Title Generation for User Generated Videos
Kuo-Hao Zeng*, Tseng-Hung Chen*, Juan Carlos Niebles*, Min Sun*
National Tsing Hua University* Stanford University*

Title Generation
--- Neural Clickbait [3] ---

Captions: A man is riding on a bike. A man does a stunt on a bike.

Video title generation aims to produce a title sentence describing the most salient event given a typical 1 minute user-generated video (UGV). While a caption describes a video as a whole and tends to be more generic.

**Contributions**

Methods 1: Highlight Sensitive Captioning
- Method 1: Highlight Sensitive Captioning (Sec. 4.2)
- Method 2: Sentence Augmentation (Sec. 4.3)

Method 2: Sentence Augmentation
- Dummy video observation

Contribution 1: Highlight Sensitive Captioning
We combine a highlight detector with video captioners [1, 2] to train models that can jointly generate titles and locate highlights.

Contribution 2: Sentence Augmentation
We propose a novel and generally applicable method to train an RNN model with both video-sentence pairs and sentence-only examples, where sentence-only examples are either the description sentences or additional sentences on the web.

Highlight Sensitive Video Captioner
We use a bidirectional single-layer LSTM model with binary classification strategy to discriminate highlight or non-highlight for each video. Higher highlight probability means higher chance to be detected as highlight clip.

Sentence Augmentation
To generate catchy or diverse titles, we propose sentence augmentation method to tackle it. It also deals with the rare word problem in the testing set.

VTW Dataset
We collected a large-scale Video Titles in the Web (VTW) dataset of 18100 automatically crawled user-generated videos and titles.

Quantitative Results
<table>
<thead>
<tr>
<th>VTW</th>
<th>S2VT (%)</th>
<th>SA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variant</td>
<td>B01+B02+B03+B04+MET+CID6</td>
<td>B01+B02+B03+B04+MET+CID6</td>
</tr>
<tr>
<td>Vanilla</td>
<td>9.3</td>
<td>9.3</td>
</tr>
<tr>
<td>HL-1</td>
<td>10.8</td>
<td>10.8</td>
</tr>
<tr>
<td>HL</td>
<td>11.4</td>
<td>11.4</td>
</tr>
<tr>
<td>Vanilla+Desc.</td>
<td>7.0</td>
<td>7.0</td>
</tr>
<tr>
<td>Desc. Aug.</td>
<td>10.5</td>
<td>10.5</td>
</tr>
<tr>
<td>Aug.</td>
<td>11.0</td>
<td>11.0</td>
</tr>
<tr>
<td>HL+Aug.</td>
<td>11.1</td>
<td>11.1</td>
</tr>
</tbody>
</table>

Human Evaluation
We ask 7 subjects to conduct the blind test. All subjects do not know which sentence is predicted by S2VT or our method. Among 1000 videos, 508 (51%) videos are selected as “ours is better”, 405 (40%) videos are selected as “S2VT is better”, and the remained 87 (9%) videos are selected as “on par”.

Typical Examples

*Ground Truth: Girl Tries to Perform Backflip from Rooftop into Pool
Baseline: Guy slides off roof and jumps into pool. Our Method: Guy jumps off roof and faceplants into pool.

20 seconds long video

*Ground Truth: BMX Rider Falls at Rail Jump
Baseline: BMX rider jumps off dirt ramp and crashes into wall.
Our Method: BMX rider falls at skatepark and fails.

20 seconds long video

*Ground Truth: Most Chin Ups While Hula Hooping
Baseline: Most consecutive two board bench presses of a fail.
Our Method: Most consecutive chin ups while hula hooping.

20 seconds long video

*Ground Truth: Jump Over Moving Car Fail
Baseline: Guy crashes into tree on zipline. Our Method: Guy fails to jump over moving car.

20 seconds long video

*Ground Truth: Guy Eats Pavement Speeding Down Road
Baseline: Guy loses control of bike on wheelie crashes into.
Our Method: Skateboarder falls off skateboard while trying to wheelie.

20 seconds long video

[3] https://twitter.com/alexjc/status/76907659239683202