Multimedia Event Detection using Visual Features

Dishan Gupta, Rahul Goutam and Amos Ng

Language Technologies Institute, Carnegie Mellon University

INTRODUCTION

What is MED

- Activity-centered happening involving people and process-driven action
- Detect occurrence of event within a video clip(NIST)

Using unsupervised feature learning

- More generalizable to different domains
- Less time-consuming
- Less expensive

Independent Subspace Analysis to learn features [4]

- Problem : High dimensionality of data
- Solution : Combine with convolution, stacking [1]

SVM for event classification

- Exponential χ^2 kernel

Dataset : TRECVID 2011 Multimedia Event Detection task

MOTIVATION

Why MED

- Explosion of multimedia content on the internet
- Video uploaded on Youtube at the rate of 30 million hrs / year
- Automatic indexing and retrieval

Challenges:

- Single event has multiple high-level concepts
- E.g. birthday party consists of combination of concepts *cake, people, cheering(audio conecpt)*
- Attempting a board trick includes skating, skiing, surfing, etc.
- Concepts shared between multiple events

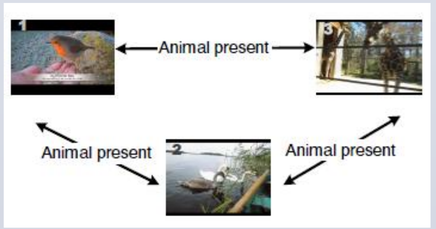

Figure 1: An illustrative graph showing the high-level feature "animal" present in three different videos. Learning such features is essential to the success of any event detector. [3]

Hierarchical Learning:

- Detecting events based on high-level concepts
- Similar to what the human brain does.
- E.g., a video containing concepts such as "man", "animal", "grass" and "food" is probably about "feeding an animal".

DATASET OVERVIEW

Figure 2: An overview of the TRECVID 2011 MED dataset[3]

Our dataset consists of 3622 video clips from the NIST TRECVID 2011 MED task dataset spanning 15 distinct events and each clip belonging to only one of them. The events are shown in Figure 2.

UNSUPERVISED FEATURE EXTRACTION [1]

ISA is a 3-layer neural network [4], with square and square-root non-linearities in the second and third layers, respectively.

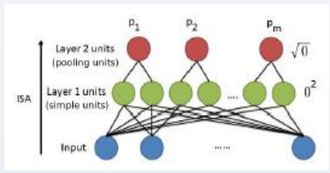

Figure 3: A standard ISA network. The pooling units take input from two simple units to represent the subspace structure of 2.[1]

Objective Function: In detail, given an input pattern x^t ,the output layer activation is given by:

$$p_i\left(x^t;W,V\right)=\sqrt{\sum_{l=1}^{k}V_{il}\left(\sum_{j=1}^{n}W_{ij}x_j^t\right)^2} \qquad (1)$$

The weights W are learned, whereas V are fixed in order to account for the subspace structure of neurons in the second layer. Thus, we are presented with the following optimization problem:

$$minimize_W \sum_{t=1}^{T}\sum_{i=1}^{m}p_i\left(x^t;W,V\right) \qquad (2)$$

$$s.t.\ WW^T=I \qquad (3)$$

The objective function (eq. 2) is convex with V fixed.

The orthonormality constraint ensures feature independence or diversity.

Large-Scale Deep Learning: We use the approach by Le et al. [1],

- ISA infeasible on high-resolution video data
- Use convolution . Apply ISA to a small input block and convolve it with the rest of the video
- For hierarchical high-level feature learning, we use stacking and construct a network of ISA layers

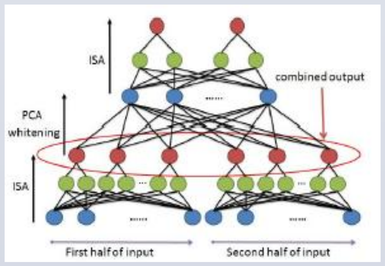

Figure 4: A stack convolutional ISA network. The convolution output of first layer ISA is input to the next layer.[1]

Model Details:

- ISA layer 1 – Block size 16x16 (spatial) x10 (temporal), 300 features
- ISA layer 2 – Block size 20x20 (spatial) x14 (temporal), 200 features
- Regularization – lambda set to 0
- Training(Unsupervised) – 50% data per event. Learn W for ISA L1 and ISA L2.
- Optimization Iterations – 2000 per ISA layer.

OPTIMIZATION

The objective in eq. (2) is convex, but normal batch gradient descent would not be able to satisfy its constraint eq. (3). Thus, during each iteration we are required to update W by projecting it into the constraint set:

$$W\rightarrow\left(WW^T\right)^{-\frac{1}{2}}W \qquad (4)$$

- Computing the inverse of the square root of a matrix takes cubic time
- convolution for scalability.

CLASSIFICATION

- Features extracted using ISA algorithm used for classification
- SVM with an exponential χ^2 kernel.

$$k\left(x,y\right)=\exp\left(-\sum_i\left(\frac{\left(x_i-y_i\right)^2}{x_i+y_i}\right)\right) \qquad (5)$$

- Training – 80% data per event.
- Testing – 20% of remaining data.

RESULTS

Event	Accuracy	Average Precision
Attempting a Board Trick	77.40%	0.32
Feeding an Animal	77.40%	0.33
Landing a Fish	82.88%	0.20
Wedding Ceremony	82.20%	0.28
Working on a Woodworking	80.13%	0.29

Table 1: Accuracy and average precision results on five events.

Mean Accuracy – 80%

Mean Averaged Precision – 0.28

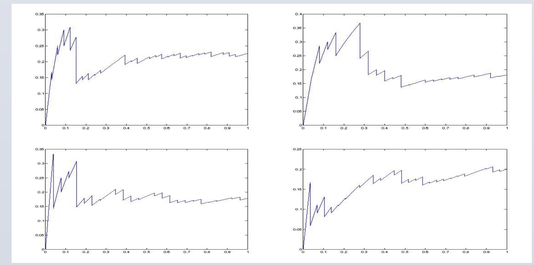

Figure 5: Precision (y-axis) vs Recall (x-axis) plots for four events.

References

[1] Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 3361–3368. IEEE, 2011.

[2] Ma, Zhigang, et al. "Knowledge adaptation for ad hoc multimedia event detection with few exemplars." *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012.

[3] Jiang, Lu, Alexander G. Hauptmann, and Guang Xiang. "Leveraging high-level and low-level features for multimedia event detection." *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012.

[4] A. Hyvarinen, J. Hurri, and P. Hoyer. *Natural Image Statistics. Springer, 2009 . pages 3361, 3362, 3363, 3364*