2016 TRECVID
Multimedia Event Detection Report
Team INF

Junwei Liang, Poyao Huang, Lu Jiang, Zhenzhong Lan, Jia Chen and Alexander Hauptmann

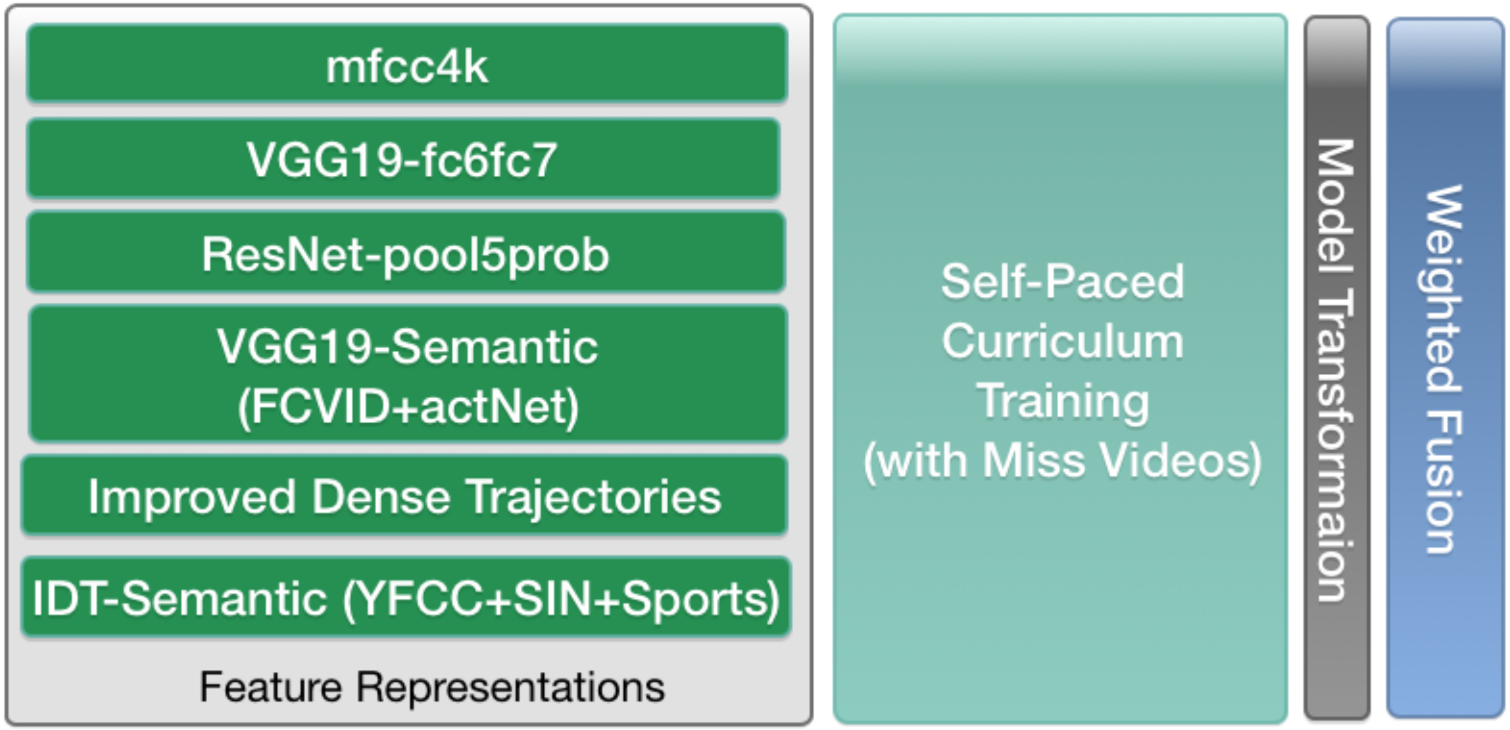
Outline

- System Overview – (10Ex, 100Ex)
  – Feature Representations
- Selected Topics
  – Learning with Miss Videos
- Final Results (MED16EvalSub)
- 0Ex System
- Conclusions

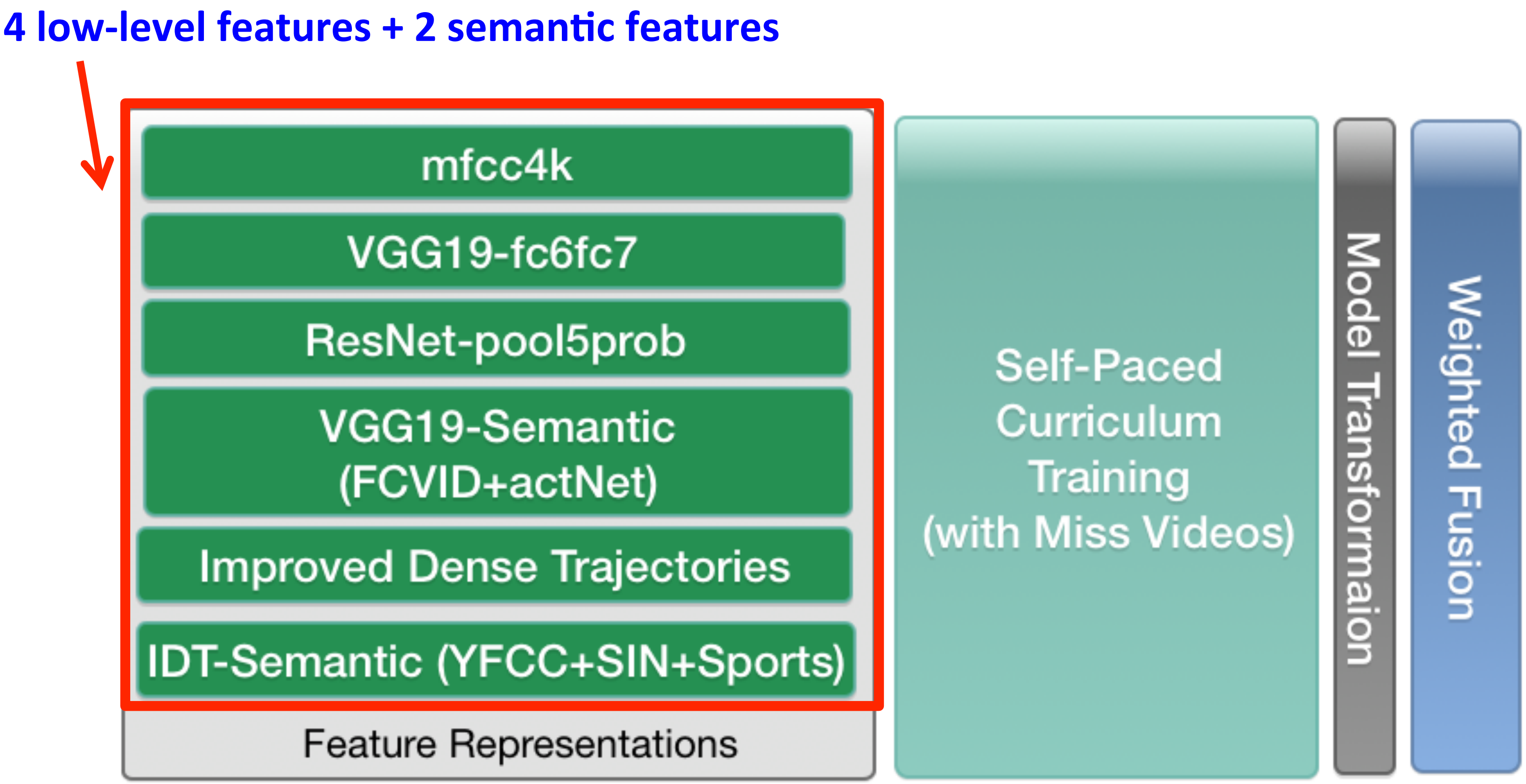
Outline

- System Overview – (10Ex, 100Ex)
  - Feature Representations
- Selected Topics
  - Learning with Miss Videos
- Final Results (MED16EvalSub)
- 0Ex System
- Conclusions

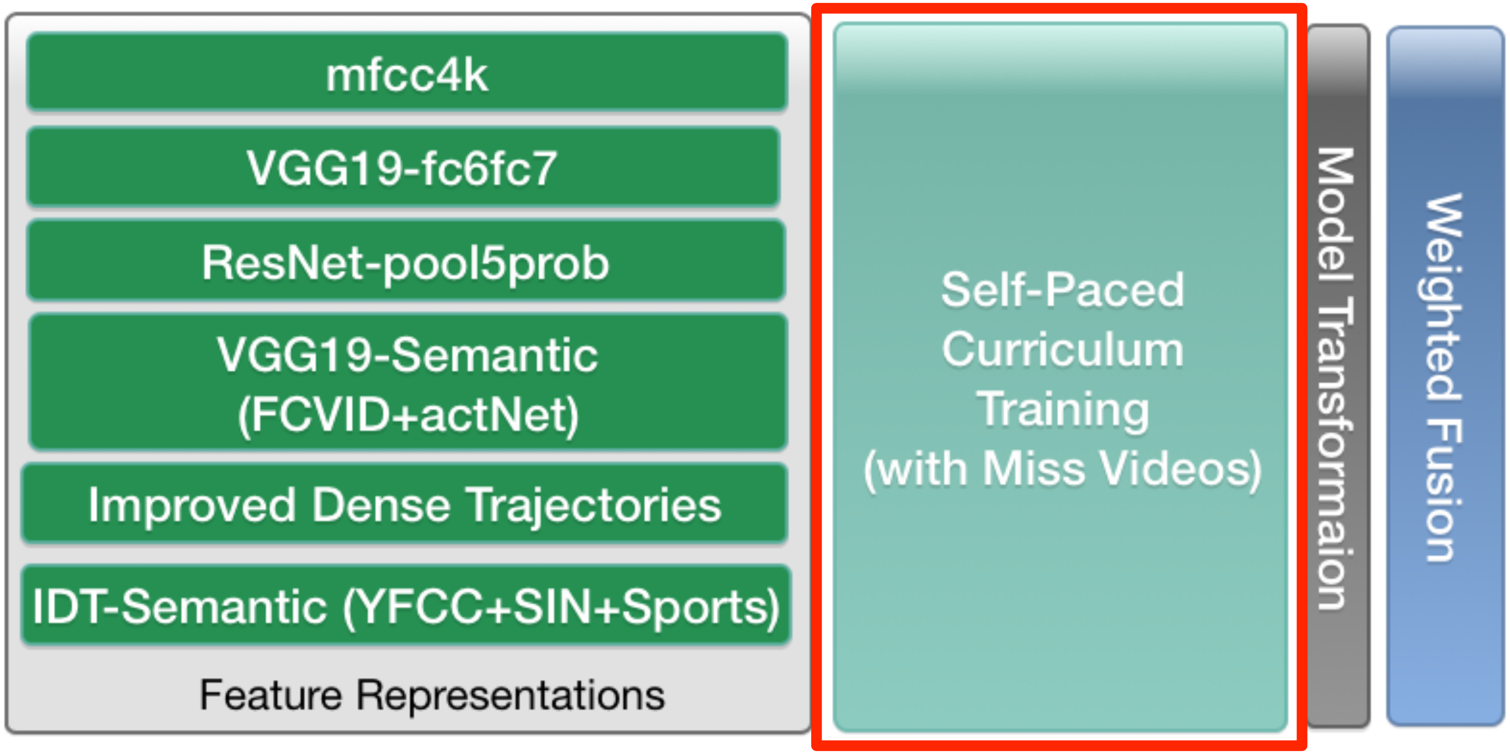
MED-System (10Ex,100Ex)



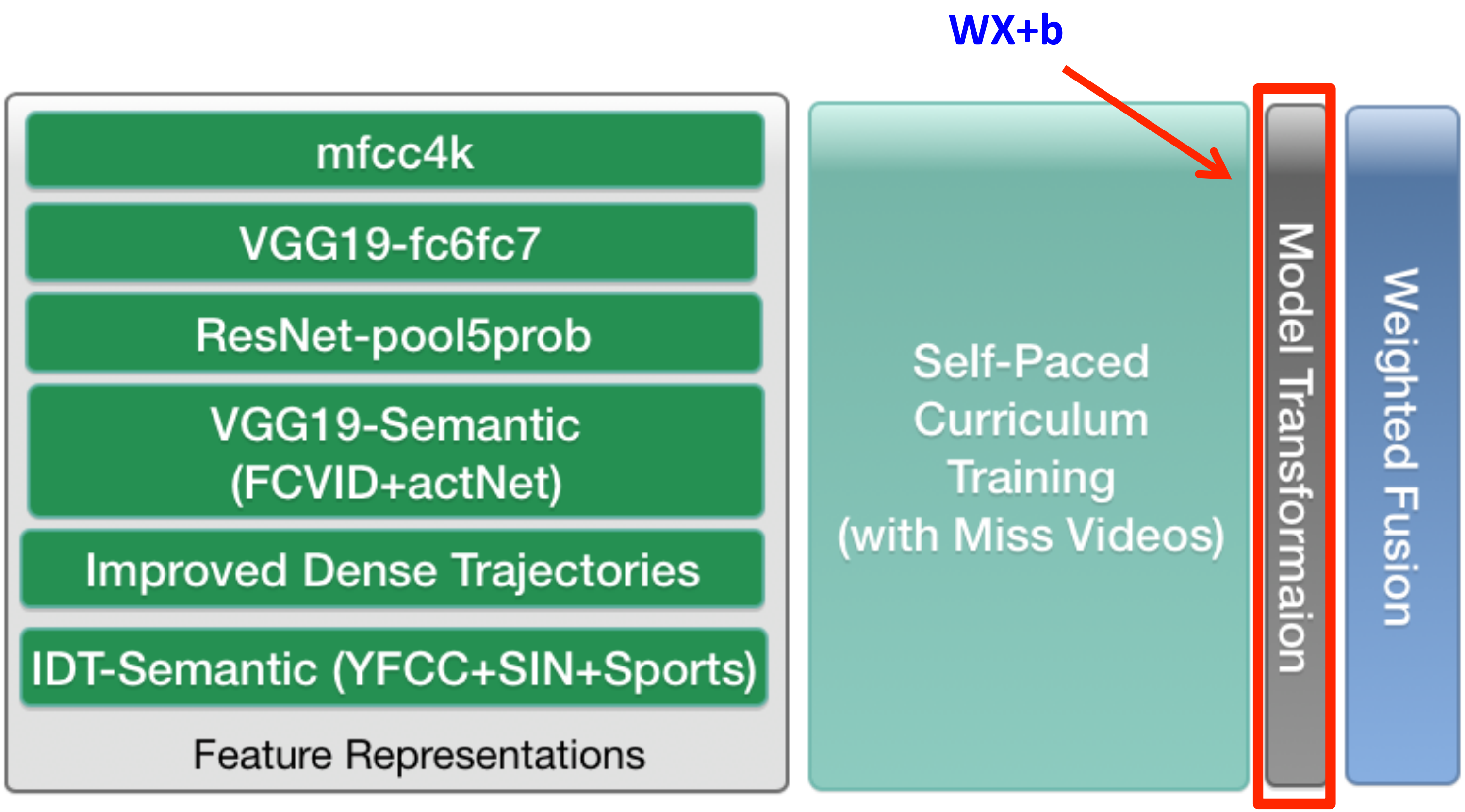
MED-System (10Ex,100Ex)



