Video to text

Cees Snoek

University of Amsterdam
The Netherlands
Large majority of slides copied from overview talk at TRECVID 2016 by Alan Smeaton, Marc Ritter and George Awad.
Why video to text?

- Robotic vision
- Assist for blinded
- Incident report for surveillance
- Multimedia search
- Movie description for blinded
- Seeing chat bot

Slide credit: Tao Mei
A typical x-media retrieval baseline

Rasiwasia et al., MM 2010
A typical image captioning baseline

Vinyals et al., CVPR 2015
Karpathy & Fei-Fei, CVPR 2015

A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.
A typical video captioning baseline

Xu et al., CVPR 2016
Overview of this new TRECVID challenge

I. Goals, data and tasks
II. Task I: Matching text to video
III. Task II: Generating text from video
GOALS, TASKS AND DATA
Goals

Measure how well an automatic system can describe a video in natural language.

Measure how well an automatic system can match high-level textual descriptions to low-level video features.

Transfer successful image captioning technology to the video domain.
Tasks

Given a set of
  2000 Twitter vine videos.
  2 sets (A and B) of text descriptions for each video.

Task I: **Matching & Ranking**
  Return for each video a ranked list of the most likely text description from each set of A and of B.

Task II: **Description Generation**
  Automatically generate a text description for each video.
Dataset Twitter vine videos

Crawled 30k+ Twitter vine video URLs.

Max video duration == 6 sec.

A subset of 2000 URLs randomly selected.

#761
White guy playing the guitar in a room

#387
An Asian young man sitting is eating something yellow
Alterative video & caption datasets

1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.

1. A woman giving speech on news channel.
2. Hillary Clinton gives a speech.
3. Hillary Clinton is making a speech at the conference of mayors.
4. A woman is giving a speech on stage.
5. A lady speak some news on TV.

1. A child is cooking in the kitchen.
2. A girl is putting her finger into a plastic cup containing an egg.
3. Children boil water and get egg whites ready.
4. People make food in a kitchen.
5. A group of people are making food in a kitchen.

1. A player is putting the basketball into the post from distance.
2. The player makes a three-pointer.
3. People are playing basketball.
4. A 3 point shot by someone in a basketball race.
5. A basketball team is playing in front of speculators.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Organizer</th>
<th>Context</th>
<th>Source</th>
<th>#Video</th>
<th>#Clip</th>
<th>#Sentence</th>
<th>#Word</th>
<th>Vocabulary</th>
<th>Duration (hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>YouCook</td>
<td>SUNY Buffalo</td>
<td>Cooking</td>
<td>Labeled</td>
<td>88</td>
<td>-</td>
<td>2,668</td>
<td>42,457</td>
<td>2,711</td>
<td>2.3</td>
</tr>
<tr>
<td>TACos</td>
<td>MP Institute</td>
<td>cooking</td>
<td>Labeled</td>
<td>123</td>
<td>7,206</td>
<td>18,227</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TACos M-L</td>
<td>MP Institute</td>
<td>cooking</td>
<td>Labeled</td>
<td>185</td>
<td>14,105</td>
<td>52,593</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>M-VAD</td>
<td>UdeM</td>
<td>movie</td>
<td>DVS</td>
<td>92</td>
<td>48,986</td>
<td>55,905</td>
<td>519,933</td>
<td>18,269</td>
<td>84.6</td>
</tr>
<tr>
<td>MPII</td>
<td>MP Institute</td>
<td>movie</td>
<td>DVS+Script</td>
<td>94</td>
<td>68,337</td>
<td>68,375</td>
<td>653,467</td>
<td>24,549</td>
<td>73.6</td>
</tr>
<tr>
<td>MSVD</td>
<td>MSR</td>
<td>multi-category</td>
<td>AMT workers</td>
<td>-</td>
<td>1,970</td>
<td>70,028</td>
<td>607,339</td>
<td>13,010</td>
<td>5.3</td>
</tr>
<tr>
<td>MSR-VTT (10K)</td>
<td>MSRA</td>
<td>20 categories</td>
<td>AMT workers</td>
<td>5,942</td>
<td>10,000</td>
<td>200,000</td>
<td>1,535,917</td>
<td>28,528</td>
<td>38.7</td>
</tr>
<tr>
<td>MSR-VTT (20K)</td>
<td>MSRA</td>
<td>20 categories</td>
<td>AMT workers</td>
<td>14,768</td>
<td>20,000</td>
<td>400,000</td>
<td>4,284,032</td>
<td>49,436</td>
<td>87.8</td>
</tr>
</tbody>
</table>

Slide credit: Tao Mei
Annotation effort

Marc Ritter’s TU Chemnitz group supported manual annotation. Each video annotated by 2 persons (A and B). In total 4000 textual descriptions were produced.

