Leveraging High-level and Low-level Features for Multimedia Event Detection

Lu Jiang, Alexander G. Hauptmann, Guang Xiang

School of Computer Science

Carnegie Mellon University

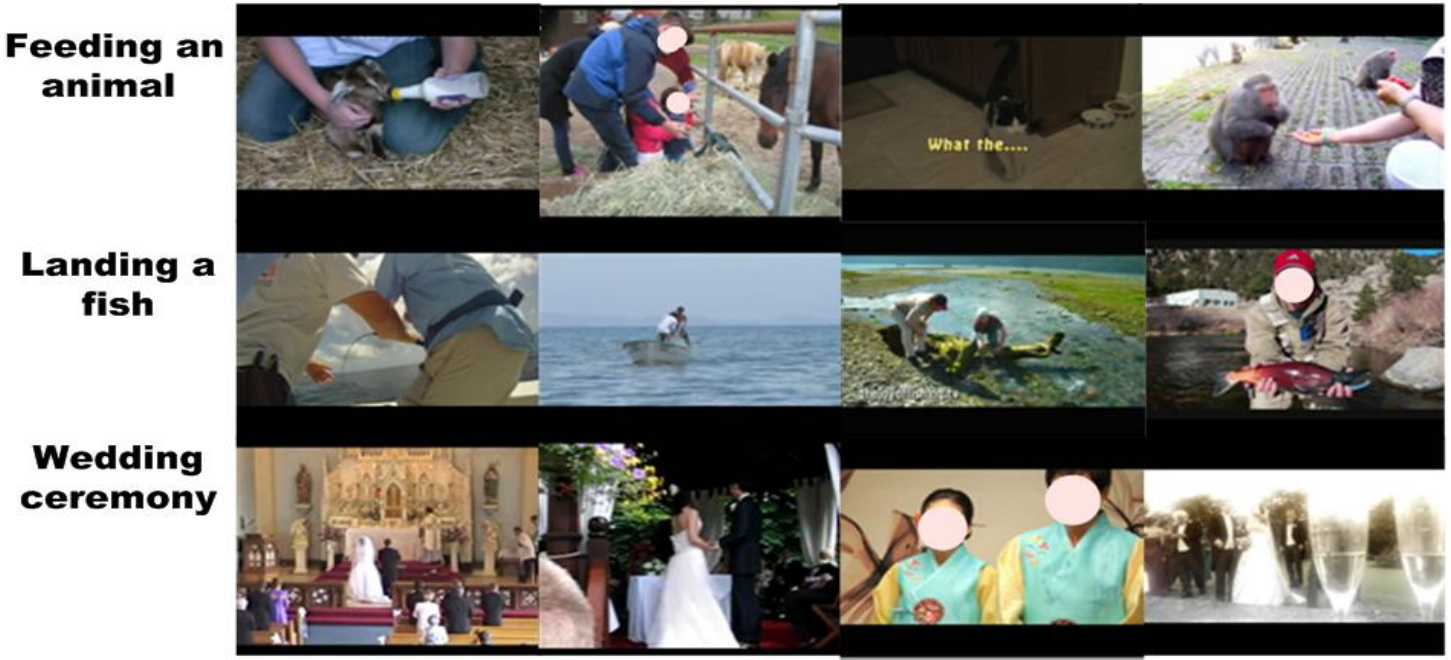
Carnegie Mellon University

Outline

- **Intuition**
- Methods
  - Graph Construction
  - Collective Classification
  - Concept Selection
- Experimental Results

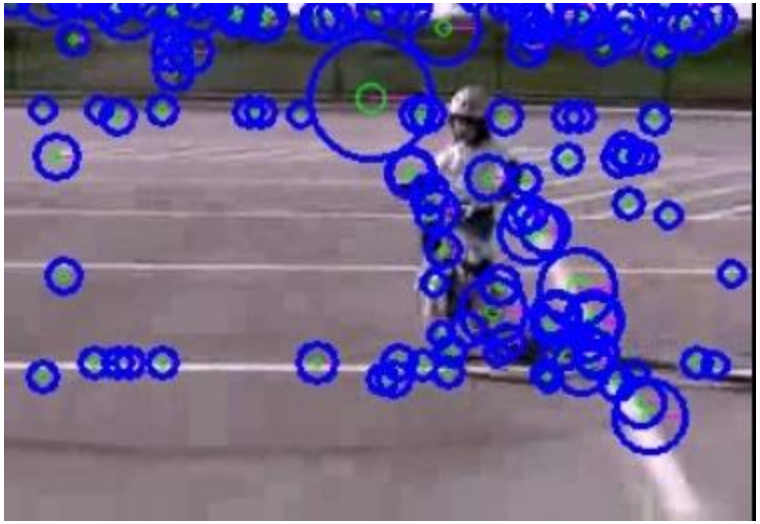
Problem

- **Multimedia Event Detection**
  - Given a collection of test videos and a list of test events, indicate whether each of the test events is present anywhere in each of the test videos.
  - Give the strength of evidence for each such judgment.



