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

Youngjae Yu   Hyungjin Ko   Jongwook Choi   Gunhee Kim

Seoul National University

Presenter: Weirui Kong, Bicheng Xu

CPSC 532L

Motivation

- **Video-to-language** tasks. E.g.,
  - Video captioning;
    "A car pulls up onto the driveway."
  - Video question answering.
    Q: He slows down in front of one ____ with a garage.
    A: house.

Introduction

- With the *Large Scale Movie Description Challenge* (LSMDC 2016), this work aims at the following **four video-to-language tasks**.

Introduction

- With the *Large Scale Movie Description Challenge* (LSMDC 2016), this work aims at the following **four video-to-language tasks**.
    - Movie Description.
        - Given a short video clip, generate a single descriptive sentence.

Introduction

- With the *Large Scale Movie Description Challenge* (LSMDC 2016), this work aims at the following **four video-to-language tasks**.
  - Movie Description.
    - Given a short video clip, generate a single descriptive sentence.
  - Movie Fill-in-the-Blank.
    - Given a video clip and a sentence with a blank in it, predict a single correct word to fill in the blank.

Introduction

- With the *Large Scale Movie Description Challenge* (LSMDC 2016), this work aims at the following **four video-to-language tasks**.
  - Movie Description.
    - Given a short video clip, generate a single descriptive sentence.
  - Movie Fill-in-the-Blank.
    - Given a video clip and a sentence with a blank in it, predict a single correct word to fill in the blank.
  - Multiple-Choice Test.
    - Given a video query and five candidate captions, find the best option.

Introduction

- With the *Large Scale Movie Description Challenge* (LSMDC 2016), this work aims at the following **four video-to-language tasks**.
  - Movie Description.
    - Given a short video clip, generate a single descriptive sentence.
  - Movie Fill-in-the-Blank.
    - Given a video clip and a sentence with a blank in it, predict a single correct word to fill in the blank.
  - Multiple-Choice Test.
    - Given a video query and five candidate captions, find the best option.
  - Movie Retrieval.
    - Given a short query sentence, search for its corresponding video.

Concept Words

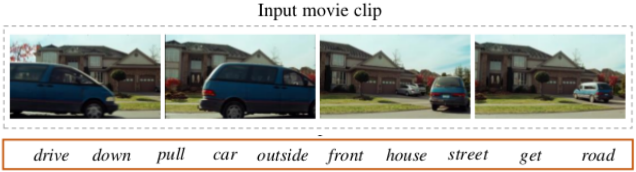
- The words that **consistently** appear across frame regions.

Concept Words

- The words that **consistently** appear across frame regions.
- Can be nouns, verbs, and adjectives.

Concept Words

- The words that **consistently** appear across frame regions.
- Can be nouns, verbs, and adjectives.
- Collected from all training caption sentences.



Trace

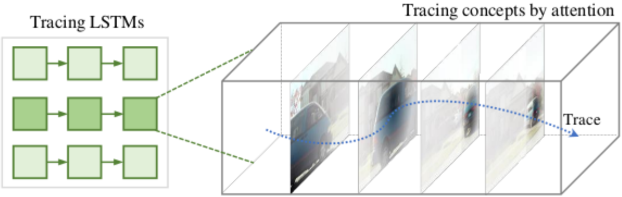
- Keep track of spatial attention over video frames.

Trace

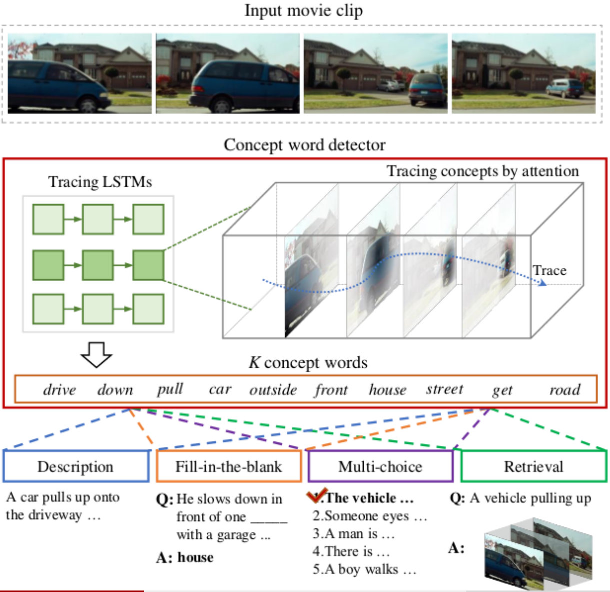
- Keep track of spatial attention over video frames.
- Spatial attentions in adjacent frames **resemble** the spatial consistency of a single concept.