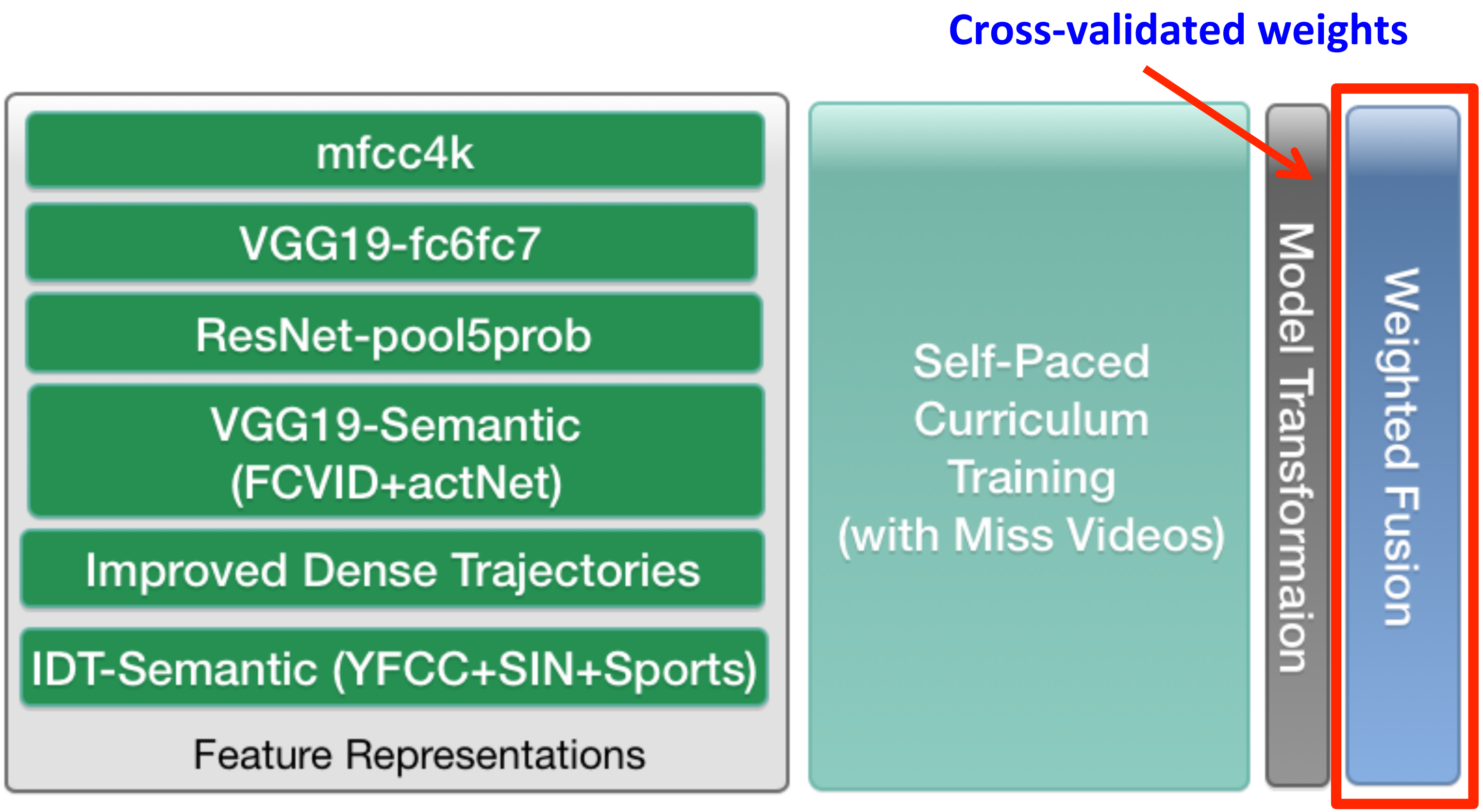
MED-System (10Ex,100Ex)



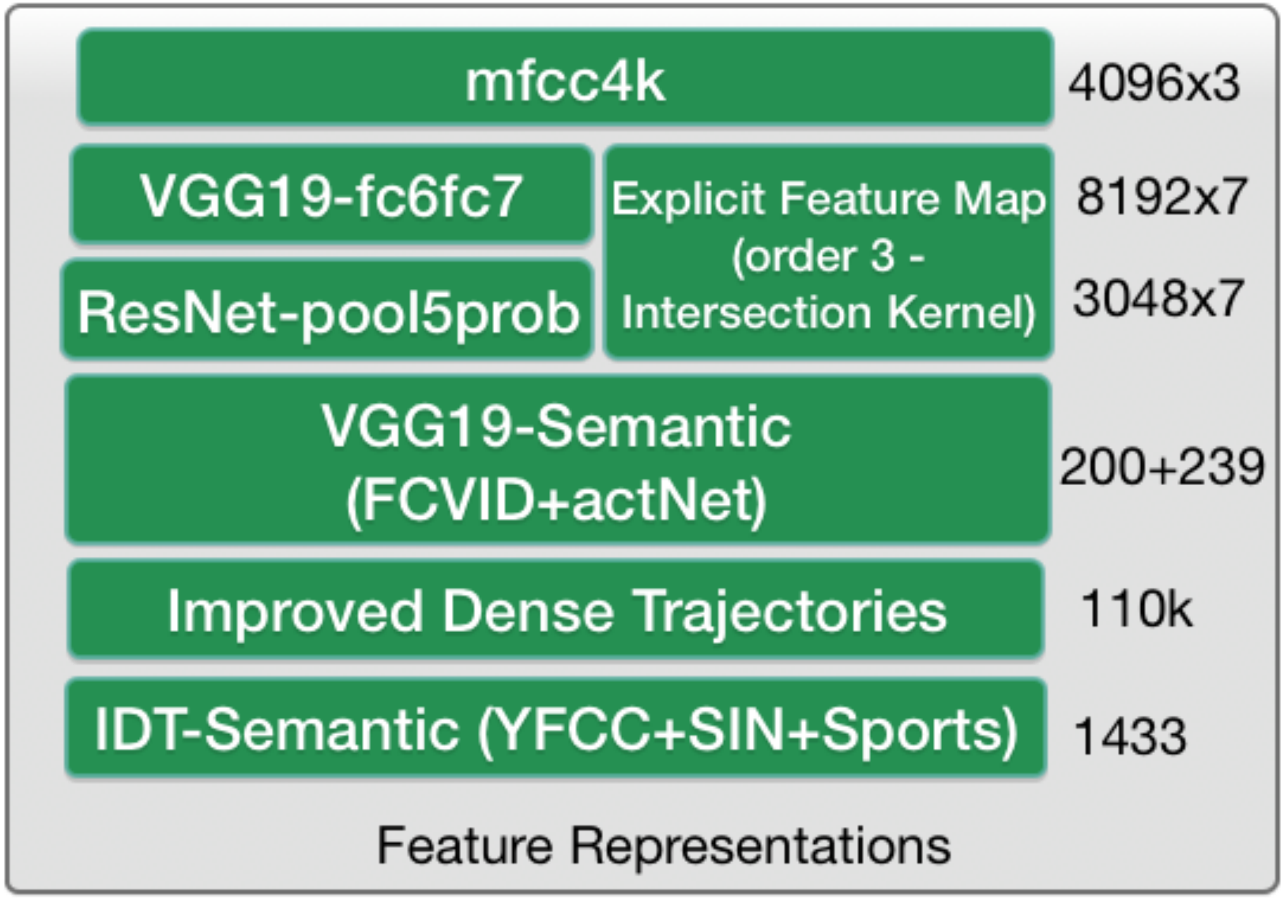
MED-System (10Ex,100Ex)



MED-System (10Ex,100Ex)

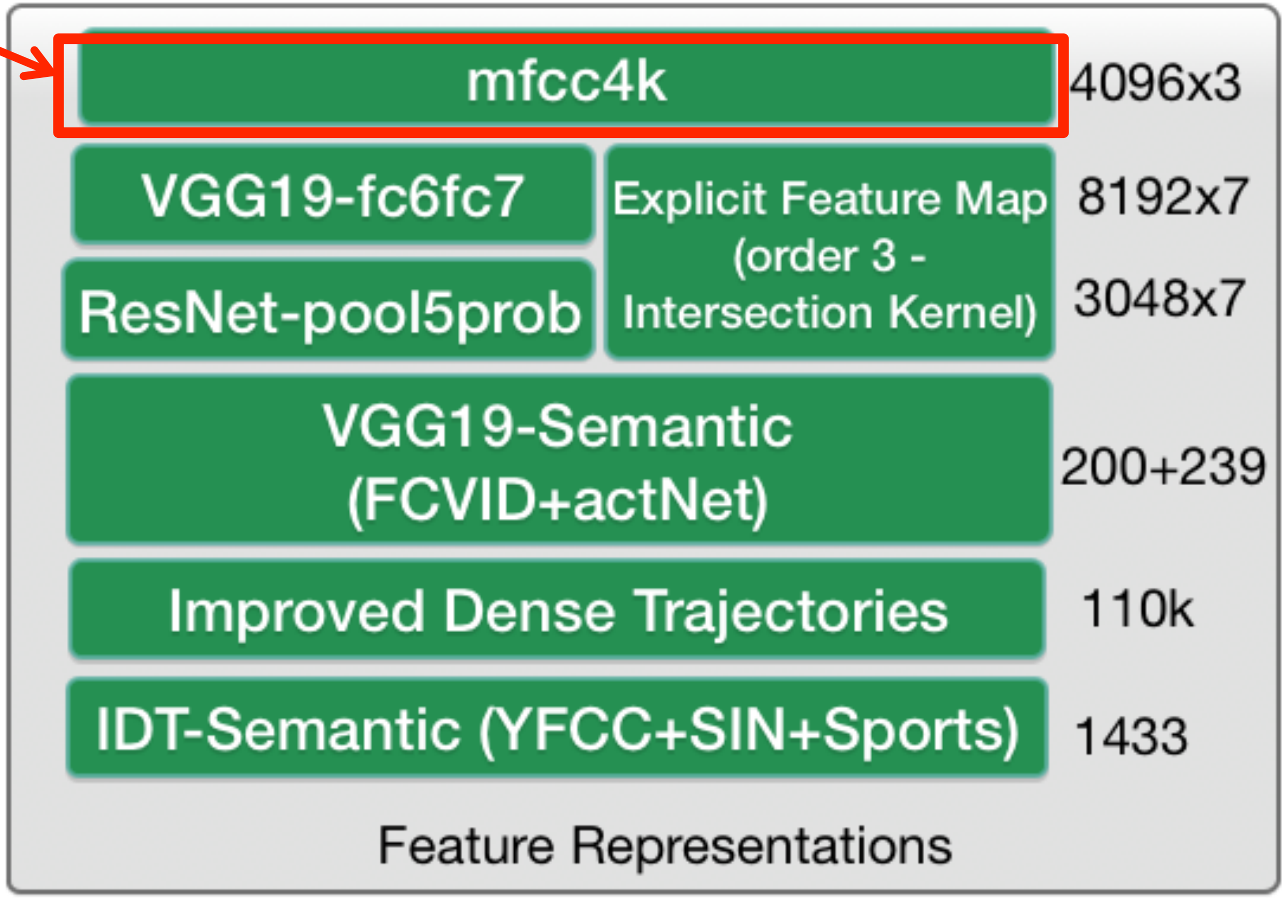


MED-System (10Ex,100Ex)

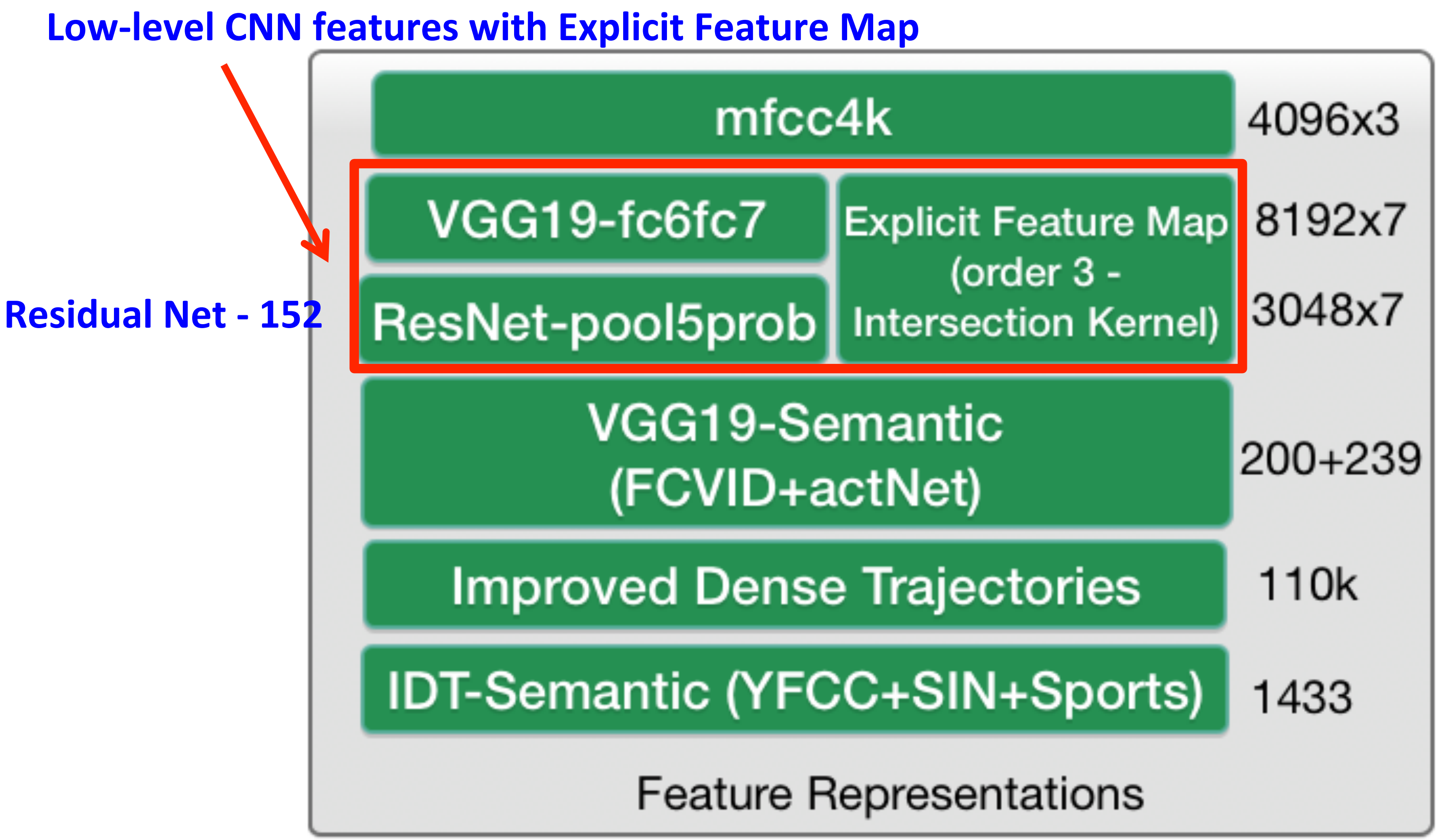


MED-System (10Ex,100Ex)

Simple Bag-of-Audio-Word method

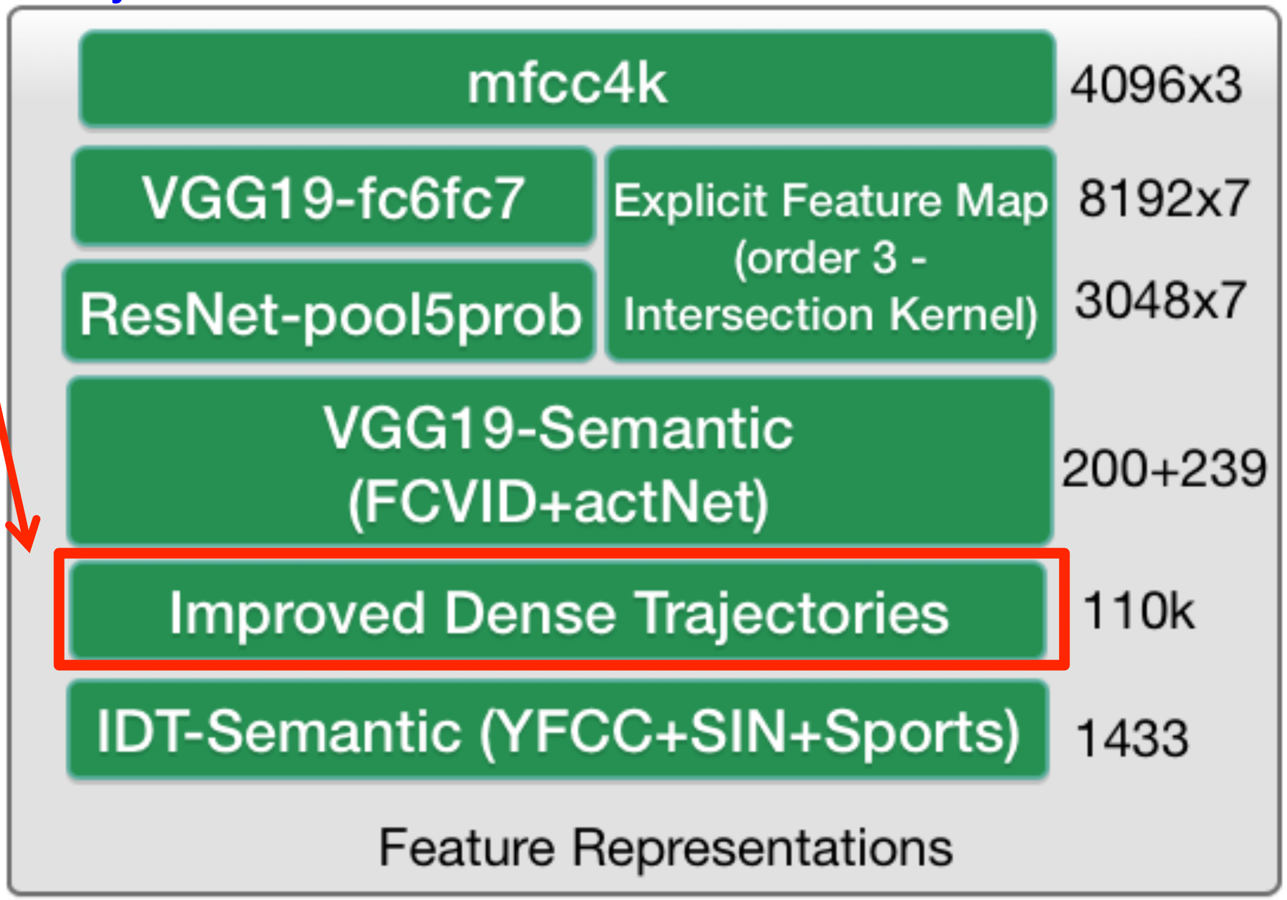


MED-System (10Ex,100Ex)



