Large-Scale Unsupervised Understanding of Multi-Modal Recipes

Anonymous ICCV submission

Paper ID ****

Abstract

The ABSTRACT

1. Introduction

Para1: Humans and Robots both need a better multi-modal recipe representation

Para2: Current computer vision and NLP fails because scaling data, limited language etc.

Para3: We are Large-Scale, Holistic, Semantic, Multi-Modal and Unsupervised.

Para4: Basic idea with pointers and contributions

2. Related Work

Para1: Activity

Para2: Recipe understanding

Para3: Robotics

Para4: Video Summarization

3. Overview

Explanation of the Figure 3 with pointers to the Sections.

4. Semantic Multi-Modal Frame Representation

4.1. Learning Atoms to Represent Multi-Modal Data

4.1.1 Learning Language Atoms

4.1.2 Learning Visual Atoms

Generating Framewise Object Proposals

Joint Proposal Clustering to Detect Visual Objects

4.2. Multi-Modal Representation via Learned Atoms

5. Unsupervised Activity Representation

5.1. BP-HMM with Orderings (BP-HMM-O)

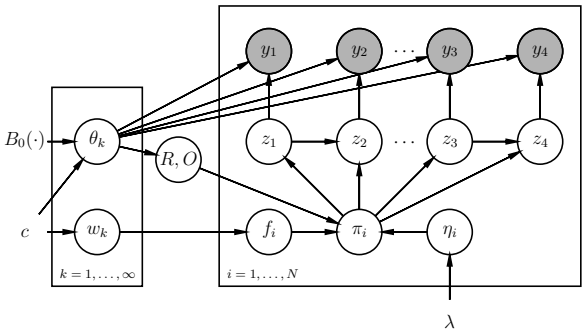

Figure 1. Graphical model for BP-HMM-O

5.2. Gibbs sampling for BP-HMM-O

Mostly deferred to supplementary, just say we ne need to sample a,b,c,etc. and we used MH for binary and Gibbs for continous/discrete ones.

6. Experiments

6.1. Dataset

We collected blah blah videos from YouTube and did blah blah... All numbers etc. These are recipes, 5 of them are evaluation set and labelled temporally and labels are matched.

6.2. Baselines

We compare against BP-HMM-O with local feauters and HMM with Semantic Features, HMM with local features, HMM with CNN feautes...

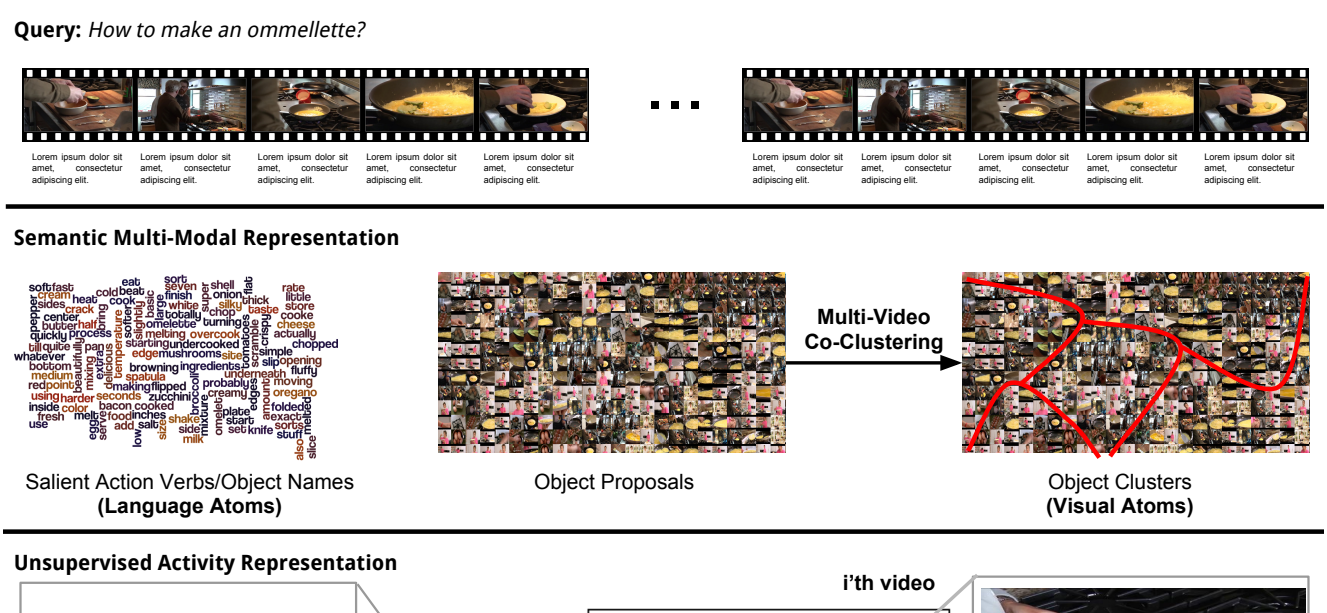

Table 1. Notation of the Paper

$y_t = [y_t^v, y_t^l]$	feature representation of t^{th} frame	I_t	t^{th} frame of the video	$x_{i,r}^p$	1 if p^{th} cluster has r^{th} proposal of i^{th} video
x^p	binary vector for p^{th} cluster	L_t	subtitle for t^{th} frame	$O^{k,k'}$	1 if $\#(z_t = k, z_{t'} = k') = 0 \forall \quad t \leq t'$
$\Theta_k = [\Theta_k^v, \Theta_k^l]$	emmition prob. of k^{th} activity	z_t	activity ID of frame t	f_i^k	1 if i^{th} video has k^{th} activity $0o.w.$
$\eta_i^{k,k'}$	$P(z_{t+1} = k' \| z_t = k)$ for i^{th} vid	$\pi_i^{k,k'}$	$\eta_i^{k,k'} \times f_i^k \times f_i^{k'}$		

6.3. Qualitative Results

6.4. Accuracy over Activity Detection

6.5. Accuracy over Activity Learning

7. Discussions and Conclusions

Discuss which recipes worked and why. Discuss the importance of semantic representation, scaling features and multi-modality.