Video Analysis

Video to Language, Highlight Detection, Video Classification

Tao Mei (tmei@microsoft.com)

Joint work with Ting Yao and Yong Rui, MSR Asia
Microsoft Project Oxford: Adding Smart to Your Applications

A portfolio of REST APIs and SDKs which enable developers to write applications which understand the content within the rapidly growing set of multimedia data.
Easy to use

Project Oxford allows you to focus on your application by easily including these services across platforms through simple REST APIs.
Video to Sentence
Video to Language

- Video description (from individual concepts to natural sentence)
  - Robotic vision
  - Movie description for blind people
  - Incident report for surveillance videos
- Video indexing
  - Learning embedding models from language-video pairs
# Image captioning competition

![Microsoft COCO logo](image)

**Overview** | **Download** | **Evaluate** | **Leaderboard** | **Challenges** | **Dataset**
--- | --- | --- | --- | --- | ---

<table>
<thead>
<tr>
<th>Table-C5</th>
<th>Table-C40</th>
<th>Table-human</th>
<th>Last update: June 8, 2015. Visit <a href="https://codaLab.org">CodaLab</a> for the latest results.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CIDEr-D</strong></td>
<td><strong>Meteor</strong></td>
<td><strong>ROUGE-L</strong></td>
<td><strong>BLEU-1</strong></td>
</tr>
<tr>
<td>Google$^{[4]}$</td>
<td>0.943</td>
<td>0.254</td>
<td>0.53</td>
</tr>
<tr>
<td>MSR Captivator$^{[5]}$</td>
<td>0.931</td>
<td>0.248</td>
<td>0.526</td>
</tr>
<tr>
<td>m-RNN$^{[13]}$</td>
<td>0.917</td>
<td>0.242</td>
<td>0.521</td>
</tr>
<tr>
<td>MSR$^{[8]}$</td>
<td>0.912</td>
<td>0.247</td>
<td>0.519</td>
</tr>
<tr>
<td>Nearest Neighbor$^{[11]}$</td>
<td>0.886</td>
<td>0.237</td>
<td>0.507</td>
</tr>
<tr>
<td>m-RNN (Baidu/ UCLA)$^{[10]}$</td>
<td>0.886</td>
<td>0.238</td>
<td>0.524</td>
</tr>
<tr>
<td>Berkeley LRCN$^{[2]}$</td>
<td>0.869</td>
<td>0.242</td>
<td>0.517</td>
</tr>
<tr>
<td>Human$^{[5]}$</td>
<td>0.854</td>
<td>0.252</td>
<td>0.484</td>
</tr>
<tr>
<td>Montreal/Toronto$^{[13]}$</td>
<td>0.85</td>
<td>0.243</td>
<td>0.513</td>
</tr>
</tbody>
</table>

[CVPR 2015 oral; arxiv @ 2014-11-17]
[arxiv @ 2015-05-07]
[arxiv @ 2015-04-25]
[CVPR 2015 poster; arxiv @ 2014-11-18]
[arxiv @ 2015-05-27]
[NIPS 2014 workshop; arxiv @ 2014-12-20]
[CVPR 2015 oral; arxiv @ 2014-11-17]
[arxiv @ 2015-02-10]
Challenges for video-to-sentence

• Video-to-sentence is still under-explored

• Learning video representation
  • visual objects (AlexNet, GoogLeNet, VGG)
  • temporal dynamics (C3D, optical flow)
  • audio (MFCC, Spectrum-SIFT)

• Deep neural network design
  • filter: 2D CNN/3D CNN
  • multi-layer RNN (LSTM)

• Sequence learning
  • sequence vs. static frames (pooling/alignment)
  • semantic relationship between entire sentence and video content
How does video-to-sentence work?