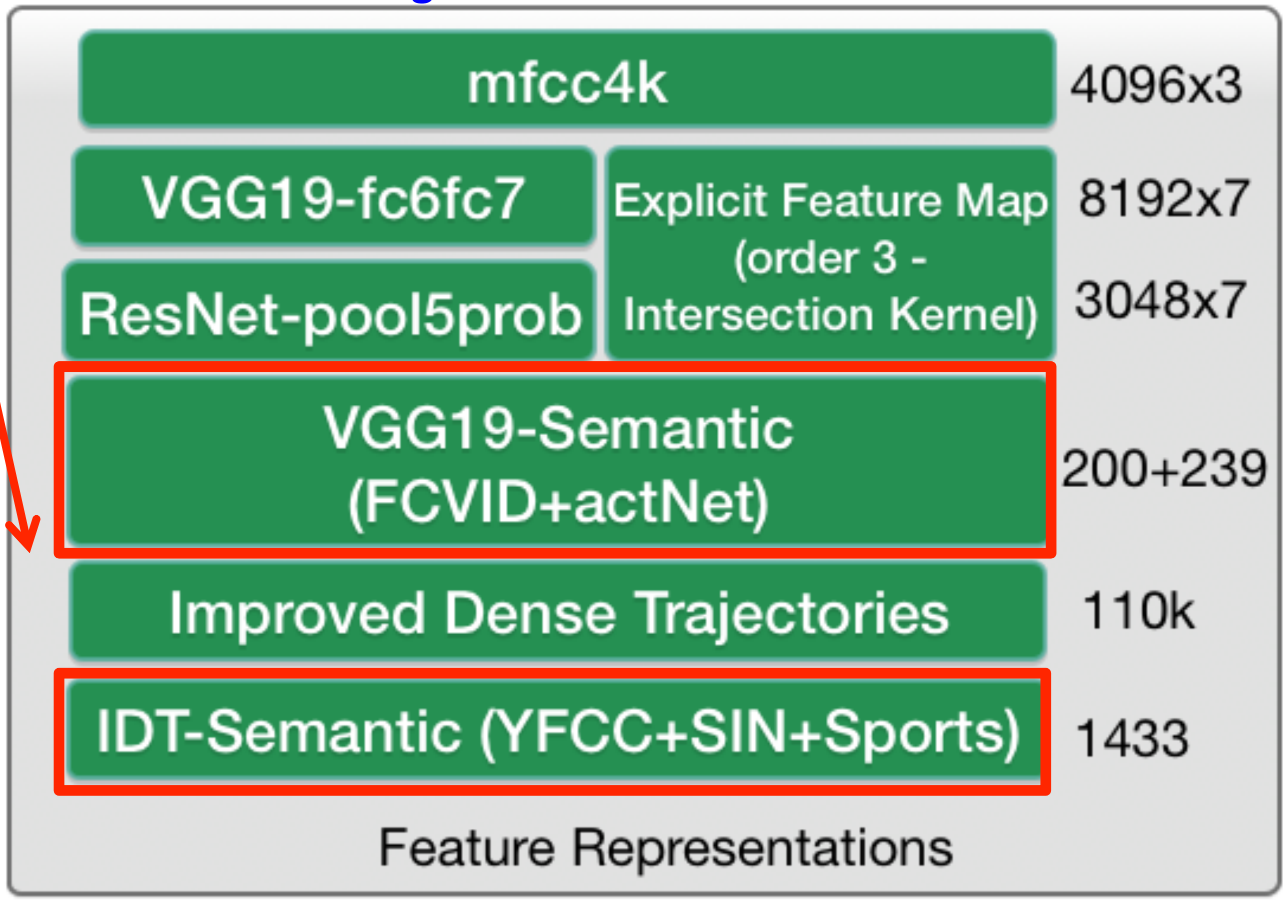
MED-System (10Ex,100Ex)

Improved Dense Trajectories



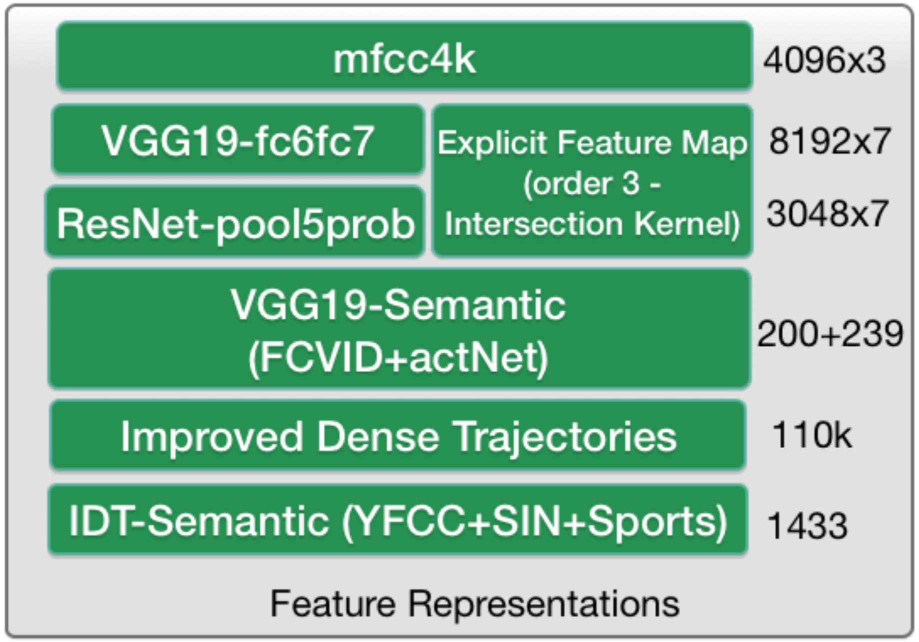
MED-System (10Ex,100Ex)

Semantic feature trained on existing video dataset



MED-System (10Ex,100Ex)

- Representations*
  - DCNN
    - ResNet > VGG
  - Kernel
    - Intersection > Chi-square (for CNN features)



* Based on experiments on MED11 TEST

Outline

- System Overview – (10Ex, 100Ex)
  – Feature Representations
- Selected Topics
  – Learning with Miss Videos
- Final Results (MED16EvalSub)
- 0Ex System
- Conclusions

MED – Learning with Miss Videos

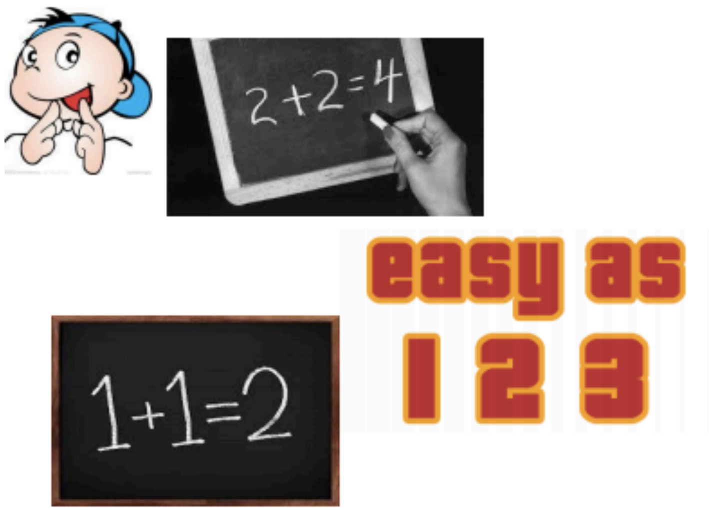
- Model Training
  - Batch Train
  - Self-Paced Curriculum Learning
    - Including Miss Videos
      - 10Ex: 10+5; 100Ex: 100+50

Self-Paced Curriculum Learning

- Curriculum Learning (Bengio et al. 2009) or self-paced learning (Kumar et al 2010) is a recently proposed learning paradigm that is inspired by **the learning process of humans and animals**.
- The samples are not learned randomly but organized in a meaningful order which illustrates from easy to gradually more complex ones.

Self-Paced Curriculum Learning

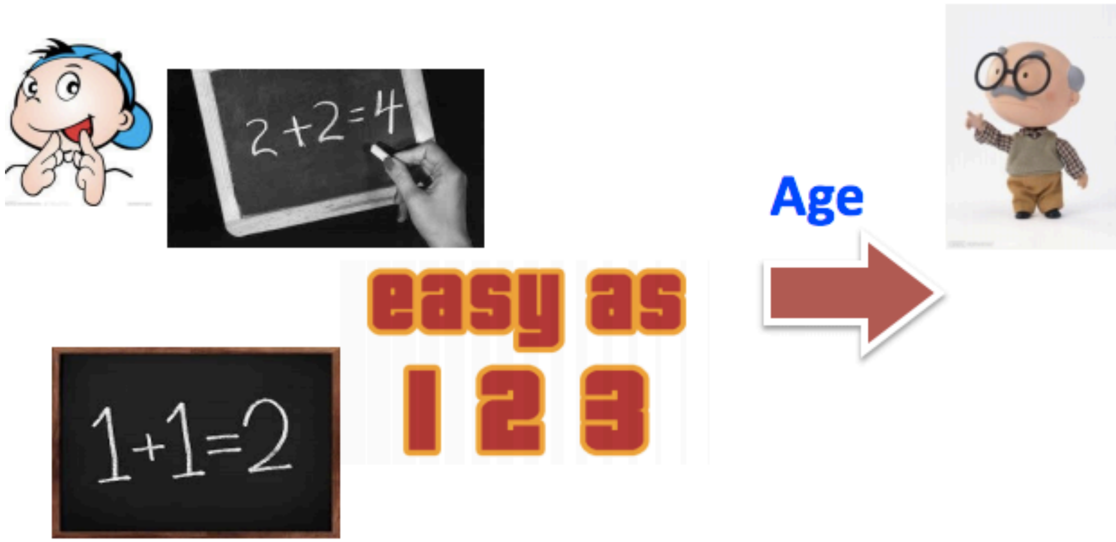
- Easy samples to complex samples.
  - Easy sample ➔ smaller loss to the already learned model.
  - Complex sample ➔ bigger loss to the already learned model.



$$\begin{aligned}
\frac{1}{g-kv}\frac{dv}{dt} &= 1 \\
\int_0^T \frac{1}{g-kv}\frac{dv}{dt}\,dt &= \int_0^T dt \\
\int_{v_0}^{v(T)} \frac{1}{g-kv}\,dv &= T \\
-\frac{1}{k}\ln|g-kv|\,\Big|_{v_0}^{v(T)} &= T \\
\ln\left|\frac{g-kv(T)}{g-kv_0}\right| &= -kT \\
\frac{g-kv(T)}{g-kv_0} &= e^{-kT}
\end{aligned}$$

Self-Paced Curriculum Learning

- Easy samples to complex samples.
  - Easy sample ➔ Positive Videos
  - Complex sample ➔ **Miss Videos**



$$
\begin{aligned}
\frac{1}{g-kv}\frac{dv}{dt} &= 1 \\
\int_0^T \frac{1}{g-kv}\frac{dv}{dt}\,dt &= \int_0^T dt \\
\int_{v_0}^{v(T)} \frac{1}{g-kv}\,dv &= T \\
-\frac{1}{k}\ln|g-kv|\,\Big|_{v_0}^{v(T)} &= T \\
\ln\left|\frac{g-kv(T)}{g-kv_0}\right| &= -kT \\
\frac{g-kv(T)}{g-kv_0} &= e^{-kT}
\end{aligned}
$$

Self-Paced Curriculum Learning

Loss Function

Regularizer

$$\min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \mathbb{E}(\mathbf{w},\mathbf{v};\lambda,\Psi) = \sum_{i=1}^{n} v_i L(y_i, g(\mathbf{x}_i,\mathbf{w})) + f(\mathbf{v};\lambda),$$

$$\text{subject to } \mathbf{v} \in \Psi$$

Biconvex Optimization Problem – Alternate Convex Search

Prior Knowledge

Latent weight variable: $\mathbf{v} = [v_1, \cdots, v_n]^T$

Model Age: λ

Curriculum Region: Ψ

Model Training - Experiments

Batch train model						
5/10 fold mean MAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.392	0.380	0.243	0.314	0.279	0.282
10Ex-incldueMiss	0.377	0.404	0.286	0.191	0.335	0.330
100Ex-noMiss	0.644	0.772	0.793	0.708	0.593	0.579
100Ex-incldueMiss	0.556	0.761	0.753	0.692	0.586	0.561

SPCL train model						
5/10 fold mean MaxMAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.465	0.506	0.391	0.487	0.392	0.483
10Ex-incldueMiss	0.464	0.575	0.505	0.446	0.437	0.494
100Ex-noMiss	0.704	0.809	0.829	0.744	0.622	0.618
100Ex-incldueMiss	0.694	0.818	0.834	0.759	0.630	0.620

Model Training - Experiments

Batch train model						
5/10 fold mean MAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.392	0.380	0.243	0.314	0.279	0.282
10Ex-incldueMiss	0.377	0.404	0.286	0.191	0.335	0.330
100Ex-noMiss	0.644	0.772	0.793	0.708	0.593	0.579
100Ex-incldueMiss	0.556	0.761	0.753	0.692	0.586	0.561

Cross-validated results on training set

SPCL train model						
5/10 fold mean MaxMAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.465	0.506	0.391	0.487	0.392	0.483
10Ex-incldueMiss	0.464	0.575	0.505	0.446	0.437	0.494
100Ex-noMiss	0.704	0.809	0.829	0.744	0.622	0.618
100Ex-incldueMiss	0.694	0.818	0.834	0.759	0.630	0.620

Model Training - Experiments

Batch train model						
5/10 fold mean MAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.392	0.380	0.243	0.314	0.279	0.282
10Ex-incldueMiss	0.377	0.404	0.286	0.191	0.335	0.330
100Ex-noMiss	0.644	0.772	0.793	0.708	0.593	0.579
100Ex-incldueMiss	0.556	0.761	0.753	0.692	0.586	0.561

Max MAP: the best MAP each run can achieve if we can find the best iteration

SPCL train model						
5/10 fold mean MaxMAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.465	0.506	0.391	0.487	0.392	0.483
10Ex-incldueMiss	0.464	0.575	0.505	0.446	0.437	0.494
100Ex-noMiss	0.704	0.809	0.829	0.744	0.622	0.618
100Ex-incldueMiss	0.694	0.818	0.834	0.759	0.630	0.620

Include Miss Videos or Not

Batch train model						
5/10 fold mean MAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.392	0.380	0.243	0.314	0.279	0.282
10Ex-incldueMiss	0.377	0.404	0.286	0.191	0.335	0.330
100Ex-noMiss	0.644	0.772	0.793	0.708	0.593	0.579
100Ex-incldueMiss	0.556	0.761	0.753	0.692	0.586	0.561

SPCL train model						
5/10 fold mean MaxMAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.465	0.506	0.391	0.487	0.392	0.483
10Ex-incldueMiss	0.464	0.575	0.505	0.446	0.437	0.494
100Ex-noMiss	0.704	0.809	0.829	0.744	0.622	0.618
100Ex-incldueMiss	0.694	0.818	0.834	0.759	0.630	0.620

AP : Better AP : Worse

Include Miss Videos or Not

Batch train model						
5/10 fold mean MAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.392	0.380	0.243	0.314	0.279	0.282
10Ex-incldueMiss	0.377	0.404	0.286	0.191	0.335	0.330
100Ex-noMiss	0.644	0.772	0.793	0.708	0.593	0.579
100Ex-incldueMiss	0.556	0.761	0.753	0.692	0.586	0.561

