Recognizing Human Actions in Videos

Jaehyun Park / 2010 CURIS Project

Significance of the Problem

- A big challenge
  - Harder than dealing with still images because of the temporal relationship between frames
- Many real-world applications
  - Analyzing sport replays
  - Security / surveillance
  - Video search

Neural Network and "Deep Learning"

- Neural Network
  - A DAG with weighted edges and an "activation" associated with neurons.
  - Weights can be "learned" by solving an optimization problem.
- Deep Learning
  - A neural network can learn more complex concepts with more layers.

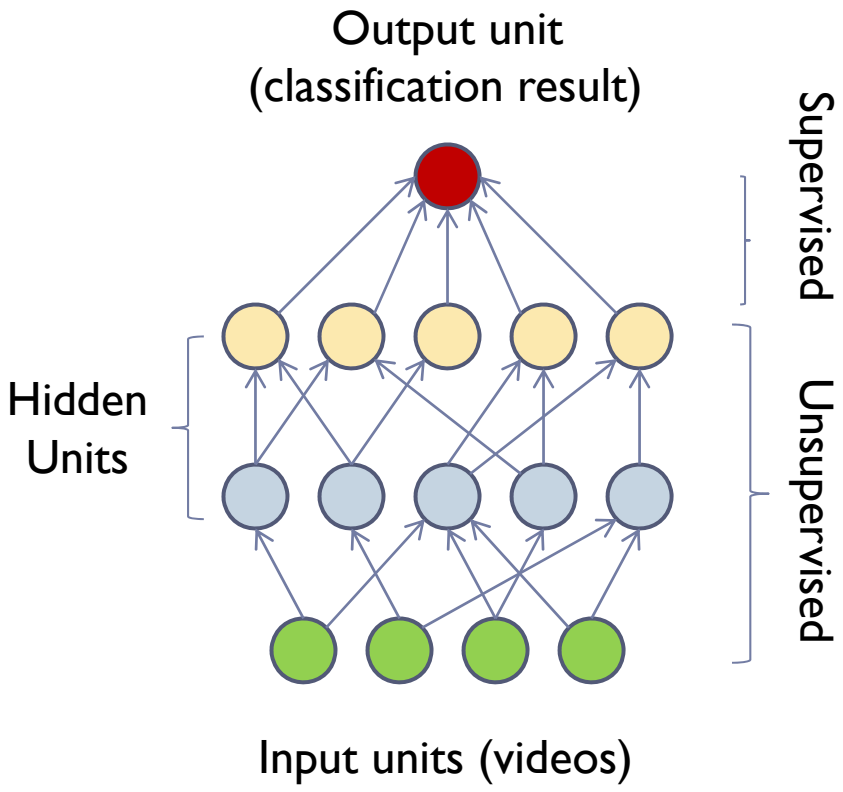


Fig 1. A neural network for classifying videos

Overview of the Method

- Training Phase
  - 1. Sample random video patches from the training data set.
  - 2. Pretrain the unsupervised layers using ISA (Independent Subspace Analysis).
  - 3. With video labels, learn the supervised layer using K-means clustering and SVM (Support Vector Machine).

- Testing Phase
  - 1. Pass each video clip through the unsupervised layers.
  - 2. Use the supervised layer to predict the action associated with the video.

ISA (Independent Subspace Analysis)

- Generalization of ICA (Independent Component Analysis)
- Learns "robust" features from the training data that can be used as indicators of certain actions.
- Maximizes the "sparseness" of the hidden activations
  - Optimization problem: Minimize $J = \sqrt{V(WX)^2}$
    (X: data, W: layer 1 weights, V: layer 2 weights)
  - The square root and square operations are component-wise.
  - Use gradient descent to learn W:

$$\frac{\partial}{\partial W} J = -((WX) \bullet F) X$$

$$F = -\frac{1}{J} V^T$$

Implementation of ISA

- Algorithm
  - Input: set of videos X
  - Output: weights of W
  - while not converged
    - Compute gradient
    - Update W
    - if change in W was small
      - Mark converged

- Usually the weights converge within 600 iterations.

