Dense-Captioning Events in Videos
ActivityNet Captions Challenge @ CVPR 2020

Ranjay Krishna
Stanford University
@ranjaykrishna
http://ranjaykrishna.com
Videos contain many overlapping complex events. Previous work to describe videos first started with labeling them with a predefined category. This video might be labeled as "Playing piano." Other work have tried to describe videos with one sentence. An example sentence would be "An elderly man is playing the piano in front of a crowd." But only having one sentence to describe a video may miss important details. For example, we miss that "The woman starts singing along with the pianist." Or that "Another man starts dancing to the music, gathering attention from the crowd."
Dense-Captioning Events

We introduce the problem of dense-captioning events, which involves describing all events that happen in the video and localizing them within in time.

- An elderly man is playing the piano in front of a crowd.
- A woman walks to the piano and briefly talks to the elderly man.
- The woman starts singing along with the pianist.
- Another man starts dancing to the music, gathering attention from the crowd.
- Eventually the elderly man finishes playing and hugs the woman, and the crowd applaud.
Evidence that people segment videos into consistent events

The ActivityNet Captions dataset connects each video to a paragraph – a set of temporally annotated sentence descriptions.

- 20k videos with 100K sentences
- Avg. 3.65 temporally localized sentences
- Avg. 13.48 words per sentence
- Avg. 94.6% content coverage for the entire paragraph; Avg. 31% content coverage per sentence

Comparison to other datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th># videos</th>
<th>Avg. length</th>
<th># sentences</th>
<th>Des.</th>
<th>Loc. Des.</th>
<th>paragraphs</th>
<th>overlapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF101 [45]</td>
<td>sports</td>
<td>13k</td>
<td>7s</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sports 1M [21]</td>
<td>sports</td>
<td>1.1M</td>
<td>300s</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Thumos 15 [15]</td>
<td>sports</td>
<td>21k</td>
<td>4s</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HMDB 51 [25]</td>
<td>movie</td>
<td>7k</td>
<td>3s</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hollywood 2 [28]</td>
<td>movie</td>
<td>4k</td>
<td>20s</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MPII cooking [40]</td>
<td>cooking</td>
<td>44k</td>
<td>600s</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ActivityNet [4]</td>
<td>human</td>
<td>20k</td>
<td>180s</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MPII MD [39]</td>
<td>movie</td>
<td>68k</td>
<td>4s</td>
<td>68,375</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>M-VAD [47]</td>
<td>movie</td>
<td>49k</td>
<td>6s</td>
<td>55,904</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MSR-VTT [55]</td>
<td>open</td>
<td>10k</td>
<td>20s</td>
<td>200,000</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MSVD [6]</td>
<td>human</td>
<td>2k</td>
<td>10s</td>
<td>70,028</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>YouCook [7]</td>
<td>cooking</td>
<td>88</td>
<td>-</td>
<td>2,688</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Charades [43]</td>
<td>human</td>
<td>10k</td>
<td>30s</td>
<td>16,129</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KITTI [12]</td>
<td>driving</td>
<td>21</td>
<td>30s</td>
<td>520</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TACoS [36]</td>
<td>cooking</td>
<td>127</td>
<td>360s</td>
<td>11,796</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TACos multi-level [37]</td>
<td>cooking</td>
<td>127</td>
<td>360s</td>
<td>52,593</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ActivityNet Captions</td>
<td>open</td>
<td>20k</td>
<td>180s</td>
<td>100k</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

More verbs and actions when compared against image captions

In comparison with image dense IMAGE captioning datasets, video captions represent events that require a temporal component to understand.

Implicit ordering of ActivityNet action labels

Camel ride
Brushing hair
Braiding hair
Playing violin
Playing ice hockey
Disc dog
Decorating the Christmas tree
Cleaning sink
Smoking a cigarette

Hurling
Waterskiing
Slacklining
Breakdancing
Longboarding
Futsal
Snatch
Clean and jerk
Powerlifting
Cumbia
Challenges for dense captioning events

Challenge 1:

Predicted proposals are short

while events vary in length

Challenges for dense captioning events

Challenge 2:
Events depend on both past and future events.

Captions need to account for context
Our original dense captioning model

- We first extract 3D conv features from a network pretrained on action prediction.
Our original dense captioning model

- We sample these features at various strides, and feed them into a hierarchical LSTM
- Whenever the proposal LSTM detects an event, we use the hidden state of the LSTM at that time step as a feature representation of the visual event
The event proposals are fed into a captioning module. Our captioning module captures temporal contexts in videos by keeping track of past and future events.
We categorize all events into two buckets: past and future, and pass the resulting hidden representation of the past, current and future event into the captioning LSTM.
Our original dense captioning model

Video features

Proposal module

Proposals

Captioning Module

Output Captions

A lady joins the man and sings along to the music.

