Multimedia Event Retrieval

Issues in Analyzing Large Scale User Generated Content

May 21, 2014

Karl Ni

Lawrence Livermore National Laboratory

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

LLNL-PRES-6554623

User Generated Content (UGC) fastest growing open source data

- In 1 month, more video content is uploaded to YouTube than all the US media companies have produced in 60 years.
- YouTube videos are uploaded 100 hours/minute.

- Organization, retrieval, and analysis requires automation and aid from machine learning.

Application: Event Detection

E001 Attempting a board trick

E002 Feeding an animal

E003 Landing a fish

E004 Wedding ceremony

…

E006 Birthday party

E007 Changing a vehicle tire

E008 Flash mob gathering

E009 Getting a vehicle unstuck

E010 Grooming an animal

E011 Making a sandwich

…

E015 Working on a sewing project

…

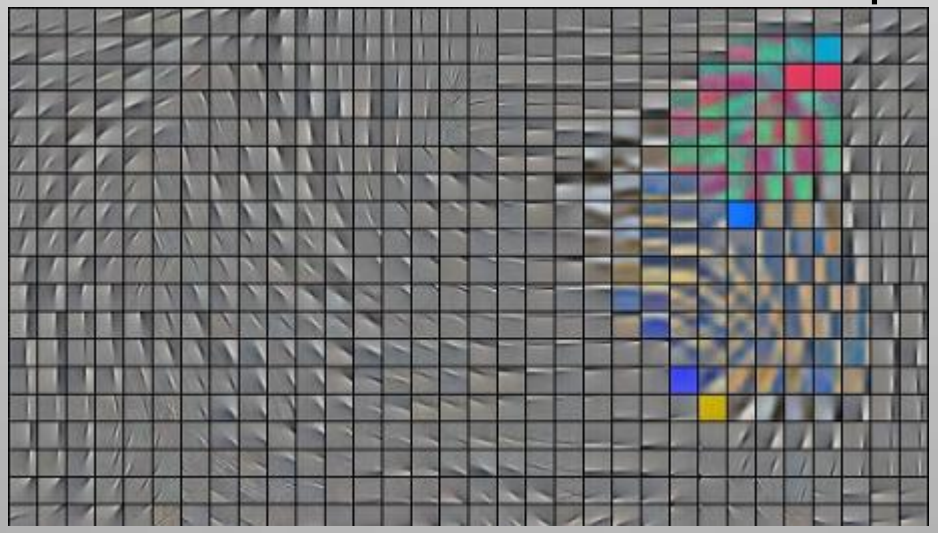
E025 Marriage proposal

