VideoStory

At TRECVID 2014

Amirhossein Habibian, Thomas Mensink, Cees Snoek

MediaMill | ISLA, University of Amsterdam

Acknowledgement

This research is supported by the STW STORY project, the Dutch national program COMMIT, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

Problem statement

Recognize and translate video events

- Learning from few examples
- Provide semantic interpretation of videos

Event Attempting bike trick

Video

Description

Recognizing events

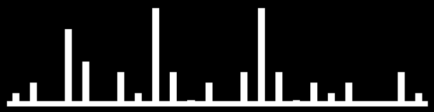
Representing videos as histograms of low-level features

Local descriptors

- Visual descriptors
  - SIFT, HOG, GIST, …
- Video descriptors
  - MBH, STIP, …
- Audio descriptors
  - MFCC, AIM, …

Feature embedding

- Bag-of-words
- VLAD
- Fisher vector
- Audio-visual BoW

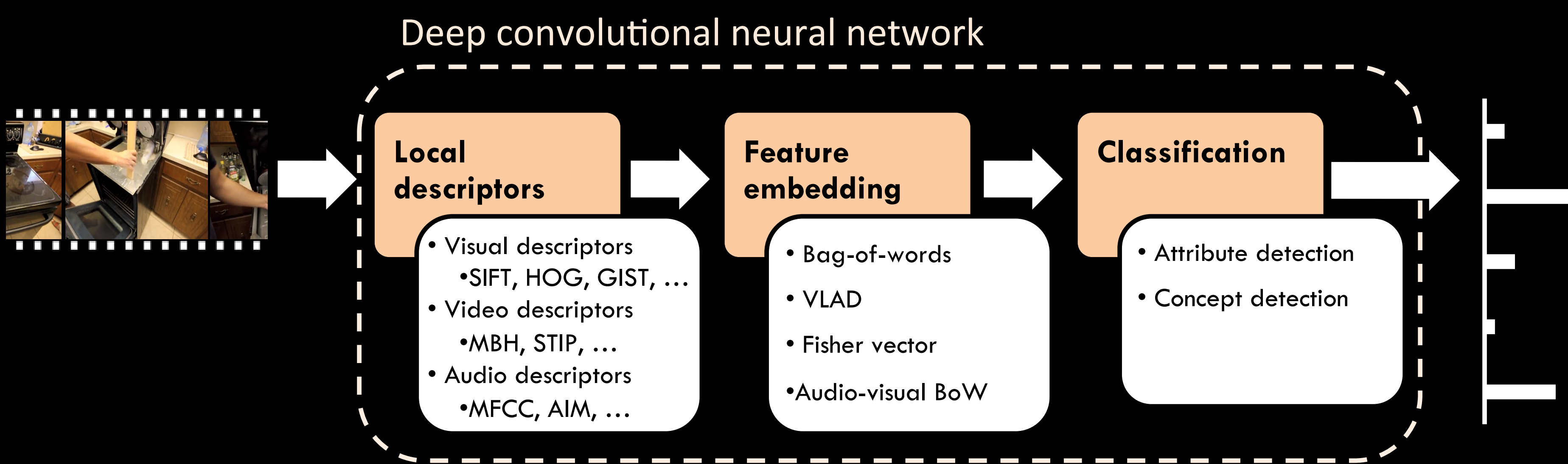


Problem: very high-dimensional and non semantically

[Jiang et al., TRECVID 2010] [Natarajan et al., CVPR 2012] [Chen et al., MM 2013]

Recognizing and translating events

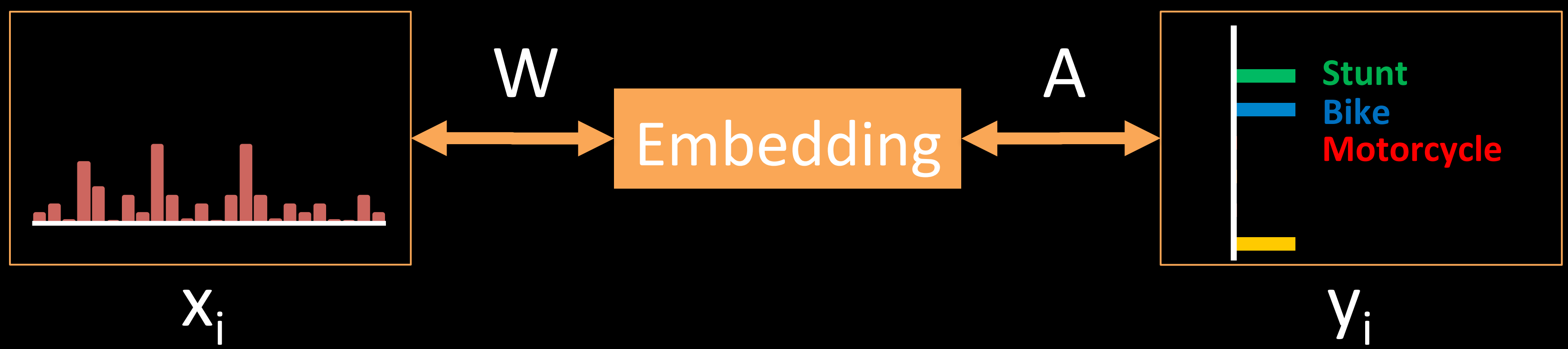
Representing videos as histograms of concept scores



