End-to-end Concept Word Detection for Video Captioning, Retrieval, and Question Answering

Youngjae Yu Seoul National University
Hyungjin Ko Seoul National University
Jongwook Choi Seoul National University
Gunhee Kim Seoul National University

SEOUL NATIONAL UNIV. VISION & LEARNING

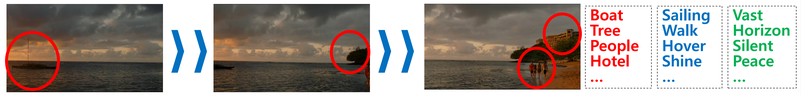
Code will be available at http://vision.snu.ac.kr/projects/lsmdc-2016

CVPR July 21-26 HONOLULU 2017

Motivation

- Representation of video is more complicated than that of image

1. More objects 2. More dynamic 3. More sentiments



- Existing CNN-RNNs are not enough to catch mid-level semantics in video
  - First detect and catch the intermediate, mid-level semantic meanings
  - Then, perform language tasks on the semantic concepts of video

Objective

Address multiple video-and-language tasks using detected concepts

- General concept word detector that can be learned in end-to-end without external sources
- Use concept words as a semantic prior to describe a video
- Evaluate four video-and-language tasks in LSMDC to prove flexibility of our framework

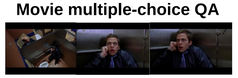
Movie description

His vanity license plate reads 732.

Movie fill-in-the-blank QA

Q1) She ___. A) **nods**
Q2) He opens the ___. A) **door**

Movie multiple-choice QA



① SOMEONE puts his arms around.
② **SOMEONEs eyes widen.**
③ He gives a faint bobble of his head.
④ With people.
⑤ Later she enters her apartment.

Movie retrieval

Query : He answers the phone

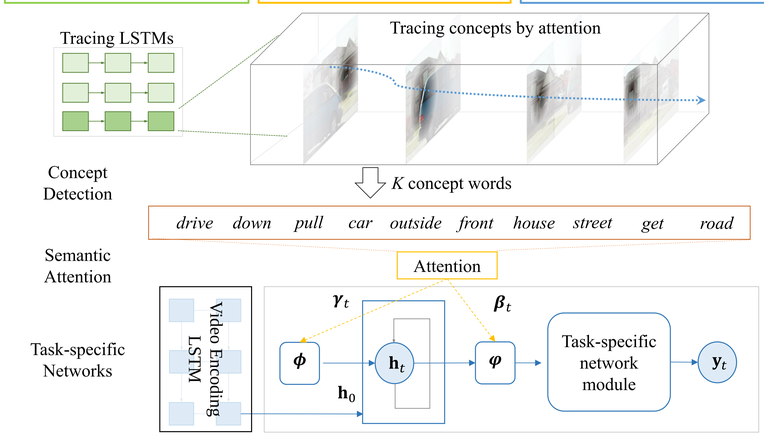
Our Solution – CT-SAN

Concept-**T**racing **S**emantic **A**ttention **N**etworks

Concept Detection
Maintain visual consistency using LSTM with attention

Semantic Attention
Use multiple word features as semantic prior in video

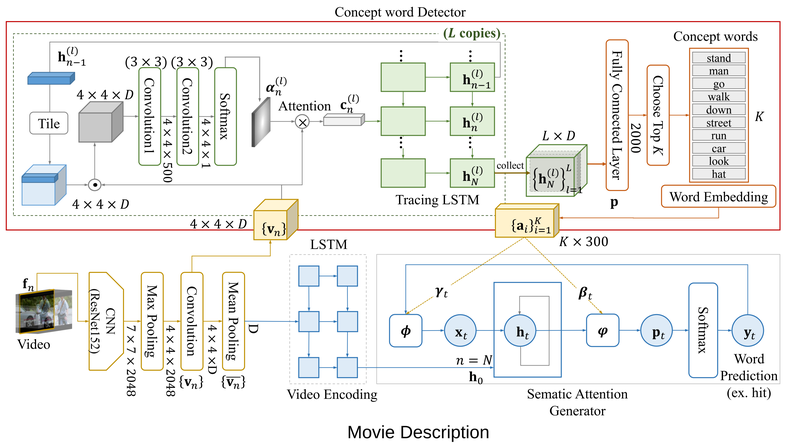
Task-specific Networks
Resolving language tasks subject to video semantic concepts