E029 Winning a race without a vehicle

E030 Working on a metal crafts project

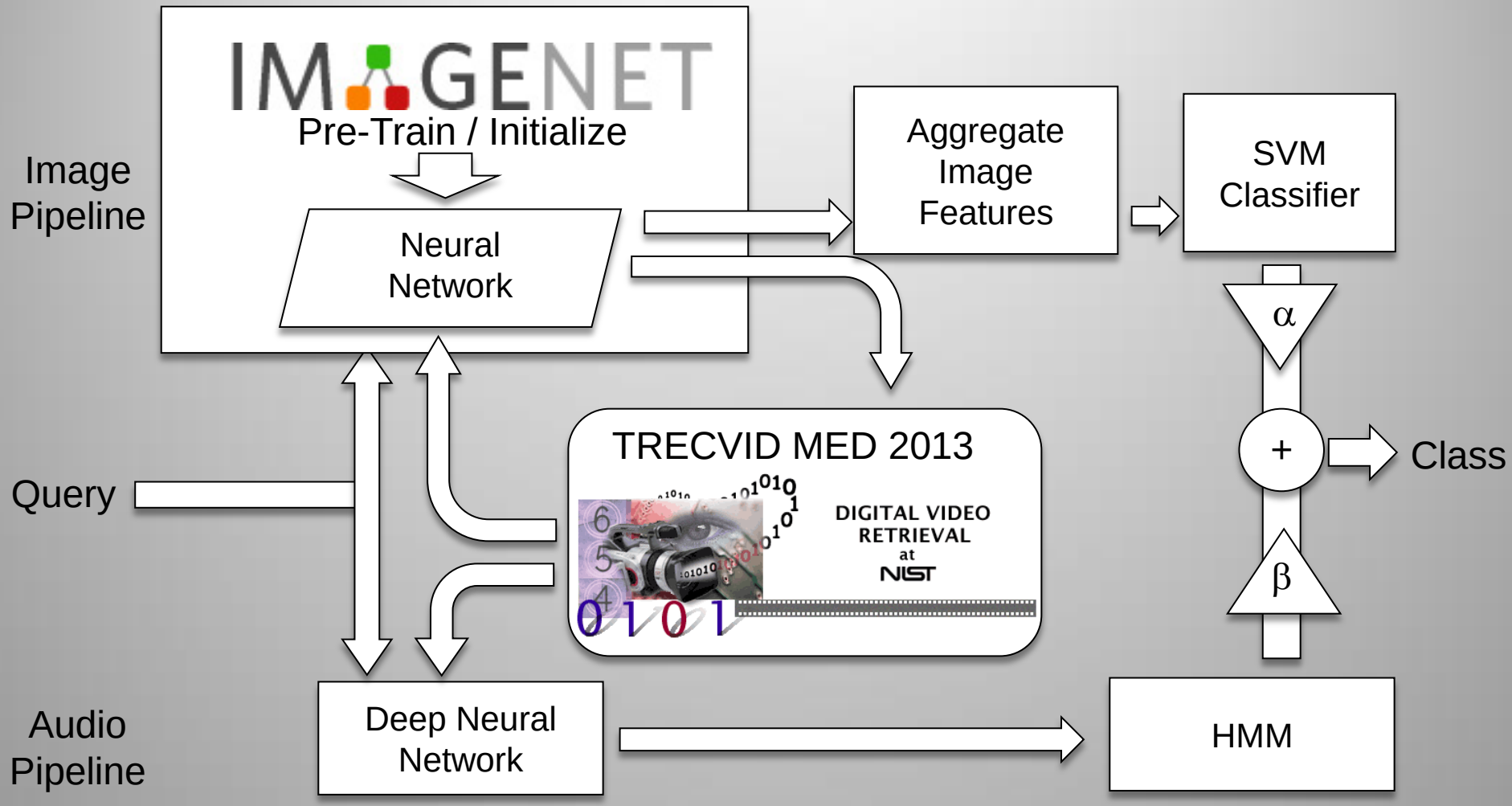
Background: Deep Learning

- To generalize, most ML algorithms add complexity by:
  - Adding more parameters (e.g., manifold learning, GMM's, probabilistic graphical models, etc.)
  - Putting more nonlinear functionality into a linear learner (e.g., SVM's, manifold learning, kPCA, k-Density estimates, etc.)
  - Recently, nonparametrics (e.g., infinite Bayesian models, etc.)
- Mid-2000's movement to do all of that but repeatedly



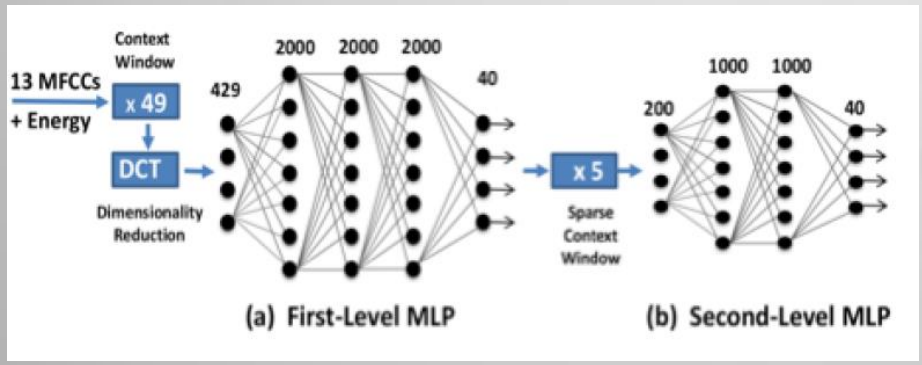
Taken from Ng. et al, '13

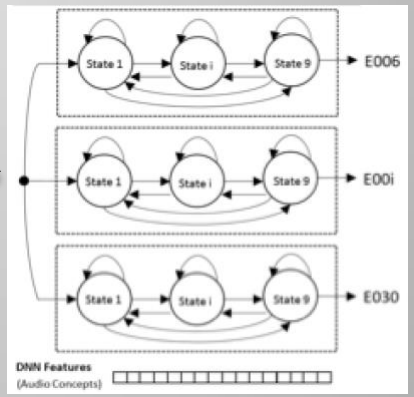
System Description



Audio Classifiers with H-DNN

Audio Deep Neural Network → HMM





- Training based on annotated datasets developed by CMU, SRI, Stanford
- Input MFCC Features + Energy Coefficients
- Two-stage for short-term and long-term predictions
- Temporal consistency improved with HMM models