Problem: define, annotate and train concept classifiers

[Smith et al., ICME 2003] [Hauptmann et al., TMM 2007] [Merler et al., TMM 2012] [Ma et al., MM 2012]

Recognition and translation by embedding



Joint space where $x_i W \approx y_i A$

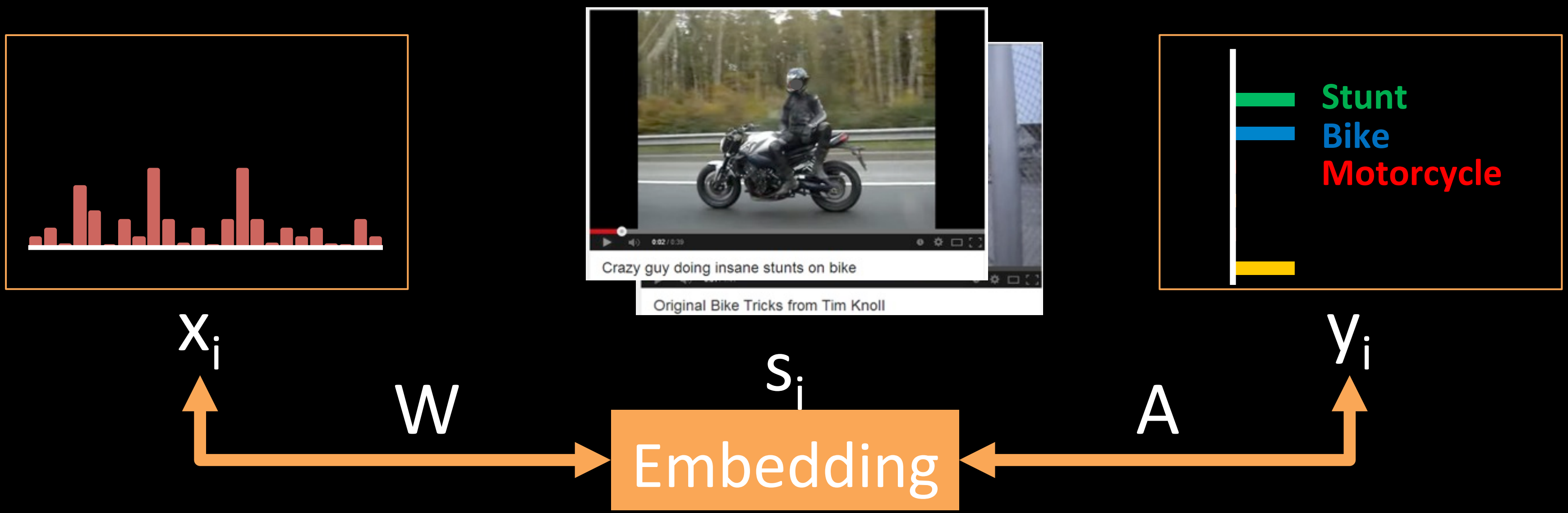
Explicitly relate training W and A from multimedia

A = Identity matrix individual term classifiers

A = Projection matrix select/group terms

[Rasiwasa et al., MM 2010] [Weston et al., IJCAI 2011] [Akata et al., CVPR 2013] [Das et al., WSDM 2013]

VideoStory: Embed the story of a video



Design criteria: learn W and A such that

Descriptiveness: preserve video descriptions

Predictability: recognize terms from video content

Key observation: Compelling forces

Crazy guy doing insane stunts on bike

Why is this important?

Grouping terms:

Number of classes is reduced

Training classifiers per group:

More positive examples available per group

We can train from freely available web data

Key contribution: Joint optimization

Jointly optimize for descriptiveness and predictability

$$L_{\mathrm{VS}}(\boldsymbol{A}, \boldsymbol{W}) = \min_{\boldsymbol{S}} L_d(\boldsymbol{A}, \boldsymbol{S}) + L_p(\boldsymbol{S}, \boldsymbol{W})$$

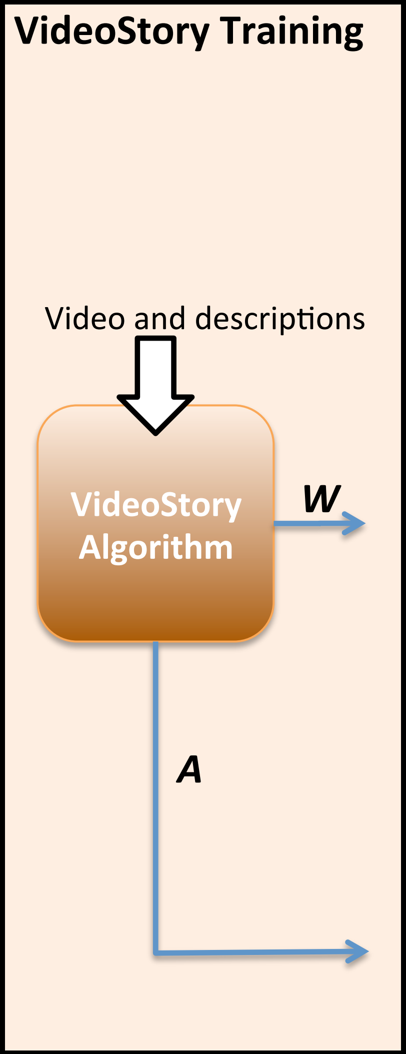
Hyperparameter: size of the embedding S

L_d Loss function for descriptiveness

L_p Loss function for predictability

VideoStory connects the two loss functions

VideoStory: Training



Set of videos and their captions

Encode video features x_i

Fisher Vectors of MBH [Wang ICCV'13]