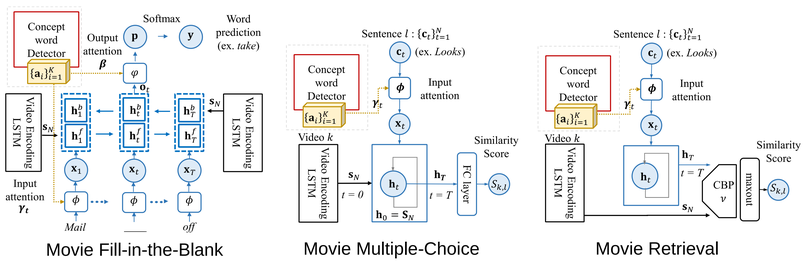
Preprocessing

- Candidates for concept words – without external sources!
  - Apply automatic POS tagging with NLTK
  - Select up to V (~2,000) nouns and verbs from the training corpus according to word frequency
- Vocabulary preprocessing
  - Collect the words that occur more than three times in the training set
  - The resulting dictionary size is $|\mathcal{V}|$ = 12,486

Architecture



- Employ *L* Concept-Tracing LSTMs, each of which can capture a concept
- Top *K* concept words are detected and used for the semantic attention
- Generate sentence using RNN decoder with semantic attention



- Fill-in-the-Blank: Bidirectional LSTM for representing the sentence
- Multiple-Choice/Retrieval: Choose the correct answer from the estimated similarity between video and sentence

⚠ See the equations in the paper!

Quantitative Results on the LSMDC2016

- Performance comparison for the movie description

Movie Description	B1	B2	B3	B4	M	R	Cr
EITanque	0.144 (4)	0.042 (5)	0.016 (3)	0.007 (2)	0.056 (7)	0.130 (7)	0.098 (2)
S2VT [24]	**0.162** (1)	**0.051** (1)	**0.017** (1)	0.007 (2)	0.070 (4)	0.149 (4)	0.082 (4)
SNUVL	0.157 (2)	0.049 (2)	0.014 (4)	0.004 (6)	0.071 (2)	0.147 (5)	0.070 (6)
sophieag	0.151 (3)	0.047 (3)	0.013 (5)	0.005 (4)	**0.075** (1)	0.152 (2)	0.072 (5)
ayush11011995	0.116 (8)	0.032 (7)	0.011 (7)	0.004 (6)	0.070 (4)	0.138 (6)	0.042 (8)
rakshithShetty	0.119 (7)	0.024 (8)	0.007 (8)	0.003 (8)	0.046 (8)	0.108 (8)	0.044 (7)
Aalto	0.070 (9)	0.017 (9)	0.005 (9)	0.002 (9)	0.033 (9)	0.069 (9)	0.037 (9)
CT-SAN	0.135 (5)	0.044 (4)	**0.017** (1)	**0.008** (1)	0.071 (2)	**0.159** (1)	**0.100** (1)

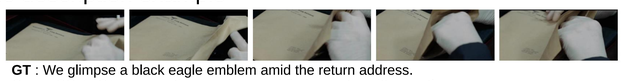
- Performance comparison for the Multi-choice, Retrieval, Fill-in-the-blank

Tasks	Multi-Choice	Movie Retrieval			
Methods	Accuracy	R@1	R@5	R@10	MedR
Aalto	39.7	–	–	–	–
SNUVL (Single)	63.1	3.8	13.6	18.9	80
EITanque	63.7	4.7	15.9	23.4	64
SNUVL (Ensemble)	65.7	3.6	14.7	23.9	50
CT-SAN (Single)	63.8	4.5	14.1	20.9	67
CT-SAN (Ensemble)	**67.0**	**5.1**	**16.3**	**25.2**	**46**

Fill-in-the-Blank	
Methods	Accuracy
amirmazaheri	34.2
SNUVL (Single)	38.0
SNUVL (Ensemble)	40.7
CT-SAN (Single)	41.9
CT-SAN (Ensemble)	**42.7**

Qualitative Results

- Movie description example



GT : We glimpse a black eagle emblem amid the return address.
Ours : SOMEONE opens the envelope and finds a note written on the page.
Concept words : *page, note, card, envelope, book, name, find, read, paper, letter*

- Fill-in-the-Blank example

Blank Sentence : He slows down in front of one _____ with a triple garage and box tree on the front lawn and pulls up onto the driveway.
Answer/Our result : house / **house**
Concept words : *drive, car, pull, down, front, outside, house, street, get, road*

- Multiple-choice example

① **Smiling and chatting, they speed down a narrow sun-dappled road in the woods.**
② Both girls turn to speak.
③ He turns and smiles.
④ As they spin around again, SOMEONE crouches by the window and raises his binoculars.
⑤ then turns back and enters the house.
Concept words : *road, drive, car, tree, house, down, pull, driveway, park, speed*

- Movie Retrieval example

Question : Throughout the cafeteria, students dance together and clap their hands

Concept words : *dance, woman, girl, hug, dress, arm, back, pose, down, show*