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

![Diagram of Tracing LSTMs and concepts by attention]
Model Overview

Input movie clip

Concept word detector

Tracing LSTMs  Tracing concepts by attention

K concept words

drive  down  pull  car  outside  front  house  street  get  road

Description

A car pulls up onto the driveway ...

Fill-in-the-blank

Q: He slows down in front of one ____ with a garage ...
A: house

Multi-choice

✓ The vehicle ...
2. Someone eyes ...
3. A man is ...
4. There is ...
5. A boy walks ...

Q: A vehicle pulling up

Retrieval

A:
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.
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.
• 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 $3 \times 3$ convolution layer.
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 $3 \times 3$ 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.
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

Tracing LSTMs

K concept words

drive  down  pull  car  outside  front  house  street  get  road
Concept Word Detector
An Attention Model for Concept Detection

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

\[ h_n^{(l)} = \text{LSTM}(c_n^{(l)}, h_{n-1}^{(l)}) \]

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

\[ \alpha_n^{(l)} = \text{softmax(Conv}(e_n^{(l)})) \]
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

Concept word Detector

Fully Connected Layer

Word Embedding

Video Encoding

Sematic Attention Generator

Word Prediction (ex. hit)
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.

---

Youngjae Yu, Hyungjin Ko, Jongwook Choi, Gunhee Kim
Seoul National University (Presenter: Weirui Kong, Bicheng Xu)
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

<table>
<thead>
<tr>
<th>Movie Description</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>M</th>
<th>R</th>
<th>Cr</th>
</tr>
</thead>
<tbody>
<tr>
<td>EITande [14]</td>
<td>0.144 (4)</td>
<td>0.042 (5)</td>
<td>0.016 (3)</td>
<td>0.007 (2)</td>
<td>0.056 (7)</td>
<td>0.130 (7)</td>
<td>0.098 (2)</td>
</tr>
<tr>
<td>S2VT [31]</td>
<td><strong>0.162 (1)</strong></td>
<td><strong>0.051 (1)</strong></td>
<td><strong>0.017 (1)</strong></td>
<td>0.007 (2)</td>
<td>0.070 (4)</td>
<td>0.149 (4)</td>
<td>0.082 (4)</td>
</tr>
<tr>
<td>SNUVL</td>
<td>0.157 (2)</td>
<td>0.049 (2)</td>
<td>0.014 (4)</td>
<td>0.004 (6)</td>
<td>0.071 (2)</td>
<td>0.147 (5)</td>
<td>0.070 (6)</td>
</tr>
<tr>
<td>sophieag</td>
<td>0.151 (3)</td>
<td>0.047 (3)</td>
<td>0.013 (5)</td>
<td>0.005 (4)</td>
<td><strong>0.075 (1)</strong></td>
<td>0.152 (2)</td>
<td>0.072 (5)</td>
</tr>
<tr>
<td>ayush11011995</td>
<td>0.116 (8)</td>
<td>0.032 (7)</td>
<td>0.011 (7)</td>
<td>0.004 (6)</td>
<td>0.070 (4)</td>
<td>0.138 (6)</td>
<td>0.042 (8)</td>
</tr>
<tr>
<td>rakshithShetty</td>
<td>0.119 (7)</td>
<td>0.024 (8)</td>
<td>0.007 (8)</td>
<td>0.003 (8)</td>
<td>0.046 (8)</td>
<td>0.108 (8)</td>
<td>0.044 (7)</td>
</tr>
<tr>
<td>Aalto</td>
<td>0.070 (9)</td>
<td>0.017 (9)</td>
<td>0.005 (9)</td>
<td>0.002 (9)</td>
<td>0.033 (9)</td>
<td>0.069 (9)</td>
<td>0.037 (9)</td>
</tr>
<tr>
<td>Base-SAN</td>
<td>0.123 (6)</td>
<td>0.038 (6)</td>
<td>0.013 (5)</td>
<td>0.005 (4)</td>
<td>0.066 (6)</td>
<td>0.150 (3)</td>
<td>0.090 (3)</td>
</tr>
<tr>
<td>CT-SAN</td>
<td>0.135 (5)</td>
<td>0.044 (4)</td>
<td><strong>0.017 (1)</strong></td>
<td><strong>0.008 (1)</strong></td>
<td>0.071 (2)</td>
<td><strong>0.159 (1)</strong></td>
<td><strong>0.100 (1)</strong></td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Fill-in-the-Blank</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple-LSTM</td>
<td>30.9</td>
</tr>
<tr>
<td>Simple-BLSTM</td>
<td>31.6</td>
</tr>
<tr>
<td>Base-SAN (Single)</td>
<td>34.5</td>
</tr>
<tr>
<td>Merging-LSTM [17]</td>
<td>34.2</td>
</tr>
<tr>
<td>Base-SAN (Ensemble)</td>
<td>36.9</td>
</tr>
<tr>
<td>SNUVL (Single)</td>
<td>38.0</td>
</tr>
<tr>
<td>SNUVL (Ensemble)</td>
<td>40.7</td>
</tr>
<tr>
<td>CT-SAN (Single)</td>
<td>41.9</td>
</tr>
<tr>
<td>CT-SAN (Ensemble)</td>
<td><strong>42.7</strong></td>
</tr>
</tbody>
</table>

- Outperform all the participants.
Quantitative Results
Movie Multiple-Choice Test & Movie Retrieval

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Multiple-Choice Methods</th>
<th>Multiple-Choice Accuracy</th>
<th>Movie Retrieval</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>MedR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aalto</td>
<td></td>
<td>39.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA-G+SA-FC7 [28]</td>
<td></td>
<td>55.1</td>
<td>3.0</td>
<td>8.8</td>
<td>13.2</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>LSTM+SA-FC7 [28]</td>
<td></td>
<td>56.3</td>
<td>3.3</td>
<td>10.2</td>
<td>15.6</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>C+LSTM+SA-FC7 [28]</td>
<td></td>
<td>58.1</td>
<td>4.3</td>
<td>12.6</td>
<td>18.9</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>Base-SAN (Single)</td>
<td></td>
<td>60.1</td>
<td>4.3</td>
<td>13.0</td>
<td>18.2</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Base-SAN (Ensemble)</td>
<td></td>
<td>64.0</td>
<td>4.4</td>
<td>13.9</td>
<td>19.3</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>SNUVL (Single)</td>
<td></td>
<td>63.1</td>
<td>3.8</td>
<td>13.6</td>
<td>18.9</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>EITanque [14]</td>
<td></td>
<td>63.7</td>
<td>4.7</td>
<td>15.9</td>
<td>23.4</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>SNUVL (Ensemble)</td>
<td></td>
<td>65.7</td>
<td>3.6</td>
<td>14.7</td>
<td>23.9</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>CT-SAN (Single)</td>
<td></td>
<td>63.8</td>
<td>4.5</td>
<td>14.1</td>
<td>20.9</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>CT-SAN (Ensemble)</td>
<td></td>
<td>67.0</td>
<td><strong>5.1</strong></td>
<td><strong>16.3</strong></td>
<td><strong>25.2</strong></td>
<td><strong>46</strong></td>
<td></td>
</tr>
</tbody>
</table>

- 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

- 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

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

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)
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
Qualitative Results

Movie Retrieval - Wrong

Q: SOMEONE cocks her head, her mouth twitching.

Concepts: smile, down, back, gaze, stare, woman, blonde, head, watch, lip

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