Trace

- Keep track of spatial attention over video frames.
- Spatial attentions in adjacent frames **resemble** the spatial consistency of a single concept.
- It can be a moving object, or an action in video clips.



Model Overview



Contributions

- A novel **end-to-end** learning approach for **detecting a list of concept words** and **attend on them** to **enhance the performance** of multiple video-to-language tasks.

Contributions

- A novel **end-to-end** learning approach for **detecting a list of concept words** and **attend on them** to **enhance the performance** of multiple video-to-language tasks.
- The proposed concept word detection and attention model can be plugged into any models of video captioning, retrieval, and question answering.

Contributions

- A novel **end-to-end** learning approach for **detecting a list of concept words** and **attend on them** to **enhance the performance** of multiple video-to-language tasks.
- The proposed concept word detection and attention model can be plugged into any models of video captioning, retrieval, and question answering.
- Win three of four tasks of LSMDC 2016.

Model

- Detail model illustration.

Concept Word Detector

Preprocessing

- The vocabulary dictionary $\mathcal{V}$.
  - $|\mathcal{V}| = 12486$.
  - The words that occur more than three times in the dataset.

Concept Word Detector

Preprocessing

- The vocabulary dictionary $\mathcal{V}$.
  - $|\mathcal{V}| = 12486$.
  - The words that occur more than three times in the dataset.
- Word Embedding.
  - Train the word2vec skip-gram embedding.
  - Obtain the word embedding matrix $E \in \mathbb{R}^{d \times |\mathcal{V}|}$.

Concept Word Detector

Preprocessing

- Video representation
  - Equidistantly sample one per ten frames from a video.

Concept Word Detector

Preprocessing

- Video representation
  - Equidistantly sample one per ten frames from a video.
  - Obtain N video frames.

Concept Word Detector

Preprocessing

- Video representation
  - Equidistantly sample one per ten frames from a video.
  - Obtain N video frames.
  - Extract the feature map of each frame from the res5c layer of **ResNet**, followed by a **max-pooling** layer and a 3×3 **convolution** layer.

Concept Word Detector

Preprocessing

- Video representation
  - Equidistantly sample one per ten frames from a video.
  - Obtain N video frames.
  - Extract the feature map of each frame from the res5c layer of **ResNet**, followed by a **max-pooling** layer and a 3×3 **convolution** layer.
  - Obtain the visual features of each frame $v_n \in \mathbb{R}^{4 \times 4 \times 500}$.

Concept Word Detector

Preprocessing

- Candidate concept words.
  - Apply automatic POS tagging to extract **nouns**, **verbs** and **adjectives** from all training captions.

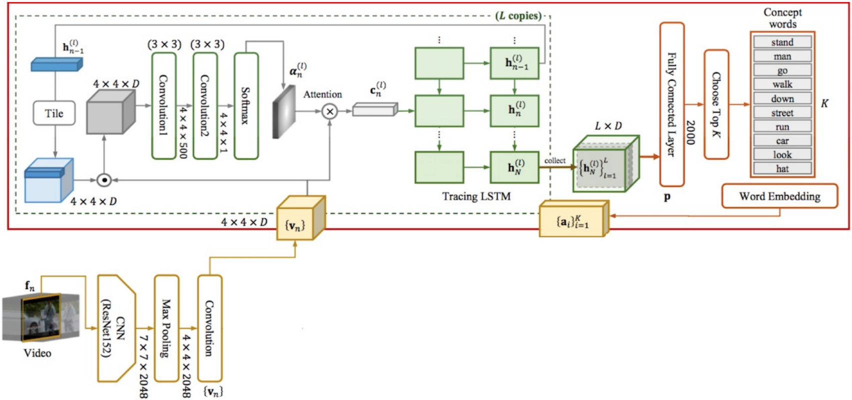
Concept Word Detector

Preprocessing

- Candidate concept words.
  - Apply automatic POS tagging to extract **nouns**, **verbs** and **adjectives** from all training captions.
  - Compute the frequencies and select the $V = 2000$ **most common words** as candidates.

