Asynchronous Temporal Fields for Action Recognition

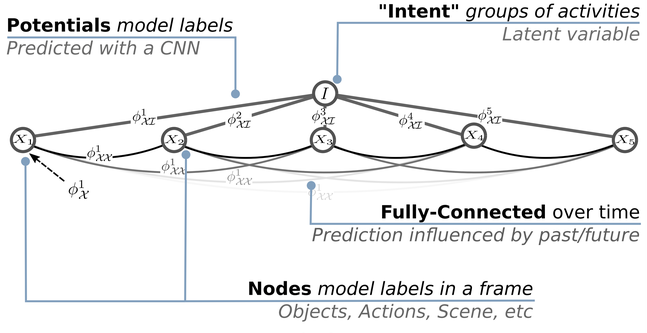
Gunnar A. Sigurdsson, Santosh Divvala, Ali Farhadi, Abhinav Gupta

Carnegie Mellon University — The Robotics Institute

AI2

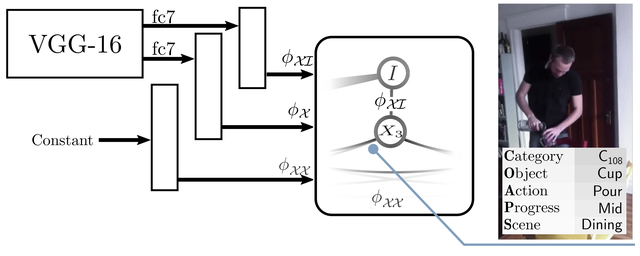
Model

- We model activites as a fully-connected CRF over time
- Nodes for each frame in the video, plus "intent"

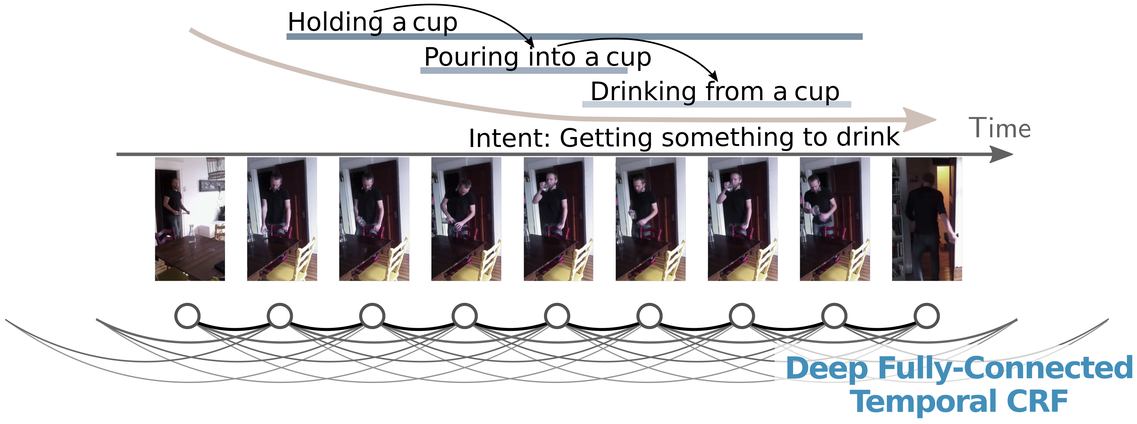


Zooming in...

- A Two-Stream network predicts the potentials



Understanding Sequences of Activities

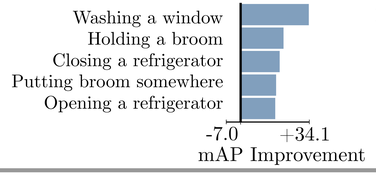


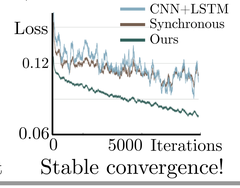
High-order Temporal Model Trained From Randomly Sampled Individual Images

Results

- Evaluation on the Charades dataset
  157 classes, 66k instances, object, verb, scene, captions, etc

Approach	mAP
RGB++	15.6
Two-Stream++	16.8
Two-Stream+LSTM	17.8
Two-Stream Extended	18.6
Ours (RGB Only)	18.3
Ours	**22.4**

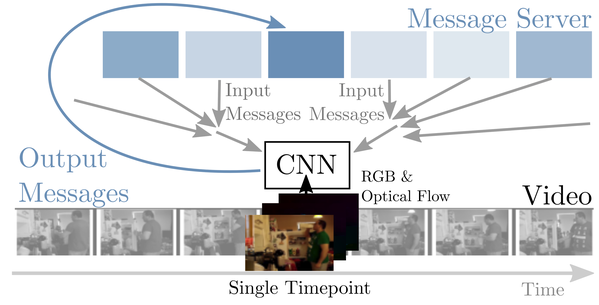




github.com/gsig

Training Method

- With *Mean-Field*, training is simple:
  1. Sample any frame
  2. Get stored "messages"
  3. Backprop using just this frame



Learned "Intent"

- We model a latent variable for groups of activities
- Learns higher-level activity concepts: