Towards Video Graph Relation-Aware Association

Xindi Shang, NUS
• Video Graph

• VidOR: a Video Graph Database

• Relation-Aware Association: a Method

• Applications
• Single image understanding tends to be mature
  • Deep neural networks
  • Large-scale annotated image database
• Video understanding
  • Not a set of images
  • Action/event is not defined in a single frame
• Video understanding
  • Observation in frames
  • Linking the observations in each frame and across frames
  • Form structure
• Structure Representation
  • What happen
  • Who participate
  • When
  • Where

• For infer high-order relations
  • Temporal relation
  • Causality relation

Video Graph

- watch
- next to
- bite
- cause
- simultaneous
- has_subject
- has_object
- observed_at
- 5-7s
- 1-10s
- 6-7s
- 1-9s
- 7-8s

Dog.1
Dog.2
Dog.3
Adult.4
OUTLINE

• Video Graph

• VidOR: a Video Graph Database

• Relation-Aware Association: a Method

• Applications
VidOR: a Video Graph Database

- 10K videos from Flickr (YFCC-100M)
  - Indoor, outdoor
  - Human centric
  - Various events
- Annotated with video graph
• 10K videos from Flickr
  • User generated
  • 30s long in average, 84 hours in total, 10K Flickr videos ≈ 1000*10K (10M) frames ≈ Size of ImageNet

• Annotating a video graph is a very difficult annotation task
  • Annotated 50,000 object entities
  • Annotated 380,000 relation instances

• How to design an annotation pipeline
  • Balance between task difficulty and management difficulty
  • Balance between cost and quality
Interactive Key-frame Generation

Tracking successful

Tracking failed
Interactive Key-frame Generation

Tracking successful

Tracking failed
• Video Graph

• VidOR: a Video Graph Database

• Relation-Aware Association: a Method

• Applications
Method to Extract Video Graph

- Visual relation extraction in videos
• Visual relation extraction in videos
Method to Extract Video Graph

- (Multiple) object tracking in videos

Relation-Aware Association

- Data association
  - Visual relation extraction
  - Multi-object tracking
  - Both just solve a single sub-problem

- Video graph extraction via relation-aware association
  - Relation recognition using the object context
  - Handle object occlusion using relation context
Multiple hypothesis association
  - handle missing detection, wrong classification of observations in frame

Figure adapted from Kim, Chanho, et al. "Multiple hypothesis tracking revisited." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.
• Video Graph
• VidOR: a Video Graph Database
• Relation-Aware Association: a Method
• Applications
• Event-centric caption

• Object-centric caption
Liang et al., Recurrent Topic-Transition GAN for Visual Paragraph Generation, *In ICCV 2017*
Video Question Answering

Q: Is there? How many? What is ?
A: Yes/No Num Other

Q: What will happen? Why falls?
A: Fall down Stumbled by a rock
Summary

• Video Graph
• VidOR: a Video Graph Database
• Relation-Aware Association: a Method
• Applications to video captioning & question answering
THANK YOU

NExT++ research is supported by the National Research Foundation, Prime Minister’s Office, Singapore under its IRC@SG Funding Initiative.

shangxin@comp.nus.edu.sg