10Ex with Batch Train: Still Better to include Miss Videos

SPCL train model						
5/10 fold mean MaxMAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.465	0.506	0.391	0.487	0.392	0.483
10Ex-incldueMiss	0.464	0.575	0.505	0.446	0.437	0.494
100Ex-noMiss	0.704	0.809	0.829	0.744	0.622	0.618
100Ex-incldueMiss	0.694	0.818	0.834	0.759	0.630	0.620

AP : Better AP : Worse

Include Miss Videos or Not

Batch train model						
5/10 fold mean MAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.392	0.380	0.243	0.314	0.279	0.282
10Ex-incldueMiss	0.377	0.404	0.286	0.191	0.335	0.330
100Ex-noMiss	0.644	0.772	0.793	0.708	0.593	0.579
100Ex-incldueMiss	0.556	0.761	0.753	0.692	0.586	0.561

100Ex with Batch Train: Miss videos confuses the classifiers

SPCL train model						
5/10 fold mean MaxMAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.465	0.506	0.391	0.487	0.392	0.483
10Ex-incldueMiss	0.464	0.575	0.505	0.446	0.437	0.494
100Ex-noMiss	0.704	0.809	0.829	0.744	0.622	0.618
100Ex-incldueMiss	0.694	0.818	0.834	0.759	0.630	0.620

AP : Better AP : Worse

Include Miss Videos or Not

Batch train model						
5/10 fold mean MAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.392	0.380	0.243	0.314	0.279	0.282
10Ex-incldueMiss	0.377	0.404	0.286	0.191	0.335	0.330
100Ex-noMiss	0.644	0.772	0.793	0.708	0.593	0.579
100Ex-incldueMiss	0.556	0.761	0.753	0.692	0.586	0.561

SPCL Train: Including miss videos is almost always better

SPCL train model						
5/10 fold mean MaxMAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.465	0.506	0.391	0.487	0.392	0.483
10Ex-incldueMiss	0.464	0.575	0.505	0.446	0.437	0.494
100Ex-noMiss	0.704	0.809	0.829	0.744	0.622	0.618
100Ex-incldueMiss	0.694	0.818	0.834	0.759	0.630	0.620

AP : Better AP : Worse

Include Miss Videos or Not

Batch train model						
5/10 fold mean MAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.392	0.380	0.243	0.314	0.279	0.282
10Ex-incldueMiss	0.377	0.404	0.286	0.191	0.335	0.330
100Ex-noMiss	0.644	0.772	0.793	0.708	0.593	0.579
100Ex-incldueMiss	0.556	0.761	0.753	0.692	0.586	0.561

10Ex SPCL Train with low-level features: improved over 25%

SPCL train model						
5/10 fold mean MaxMAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.465	0.506	0.391	0.487	0.392	0.483
10Ex-incldueMiss	0.464	0.575	0.505	0.446	0.437	0.494
100Ex-noMiss	0.704	0.809	0.829	0.744	0.622	0.618
100Ex-incldueMiss	0.694	0.818	0.834	0.759	0.630	0.620

AP : Better AP : Worse

Comparing BatchTrain and SPCL

Batch train model						
5/10 fold mean MAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.392	0.380	0.243	0.314	0.279	0.282
10Ex-incldueMiss	0.377	0.404	0.286	0.191	0.335	0.330
100Ex-noMiss	0.644	0.772	0.793	0.708	0.593	0.579
100Ex-incldueMiss	0.556	0.761	0.753	0.692	0.586	0.561

SPCL train model						
5/10 fold mean MaxMAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.465	0.506	0.391	0.487	0.392	0.483
10Ex-incldueMiss	0.464	0.575	0.505	0.446	0.437	0.494
100Ex-noMiss	0.704	0.809	0.829	0.744	0.622	0.618
100Ex-incldueMiss	0.694	0.818	0.834	0.759	0.630	0.620

Comparing BatchTrain and SPCL

Batch train model						
5/10 fold mean MAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.392	0.380	0.243	0.314	0.279	0.282
10Ex-incldueMiss	0.377	0.404	0.286	0.191	0.335	0.330
100Ex-noMiss	0.644	0.772	0.793	0.708	0.593	0.579
100Ex-incldueMiss	0.556	0.761	0.753	0.692	0.586	0.561

SPCL outperforms BatchTrain on all features - 10Ex

SPCL train model						
5/10 fold mean MaxMAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.465	0.506	0.391	0.487	0.392	0.483
10Ex-incldueMiss	0.464	0.575	0.505	0.446	0.437	0.494
100Ex-noMiss	0.704	0.809	0.829	0.744	0.622	0.618
100Ex-incldueMiss	0.694	0.818	0.834	0.759	0.630	0.620

Comparing BatchTrain and SPCL

Batch train model						
5/10 fold mean MAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.392	0.380	0.243	0.314	0.279	0.282
10Ex-incldueMiss	0.377	0.404	0.286	0.191	0.335	0.330
100Ex-noMiss	0.644	0.772	0.793	0.708	0.593	0.579
100Ex-incldueMiss	0.556	0.761	0.753	0.692	0.586	0.561

SPCL outperforms BatchTrain on all features - 100Ex

SPCL train model						
5/10 fold mean MaxMAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.465	0.506	0.391	0.487	0.392	0.483
10Ex-incldueMiss	0.464	0.575	0.505	0.446	0.437	0.494
100Ex-noMiss	0.704	0.809	0.829	0.744	0.622	0.618
100Ex-incldueMiss	0.694	0.818	0.834	0.759	0.630	0.620

Comparing BatchTrain and SPCL

Batch train model						
5/10 fold mean MAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.392	0.380	0.243	0.314	0.279	0.282
10Ex-incldueMiss	0.377	0.404	0.286	0.191	0.335	0.330
100Ex-noMiss	0.644	0.772	0.793	0.708	0.593	0.579
100Ex-incldueMiss	0.556	0.761	0.753	0.692	0.586	0.561

SPCL train model						
5/10 fold mean MaxMAP	mfcc4k	VGG-fc6fc7	resNet-pool5prob	s-vgg	IDT	s-IDT
10Ex-noMiss	0.465	0.506	0.391	0.487	0.392	0.483
10Ex-incldueMiss	0.464	0.575	0.505	0.446	0.437	0.494
100Ex-noMiss	0.704	0.809	0.829	0.744	0.622	0.618
100Ex-incldueMiss	0.694	0.818	0.834	0.759	0.630	0.620

The weights of late fusion are calculated from cross-validation result (this table)

Outline

- System Overview – (10Ex, 100Ex)
  - Feature Representations
- Selected Topics
  - Learning with Miss Videos
- Final Results (MED16EvalSub)
- 0Ex System
- Conclusions

Final Results – MED16EvalSub

- Test Set
  - Pre-specified Events
  - MED16EvalSub
  - 32000 (16000 HAVIC + 16000 YFCC100M)

YFCC Resources

- YFCC100M video collection:
  - raw and resized videos
  - key-frames
  - video-level and shot-level DCNN features
  - Extracted concepts
  - API to content-based video engine.

https://sites.google.com/site/videosearch100m/

Final Results – MED16EvalSub

	MeanxInfAP	E024	E037
BatchTrain_010Ex	33.6	8.8	20.5
SPCL_010Ex	33.9	13.0	21.7
BestRun_010Ex*	38.5	19.2	24.5
BatchTrain_100Ex	46.4	20.0	33.0
SPCL_100Ex	47.3	24.8	36.6
BestRun_100Ex*	47.5	16.4	31.2

* Excluding our runs

Final Results – MED16EvalSub

	MeanxInfAP	E024	E037
BatchTrain_010Ex	33.6	8.8	20.5
SPCL_010Ex	**33.9**	13.0	21.7
BestRun_010Ex*	38.5	19.2	24.5
BatchTrain_100Ex	46.4	20.0	33.0
SPCL_100Ex	**47.3**	24.8	36.6
BestRun_100Ex*	47.5	16.4	31.2

SPCL performs OK on 100Ex, badly on 10Ex

* Excluding our runs

Final Results – MED16EvalSub

	MeanxInfAP	E024	E037
BatchTrain_010Ex	33.6	8.8	20.5
SPCL_010Ex	**33.9**	13.0	21.7
BestRun_010Ex*	38.5	19.2	24.5
BatchTrain_100Ex	46.4	20.0	33.0
SPCL_100Ex	**47.3**	24.8	36.6
BestRun_100Ex*	47.5	16.4	31.2

SPCL performs slightly better than BatchTrain
(How to find the best iteration model?)
(Now we use Iteration 10/30 model)

* Excluding our runs

Final Results – MED16EvalSub

	MeanxInfAP	E024	E037
BatchTrain_010Ex	33.6	8.8	20.5
SPCL_010Ex	33.9	13.0	21.7
BestRun_010Ex*	38.5	19.2	24.5
BatchTrain_100Ex	46.4	20.0	33.0
SPCL_100Ex	47.3	**24.8**	**36.6**
BestRun_100Ex*	47.5	16.4	31.2

* Excluding our runs

Selected Events where SPCL is better than the other runs

Final Results – MED16EvalSub

	MeanxInfAP	E024	E037
BatchTrain_010Ex	33.6	8.8	20.5
SPCL_010Ex	33.9	13.0	21.7
BestRun_010Ex*	38.5	19.2	24.5
BatchTrain_100Ex	46.4	20.0	33.0
SPCL_100Ex	47.3	24.8	36.6
BestRun_100Ex*	47.5	16.4	31.2

* Excluding our runs

Selected Events where
SPCL performs better than BatchTrain

Final Results – MED16EvalSub

	MeanxInfAP	E022	E028	E036
BatchTrain_010Ex	33.6	15.8	40.0	48.2
SPCL_010Ex	33.9	13.7	47.3	52.5
BestRun_010Ex*	38.5	18.3	47.0	38.1
BatchTrain_100Ex	46.4	40.1	58.7	54.2
SPCL_100Ex	47.3	41.0	57.5	50.9
BestRun_100Ex*	47.5	39.4	52.0	47.1

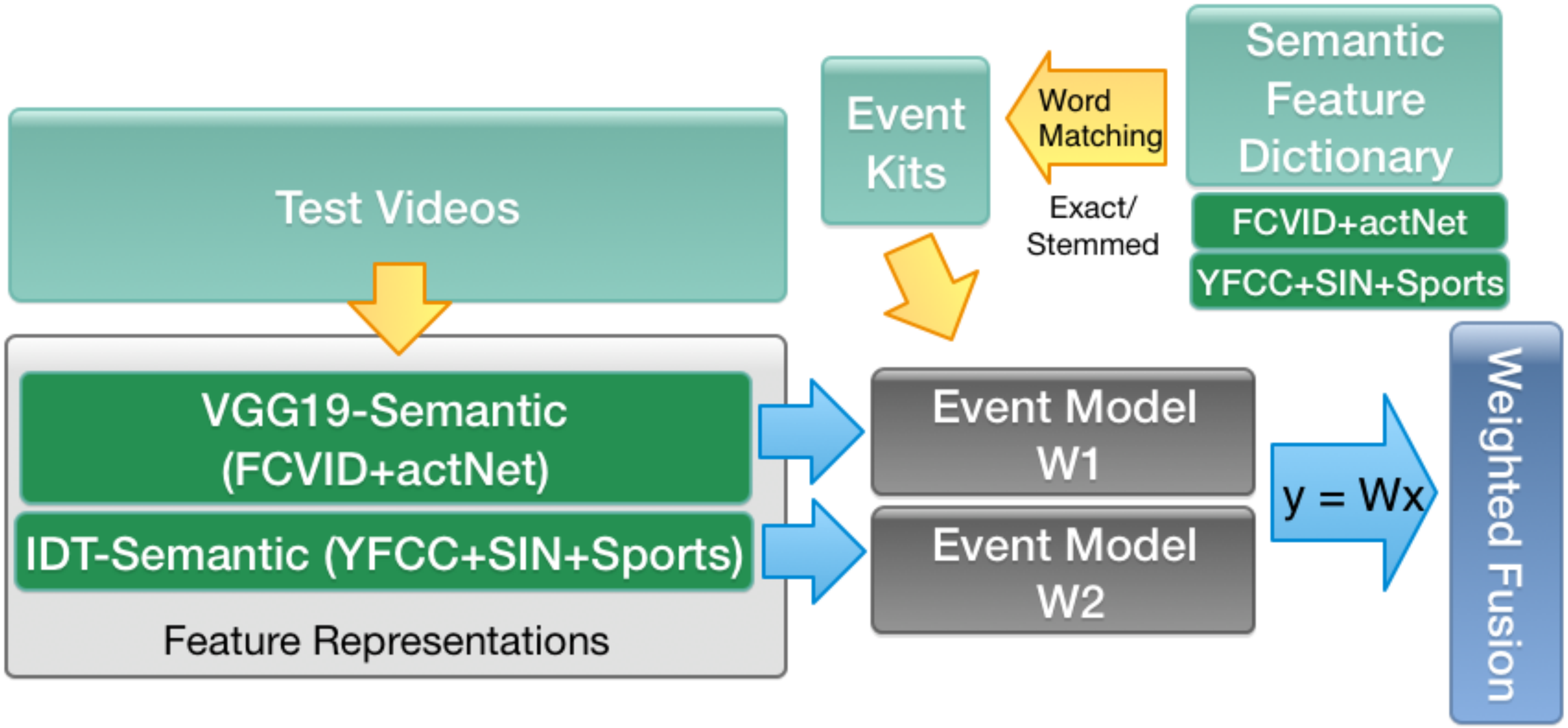
* Excluding our runs

But sometimes
SPCL is worse than BatchTrain
(Important to find the best model in SPCL)

Outline

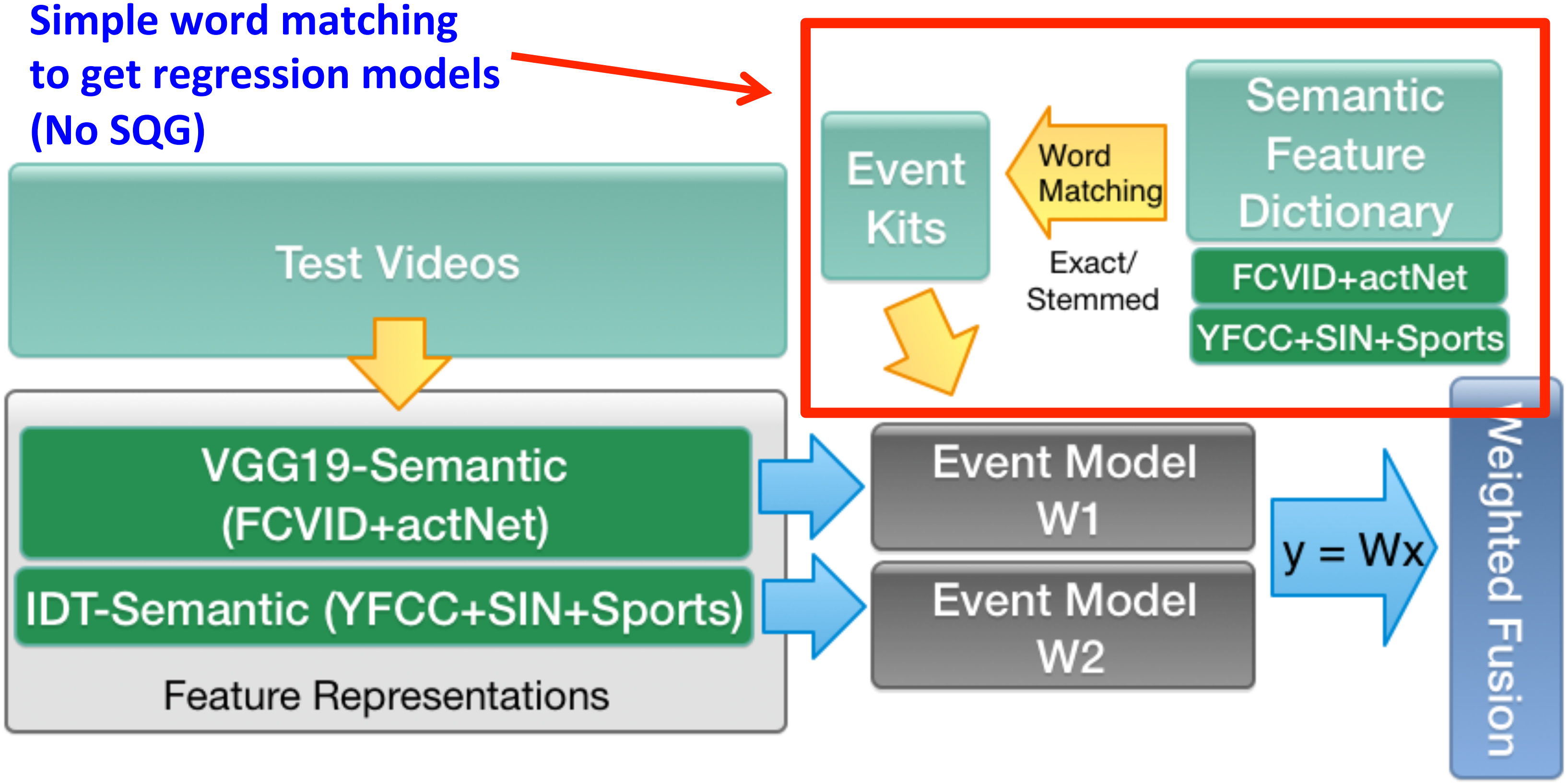
- System Overview – (10Ex, 100Ex)
  – Feature Representations
- Selected Topics
  – Learning with Miss Videos
- Final Results (MED16EvalSub)
- 0Ex System
- Conclusions

