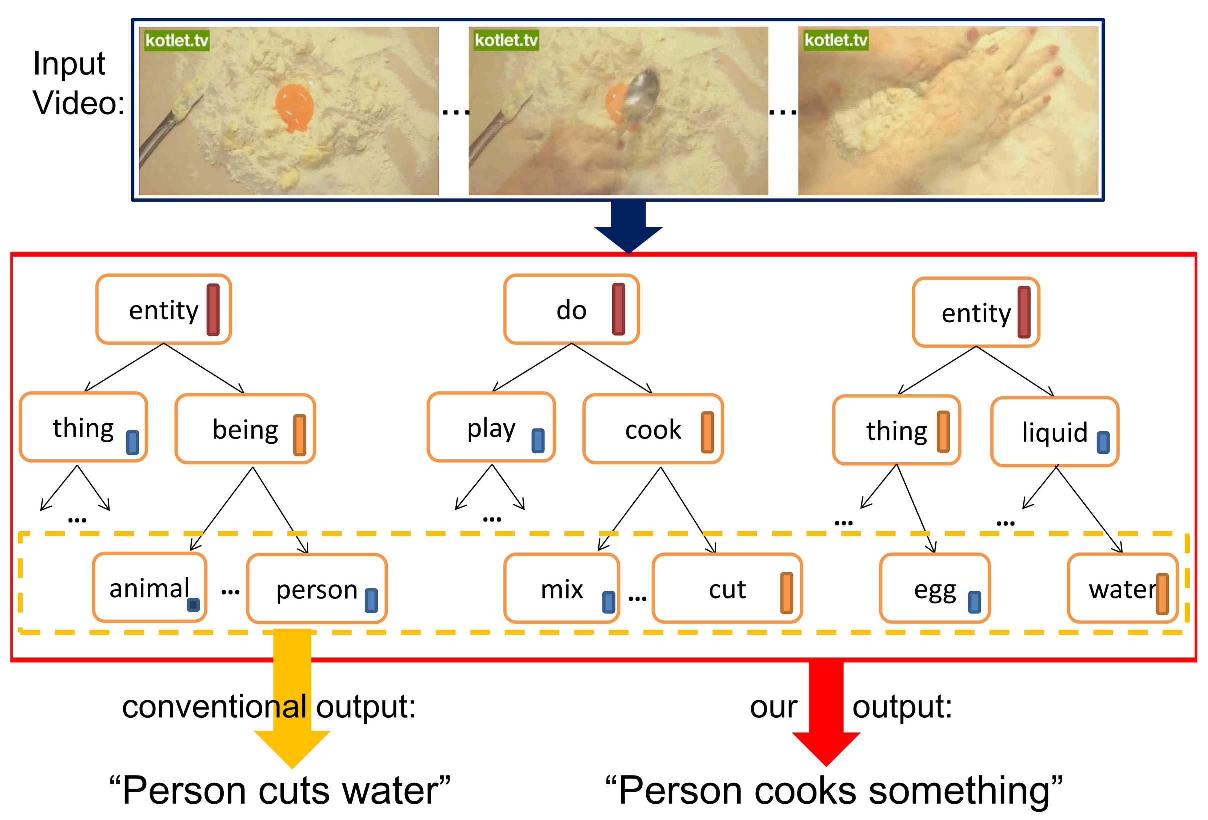
YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition

Sergio Guadarrama[1], Niveda Krishnamoorthy[2], Girish Malkarnenkar[2], Subhashini Venugopalan[2], Raymond Mooney[2], Trevor Darrell[1], Kate Saenko[3]

[1] UC-Berkeley [2] UT Austin [3] UMass Lowell

Goals

Given a short YouTube video, output a natural language sentence that describes the main activity in the video.
When the model is not confident enough it should produce a less specific answer, but not over generalize.



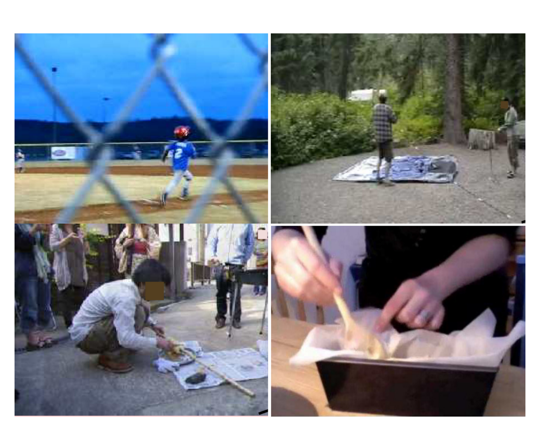
Humans: "A woman is mixing an egg", "Someone is making dough"

Conventional methods try to predict a caption composed of the most visually likely objects and actions (leaf nodes), whereas our method can predict a less specific phrase that is nonetheless visually plausible and informative. The bars inside nodes indicate the posterior probability of the node given the input video (more red and taller indicates higher probability).