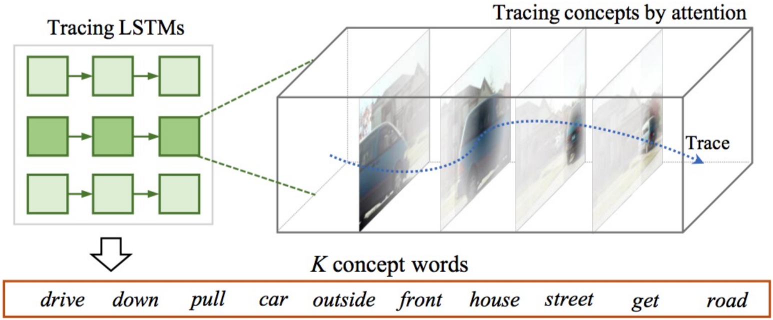
Concept Word Detector

Model Overview



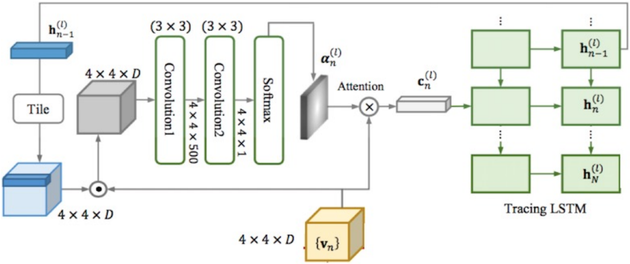
Concept Word Detector

Tracing LSTMs



Concept Word Detector

An Attention Model for Concept Detection



$$c_n^{(l)} = \alpha_n^{(l)} \otimes v_n, \text{ where } A \otimes B = \textstyle\sum_{j,k} A_{(j,k)} \cdot B_{(j,k,:)}$$

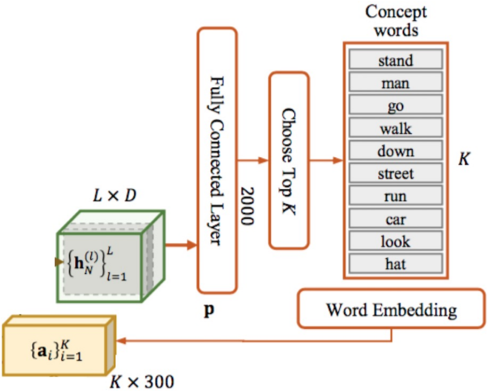
$$h_n^{(l)} = LSTM(c_n^{(l)}, h_{n-1}^{(l)})$$

$$e_n^{(l)}(j,k) = v_n(j,k) \odot h_{n-1}^{(l)}, \text{ where } \odot \text{ is elementwise product}$$

$$\alpha_n^{(l)} = softmax(Conv(e_n^{(l)}))$$

Concept Word Detector

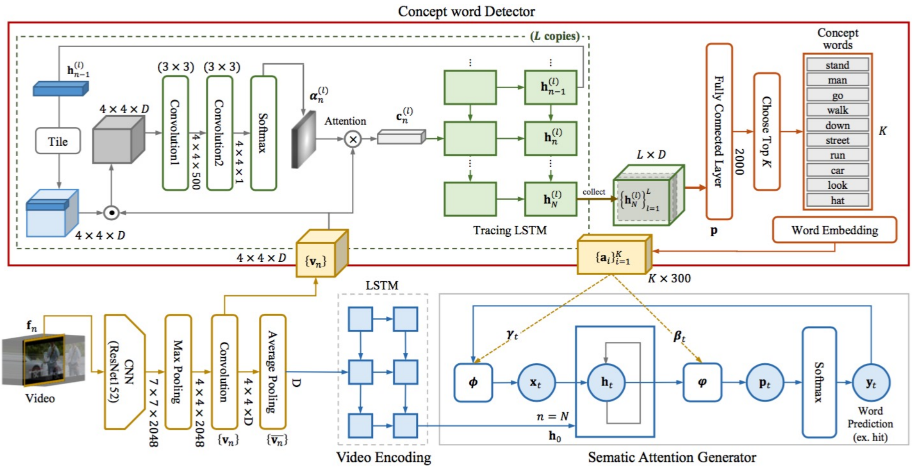
An Attention Model for Concept Detection



The concept confidence vector p: $p = \sigma(W_p[h_N^{(1)}; \cdots; h_N^{(L)}] + b_p) \in \mathbb{R}^V$

The cross entropy loss: $\mathcal{L} = -\frac{1}{V}\sum_{i=1}^{V}[p_i^* \log(p_i) + (1 - p_i^*)\log(1 - p_i)]$

Model for Video Description



Experiments

- Experiments on the **four** tasks of LSMDC 2016.

Dataset

- Four video-to-language tasks.
  - Movie Description.
  - Movie Fill-in-the-Blank.
  - Multiple-Choice Test.
  - Movie Retrieval.
- Contain a parallel corpus of **118,114** sentences, and **118,081** video clips sampled from **202** movies.