MED-pipeline (0Ex)



MED-pipeline (0Ex)

Simple word matching
to get regression models
(No SQG)



It performs well if the event kit text is in the dictionary
(E037 Parking a vehicle -> ParkingCars FCVID)

Outline

- System Overview – (10Ex, 100Ex)
  - Feature Representations
- Selected Topics
  - Learning with Miss Videos
- Final Results (MED16EvalSub)
- 0Ex System
- Conclusions

Conclusions

- We present a 10/100 Ex system trained with miss video using self-paced curriculum learning.
- In the future, we will find better way to get model from SPCL iterations (the model before overfitting to noise)

informedia
digital video understanding
SEARCH summarize visualize retrieve

2016 TRECVID
Ad-hoc Video Search - Report
Team INF

Junwei Liang, Poyao Huang, Lu Jiang, Zhenzhong Lan, Jia Chen and Alexander Hauptmann

Carnegie Mellon University

Outline

- System Overview
- Selected Topics
  - Webly-Labeled Learning
  - Experimental Results
    - FCVID and YFCC
    - AVS Extra
- Conclusions

Outline

- System Overview
- Selected Topics
  – Webly-Labeled Learning
  – Experimental Results
    - FCVID and YFCC
    - AVS Extra
- Conclusions

System Overview

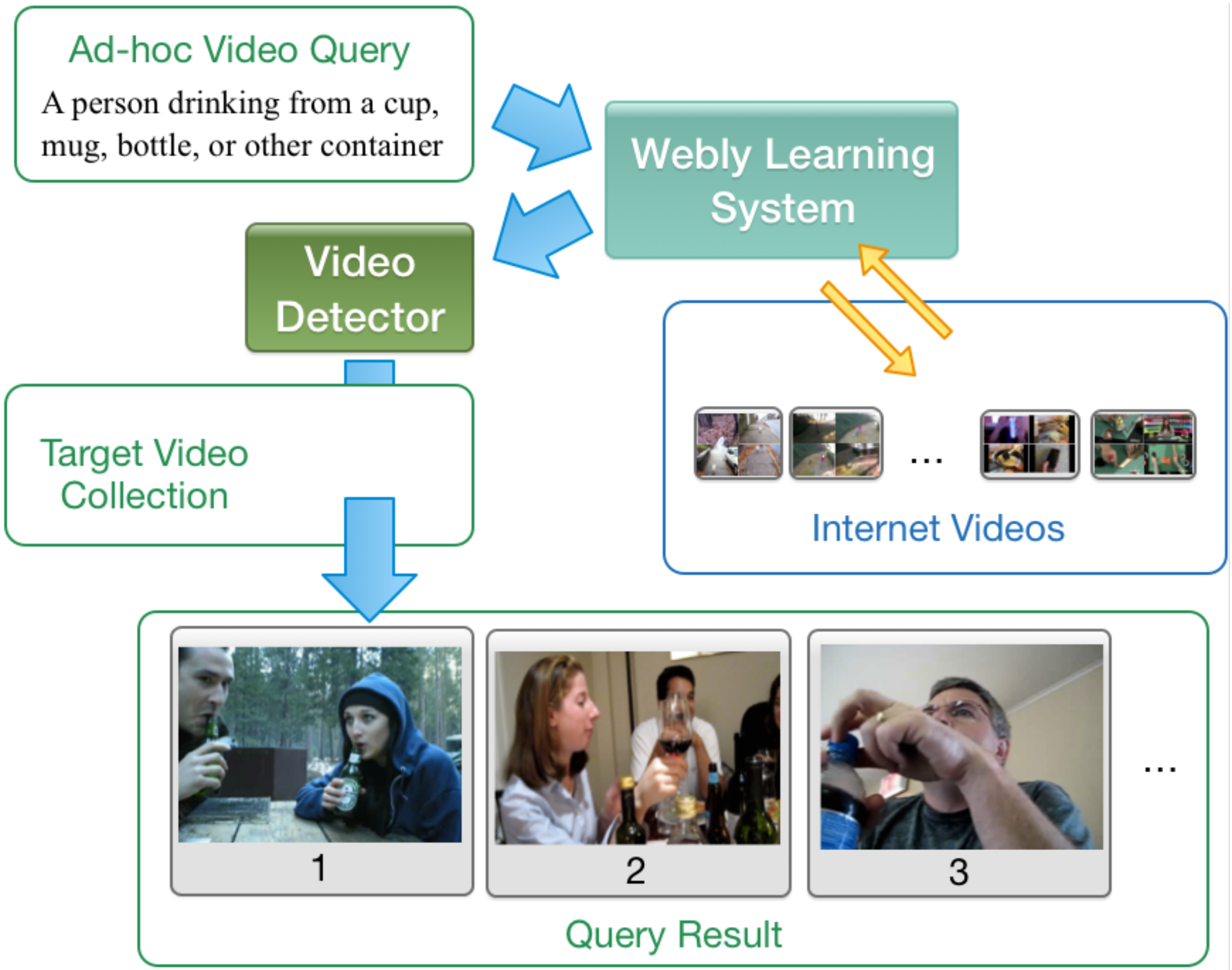
- Task
  - Given a text query, find relevant video shots in 116,097 shots (> 3sec)
  - Queries:

    01 Find shots of a person playing guitar outdoors
    …
    03 Find shots of a person playing drums indoors
    …
    28 Find shots of a person wearing a helmet
    29 Find shots of a person lighting a candle
    …

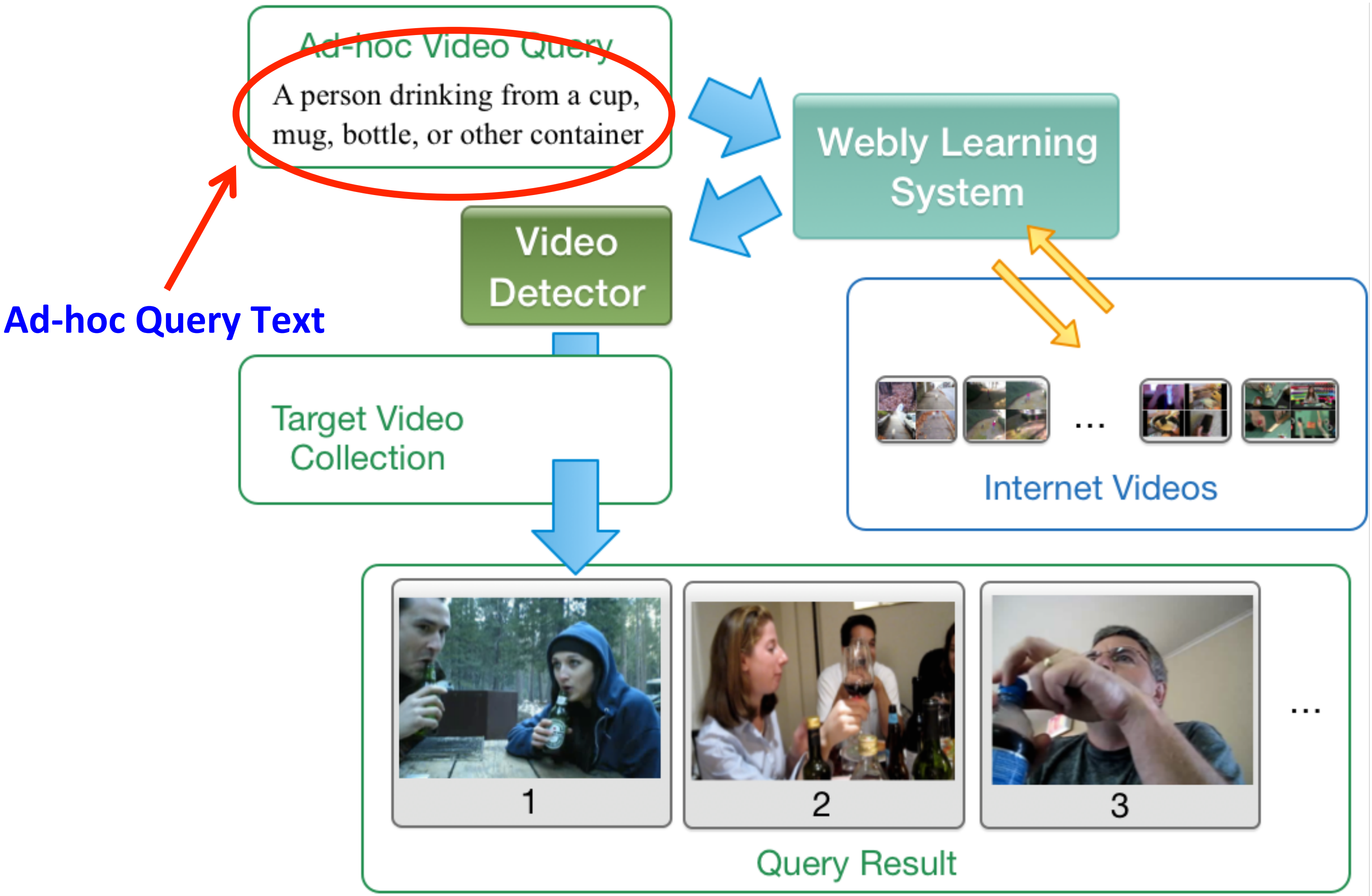
System Overview

- System Type
  - F: Fully Automatic
  - E: Used only training data collected automatically using only the official query textual description. (No annotation Run)

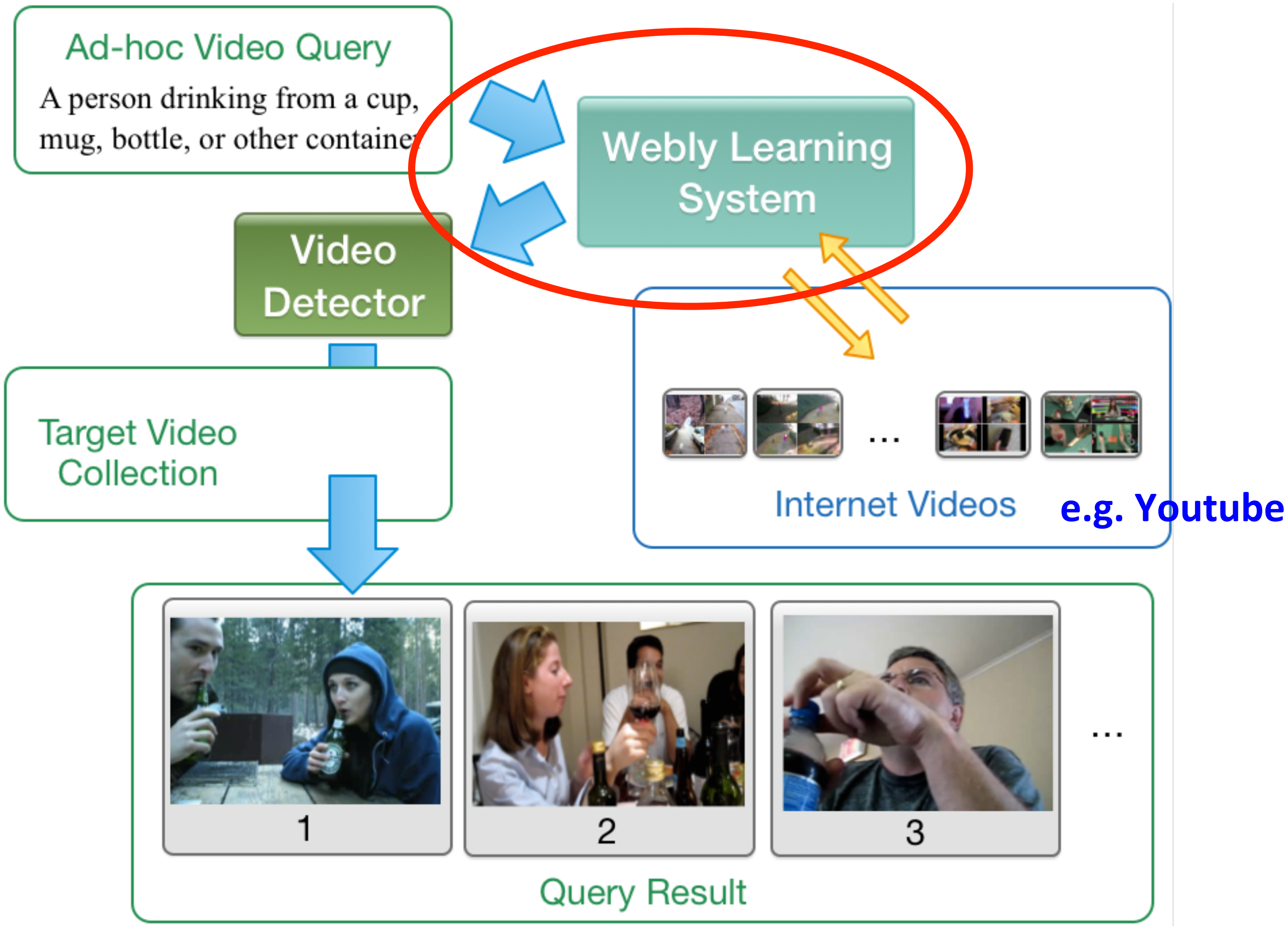
System Overview



System Overview



System Overview



Outline

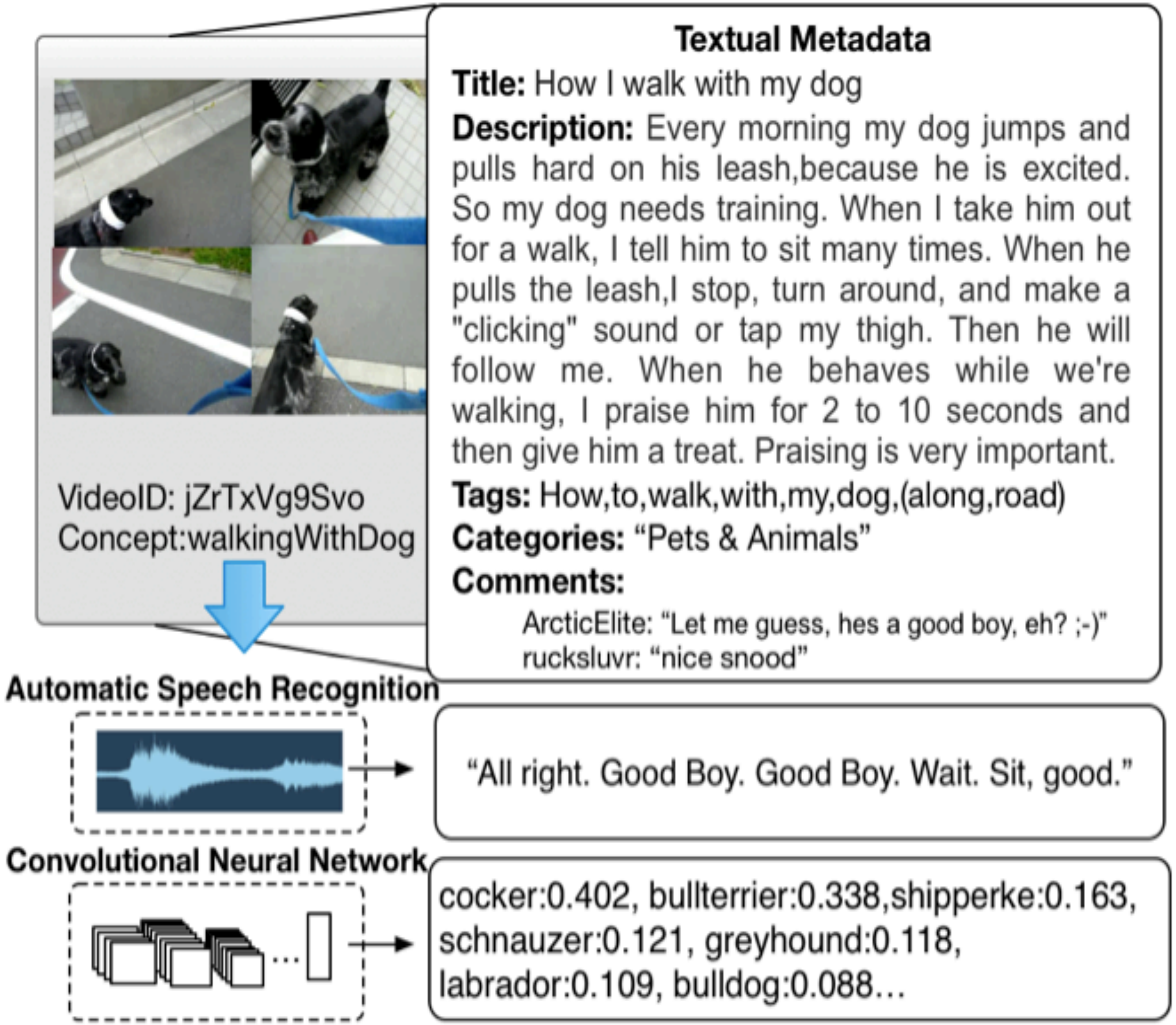
- System Overview
- Selected Topics
  - Webly-Labeled Learning
  - Experimental Results
    - FCVID and YFCC
    - AVS Extra
- Conclusions