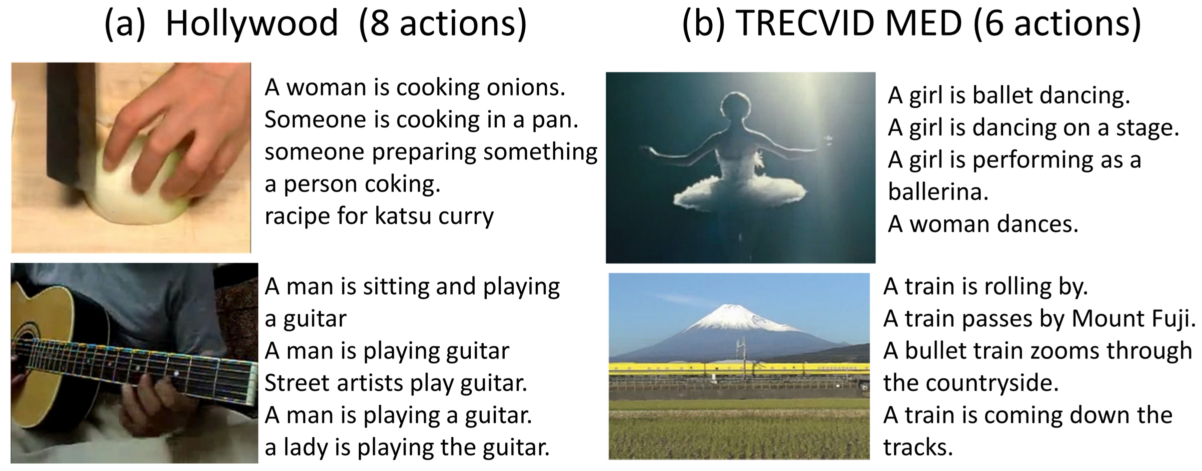
YouTube Dataset

We use the YouTube dataset collected by (Chen and Dolan, ACL 2011) consisting of 1970 videos and around 41 sentences on average per video, see (c) below

(a) Hollywood (8 actions)



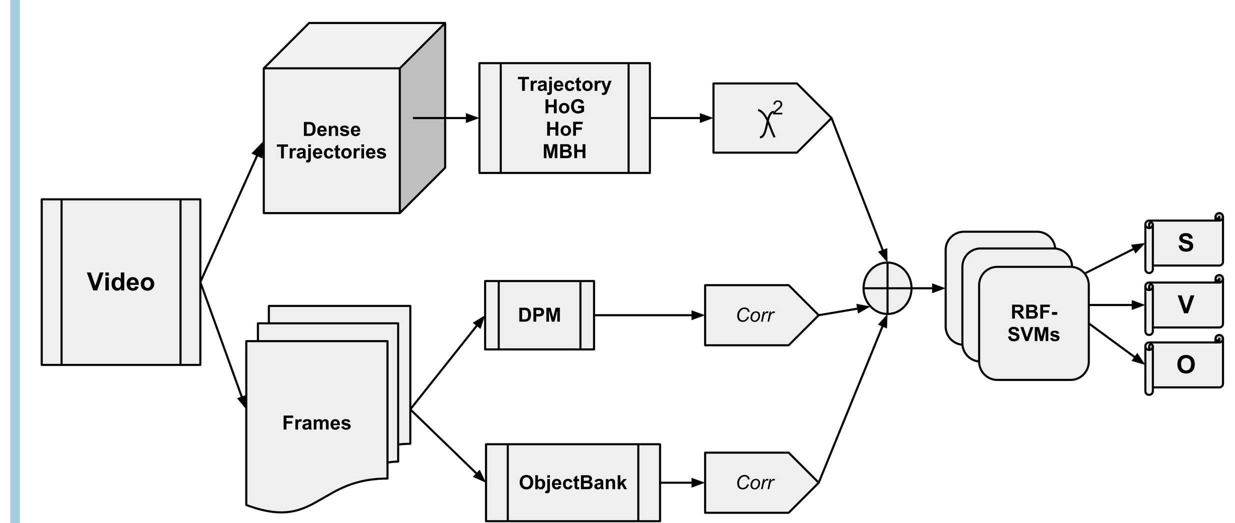
(b) TRECVID MED (6 actions)



(c) YouTube (218 actions)

This new dataset (c) contains many more actions than the other previously used activity datasets (a-b).