Features

- Low-level: SIFT, Color SIFT and MoSIFT.
  - Capture local appearance and texture statistics of objects.
  - Better performance in classification but less interpretable.
- High-level: Semantic Concepts and Object Bank
  - Estimate the probability of observing an object or concept .
  - Consistent with human's understanding but less effective than low-level features.



Low-level features:

- A collection of Interest points

High-level features:

- Outdoor, Person, Bike

Carnegie Mellon University

Classical Fusion

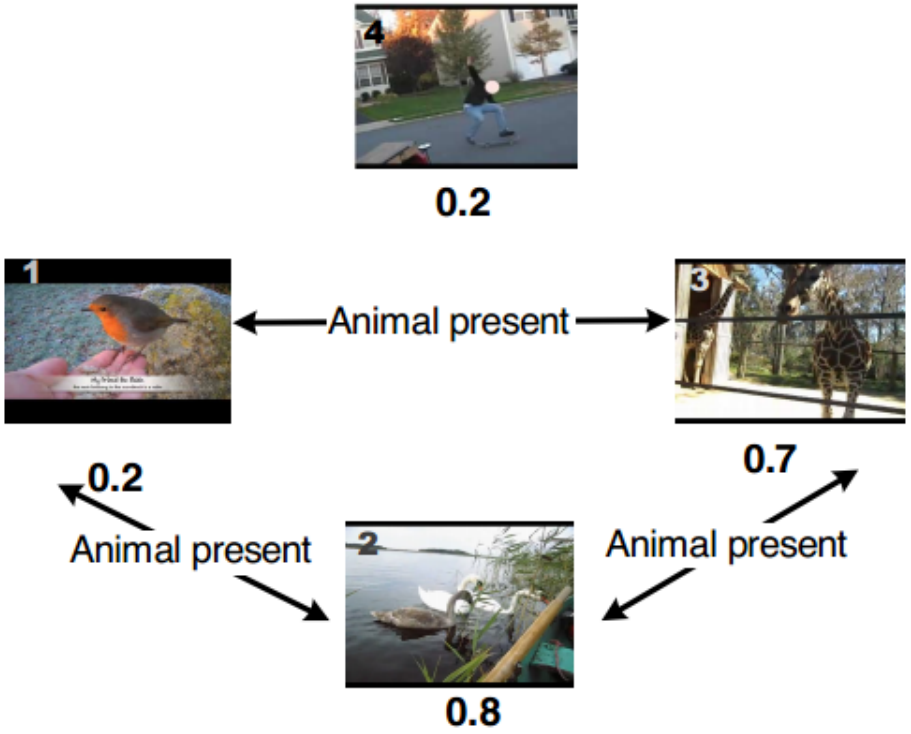
- Early fusion:
  - Concatenate the feature space and then perform classification.
  - Loses any semantic meaning during the fusion.
- Late fusion:
  - Average the classification results.
  - Preserves some semantics.

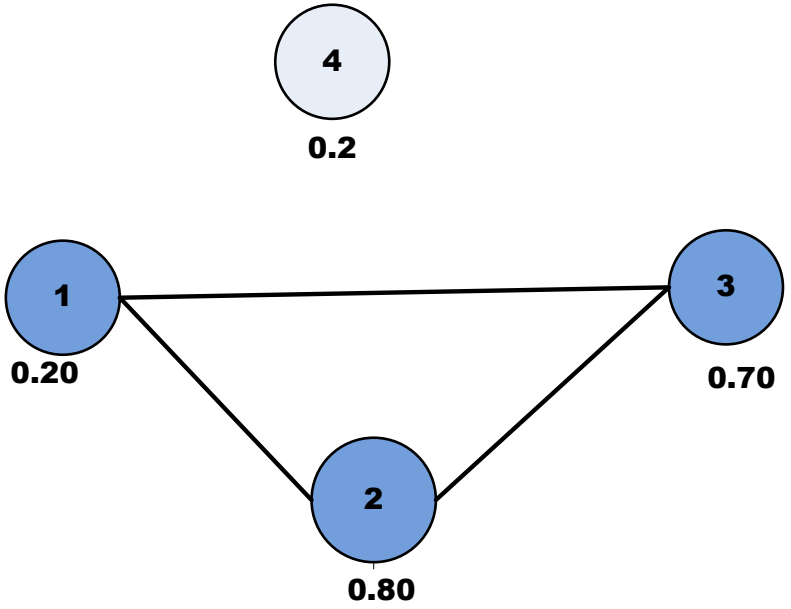
Carnegie Mellon University

Our Fusion

- Train a local classifier with low-level features to capture the general idea of a video.
- Construct a set of graphs in which two videos about the same event tends to be linked together.
- Diffuse the local classification score through the graphs to obtain the final prediction.