Tasks

- Movie Description.
  - Given a short video clip, generate a single descriptive sentence.
  - **Evaluation metrics:** BLEU-1,2,3,4, METEOR, ROUGE-L, and CIDEr.

Tasks

- Movie Description.
  - Given a short video clip, generate a single descriptive sentence.
  - **Evaluation metrics:** BLEU-1,2,3,4, METEOR, ROUGE-L, and CIDEr.
- Movie Fill-in-the-Blank.
  - Given a video clip and a sentence with a blank in it, predict a single correct word to fill in the blank.
  - **Evaluation metric:** percentage of predicted words that match with GTs (prediction accuracy).

Tasks

- Multiple-Choice Test.
  - Given a video query and **five** candidate captions, find the best option.
  - The correct answer is the GT caption of the query video.
  - Four other distractors are randomly chosen from the other captions that have **different activity-phrase labels** from the correct answer.
  - **Evaluation metric:** percentage of correctly answered test questions out of 10,053 public-test data.

Tasks

- Multiple-Choice Test.
  - Given a video query and **five** candidate captions, find the best option.
  - The correct answer is the GT caption of the query video.
  - Four other distractors are randomly chosen from the other captions that have **different activity-phrase labels** from the correct answer.
  - **Evaluation metric:** percentage of correctly answered test questions out of 10,053 public-test data.
- Movie Retrieval.
  - Given a short query sentence, search for its corresponding video out of 1,000 candidate videos, sampled from the public-test data.
  - **Evaluation metrics:** Recall@1/5/10, and Median Rank (MedR).

Quantitative Results

Movie Description

Movie Description	B1	B2	B3	B4	M	R	Cr
EITanque [14]	0.144 (4)	0.042 (5)	0.016 (3)	0.007 (2)	0.056 (7)	0.130 (7)	0.098 (2)
S2VT [31]	**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)
Base-SAN	0.123 (6)	0.038 (6)	0.013 (5)	0.005 (4)	0.066 (6)	0.150 (3)	0.090 (3)
CT-SAN	0.135 (5)	0.044 (4)	**0.017** (1)	**0.008** (1)	0.071 (2)	**0.159** (1)	**0.100** (1)

- Ranks (5,4,1,1)-th in the BLUE language metrics.
- Ranks (2,1,1)-th in the other language metrics.

Quantitative Results

Movie Fill-in-the-Blank

Fill-in-the-Blank	Accuracy
Simple-LSTM	30.9
Simple-BLSTM	31.6
Base-SAN (Single)	34.5
Merging-LSTM [17]	34.2
Base-SAN (Ensemble)	36.9
SNUVL (Single)	38.0
SNUVL (Ensemble)	40.7
CT-SAN (Single)	41.9
CT-SAN (Ensemble)	**42.7**

- Outperform all the participants.

Quantitative Results

Movie Multiple-Choice Test & Movie Retrieval

Tasks	Multiple-Choice	Movie Retrieval			
Methods	Accuracy	R@1	R@5	R@10	MedR
Aalto	39.7	–	–	–	–
SA-G+SA-FC7 [28]	55.1	3.0	8.8	13.2	114
LSTM+SA-FC7 [28]	56.3	3.3	10.2	15.6	88
C+LSTM+SA-FC7 [28]	58.1	4.3	12.6	18.9	98
Base-SAN (Single)	60.1	4.3	13.0	18.2	83
Base-SAN (Ensemble)	64.0	4.4	13.9	19.3	74
SNUVL (Single)	63.1	3.8	13.6	18.9	80
EITanque [14]	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**

- Rank 1st.
- Benefit from the concept word detector.

Qualitative Results

Movie Description - Correct

GT : The sun sets behind the watery horizon as the foursome continues along the shore toward a distant resort.
Ours : The sun shines as the sun sets to the horizon.
Concepts: *cloud, sky, sun, horizon, vast, shore, distance, light, boat, white*

(a)

Qualitative Results

Movie Description - Wrong

GT : We can see awards line a shelf in his office.
Ours : The clock shows a minute, then the screen shows a map of the mothership.
Concepts : *read, screen, office, clock, row, red, show, name, map, down*

(b)

- The concept words relevant to the GT sentence are well detected such as office or clock.

Qualitative Results

Movie Fill-in-the-Blank - Correct

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 : house **Our result** : house

Concepts : *drive, car, pull, down, front, outside, house, street, get, road*

(c)

Qualitative Results

Movie Fill-in-the-Blank - Wrong

Blank Sentence : People _____ down the path and hide behind the pile of pumpkins.