Encode video descriptions y_i

Bag-of-words of terms

Train using *Stochastic Gradient Descent*

YouTube46K dataset

Videos and title descriptions from YouTube

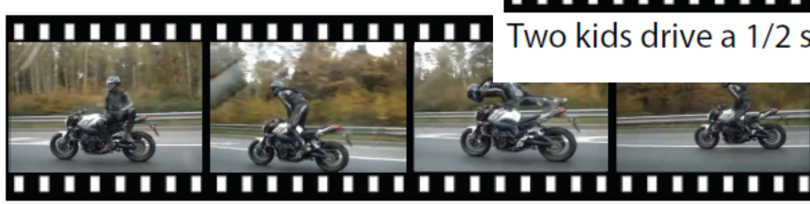
46K videos, 19K unique terms in descriptions

Seeded from video event descriptions

Filters to remove low quality videos

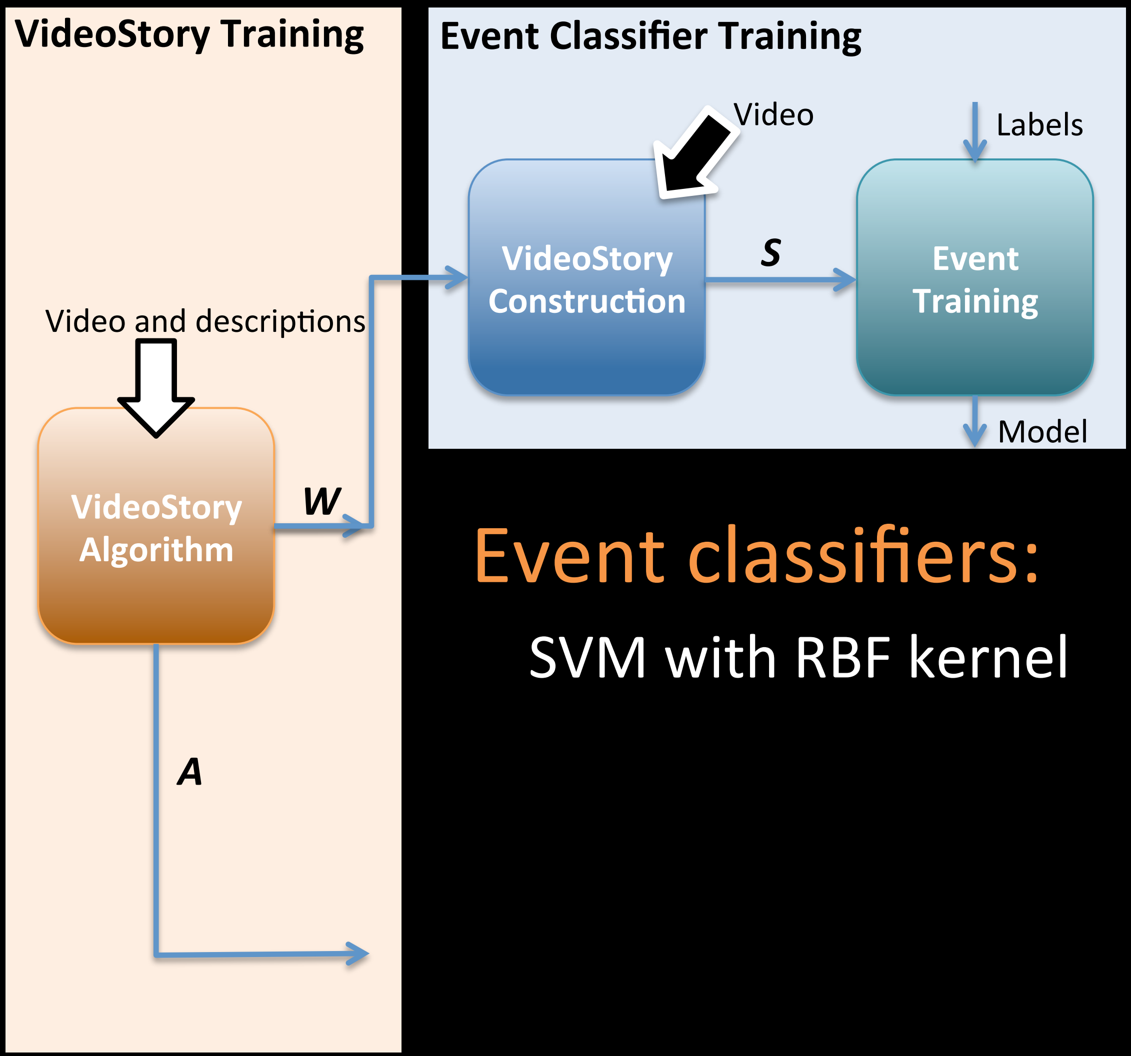

Cute tabby cat gives her dog a bath


Two kids drive a 1/2 size Jeep through mud


Crazy guy doing insane stunts on bike.