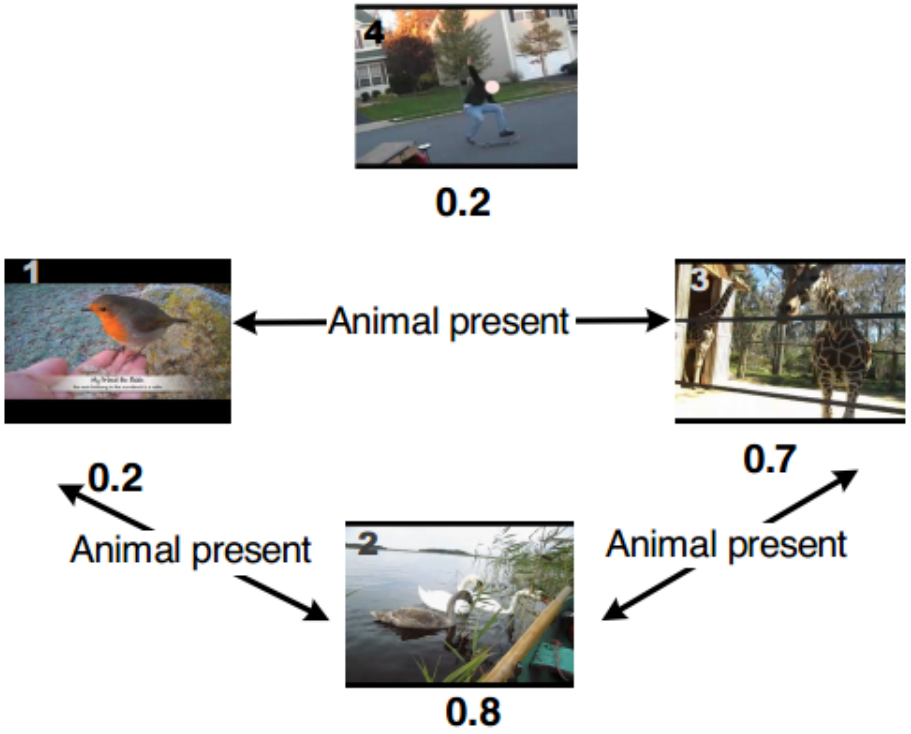
An Illustrative Example

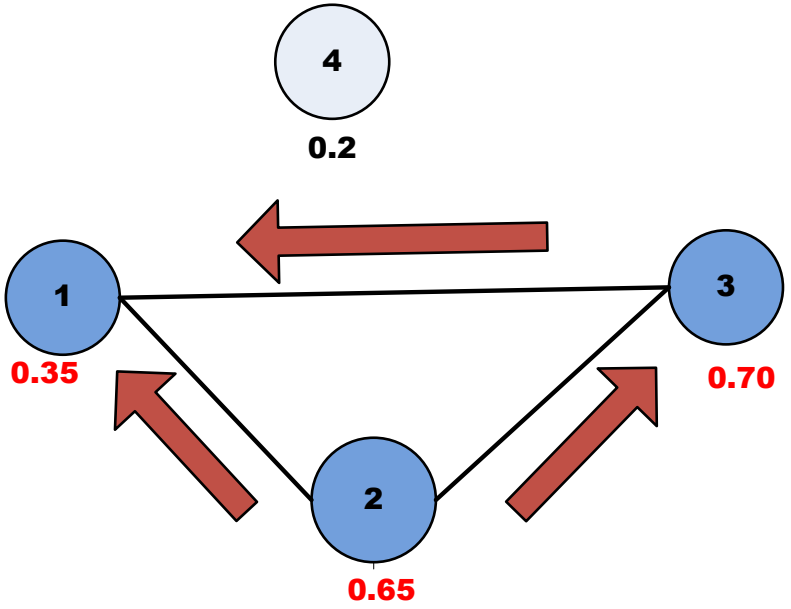




- Each video clip is a node.
- Two video clips are linked if the high-level feature concept is present in both clips.
- Each video has a local classification score (shown below each node) provided by the local classifier.