Answer : hurry **Our result** : run

Concepts : *tree, down, towards, run, walk, people, stone, house, forest, river*

(d)

- A near-miss case where the model also predicts a plausible answer.

Qualitative Results

Movie Multiple-Choice Test - Correct

Candidate Sentences

① SOMEONE slams SOMEONEs head against the trunk.

② **Now, the car speeds down an empty road lined with tall evergreens that just into the pale blue sky. (GT Answer)**

③ SOMEONE sets hers down and smiles.

④ Now she lies on top of him.

⑤ As SOMEONE gazes after them, SOMEONE approaches.

Concepts : *car, drive, road, pull, down, street, house, get, speed, front*

(e)

Qualitative Results

Movie Multiple-Choice Test - Wrong

Candidate Sentences

① SOMEONE glares at SOMEONE, his lips curved into a frown.

② SOMEONE follows, looking dazed. (GT Answer)

③ The kid walks into the garage and sees him.

④ He comes towards her and pulls up a chair.

⑤ He walks down the hall past an open doorway and starts to go upstairs.

Concepts : *room, hall, back, walk, down, stand, go, step, smile, see*

(f)

- The chosen answer is overlapped with some of detected words.

Qualitative Results

Movie Retrieval - Correct

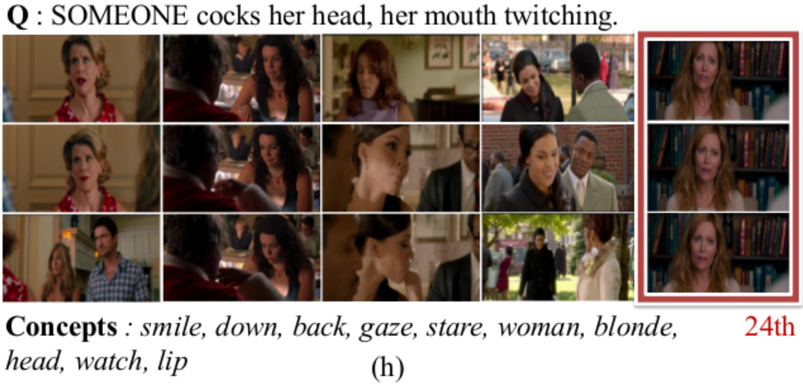
Q : They notice SOMEONE swimming.

Concepts: *water, pool, back, watch, down, stare, arm, smile, gaze, boy*

(g)

Qualitative Results

Movie Retrieval - Wrong



- Fail to catch rare word like *twitch* and *cocks*. Miss to catch subtle movement of mouth.

Conclusion

- Propose an **end-to-end** trainable approach for **detecting a list of concept words** that can be used as semantic priors for **multiple** video-to-language models.

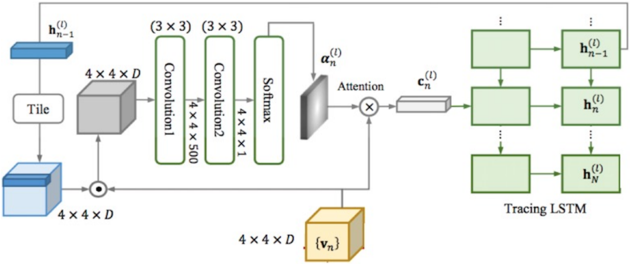
Conclusion

- Propose an **end-to-end** trainable approach for **detecting a list of concept words** that can be used as semantic priors for **multiple** video-to-language models.
- Develop a semantic attention mechanism that **effectively exploits** the discovered concept words.

Conclusion

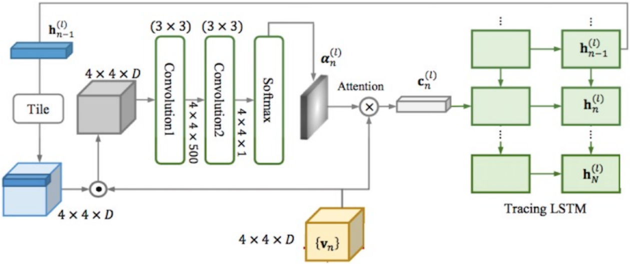
- Propose an **end-to-end** trainable approach for **detecting a list of concept words** that can be used as semantic priors for **multiple** video-to-language models.
- Develop a semantic attention mechanism that **effectively exploits** the discovered concept words.
- Win three tasks in LSMDC 2016.

Potential Improvements



- The update of the attention weights is hard to interpret.

Potential Improvements



- The update of the attention weights is hard to interpret.
- The design of the tracing LSTM seems not so intuitive.

Questions?

- Thanks for your attention!