Vision Pipeline

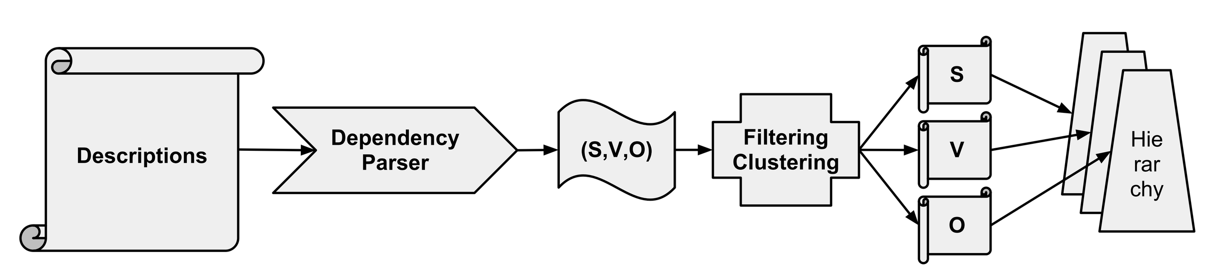


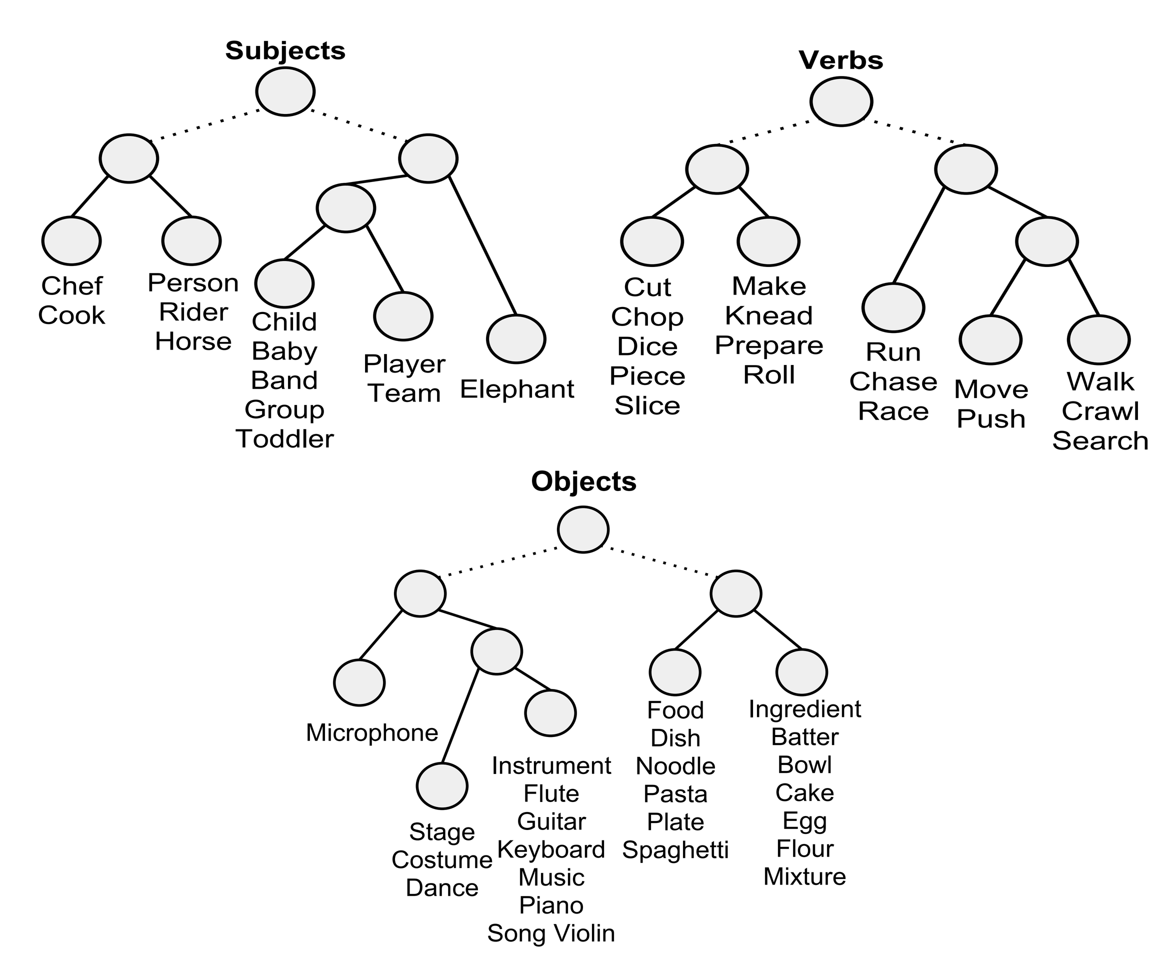
$$K(x_i, x_j) = \exp\left(-\sum_c \frac{1}{A^c} D_c(x_i^c, x_j^c)\right) \quad (1)$$