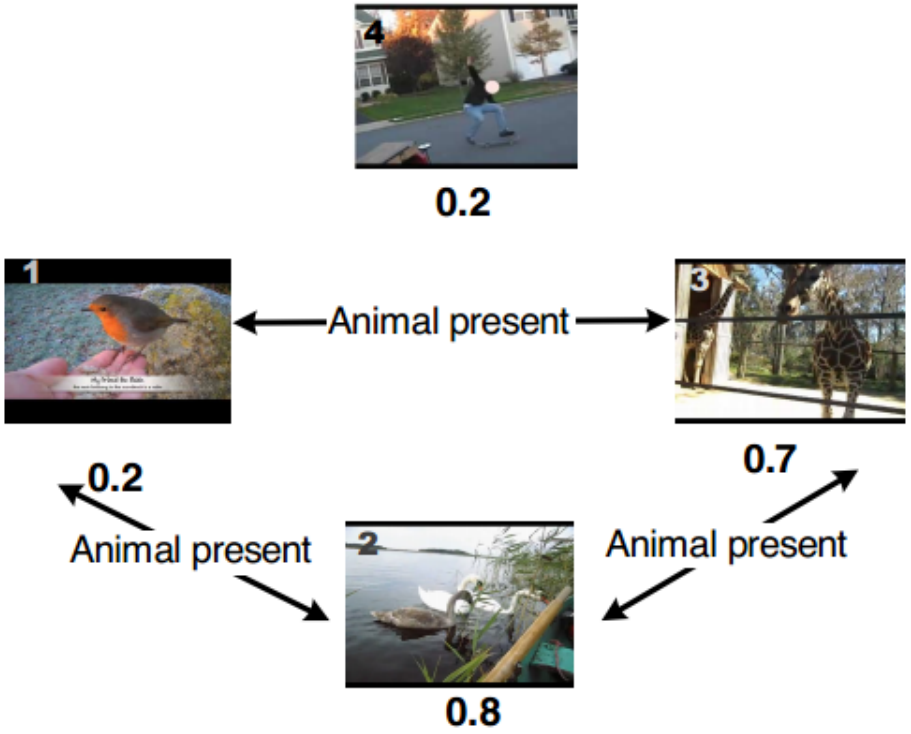
An Illustrative Example

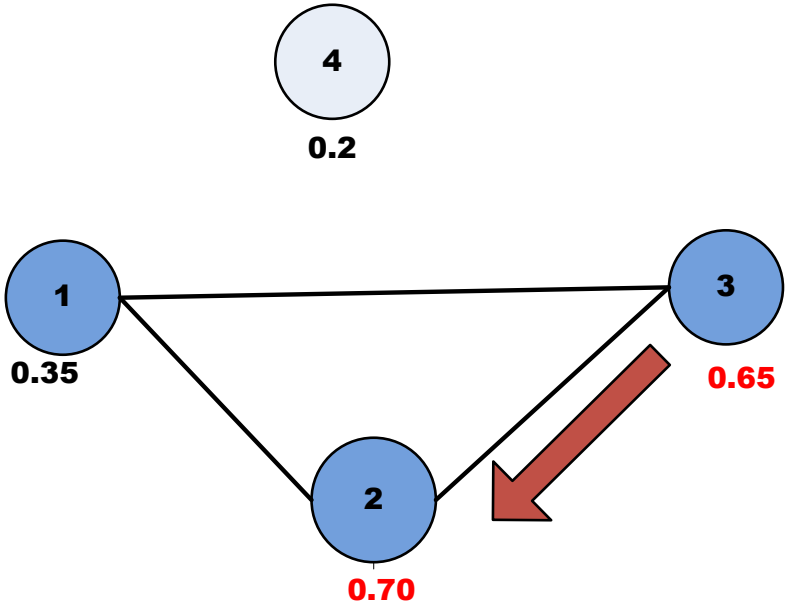




- The local classification score is diffused through the graph until it converges.