Webly Labeled Learning

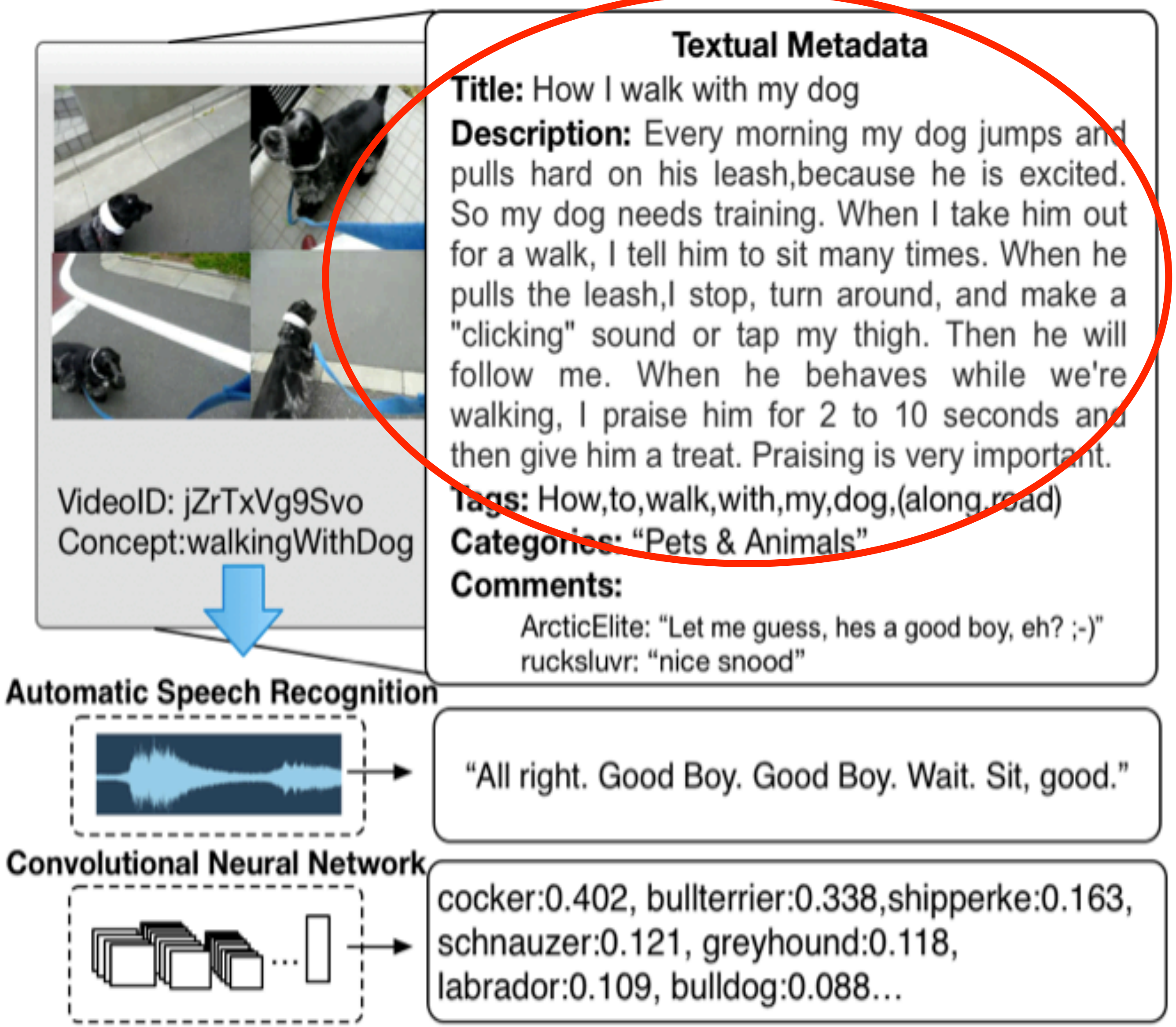
- Learn from webly-labeled* video data
  - Virtually unlimited data
  - No need for manual annotation
  - But very noisy

*_Webly_ stands for typically useful but often unreliable information in web content

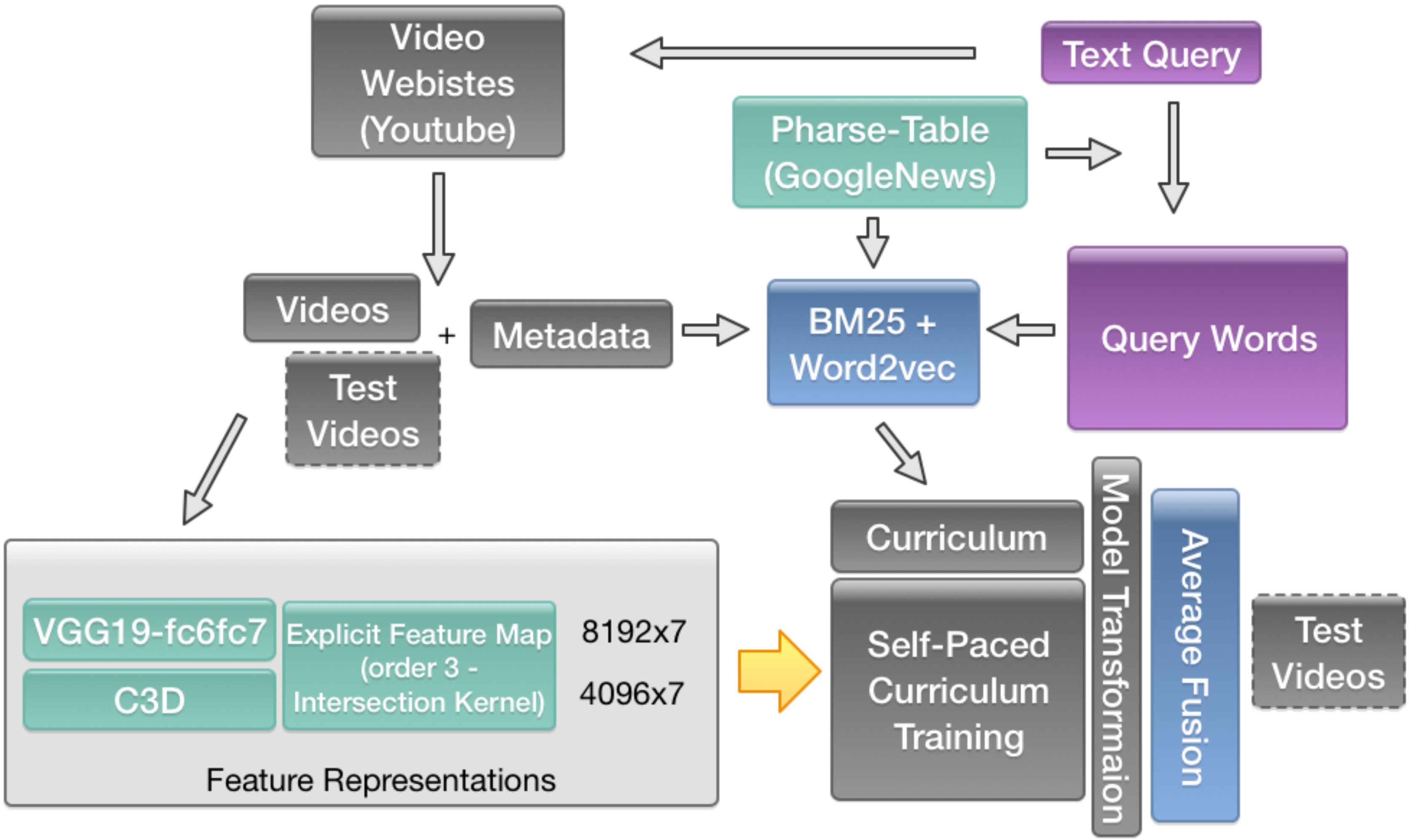
Webly Labeled Video :



Webly Labeled Video :