Available for download: www.mediamill.nl

VideoStory: Event classifier training



Event classifiers:

SVM with RBF kernel

Datasets for evaluation

TRECVID Multimedia Event Detection 2013

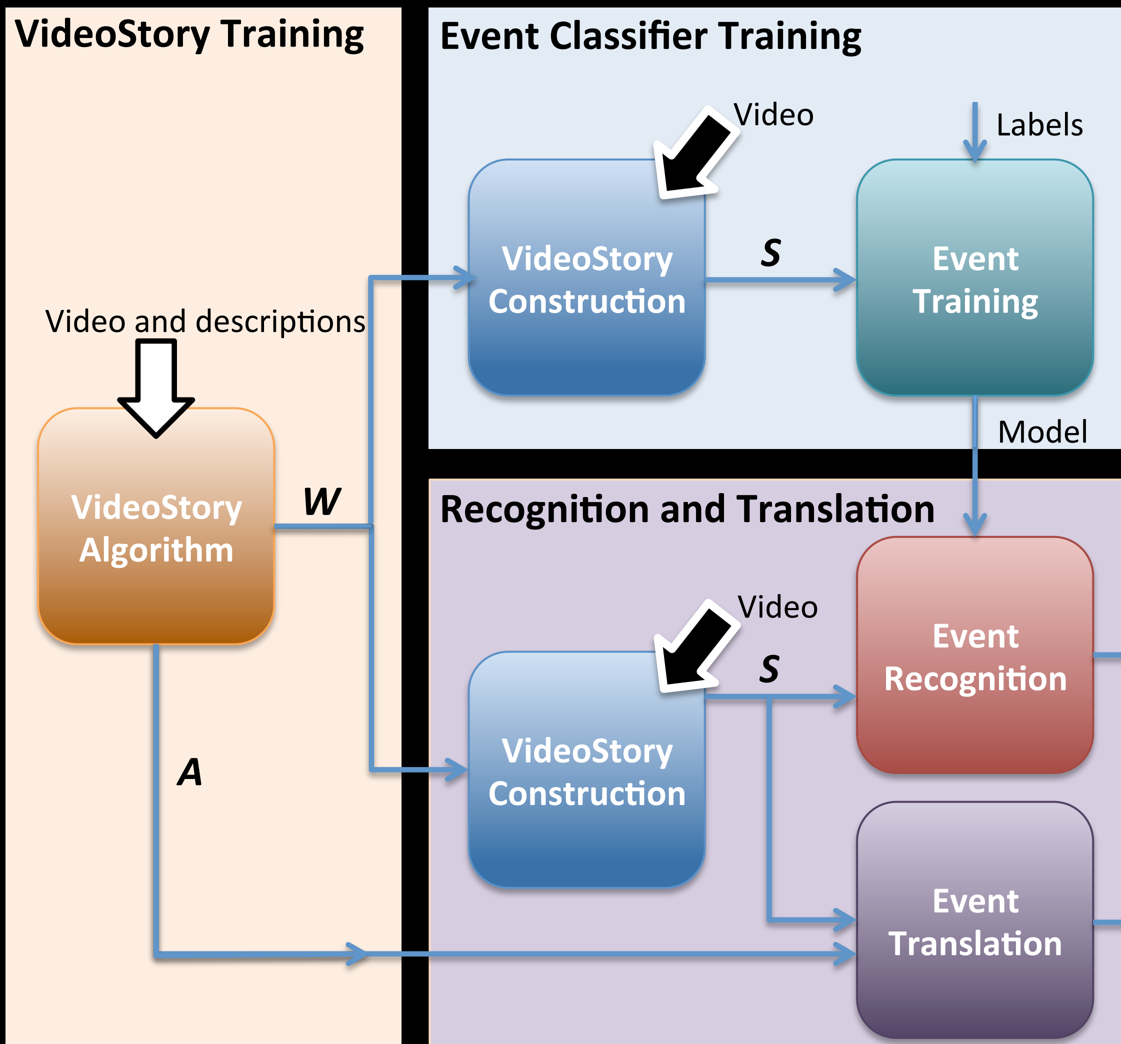
56K videos - 20 events - 10 positives train videos

Columbia Consumer Video

9K videos - 15 events - 10 positives train videos

[Jiang et al. ICMR 2011][Strassel et al. LREC 2012]

VideoStory: Recognition and translation



Evaluation:

Event score

Mean Average Precision

Description

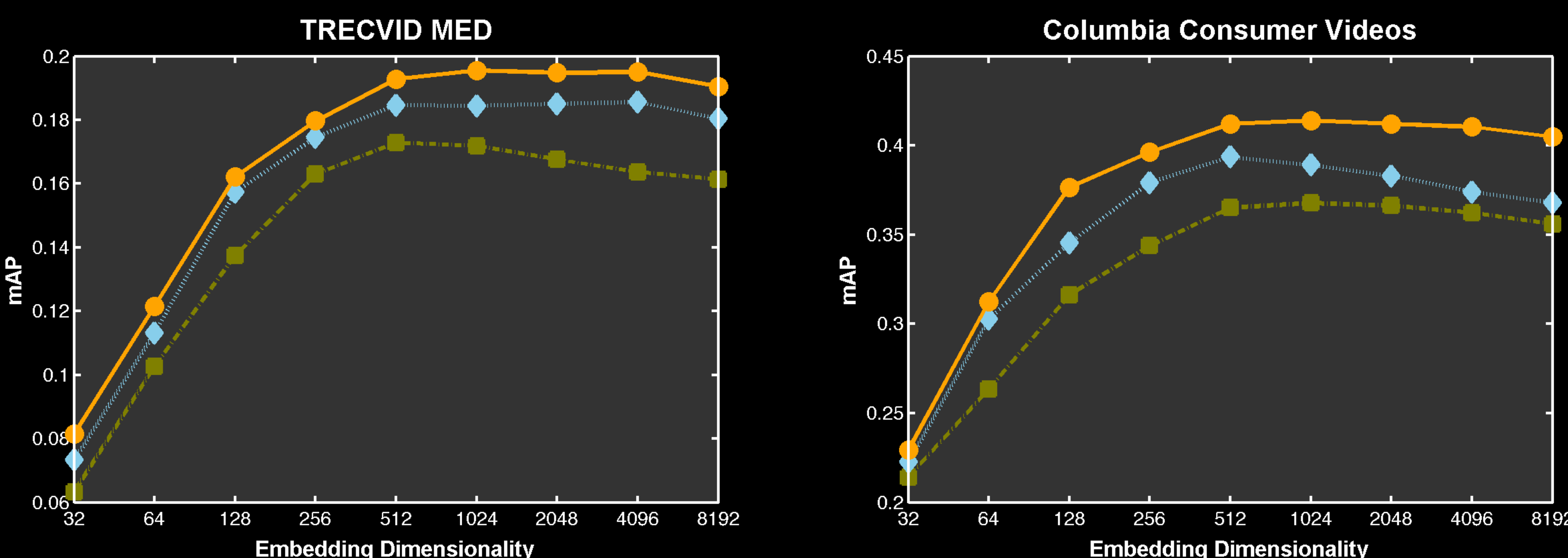
Rouge-1

Experiment 1: Effect of Embedding

Frequent terms: train classifier for most frequent terms

Grouping first: first descriptiveness; then predictability

VideoStory: joint descriptiveness and predictability



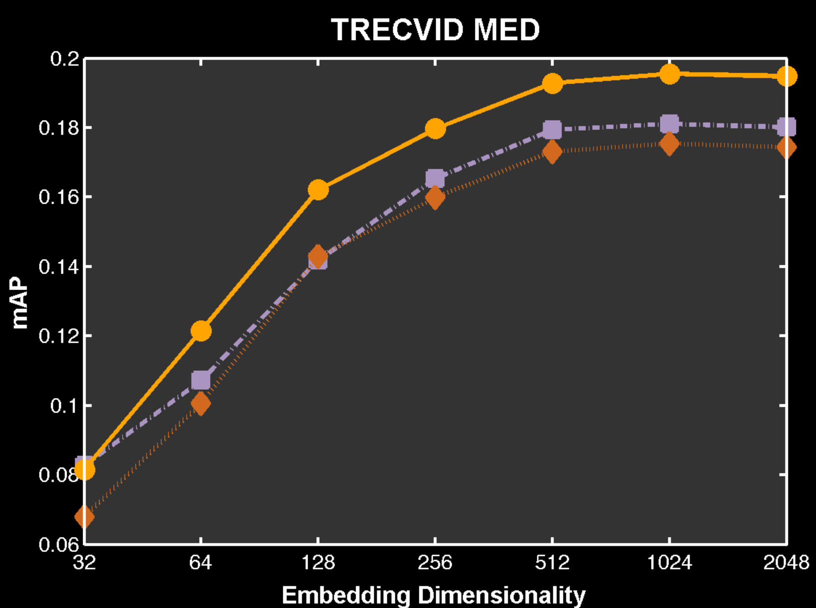
VideoStory outperforms other embeddings

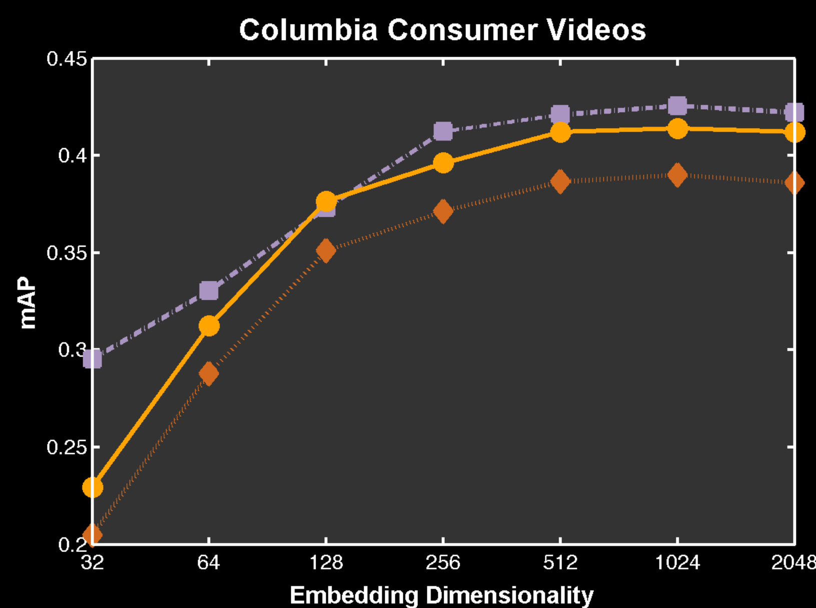
Experiment 2: Story Quality vs. Quantity

