis represents the percentage of labeled training examples, and the y-axis represents classification accuracy. The legend indicates the models compared: Co-training, SVM Early Fusion, SVM Late Fusion, SVM Video View, SVM Text View. Co-training consistently outperforms the other models across all labeled data percentages.](image_url)
Results
Co-training v. Supervised SVM

![Graph showing classification accuracy vs. percentage of labeled training examples for different methods, including Co-training, SVM Early Fusion, SVM Late Fusion, SVM Video View, and SVM Text View. The graph indicates that Co-training and SVM Late Fusion generally perform better than the other methods, especially at lower percentages of labeled examples.](image-url)
Results

Co-training v. Supervised SVM

![Graph showing classification accuracy vs. percentage of labeled training examples.](image)

- Co-training
- SVM Early Fusion
- SVM Late Fusion
- SVM Video View
- SVM Text View

SVM Early Fusion
Results

Co-training v. Supervised SVM

![Graph showing comparison between Co-training and Supervised SVM methods]
Results
Co-training v. Supervised SVM

Classification Accuracy

SVM Video View

Percentage of Labeled Training Examples
Results
Co-training v. Supervised SVM

![Graph showing comparison between Co-training and Supervised SVM]
What if test Videos have no captions?

- During training
  - Video has associated text caption

- During Testing
  - Video with no text caption

- Real life situation

- Co-training can exploit text captions during training to improve video classifier
Results

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

![Graph showing classification accuracy comparison between Co-training Video View and SVM Video View. The graph indicates a difference of approximately 2%.](image)
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

[More information on http://www.cs.utexas.edu/users/ml/co-training]
Questions?
References

- Bekkerman et al. Multi-way distributional clustering, ICML 2005
- Blum and Mitchell, Combining labeled and unlabeled data with co-training, COLT 1998
- Laptev, On space-time interest points, IJCV 2005
- Weka Data Mining Tool (Witten and Frank)
Dataset Details

- **Image:**
  - $k=25$ for k-Means
  - Number of textual features - 363

- **Video:**
  - Most clips 20 to 40 frames
  - $k=200$ in k-Means
  - Number of textual features - 381
Feature Details

- **Image Features**
  - Texture features - Gabor filters with 3 scales and 4 orientations
  - Color - Mean, Standard deviation & Skewness of per-channel RBG and Lab color pixel values

- **Video Features**
  - Maximizes a normalized spatio-temporal Laplacian operation over both spatial and temporal scales
  - HoG - 3x3x2 spatio-temporal blocks, 4-bin HoG descriptor for every block => 72 element descriptor
Methodology Details

- Batch size = 5 in Co-training
- Thresholds for image experiments
  - image view = 0.65
  - text view = 0.98
- Thresholds for video experiments
  - image view = 0.6
  - text view = 0.9
- Experiments evaluated using two-tailed paired t-test with 95% confidence level