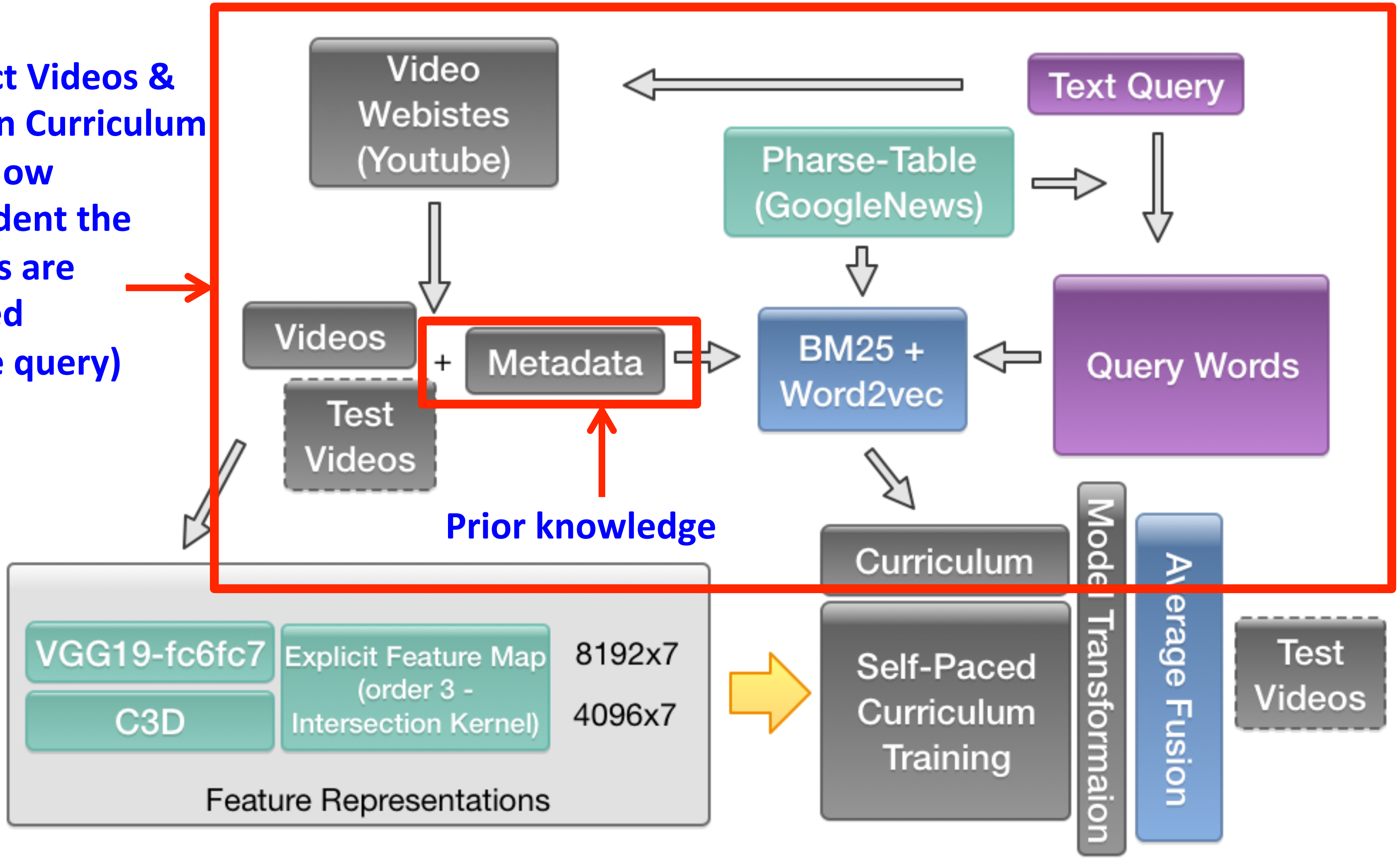


AVS Webly Learning Pipeline

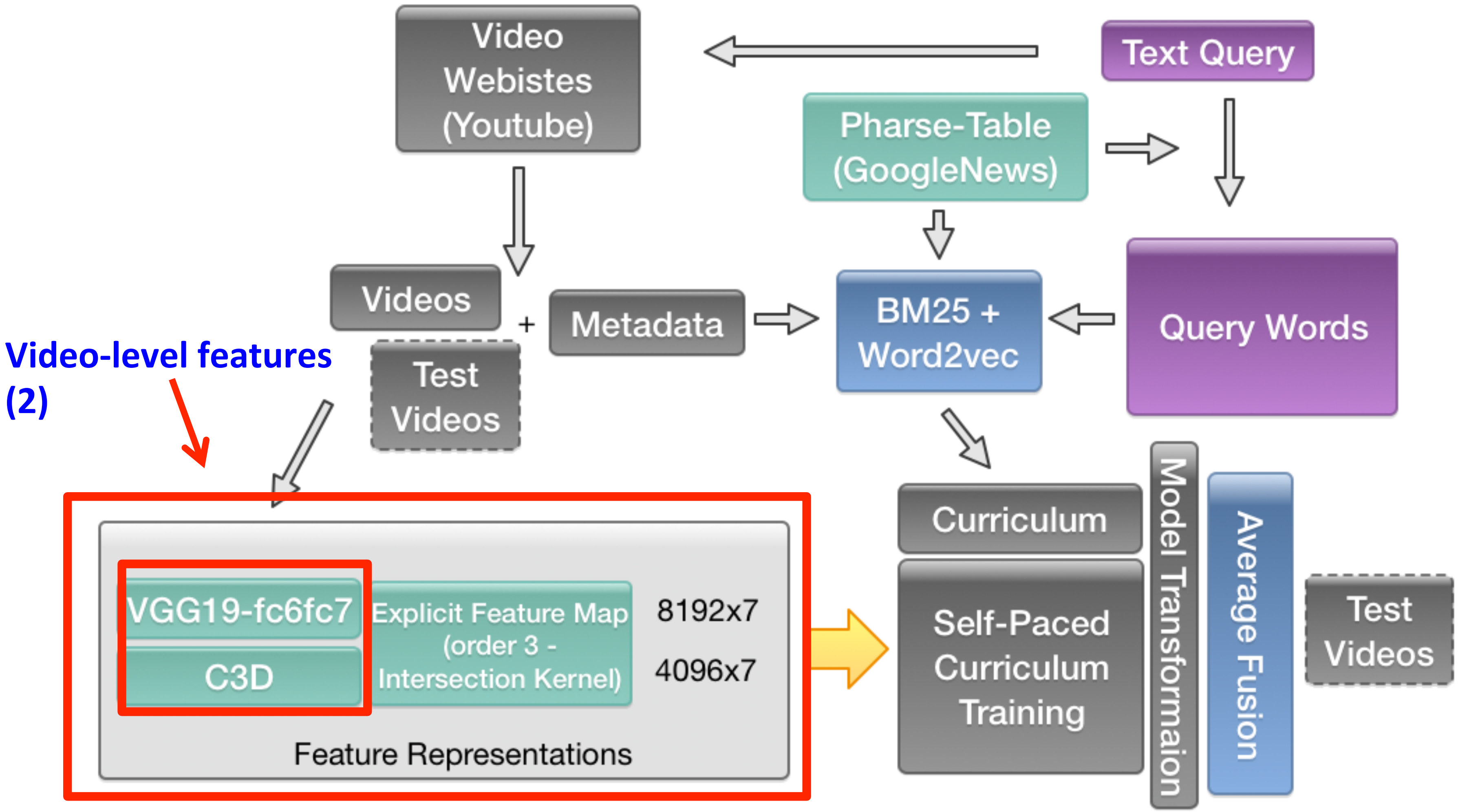


AVS Webly Learning Pipeline

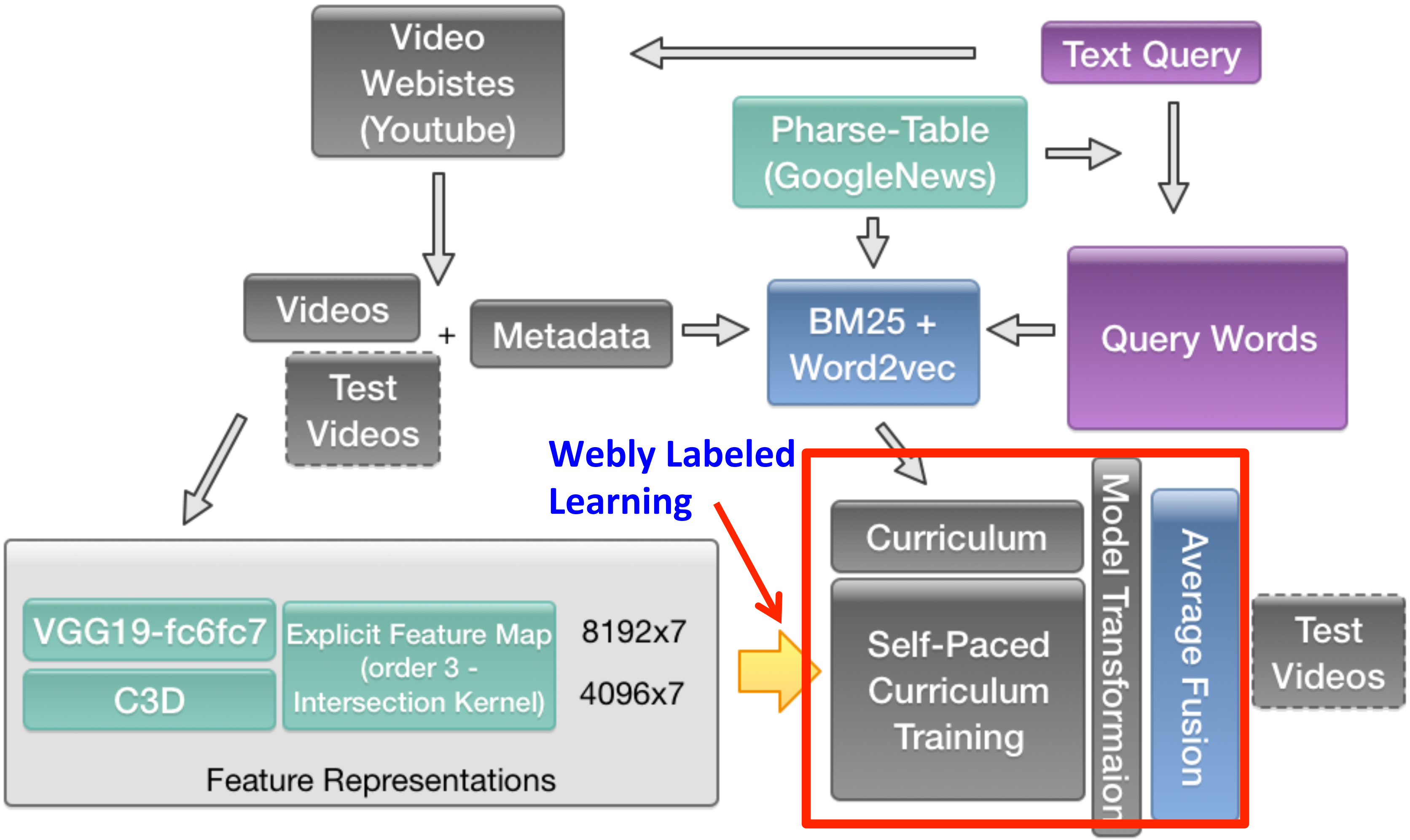
Collect Videos & Design Curriculum (i.e. How Confident the videos are related to the query)



AVS Webly Learning Pipeline



AVS Webly Learning Pipeline

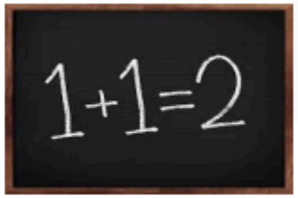


WEbly-Labeled Learning

- Curriculum Learning (Bengio et al. 2009) or self-paced learning (Kumar et al 2010) is a recently proposed learning paradigm that is inspired by **the learning process of humans and animals**.
- The samples are not learned randomly but organized in a meaningful order which illustrates from easy to gradually more complex ones.

WEbly-Labeled Learning

- Easy samples to complex samples.
  – Easy sample ➔ smaller loss to the already learned model.
  – Complex sample ➔ bigger loss to the already learned model.



$$
\begin{aligned}
\frac{1}{g-kv}\frac{dv}{dt} &= 1 \\
\int_0^T \frac{1}{g-kv}\frac{dv}{dt}\,dt &= \int_0^T dt \\
\int_{v_0}^{v(T)} \frac{1}{g-kv}\,dv &= T \\
-\frac{1}{k}\ln|g-kv|\,\Big|_{v_0}^{v(T)} &= T \\
\ln\left|\frac{g-kv(T)}{g-kv_0}\right| &= -kT \\
\frac{g-kv(T)}{g-kv_0} &= e^{-kT}
\end{aligned}
$$

WEbly-Labeled Learning

$$\min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \mathbb{E}(\mathbf{w},\mathbf{v};\lambda,\Psi) = \sum_{i=1}^{n} v_i L(y_i, g(\mathbf{x}_i,\mathbf{w})) + f(\mathbf{v};\lambda),$$

$$\text{subject to } \mathbf{v} \in \Psi$$

Latent weight variable: $\mathbf{v} = [v_1, \cdots, v_n]^T$
Model Age: λ
Curriculum Region: Ψ

WEbly-Labeled Learning

Loss Function

Regularizer

$$\min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \mathbb{E}(\mathbf{w},\mathbf{v};\lambda,\Psi) = \sum_{i=1}^{n} v_i L(y_i, g(\mathbf{x}_i,\mathbf{w})) + f(\mathbf{v};\lambda),$$

$$\text{subject to } \mathbf{v} \in \Psi$$

Webly Labeled Prior Knowledge

Latent weight variable: $\mathbf{v} = [v_1, \cdots, v_n]^T$

Model Age: λ

Curriculum Region: Ψ

WEbly-Labeled Learning

Loss Function

Regularizer

$$\min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \mathbb{E}(\mathbf{w},\mathbf{v};\lambda,\Psi)=\sum_{i=1}^{n} v_i L(y_i,g(\mathbf{x}_i,\mathbf{w}))+f(\mathbf{v};\lambda),$$

$$\text{subject to } \mathbf{v}\in\Psi$$

Biconvex Optimization Problem – Alternate Convex Search

Webly Labeled Prior Knowledge

Latent weight variable: $\mathbf{v} = [v_1, \cdots, v_n]^T$
Model Age: λ
Curriculum Region: Ψ

Algorithm

$$\min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \mathbb{E}(\mathbf{w},\mathbf{v};\lambda,\Psi)=\sum_{i=1}^{n} v_i L(y_i,g(\mathbf{x}_i,\mathbf{w}))+f(\mathbf{v};\lambda),$$

subject to $\mathbf{v} \in \Psi$

Algorithm 1: Webly-labeled Learning (WELL).

---

input : Input dataset $\mathcal{D}$, curriculum region Ψ, self-paced function f and a step size μ

output: Model parameter $\mathbf{w}$

```
Initialize v*, λ in the curriculum region;
while not converged do
    Update w* = arg min_w E(w, v*; λ, Ψ);
    Update v* = arg min_v E(w*, v; λ, Ψ);
    if λ is small then increase λ by the step size μ;
end
return w*
```

Algorithm

$$\min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \mathbb{E}(\mathbf{w},\mathbf{v};\lambda,\Psi) = \sum_{i=1}^{n} v_i L(y_i, g(\mathbf{x}_i,\mathbf{w})) + f(\mathbf{v};\lambda),$$

subject to $\mathbf{v} \in \Psi$

Algorithm 1: Webly-labeled Learning (WELL).

input : Input dataset $\mathcal{D}$, curriculum region Ψ, self-paced function f and a step size μ

output: Model parameter $\mathbf{w}$

Initialize $\mathbf{v}^*$, λ in the curriculum region;
while *not converged* **do**
  Update $\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathbb{E}(\mathbf{w}, \mathbf{v}^*; \lambda, \Psi)$;
  Update $\mathbf{v}^* = \arg\min_{\mathbf{v}} \mathbb{E}(\mathbf{w}^*, \mathbf{v}; \lambda, \Psi)$;
  if λ *is small* **then** increase λ by the step size μ;
end
return $\mathbf{w}^*$

Algorithm

$$\min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \mathbb{E}(\mathbf{w},\mathbf{v};\lambda,\Psi)=\sum_{i=1}^{n} v_i L(y_i,g(\mathbf{x}_i,\mathbf{w}))+f(\mathbf{v};\lambda),$$

$$\text{subject to } \mathbf{v}\in\Psi$$

Algorithm 1: Webly-labeled Learning (WELL).

```
input  : Input dataset D, curriculum region Ψ,
         self-paced function f and a step size μ
output: Model parameter w

Initialize v*, λ in the curriculum region;
while not converged do
    Update w* = arg min_w E(w, v*; λ, Ψ);
    Update v* = arg min_v E(w*, v; λ, Ψ);
    if λ is small then increase λ by the step size μ;
end
return w*
```

Outline

- System Overview
- Selected Topics
  - Webly-Labeled Learning
  - Experimental Results
    - FCVID and YFCC (*)
    - AVS Extra
- Conclusions

* Liang, Junwei, Lu Jiang, Deyu Meng, and Alexander Hauptmann. "Learning to detect concepts from webly-labeled video data." IJCAI, 2016.

Outline

- System Overview
- Selected Topics
  - Webly-Labeled Learning
  - Experimental Results
    - FCVID and YFCC (-)
    - AVS Extra
- Conclusions

AVS – Extra Experiments

	MeanxInfAP	505	509	511
IACC.3_VGG	0.003	-	-	-
BatchTrain_VGG_top1000	0.016	0.002	0.099	0.033
C3D_top1000	0.024	0.003	**0.123**	0.040
VGG_top1000*	0.024	0.020	0.030	0.080
VGG_top500	0.029	**0.021**	0.044	0.088
C3D+VGG_top1000*	0.040	0.013	0.117	**0.109**
Best System (F)**	0.054	**0.002**	**0.036**	**0.025**

* The system runs that we submitted

** Excluding our system runs

AVS – Extra Experiments

Only learning from IACC.3 metadata - failed

	MeanxInfAP	505	509	511
IACC.3_VGG	0.003	-	-	-
BatchTrain_VGG_top1000	0.016	0.002	0.099	0.033
C3D_top1000	0.024	0.003	**0.123**	0.040
VGG_top1000*	0.024	0.020	0.030	0.080
VGG_top500	0.029	**0.021**	0.044	0.088
C3D+VGG_top1000*	0.040	0.013	0.117	**0.109**
Best System (F)**	0.054	**0.002**	**0.036**	**0.025**

* The system runs that we submitted

** Excluding our system runs

AVS – Extra Experiments

Better than simple batch train 50%

	MeanxInfAP	505	509	511
IACC.3_VGG	0.003	-	-	-
BatchTrain_VGG_top1000	0.016	0.002	0.099	0.033
C3D_top1000	0.024	0.003	**0.123**	0.040
VGG_top1000*	0.024	0.020	0.030	0.080
VGG_top500	0.029	**0.021**	0.044	0.088
C3D+VGG_top1000*	0.040	0.013	0.117	**0.109**
Best System (F)**	0.054	**0.002**	**0.036**	**0.025**

* The system runs that we submitted
** Excluding our system runs

AVS – Extra Experiments

Combining C3D and VGG improved 67%

	MeanxInfAP	505	509	511
IACC.3_VGG	0.003	-	-	-
BatchTrain_VGG_top1000	0.016	0.002	0.099	0.033
C3D_top1000	0.024	0.003	**0.123**	0.040
VGG_top1000*	0.024	0.020	0.030	0.080
VGG_top500	0.029	**0.021**	0.044	0.088
C3D+VGG_top1000*	0.040	0.013	0.117	**0.109**
Best System (F)**	0.054	**0.002**	**0.036**	**0.025**

* The system runs that we submitted

** Excluding our system runs

AVS – Extra Experiments

Selected queries where our system significantly outperforms the rest

	MeanxInfAP	505	509	511
IACC.3_VGG	0.003	-	-	-
C3D_top1000	0.024	0.003	**0.123**	0.040
VGG_top1000*	0.024	0.020	0.030	0.080
VGG_top500	0.029	**0.021**	0.044	0.088
C3D+VGG_top1000*	0.040	0.013	0.117	**0.109**
Best System (F)**	0.054	**0.002**	**0.036**	**0.025**

* The system runs that we submitted
** Excluding our system runs

AVS – Extra Experiments

Selected queries where our system performs very badly (about 14 out of 30 are under 0.01)

	MeanxInfAP	506	513	522
IACC.3_VGG	0.003	-	-	-
C3D_top1000	0.024	0.002	0.000	0.000
VGG_top1000*	0.024	0.016	0.000	0.006
VGG_top500	0.029	**0.032**	0.000	**0.010**
C3D+VGG_top1000*	0.040	0.017	0.000	0.002
Best System (F)**	0.054	**0.435**	**0.176**	**0.229**

* The system runs that we submitted

** Excluding our system runs

AVS – Extra Experiments

506 Find shots of the 43rd president George W. Bush sitting down talking with people indoors
- Not enough data

	MeanxInfAP	506	513	522
IACC.3_VGG	0.003	-	-	-
C3D_top1000	0.024	0.002	0.000	0.000
VGG_top1000*	0.024	0.016	0.000	0.006
VGG_top500	0.029	**0.032**	0.000	**0.010**
C3D+VGG_top1000*	0.040	0.017	0.000	0.002
Best System (F)**	0.054	**0.435**	**0.176**	**0.229**

* The system runs that we submitted
** Excluding our system runs

AVS – Extra Experiments

513 Find shots of military personnel interacting with protesters

	MeanxInfAP	506	513	522
IACC.3_VGG	0.003	-	-	-
C3D_top1000	0.024	0.002	0.000	0.000
VGG_top1000*	0.024	0.016	0.000	0.006
VGG_top500	0.029	**0.032**	0.000	**0.010**
C3D+VGG_top1000*	0.040	0.017	0.000	0.002
Best System (F)**	0.054	**0.435**	**0.176**	**0.229**