The outputs are over the leafs of the Hierarchies

Learning Hierarchies

Language Pipeline





Small portions of the Hierarchies learned over Subjects, Verbs and Objects

Defining Semantic Accuracy

Given a Hierarchy of labels and a matching function μ_{L_t} the accuracy $\phi_H(f)$ over a hierarchy H with respect to a ground truth set leaf nodes $L_t \subset L$ is defined by:

$$\mu_{L_t}(v) = \max_{l \in L_t} \{s_t(v, l)\} \quad (2)$$

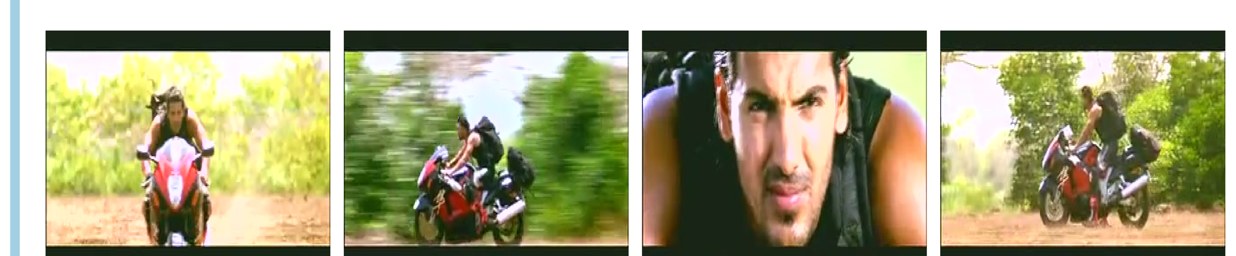
$$s_{\mathrm{WUP}}(v, l) = \frac{2 \cdot \mathrm{depth}(lcs)}{\mathrm{depth}(v) + \mathrm{depth}(l)} \quad (3)$$

$$\phi_H(f) = \mathbb{E}[\mu_{L_t}(f(X))] \quad (4)$$

Qualitative Results

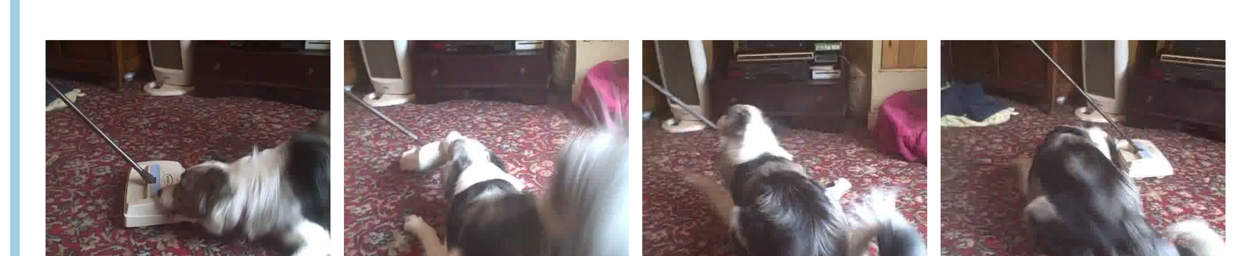
GT: *A woman is mixing some egg with flour.*
FL: A person cuts the water.
OU: **A person cooks something.**
HE: A person does something.

GT: *A cat is playing with a ferret.*
FL: A person plays a water.
OU: **An animal plays something.**
HE: An animal does something.



GT: *A man is riding a motorcycle.*
FL: A person rides a person.
OU: **A person rides a vehicle.**
HE: The person does something.