Representation
Attention
Context
LSTM

h_i

past

future

a lady joins

...
Example model predictions

A man is speaking to the camera while standing in the field with a dog.
Example model predictions
Impact of ActivityNet Captions

Activity-Net Entities

Extends ActivityNet-Captions with 158k bounding box annotations for each noun phrase, on ~15K videos

Dataset

Impact of ActivityNet Captions

Something–something dataset

Generalizes ActivityNet–Captions by removing objects and “cultural” labels.

Impact of ActivityNet Captions

ActivityNet-QA
Auto-generated QA from ActivityNet-Captions.

Models proposed for dense captioning

Transformers
Hierarchical models
Grounding objects
Dual language-vision fusion

Li, et al. "Jointly localizing and describing events for dense video captioning." CVPR 2018
Wang, et al. "Video captioning via hierarchical reinforcement learning." CVPR 2018
ActivityNet Challenge

CVPR 2020 workshop on activity recognition
Participation at a glance

44 Entries

15 Teams
Evaluation Metric

We combine the above metrics and report average METEOR for events that have temporal Intersection over Union (t-IoU) greater than the following thresholds: 0.3, 0.5, 0.7 and 0.9.
Common themes across submissions

• Multi-model features
  • Using RGB, audio, motion, spatio-temporal features.

• Upgrading to newer encoders:
  • ResNeXt encoders, BERT language modeling

• New proposal networks and variants:
  • Bidirectional-SST + Pointer Networks

• REINFORCE on METEOR
  • Most submissions reported using REINFORCE to improve performance on metrics

• Re-ranking generated captions
  • based on diversity and other metrics
Potential directions of further improvement

• Using sub-word generation:
  • Utilizing wordpiece tokenization for captioning models
  • Handling lower frequency words or rare concepts

• Better evaluation metrics:
  • METEOR, CIDEr, etc. do not correlate with human judgment.

• Compositionality of events:
  • Events are hierarchical, learning the compositionality could lead to more data efficient models. (Ji et al. Action Genome)

• Proxy language and vision tasks
  • Learning other vision or language tasks concurrently

• Leveraging commonsense knowledge
  • Using external knowledge bases like freebase or FrameNet can contextualize actions and events
## Overall Results

<table>
<thead>
<tr>
<th>Name</th>
<th>Team</th>
<th>Final Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Ranjay Krishna</td>
<td>Stanford University</td>
</tr>
</tbody>
</table>

Baseline: Krishna et al. Dense Captioning Events in Videos, ICCV 2017
## Overall Results

<table>
<thead>
<tr>
<th>Name</th>
<th>Team</th>
<th>Final Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Ranjay Krishna</td>
<td>4.82</td>
</tr>
<tr>
<td>Third</td>
<td>XC Wang</td>
<td>8.59</td>
</tr>
</tbody>
</table>

Baseline: Krishna et al. Dense Captioning Events in Videos, ICCV 2017
## Overall Results

<table>
<thead>
<tr>
<th>Name</th>
<th>Team</th>
<th>Final Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Second</strong></td>
<td>T. Wang &amp; Mingjing Yu</td>
<td>SYSU</td>
</tr>
<tr>
<td><strong>Third</strong></td>
<td>XC Wang</td>
<td>ByteDance</td>
</tr>
<tr>
<td><strong>Baseline</strong></td>
<td>Ranjay Krishna</td>
<td>Stanford University</td>
</tr>
</tbody>
</table>

Baseline: Krishna et al. Dense Captioning Events in Videos, ICCV 2017
# Overall Results

<table>
<thead>
<tr>
<th>Name</th>
<th>Team</th>
<th>Final Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First</strong></td>
<td>Yuqing Song</td>
<td>9.89</td>
</tr>
<tr>
<td><strong>Team</strong></td>
<td>Renmin University</td>
<td></td>
</tr>
<tr>
<td><strong>Score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Second</strong></td>
<td>T. Wang &amp; Mingjing Yu</td>
<td>9.28</td>
</tr>
<tr>
<td><strong>Team</strong></td>
<td>SYSU</td>
<td></td>
</tr>
<tr>
<td><strong>Score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Third</strong></td>
<td>XC Wang</td>
<td>8.59</td>
</tr>
<tr>
<td><strong>Team</strong></td>
<td>ByteDance</td>
<td></td>
</tr>
<tr>
<td><strong>Score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Baseline</strong></td>
<td>Ranjay Krishna</td>
<td>4.82</td>
</tr>
<tr>
<td><strong>Team</strong></td>
<td>Stanford University</td>
<td></td>
</tr>
<tr>
<td><strong>Score</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Baseline: Krishna et al. Dense Captioning Events in Videos, ICCV 2017
Dense-Captioning Events in Videos
ActivityNet Captions Challenge @ CVPR 2020

Ranjay Krishna
Stanford University
@ranjaykrishna
http://ranjaykrishna.com