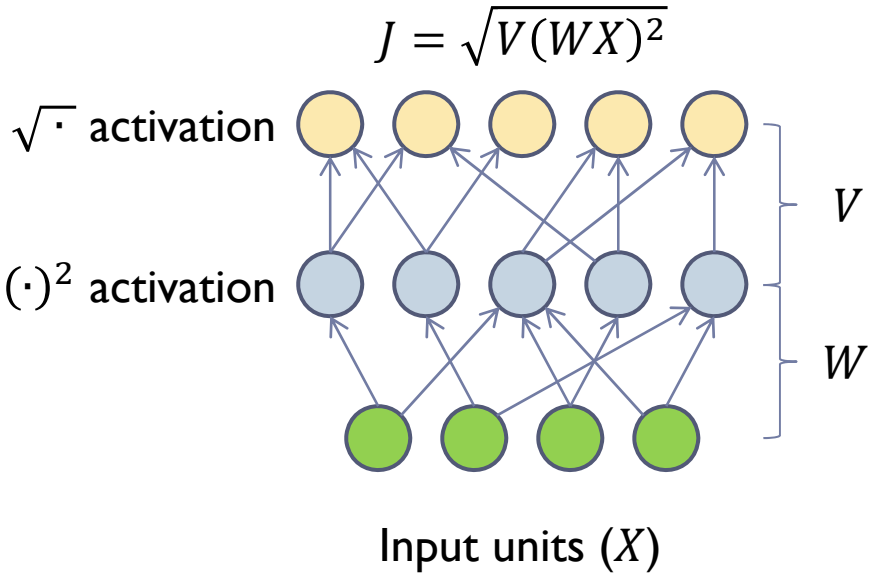
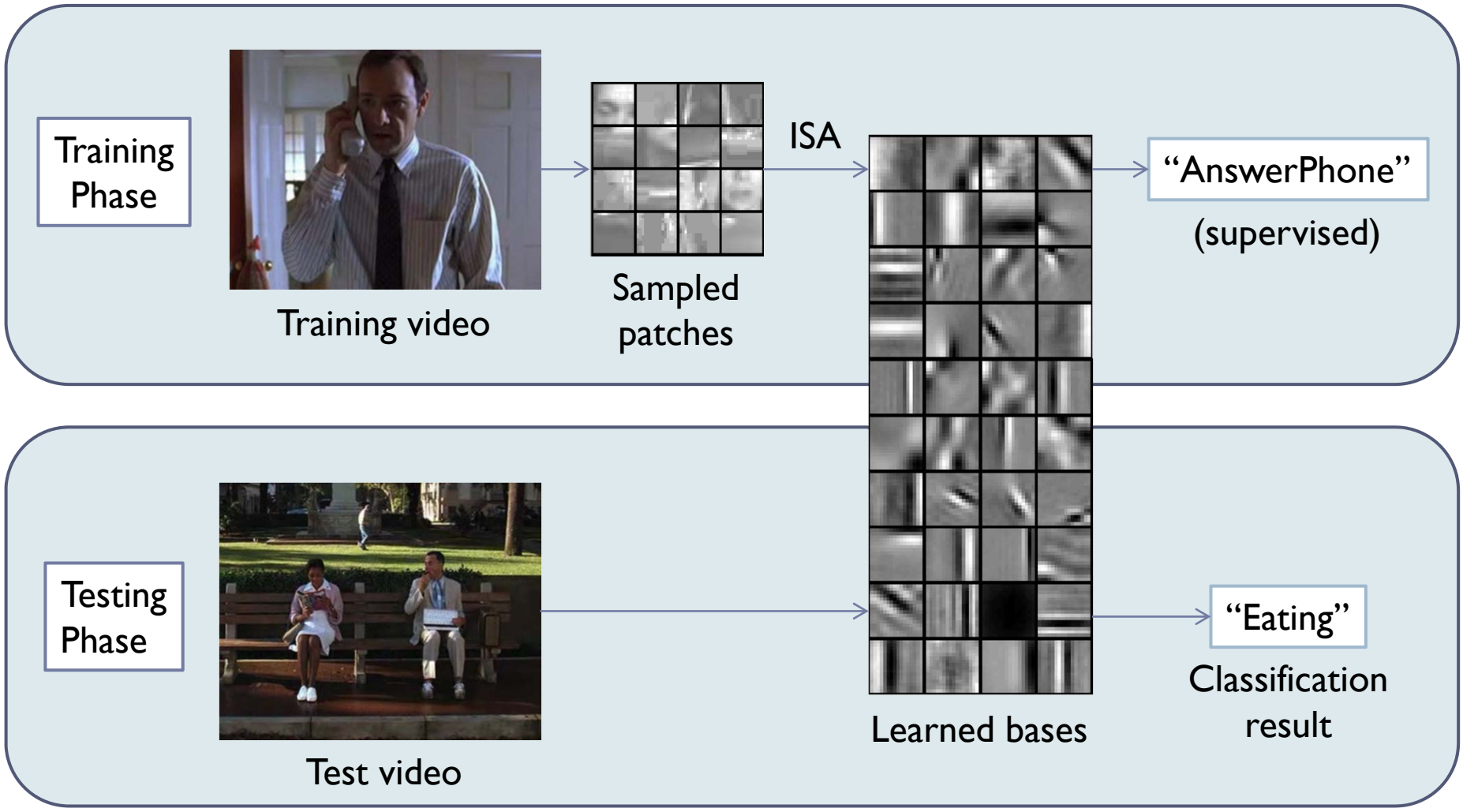


Fig 2. Visualization of the unsupervised layer

Flow Diagram



Results

Method	Accuracy
Our method	**50%**
Harris3D + HOG/HOF [Laptev et al 2003, 2004]	45%
Cuboids + HOG/HOF [Dollar et al 2005, Laptev 2004]	46%
Hessian + HOG/HOF [Laptev 2004, Williems et al 2008]	46%
Dense + HOG / HOF [Laptev 2004]	47%

Table 1. Performance on Hollywood2 dataset

Method	Accuracy
Our method	**87%**
Cuboids + HOG3D [Klaser 2008, Dollar et al 2005]	83%
Hessian + HOG/HOF [Laptev 2004, Williems 2008]	79%
Dense + HOG3D [Klaser et al 2008]	86%
Dense + HOF [Laptev 2004]	83%

Table 2. Performance on UCF dataset

Discussion

- Nature of videos
  - ISA learns basic elements of videos: still / moving edges, changes in color, etc.
  - Temporal information is incorporated in the bases.

- Deep Learning
  - Stacking layers helps classifying videos.
  - Complex features cannot be learned by simply increasing the number of neurons; the structure of the network is also a crucial factor.

Future Work

- Getting a good result on more challenging datasets
  - Practical problems arise.
    e.g. videos are too big to load into the memory.
  - Hollywood2: 20-hour videos from various movies
  - TRECVID: 99-hour surveillance videos recorded in London Gatwick Airport

- Stacking more layers
  - There are very few papers describing an architecture with more than four layers.