Expert10K: 10K TRECVID videos with expert descriptions

YouTube10K: 10K random subset of YouTube46K dataset

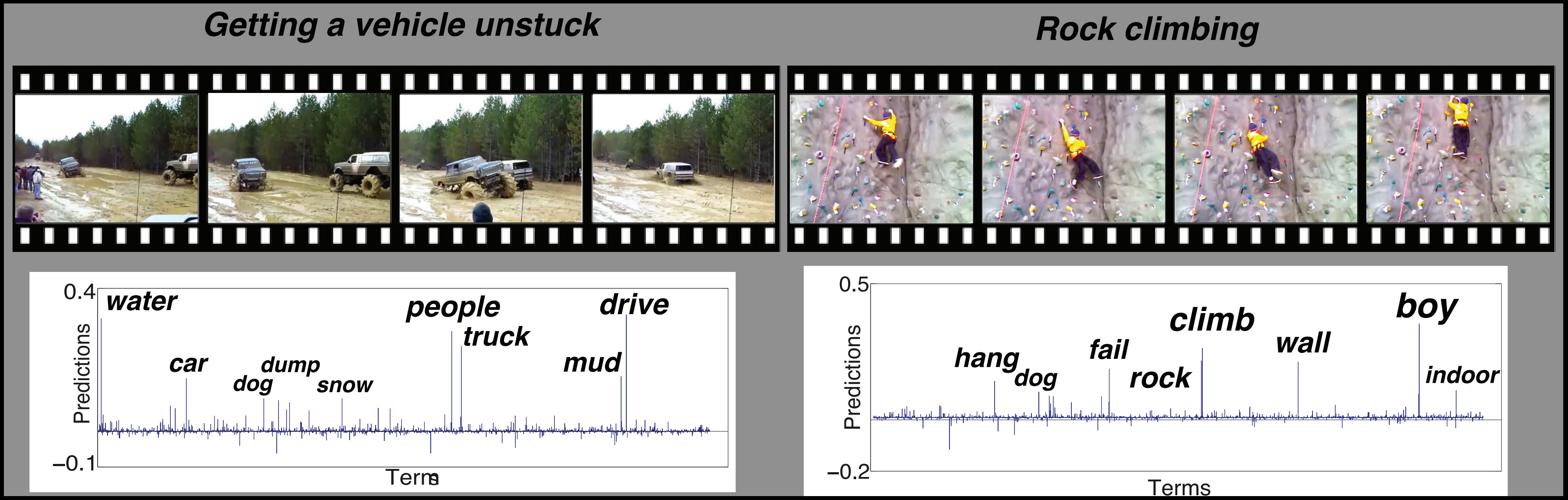
YouTube46K: 46K YouTube videos and descriptions