* The system runs that we submitted
** Excluding our system runs

AVS – Extra Experiments

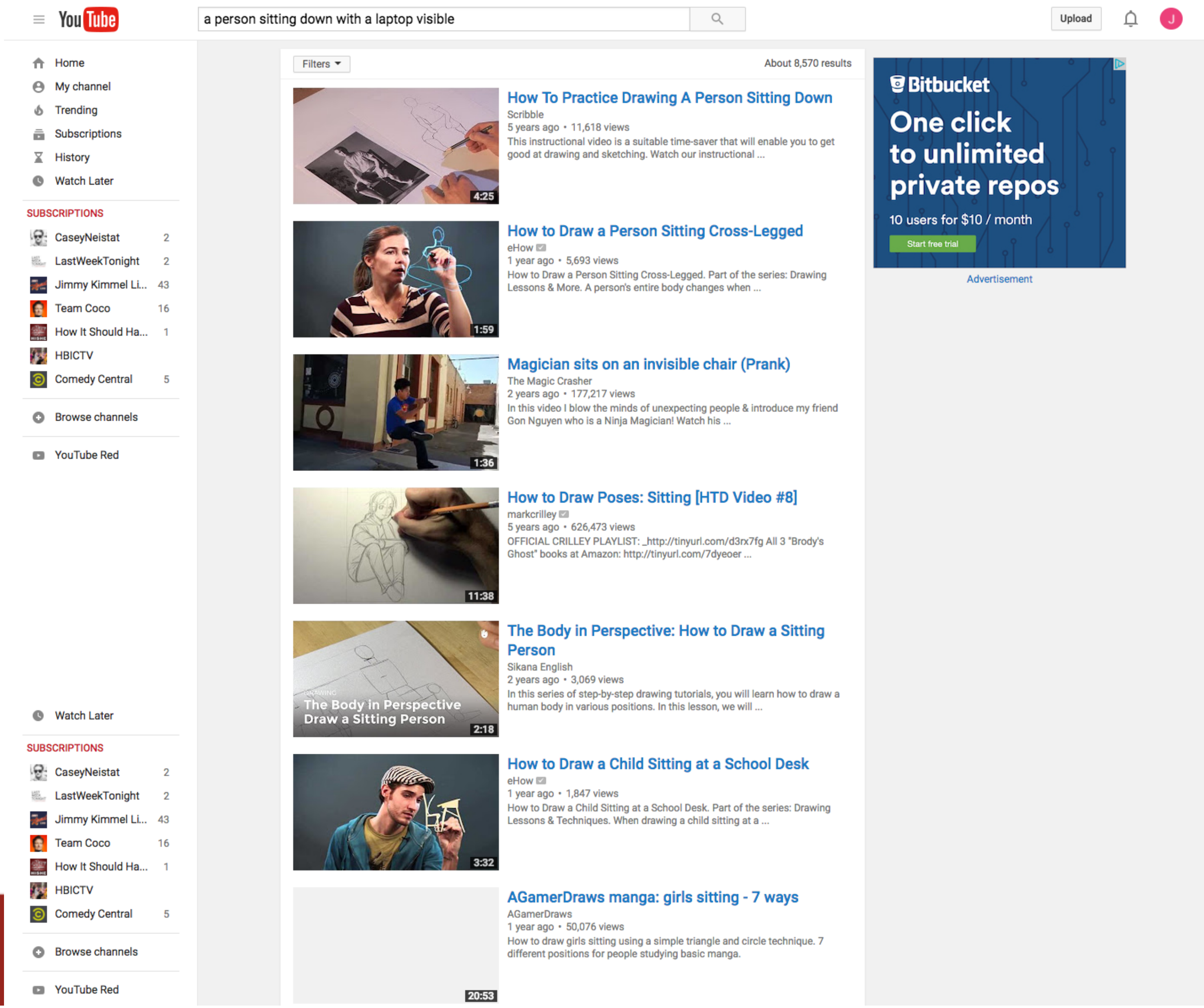
522 Find shots of a person sitting down with a laptop visible

- **Not good for retrieval based on textual metadata**

	MeanxInfAP	506	513	522
IACC.3_VGG	0.003	-	-	-
C3D_top1000	0.024	0.002	0.000	0.000
VGG_top1000*	0.024	0.016	0.000	0.006
VGG_top500	0.029	**0.032**	0.000	**0.010**
C3D+VGG_top1000*	0.040	0.017	0.000	0.002
Best System (F)**	0.054	**0.435**	**0.176**	**0.229**

* The system runs that we submitted
** Excluding our system runs

A person sitting down with a laptop



Outline

- System Overview
- Selected Topics
  - Webly-Labeled Learning
  - Experimental Results
    - FCVID and YFCC (-)
    - AVS Extra
- Conclusions & Future Work

Conclusion & Future Work

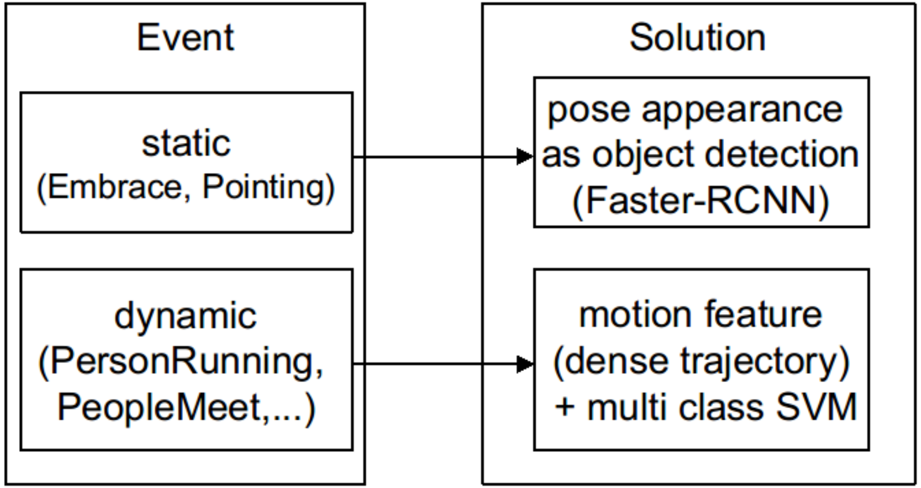
- We present a Webly-Labeled Learning framework for video detector learning
- It utilizes prior knowledge from the Internet to allow fully automatic video query with no annotation
- In the future, we will incorporate SQG and object detection for certain type of queries

INF@TREC 2016: Surveillance Event Detection

Jia Chen[1], Jiande Sun[2], Yang Chen[3], Alexandar Hauptmann[1]
[1]Carnegie Mellon University
[2]Shandong University
[3]Zhejiang University

System overview

- Mixed strategy approach
  - ‘Static’ actions primarily defined by key poses
    - Embrace, Pointing, Cell2Ear
  - ‘Dynamic’ action primarily defined by motions
    - Running, People meeting, ...



Static action

- Object detection for pose overall appearance
- One model for all cameras (camera irrelevant)
- Train data
  - manually label the bounding box for the corresponding people involved in the event
  - Embrace (1,853 bounding boxes)
  - Pointing (2,518 bounding boxes)
  - Cell2Ear (1,391 bounding boxes)

Pose modeling

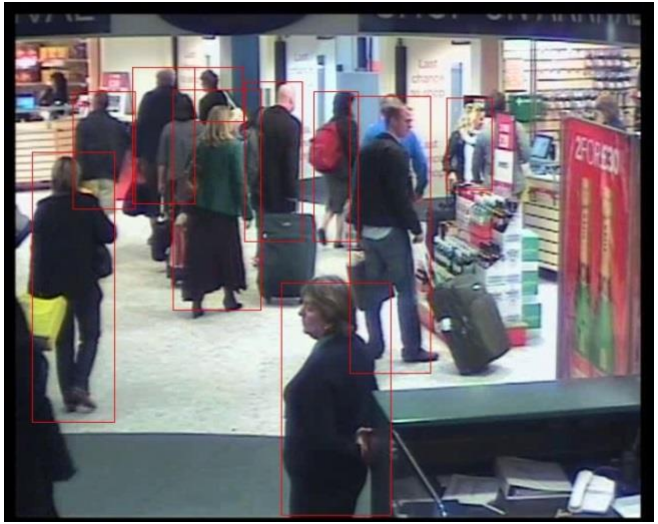
- Overall appearance vs key point skeleton

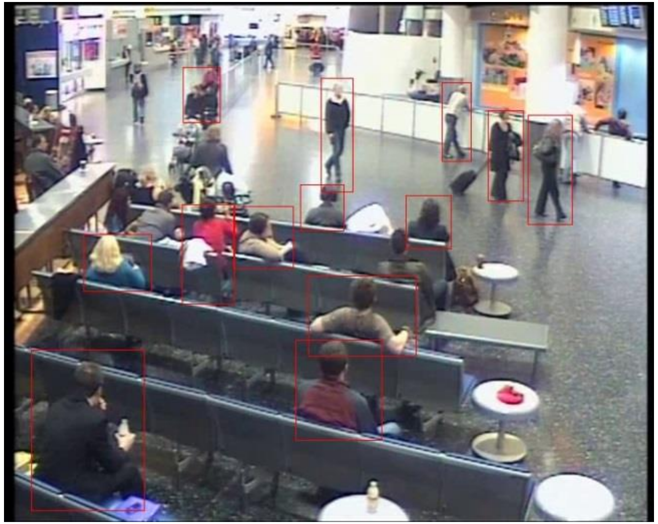
overall appearance

key point skeleton

Unsupervised data generation for hard negative class

- Other poses are used as hard negatives
- Automatically generate labels for this negative class using a pre-trained person detector





Prediction in test stage

- predict pose on images per 10 frames (0.4s)
- threshold the score at 0.1
- average pooling score in sliding windows
  - width: 50 frames
  - stride: 50 frames

Dynamic actions (from 2015)

- Raw feature extraction
  - dense trajectory and improved dense trajectory
- Feature Encoding
  - fish vector and spatial fish vector

tra		hog		hof		MBHx		MBHy	
sfv	fv	sfv	fv	sfv	fv	sfv	fv	sfv	fv

- SVM as multi-class classifier (one model for one camera)
- Score fusion

Performance

- Object detection metric
  - AP is much lower than that on object detection dataset (>=0.8), e.g. MSCOCO
  - Embrace/Pointing/Cell2Ear pose is more fine-grained and much harder than person detection
  - Ratio of pos/neg in SED test data much smaller than 1:6 (1:921)

	mAP (1:6)
Embrace	0.425
Pointing	0.263
Cell2Ear	0.024

Performance

- Event detection metric*
  - promising performance on PMiss for Embrace
  - promising performance on RFA for Cell2Ear
  - mediocre performance of Pointing on actualRFA and actual PMiss leads to worst performance on actual DCR

	actualDCR	minDCR	actualRFA	actualPMiss	#CorDet
Cell2Ear	0.9901	0.9308	5.57	0.962	12
Embrace	0.7335	0.7006	40.93	0.529	139
Pointing	0.9648	0.9550	22.33	0.853	254

*Evaluated on Eev08

Embrace case study (true positive)

predict score: 1.00

predict score: 0.71

Embrace case study (false positive)

predict score: 1.00

fusion with motion feature
will help solve such cases

predict score: 0.95

3d information will help solve
such cases

Pointing case study (true positive)

predict score: 1.00

predict score: 0.87

Pointing case study (false positive)

predict score: 0.96

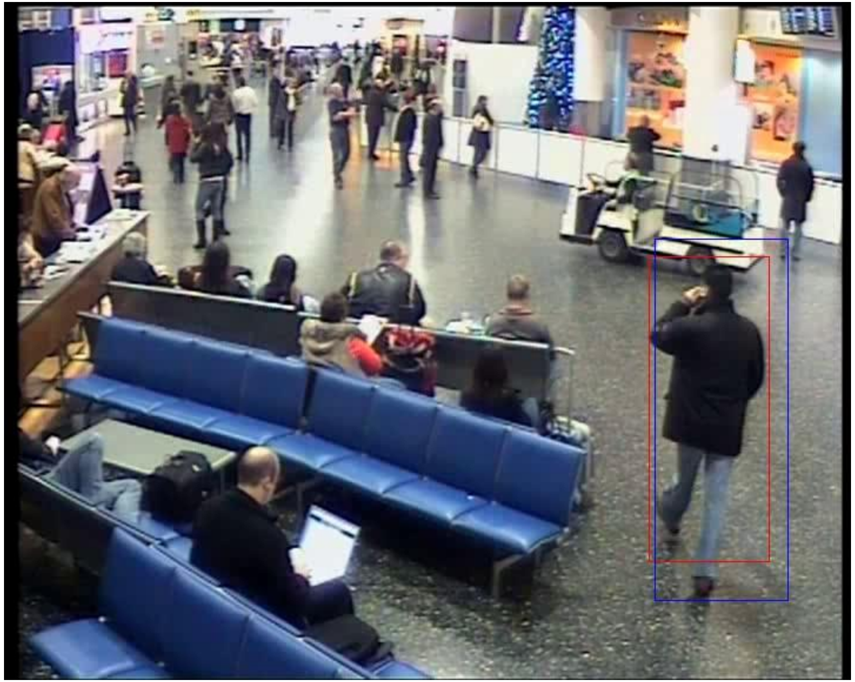
need key point information to guide the model to attend to certain regions (e.g. palm, elbow and shoulder)

predict score: 0.95

need additional motion information to solve such cases

Cell2Ear case study (true positive)

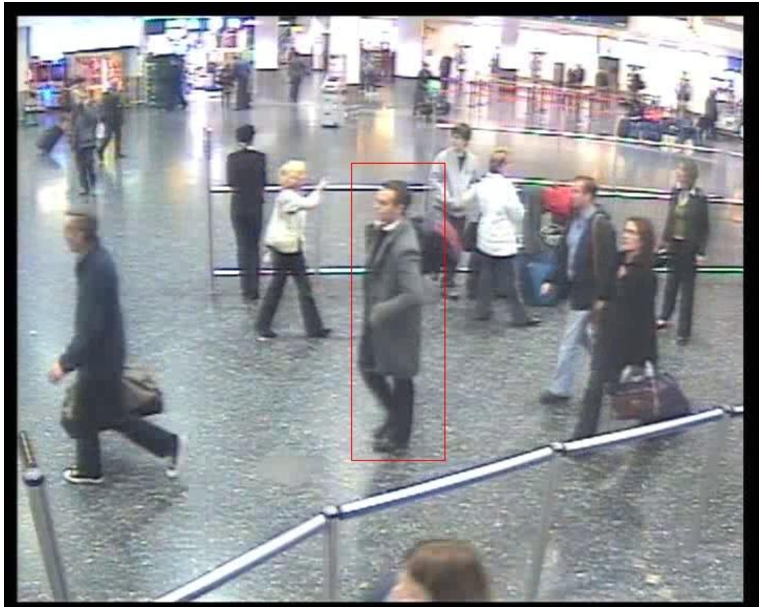
predict score: 0.49



predict score: 0.25

Cell2Ear case study (false positive)

predict score: 0.88



need additional motion information to solve such cases

predict score: 0.88

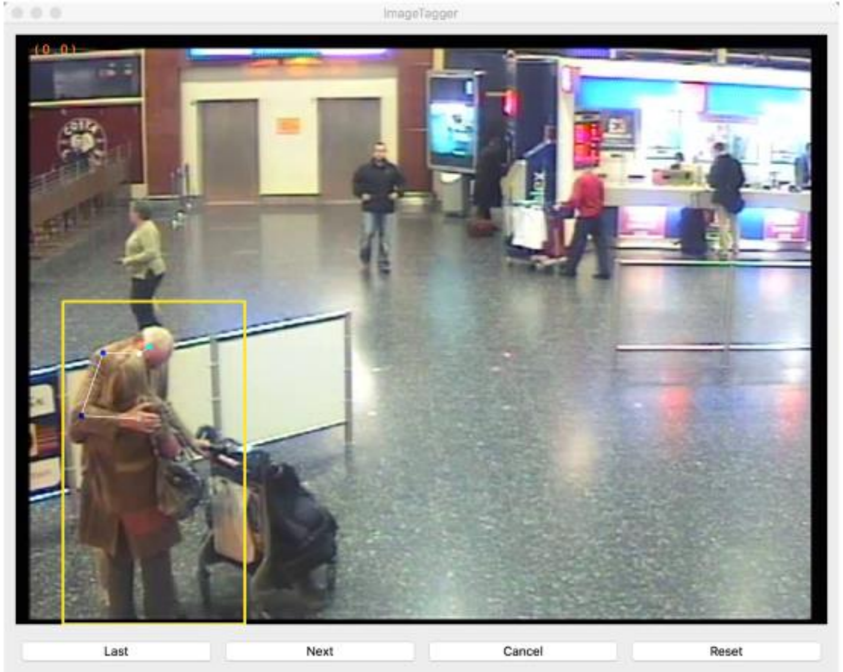
need key point information to guide the model to attend to certain regions (e.g. palm, elbow and shoulder)

Preliminary experiment to verify the need of skeleton key-points

- sample 900 images
  - Embrace: 150 (100 for train and 50 for test)
  - Pointing: 150 (100 for train and 50 for test)
  - Cell2Ear: 150 (100 for train and 50 for test)
  - Other: 450 (100 for train and 150 for test)

Preliminary experiment to verify the need of skeleton key-points

- Manually label the key point pose
  - Head, neck, L shoulder, R shoulder, L elbow, Relbow, L palm, R plam



Preliminary experiment to verify the need of skeleton key-points

- keypoint information alone performs 10% over appearance information alone
- keypoint position + global appearance fail to improve over key point position alone (need attention-based approach)

feature	accuracy (%)
keypoint position	66.0
appearance	56.3
keypoint position + appearance	59.3

Conclusion and future work

- Pose based approach for static action type is promising
- Need key-point poses for better performance
- Combining pose detection with motion
  - Using pose detection with motion features can solve some of the hard cases in single frame key pose detection alone
- 3-D reconstruction is necessary for interaction events such as Embrace