System Description

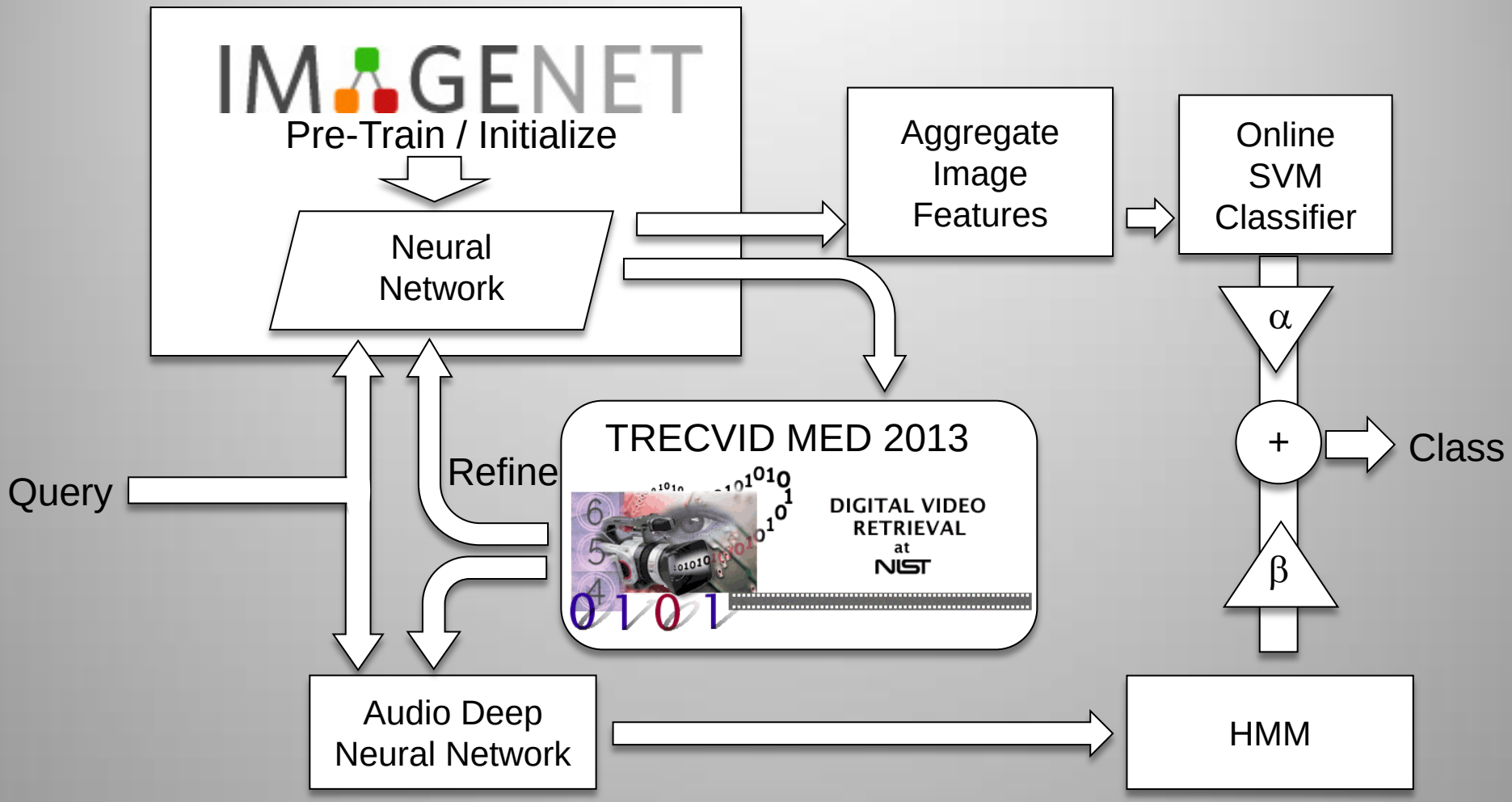
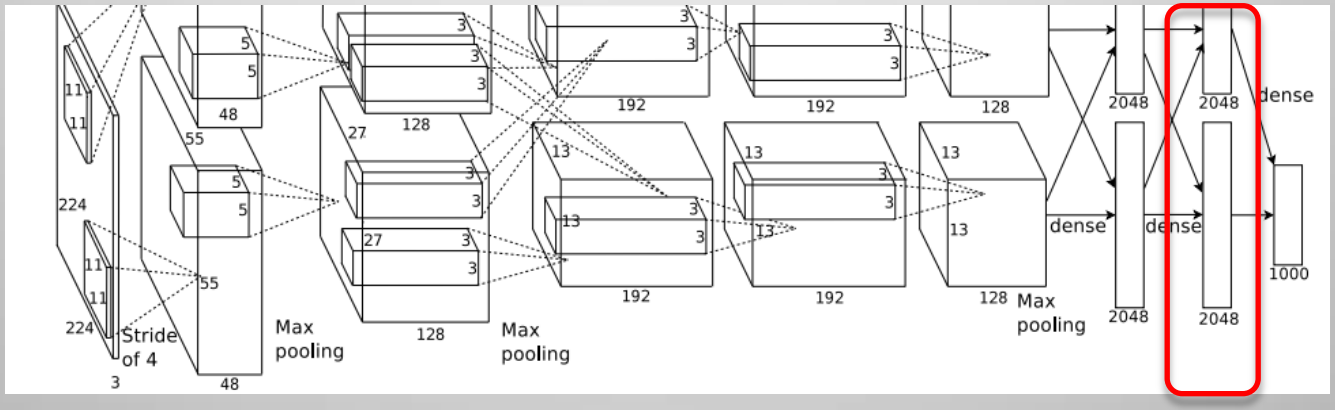