Web supervision on par with expert provided descriptions

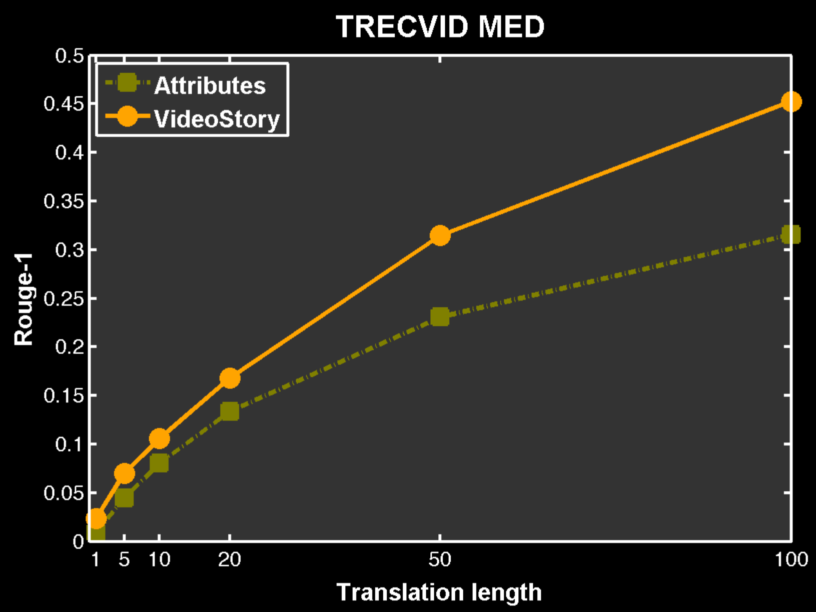
Experiment 3: VideoStory translation



Experiment 3: VideoStory translation

Evaluate on TRECVID MED

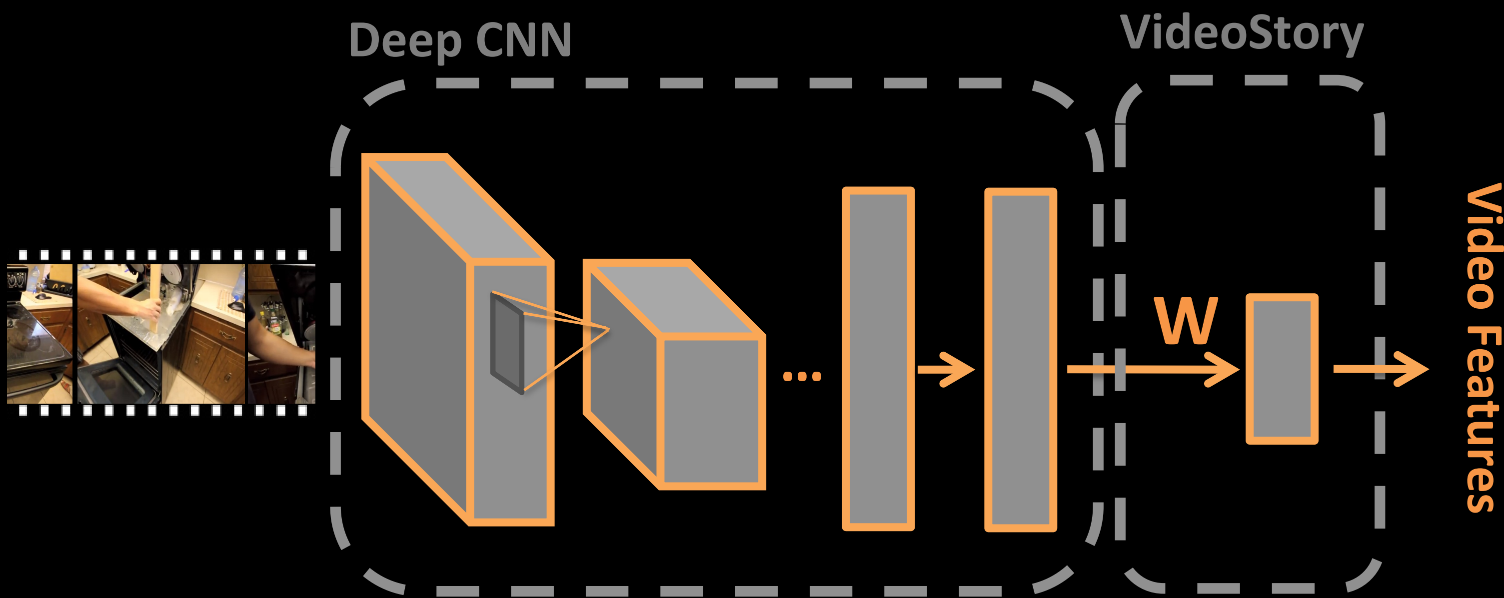
- Ground-truth: provided descriptions
- Measure with ROUGE-1



VideoStory outperforms predefined attributes

VideoStory at TRECVID

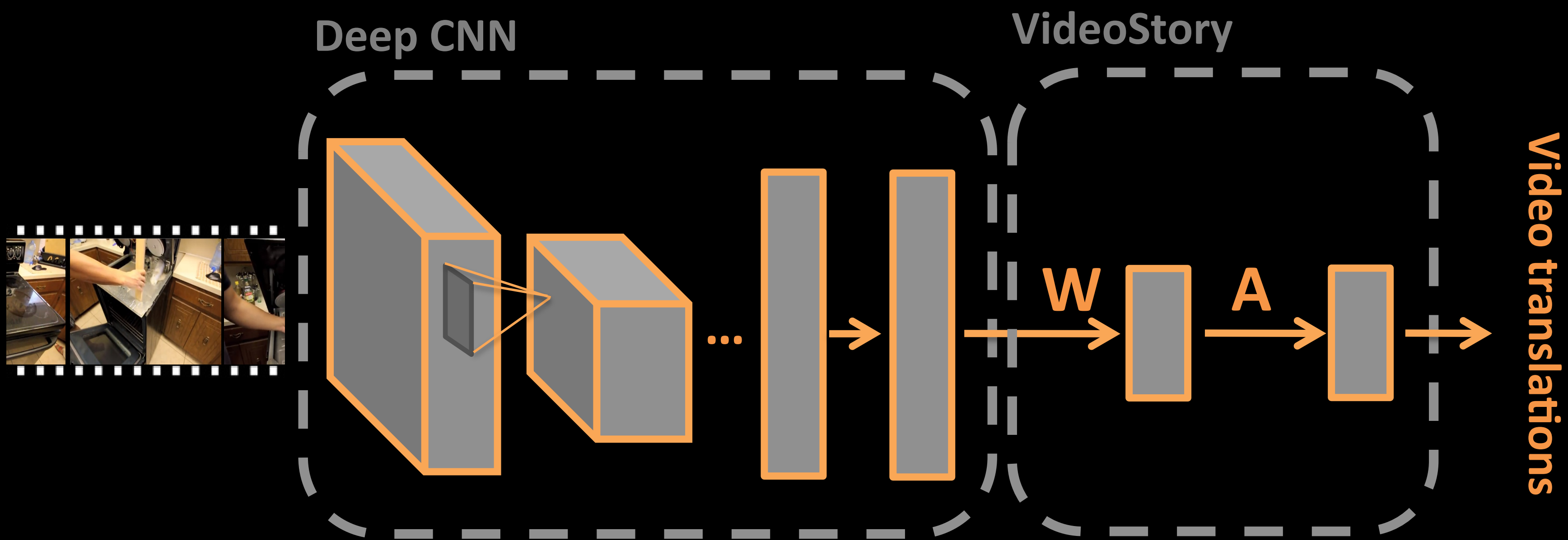
VideoStory recognition at TRECVID 2014



Features for training event classifiers

- Example based event search (10Ex and 100Ex)
- We train SVM with RBF kernel

VideoStory translation at TRECVID 2014



Translations for matching with event definition

- Text based event search (0Ex)
- We use cosine similarity

Computational efficiency

Fast feature computation

- Convolution and multiplication over pixel values
- ~54 secs per video mostly spent on video decoding

