Video Relation Inference

Xindi Shang, NUS

NExT++ Workshop 2018 @ Nanjing
Video Relation

- Context & knowledge
  - happy with family
  - mother
- Generate description
- Multimedia QA
- Predict future activity

- Video Relations:
  - Hold
  - Watch
  - Watch
  - In front
  - Hug
• Visual relation provides evidences for high-level semantic relation
• Visual Relation Detection (VRD)
Technical Overview of VRD

- Video Object Tracking
- Image Object Detection
- Video Feature Engineering
- Action Recognition
- Time-Series Modeling

- Video Object Detection
- Relation Prediction
- Temporal Localization
• 10K videos from social platform
  • 30s long in average

• Object entity annotation
  • 80 categories: adult, child, dog, chair, ball, laptop, etc.
  • Annotated 2,000,000 bounding boxes to localize the objects

• Relation annotation
  • 50 categories: watch, hold, lean_on, push, lift, kiss, grab, get_on, etc.
  • Annotated 100,000 relation instances
Dataset Samples
Dataset Samples
Dataset Samples
Dataset Samples
Dataset Samples
Video Relation Inference

Visual relations

VRD classifier

Model trained on the dataset

Video captioning
Video QA
...

Generic relation feature extractor

Appearance feature
Motion feature
...

videos

© NExT++. Copyright Reserved.
• Directly utilize relation triplets

“Exploring Visual Relationship for Image Captioning”, Yao Ting, Pan Yingwei, Li Yehao and Mei Tao, ECCV 2018
Summary

• Understanding relations is critical towards holistic video content understanding
• We made 10K video relation dataset to improve current VRD model
• In future work, use VRD to improve video captioning and QA

• https://lms.comp.nus.edu.sg/research/VidVRD.html
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