

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

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

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

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...

108

Query: How to make an ommellette?

162

109

163

110

164

111

165

112

166

113

167

114

168

Semantic Multi-Modal Representation

169

115

170

116

171

117

172

118

173

119

174

120

175

121

176

122

177

123

178

124

Unsupervised Activity Representation

179

125

180

126

181

127

182

128

183

129

184

130

185

131

186

132

187

133

188

134

189

135

190

136

191

137

192

138

193

139

194

140

195

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 $\#\{z_t = k, z_{t'} = k'\} = 0 \forall t \leq t'$
$\Theta_k = [\Theta_k^v, \Theta_k^l]$	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 $Oo.w.$
$\eta_i^{k,k'}$	$P(z_{t+1} = k' z_t = k)$ for i^{th} vid	$\pi_i^{k,k'}$	$\eta_i^{k,k'} \times f_i^k \times f_i^{k'}$		

6.3. Qualitative Results

200

6.4. Accuracy over Activity Detection

201

6.5. Accuracy over Activity Learning

202

7. Discussions and Conclusions

203

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

204

205

206

207

208

209

210

211

212

213

214

215