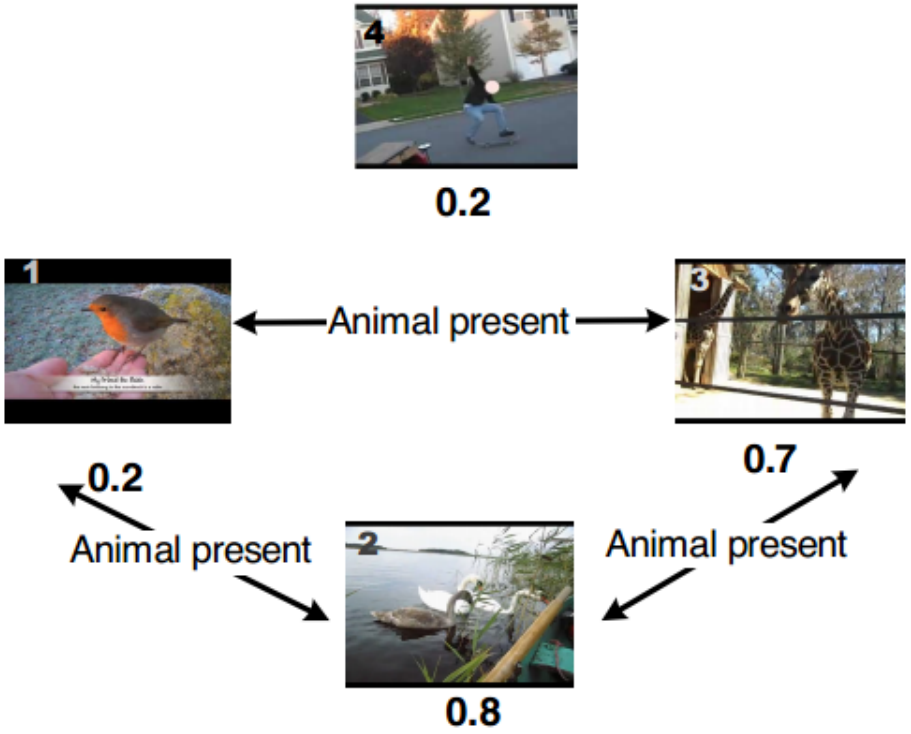
An illustrative example

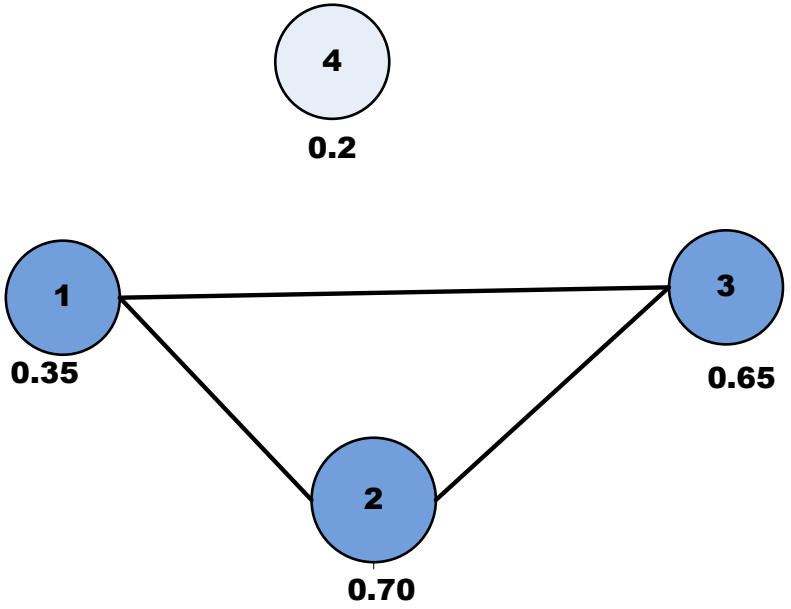




- The local classification score is diffused through the graph until it converges.

An illustrative example





- The local classification score is diffused through the graph until it converges.

Carnegie Mellon University

Outline

- Intuition
- Methods
  - **Graph Construction**
  - Collective Classification
  - Concept Selection
- Experiment Results

Carnegie Mellon University

Graph Construction

- Construct an undirected weighted graph for each high-level feature. The weight is calculated from:

$$w_{ij} = \begin{cases} |z_i - z_j| & z_i > \delta \wedge z_j > \delta \\ \text{inf} & \text{otherwise} \end{cases}$$

where z_i is the high-level concept score for the i^{th} video and δ is the threshold for the graph.

- We developed and evaluated two methods to learn the threshold: an aggressive and a cautious version.

Carnegie Mellon University

Aggressive Thresholding

- Learning the threshold by maximizing the mutual information. Suppose Z_{tr} and Y_{tr} are two random variables for the high-level feature score and the label:

$$\delta = \arg\max_{\delta} H(Y_{tr}) - H(Y_{tr}|Z_{tr};\delta)$$

- Given a dataset the first term is fixed so we have

$$\delta = \arg\min_{\delta} H(Y_{tr}|Z_{tr};\delta)$$

- The loss function is a logarithmic-like function and we call it aggressive since it penalizes incorrect edges with a logarithmic loss.