- **Language template-based model** [UT Austin’14, SUNY-Byffalo’15]
  - SVO detection -> template-based sentence generation

```
Predicting the best words for describing:
Subject (S) - Verb (V) - Object (O)

Generating sentence using template:
“determiner (a/the) - Subject - Verb (tense) - Preposition (optional) - determiner (a/the) - Object (optional)”

S: man, V: play, O: guitar

Sentence: “a man is playing a guitar”
```
How does video-to-sentence work?

- **RNN-based model** [UC Berkeley‘14’15, UdeM’15]
  - decoding (temporal) video representation into sequence of words

- **UC Berkeley’14**: AlexNet + mean pooling + B
- **UdeM’15**: (GoogLeNet + 3D CNN) + soft-attention + B
- **UC Berkeley’15**: (VGG + Optical Flow) + sequence encoding-decoding
- **MSRA**: (VGG 2D CNN + 3D CNN) + mean pooling + A + joint learning
Our work: joint embedding and translating

• Key issues in sentence generation
  • relevance: relationship between sentence (S, V, O) semantics and video content
  • coherence: sentence grammar

LSTM: a man is playing a guitar
LSTM-E: a man is playing a piano

LSTM: a man is dancing
LSTM-E: a group of people are dancing

• Joint learning: relevance + coherence
  • Holistically looking at both entire sentence semantics and video content
  • Learning powerful video representation: 2D CNN (visual) + C3D (motion)
$E(\mathcal{V}, S) = (1 - \lambda) \times \|T_V \mathcal{V} - T_S S\|_2^2 - \lambda \times \sum_{t=1}^{N_s} \log Pr_t(w_t)$
Evaluations

- Dataset ([MSR Video Description Corpus](https://github.com/MSR-VisionLab/VideoDescription), a.k.a. YouTube2Text)
  - 1,970 Youtube video snippets (1,200 training, 100 validation, 670 testing)
  - 10-25 sec for each clip
  - ~40 human-generated sentences for each clip (by AMT)
  - dictionary: 15,903 -> 7,000; 45 S-groups, 218 V-groups, 241 O-groups
- Training: 12 hrs in one single CPU; testing: ~5 sec per clip
## Performance

The accuracy of S-V-O triplet prediction.

<table>
<thead>
<tr>
<th>Model</th>
<th>Team</th>
<th>Subject%</th>
<th>Verb%</th>
<th>Object%</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGM</td>
<td>UT Austin, COLING (2014/08)</td>
<td>76.42</td>
<td>21.34</td>
<td>12.39</td>
</tr>
<tr>
<td>CRF</td>
<td>SUNY-Buffalo, AAAI (2015/01)</td>
<td>77.16</td>
<td>22.54</td>
<td>9.25</td>
</tr>
<tr>
<td>CCA</td>
<td>Stanford, CVPR (2010/06)</td>
<td>77.16</td>
<td>21.04</td>
<td>10.99</td>
</tr>
<tr>
<td>JEM</td>
<td>SUNY-Buffalo, AAAI (2015/01)</td>
<td>78.25</td>
<td>24.45</td>
<td>11.95</td>
</tr>
<tr>
<td>LSTM</td>
<td>UC Berkeley, NAACL (2014/12)</td>
<td>71.19</td>
<td>19.40</td>
<td>9.70</td>
</tr>
<tr>
<td>LSTM-E</td>
<td>MSRA, arxiv (2015/05)</td>
<td>80.45</td>
<td>29.85</td>
<td>13.88</td>
</tr>
</tbody>
</table>

The performance of sentence generation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Team</th>
<th>METEOR%</th>
<th>BLEU@4%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>UC Berkeley, NAACL (2014/12)</td>
<td>26.9</td>
<td>31.2</td>
</tr>
<tr>
<td>SA</td>
<td>UdeM, arxiv (2015/02)</td>
<td>29.6</td>
<td>42.2</td>
</tr>
<tr>
<td>S2VT</td>
<td>UC Berkeley, arxiv (2015/05)</td>
<td>29.8</td>
<td>--</td>
</tr>
<tr>
<td>LSTM-E</td>
<td>MSRA, arxiv (2015/05)</td>
<td>31.0</td>
<td>45.3</td>
</tr>
</tbody>
</table>
Video-to-Sentence results (within YouTube2Text)