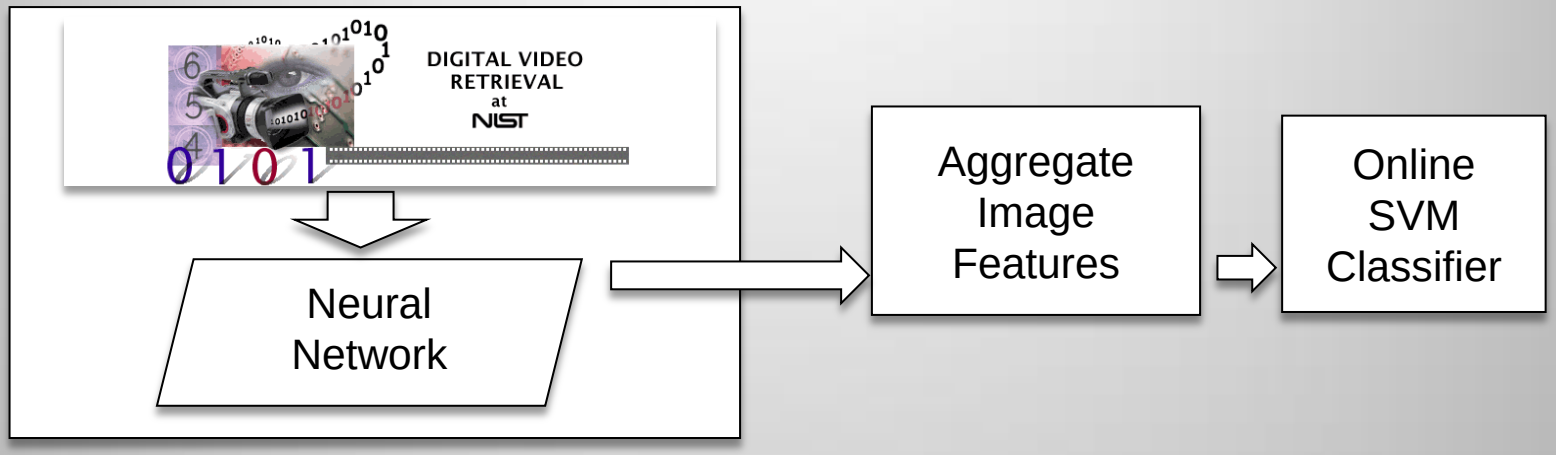
Image Classifiers with D-NN

- ImageNET 2012: Object detection on 1,000 classes
  - Competition winner: Alex Krizhevsky
  - Deep Neural Network Architecture.