Carnegie Mellon University

Cautious Thresholding

- Learning the threshold minimizing the following loss function:

$$\delta = \arg\min_{z_i \in \mathbf{Z_{tr}}\, \delta} e^{\epsilon_{<0}(z_i;\delta)} - \epsilon_{>0}(z_i;\delta)$$

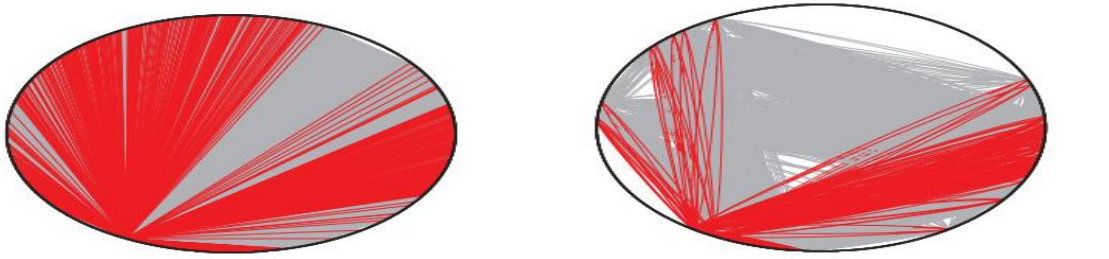
- where $\epsilon_{<0}(z_i;\delta)$ counts the number of incorrectly connected samples and imposes a significant penalty when $\epsilon_{<0}(z_i;\delta)$ becomes larger.
- The loss function is an exponential like function.

Comparison

- Aggressive threshold: higher recall and lower precision.
- Cautious threshold: higher precision but lower recall.
- Both are good In terms of selecting the concepts related to a given event:

	Parkour		Wedding Ceremony	
Rank	log-loss	exp-loss	log-loss	exp-loss
1	windows	walk/running	dresses	adult
2	walk/running	building	asian people	person
3	body parts	suburban	adult	body parts
4	road	outdoor	talking	male person
5	streets	face	adult female	standing

- Significantly different in their generated graphs:



(a) log-loss (b) exp-loss

Carnegie Mellon University

Outline

- Intuition
- Methods
  - Graph Construction
  - **Collective Classification**
  - Concept Selection
- Experiment Results

Carnegie Mellon University

Collective Classification

- A classification method in networked data where i.i.d. assumption is not expected to hold.
  - In a citation network, predict a paper's topic by the topics of the paper in its reference.
  - In a social network, predict a person's interest by those of his/her friends.
- The score of a node is determined by the weighted combination of those of its neighbors. Because of the recursive definition, the score of each vertex must be inferred simultaneously. This is called collective classification.
- Two common approaches:
  Loopy Belief Propagation and Gibbs Sampling.

Carnegie Mellon University

Gibbs Sampling

- The score of a node is updated by

$$P(y_i|\mathcal{N}_i, h_l) = \frac{1}{Z}\prod_{v_j \in \mathcal{N}_i} P(y_j|\mathcal{N}_j, h_l)^{w_{ij}}$$

where $\mathcal{N}_i$ denotes the neighbors of v_i ; $h_l(v_i)$ denotes the local classification score; w_{ij} is the edge weight.

- Gibbs Sampling:
  - Assign each node's score with its local classification score.
  - For each iteration
    - Generate a random ordering.
    - Update each node's score according to the ordering.
    - Average the score with the scores obtained in previous iterations.
  - The scores collected at the beginning are very inaccurate and thus will be discarded (burn-in period).

Markov Random Walk

- Introducing a damping factor:
  - jump to any of v_i's neighbors with probability d.
  - jump to v_i's local classification score with probability (1-d)
- Update a node's score according to:

$$P(y_i|\mathcal{N}_i, h_l) = \frac{1}{Z'}((1-d)h_l(v_i) + d\sum_{v_j \in \mathcal{N}_i} \frac{w_{ij}P(y_j|\mathcal{N}_j, h_l)}{\sum_{v_k \in \mathcal{N}_j} w_{jk}}$$

- In each iteration update a node's score according to the above function until it finally converges.
- We can prove its convergence since all component in the constructed graphs are fully connected sub-graphs.

Carnegie Mellon University

Comparison

- Random walk is expected to be much more efficient than Gibbs Sampling, since it usually converges in less than 20 iterations whereas Gibbs sampling usually takes more than a few thousands of iterations to converge.
- Given insufficient numbers of iteration (around 2000), Gibbs sampling may not converge. However, the convergence of random walk in this problem is guaranteed.