Human: a kitten is playing with his toy
LSTM: a cat is playing with a **mirror**
LSTM-E: a kitten is playing with a **toy**

Human: a man is singing on the stage
LSTM: a man is playing a **flute**
LSTM-E: a man is singing

Human: a group of people are dancing
LSTM: a **man** is dancing
LSTM-E: a group of people are dancing

Human: a person is playing a piano keyboard
LSTM: a man is playing a **guitar**
LSTM-E: a man is playing a **piano**

Human: a man is talking on a cell phone
LSTM: a **woman** is talking
LSTM-E: a **man** is talking on a phone

Human: a man is riding his motorcycle
LSTM: a man is riding a **car**
LSTM-E: a man is riding a **motorcycle**
A car is running

A man is cutting a piece of meat

A man is performing on a stage

A man is riding a bike

A man is singing

A panda is walking

A woman is riding a horse

A man is flying in a field
What if applying image captioning tech to video?

**Video-to-sentence:**

LSTM-E: a group of people are dancing

**Image-to-sentence (keyframe-based):**

- a group of people are jumping up on a stage look on a horse
- the two people are standing in the front of their heads
- a group of people standing around next to each other

http://deeplearning.cs.toronto.edu/i2t
Highlight detection
Example: parkour (highlight + timelapse 4X + music)
Example: GoPro video
Video classification
Action recognition from video

- Examples of video categories (CCV-20)
Action recognition

surfing: 0.1114
swimming: 0.0568
kayaking: 0.0534
Framework

- Multi-granular spatiotemporal architecture
  - deep feature learning representation for video
  - multi-granular streams (frame + optical flow + clip + video)
  - relative importance learning for each component
### THUMOS Challenge 2015

In conjunction with CVPR’15

<table>
<thead>
<tr>
<th>Rank</th>
<th>Entry</th>
<th>Run1</th>
<th>Run2</th>
<th>Run3</th>
<th>Run4</th>
<th>Run5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U. of Tech., Sydney &amp; CMU</td>
<td>0.7384</td>
<td>0.7137</td>
<td>0.7011</td>
<td>0.6913</td>
<td>0.647</td>
</tr>
<tr>
<td>2</td>
<td>MSR Asia (MSM)</td>
<td>0.6861</td>
<td>0.6869</td>
<td>0.6878</td>
<td>0.6886</td>
<td>0.6897</td>
</tr>
<tr>
<td>3</td>
<td>Zhejiang University</td>
<td>0.6876</td>
<td>0.6643</td>
<td>0.6859</td>
<td>0.6809</td>
<td>0.5625</td>
</tr>
<tr>
<td>4</td>
<td>INRIALEAR</td>
<td>0.6814</td>
<td>0.6811</td>
<td>0.5395</td>
<td>0.6739</td>
<td>0.6793</td>
</tr>
<tr>
<td>5</td>
<td>CUHK &amp; Shenzhen Inst. Adv. Tech.</td>
<td>0.4894</td>
<td>0.5746</td>
<td>0.6803</td>
<td>0.6576</td>
<td>0.6604</td>
</tr>
<tr>
<td>6</td>
<td>University of Amsterdam</td>
<td>0.6798</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>Tianjin University</td>
<td>0.6666</td>
<td>0.6551</td>
<td>0.6324</td>
<td>0.5514</td>
<td>0.5337</td>
</tr>
<tr>
<td>8</td>
<td>USC &amp; Tsinghua U.</td>
<td>0.6354</td>
<td>0.6398</td>
<td>0.6346</td>
<td>0.5639</td>
<td>0.6357</td>
</tr>
<tr>
<td>9</td>
<td>MIL - U. Tokyo</td>
<td>0.6158</td>
<td>0.6172</td>
<td>0.6174</td>
<td>0.6087</td>
<td>0.4886</td>
</tr>
</tbody>
</table>