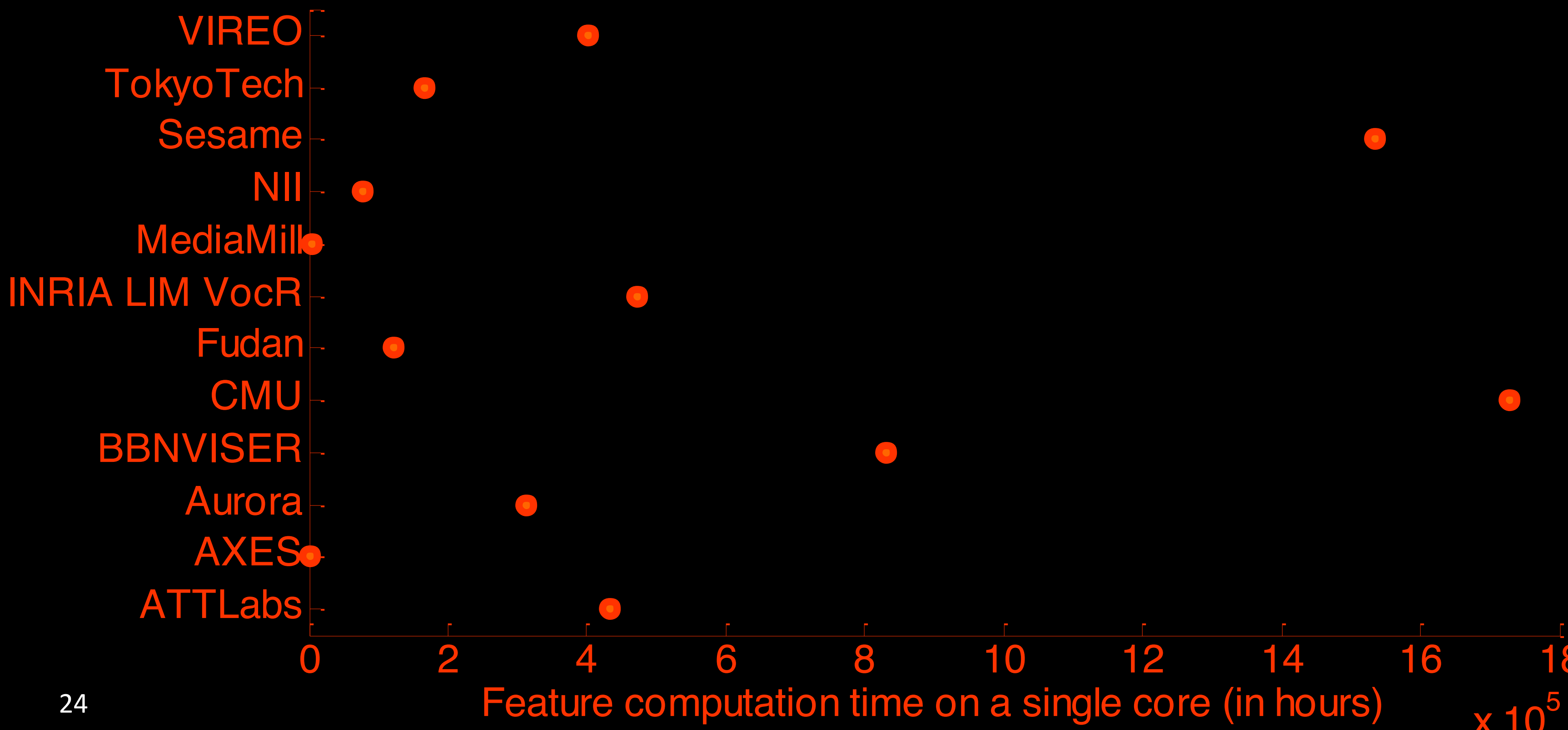
Fast event classifier train and test

- Training < 60 s per event
- Classifying one test video only 0.015 s
- 1K-dimensional video representation

Efficiency: feature computation

Time to compute the features for MED14Full

Takes 259 hours on a single machine with 16 cores



Event recognition accuracy

		MED 13	MED 14
100Ex	CNN features	.350	.280
	VideoStory - CNN	.389	.300
010Ex	CNN features	.198	.167
	VideoStory - CNN	.243	.200
000Ex	VideoStory - CNN	.124	.037

Competitive accuracy with a <u>single</u> feature only

Conclusions

VideoStory a semantic multimedia embedding

- Jointly optimizes descriptiveness & predictability
- Training event classifiers from few examples
- Translate videos to textual description

Effectively and efficiently recognizes events

Adds meaning to deep convolutional networks

Thank you!