Carnegie Mellon University

Outline

- Intuition
- Methods
  - Graph Construction
  - Collective Classification
  - **Concept Selection**
- Experiment Results

Carnegie Mellon University

Concept Selection

- For an event not all high-level features are helpful. Therefore we build a forward-wrapper to select the helpful ones for each event.
- The idea is to apply a greedy search:
  - First select the best concept and put it into the subset.
  - Select the next best concept that works best with the concepts already in the subset.
- The final prediction of a sample is a linear interpolation of the collective classification score and the local classification score.

Carnegie Mellon University

Outline

- Intuition
- Methods
  - Graph Construction
  - Collective Classification
  - Concept Selection
- **Experiment Results**

Carnegie Mellon University

Setup

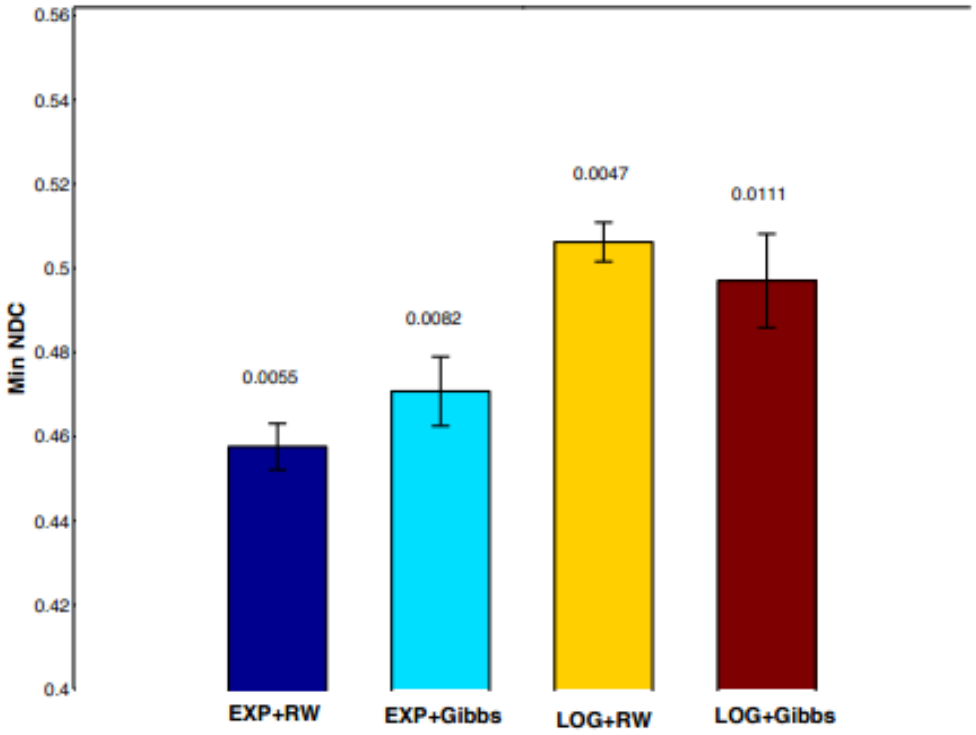
- Development set in TRECVID 2011 evaluation which consists of 15 predefined events and 2049 video clips.
- Low-level feature
  - Color SIFT (CSIFT).
- High-level feature
  - 346 visual concepts in TRECVID 2011 Semantic Indexing contest.
- Local Classifier
  - SVM with Chi-square kernel
  - Multi-class classification is achieved using one-versus-all SVM.
- Evaluation Criteria
  - Averaged Minimum NDC (Normalized Detection Cost) on 5 fold cross-validation sets.

Comparison with Baseline Method

ID	LL	HL	EF-FC	EF-KF	LF	CCA	FFCC
1	0.576	0.490	0.492	0.508	0.497	0.633	**0.454**
2	0.872	0.786	0.810	0.828	0.792	0.875	**0.743**
3	0.588	0.526	0.476	0.498	0.502	0.629	**0.408**
4	0.430	0.360	0.353	0.331	0.339	0.422	**0.250**
5	0.722	0.719	0.682	0.658	**0.652**	0.803	0.688
6	0.673	0.553	0.575	0.568	0.539	0.562	**0.481**
7	0.790	0.700	0.635	0.657	0.683	0.749	**0.542**
8	0.455	0.396	0.378	**0.343**	0.391	0.457	0.355
9	0.547	0.389	0.376	0.407	0.415	0.562	**0.347**
10	0.807	0.701	0.680	0.707	0.675	0.562	**0.611**
11	0.682	0.787	0.772	0.746	0.697	0.803	**0.566**
12	0.624	0.488	0.483	0.524	**0.472**	0.655	0.493
13	0.598	0.450	0.438	0.443	0.458	0.542	**0.429**
14	0.544	0.556	0.459	0.443	0.495	0.544	**0.408**
15	0.724	0.653	0.646	0.683	0.690	0.721	**0.543**
AVG	0.642	0.570	0.550	0.556	0.553	0.647	**0.488**

- LL for low-level features,
- HL high-level features,
- EF-FC early fusion by feature concatenation,
- EF-KF early fusion by kernel fusion,
- LF late fusion
- CCA: CCA feature space projection+ LF SVM
- FFCC for the proposed method.