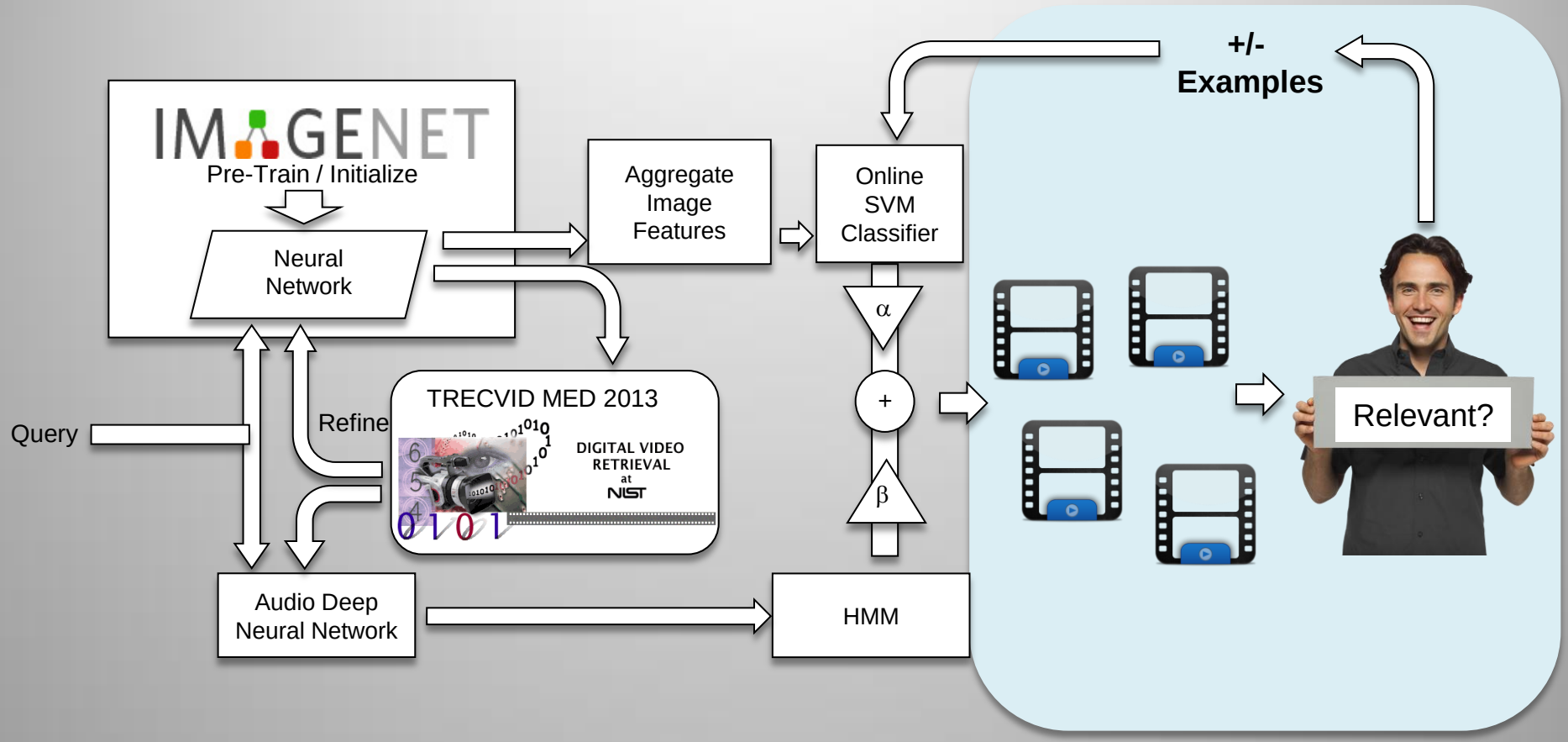
- Image Network will get you close, but…
  - Specifically targeted at nouns and objects
  - Discrete concepts produce a largely noisy result
  - Fewer output options means limited discrimination
  - Effectively make a “hard” decision before the real result

From ImageNET to TRECVid



- Newly tuned TRECVid *image* classifier: apply 2, use 1 layer:
  - L8: Reconstruction ICA Objective Layer: 4096 neurons
  - (Tune L9): Softmax regression layer for multiple instance learning
  - Sigmoid non-linearity (in addition to ReLU) onto L8
- Sum over all frames (L8) to obtain histogram-like features
- Use an SVM Classifier to make the final decision

Classification & Relevance Feedback



Quantitative Comparisons

- Corpora
  - TRECVID-2013 MED Dataset, ~150k Videos from YouTube, Some Hollywood 2 Datasets
- Image Feature-Based Comparisons
  - 8% Improvement Over SIFT-BoW Based Approach [TRECVID MED 13, Competition Results]
- HMM-HDNN Approach
  - 0.68% average precision improvement over MFCC-Based Approach [Elizalde et al]
- Multi-modal Approach
  - 4.23% Improvement over SIFT + MFCC-Based Approach
  - -25.3% based on comparative Amazon Turk classification

The Video LDRD Research Team

- Program Manager
  - Doug Poland
- LLNL
  - Carmen Carrano
  - Karl Ni
  - David Buttler
- ICSI Berkeley
  - Gerald Friedland
  - Benjamin Elizalde
  - Xiao-Yong Wei
  - Damian Borth
- Collaborators
  - Livermore Brain Program
  - Livermore Computing
  - Stanford
    - Adam Coates
    - Brody Huval

Questions?