GT: *A toy train runs into a toy car.*
FL: A car rides the motorbike.
OU: **A car rides the vehicle.**
HE: Someone does something.



GT: *A dog is attacking a vacuum.*
FL: A dog plays a water.
OU: **An animal does something with the instrument**
HE: An animal does something.

GT: *A baby panda is climbing a step.*
FL: The cat plays with the water.
OU: **An animal plays an instrument.**
HE: An animal does something.

Binary 0-1 Accuracy

Method	0-1 Loss S%	V%	O%
Prior	78.36	13.43	6.12
FL / HE	78.51	22.09	12.84
OU	**80.90**	**29.10**	**17.01**

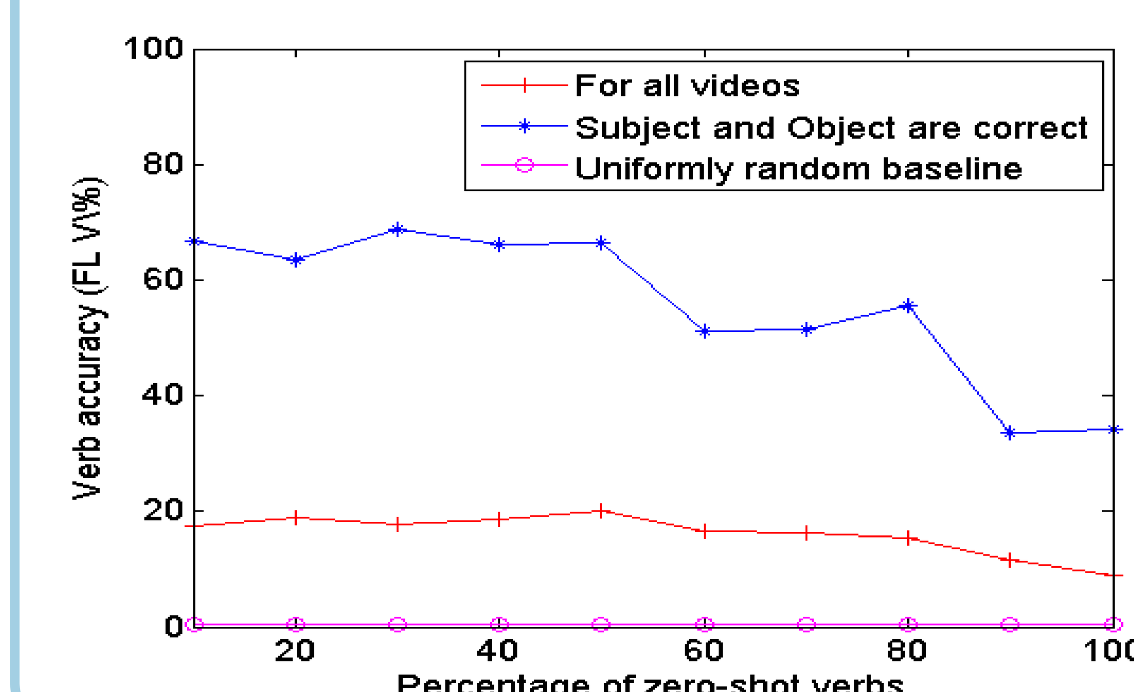
Prior:Most Frequent triplet, FL:Flat classifiers, HE: Hedging your bets, OU:first level of our semantic hierarchies.

Comparison of WUP Similarity

Alg	WUP Similarity: Most Common S%	V%	O%	Valid Answer S%	V%	O%
FL	88.94	43.56	36.77	93.28	59.52	51.91
HE	78.13	31.29	23.37	81.03	45.71	28.45
OU	**92.57**	**46.83**	**46.66**	**93.72**	**61.19**	**58.41**

FL:Flat classifiers, HE: Hedging your bets, OU:Our method.

Zero-shot Activity Recognition



Human Evaluation

We use Amazon Mechanical Turk to compare the methods by evaluating them on a video retrieval task.

Retrieval Method	FL	HE	**OU**	Ground Truth
Average Rating	1.81	1.54	**1.99**	3.90

The differences in the ratings of the three systems are statistically significant.

Conclusions

We presented a system that takes a short video clip "in-the-wild" and outputs a brief sentence that sums up the main activity in the video, such as the actor, the action and its object.
The semantic hierarchies learned from the data help to choose an appropriate level of generalization, and a prior learned from web-scale natural language corpora penalizes unlikely combinations of actors/actions/objects.