Comparison of Proposed Methods



We repeat each experiments 5 times with different cross validation partitions.

RW: Random Walk

Gibbs: Gibbs Sampling,

EXP: Exponential loss function

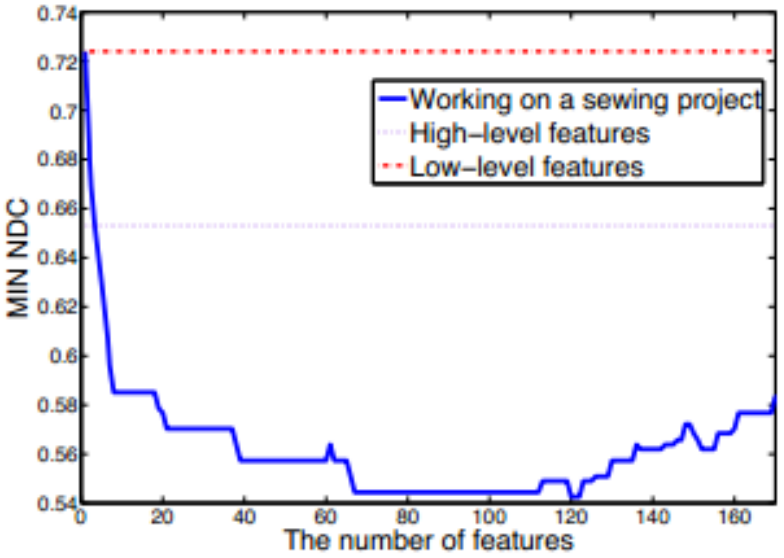
LOG: Logarithmic function

Observations:

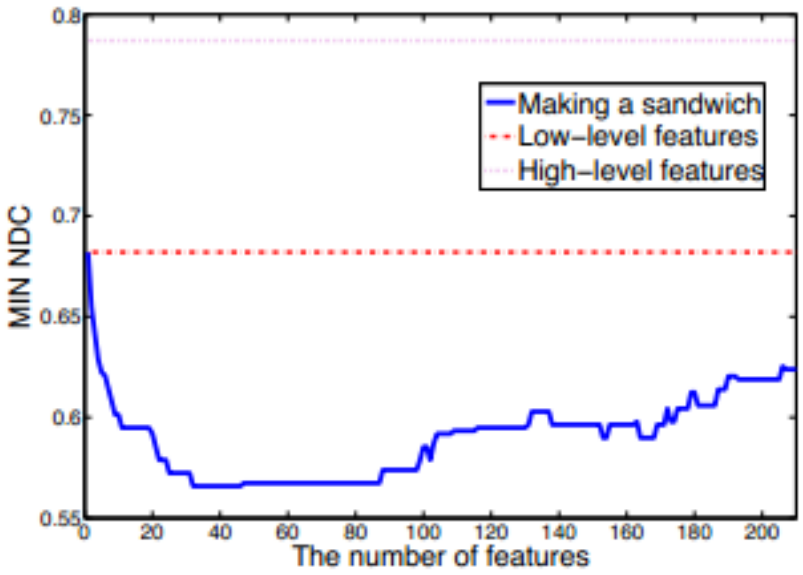
1. Random walk with exponential loss function yields the best result.
2. The variance of random walk is smaller than that of Gibbs Sampling.
3. The EXP function seems to be better than LOG function.

How many concepts to incorporate?

- Using our forward wrapper, we greedily select the best subset of the concepts.



(a) working on a sewing project



(b) Making a sandwich

- the size of the optimal subset for all events is less than 125, and the number varies in different events.
- The mean is 68.7 and standard deviation is 30.4.

Interpreting the Result for "Wedding Ceremony"

CSIFT: 0.359.

-0.076, -0.062 because of the presence of "kitchen" and "anchorperson" ;

-0.089 because of other 27 concepts like: "room" and "3 or more people" ;

Final score after change : 0.132.

CSIFT: 0.410.

+0.094, +0.049, +0.034 because of the presence of "adult" , "legs" and "child" ;

+0.034 because of the absence of the concept "bird" ;

+0.215 because of other 25 concepts;

Final score after change : 0.836.

THANK YOU.

QUESTIONS?