NIST asked annotators to combine 4 facets if applicable:

- **Who** is the video describing (objects, persons, animals, ...etc)?
- **What** are the objects and beings doing (actions, states, events, ...)?
- **Where** (locale, site, place, geographic, ...etc)?
- **When** (time of day, season, ...etc)?
## Annotation UI Overview

**Video DB ID:** 2017  
**GoTo:** 1  
**Annotated:** 0 of 500  
**Current vine:** 1 of 500

<table>
<thead>
<tr>
<th>1 - Where</th>
<th>2 - When</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3 - Who</th>
<th>4 - What</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Description**

- [Video Player]  
- [Save]  
- [Delete]  
- [Toolips on/off]  
- [Next]
Annotation challenges

- Bad video quality
- A lot of simple scenes/events with repeating plain descriptions
- A lot of complex scenes containing too many events to be described
- Clips sometimes appear too short for a convenient description
- Audio track relevant for description but has not been used to avoid semantic distractions
- Non-English Text overlays/subtitles hard to understand
- Cultural differences in reception of events/scene content
- Finding a neutral scene description appears as a challenging task
- Well-known people in videos may have influenced (inappropriately) the description of scenes
- Specifying time of day (frequently) impossible for indoor-shots
- Description quality suffers from long annotation hours
- Some offline vines were detected
- A lot of vines with redundant or even identical content
Annotation process

4900 Vines imported | 100 offline Vines deleted

5000 Vines

1st annotation

750 redundant Vines deleted

4000 Vines

2nd annotation

300 redundant Vines deleted

900 Vines

400 annotations

3600 annotations

final export

2000 annotated Vines

2 heterogeneous annotations per Vine

4000 annotations

exported XML (training)

exported XML (final)
## Annotation statistics

<table>
<thead>
<tr>
<th>UID</th>
<th># annotations</th>
<th>avg (sec)</th>
<th>max (sec)</th>
<th>min (sec)</th>
<th># time (hh:mm:ss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>700</td>
<td>62.16</td>
<td>239.00</td>
<td>40.00</td>
<td>12:06:12</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
<td>84.00</td>
<td>455.00</td>
<td>13.00</td>
<td>11:40:04</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>56.84</td>
<td>499.00</td>
<td>09.00</td>
<td>07:53:38</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>81.12</td>
<td>491.00</td>
<td>12.00</td>
<td>11:16:00</td>
</tr>
<tr>
<td>4</td>
<td>500</td>
<td>234.62</td>
<td>499.00</td>
<td>33.00</td>
<td>32:35:09</td>
</tr>
<tr>
<td>5</td>
<td>500</td>
<td>165.38</td>
<td>493.00</td>
<td>30.00</td>
<td>22:58:12</td>
</tr>
<tr>
<td>6</td>
<td>500</td>
<td>57.06</td>
<td>333.00</td>
<td>10.00</td>
<td>07:55:32</td>
</tr>
<tr>
<td>7</td>
<td>500</td>
<td>64.11</td>
<td>495.00</td>
<td>12.00</td>
<td>08:54:15</td>
</tr>
<tr>
<td>8</td>
<td>200</td>
<td>82.14</td>
<td>552.00</td>
<td>68.00</td>
<td>04:33:47</td>
</tr>
<tr>
<td>total</td>
<td>4400</td>
<td>98.60</td>
<td>552.00</td>
<td>09.00</td>
<td>119:52:49</td>
</tr>
</tbody>
</table>
### Caption examples

<table>
<thead>
<tr>
<th>Partition A</th>
<th>Partition B</th>
</tr>
</thead>
<tbody>
<tr>
<td>a dog jumping onto a couch</td>
<td>a dog runs against a couch indoors at daytime</td>
</tr>
<tr>
<td>in the daytime, a driver let the steering wheel of car and slip on the slide above his car in the street</td>
<td>on a car on a street the driver climb out of his moving car and use the slide on cargo area of the car</td>
</tr>
<tr>
<td>an asian woman turns her head</td>
<td>an asian young woman is yelling at another one that poses to the camera</td>
</tr>
<tr>
<td>a woman sings outdoors</td>
<td>a woman walks through a floor at daytime</td>
</tr>
<tr>
<td>a person floating in a wind tunnel</td>
<td>a person dances in the air in a wind tunnel</td>
</tr>
</tbody>
</table>
Task 1

MATCHING TEXT TO VIDEO
Task 1: Matching & Ranking

Person reading newspaper outdoors at daytime

Person playing golf outdoors in the field

Three men running in the street at daytime

Two men looking at laptop in an office

x 2000

x 2000 type A  ... and ...  X 2000 type B
Mean inverted rank

The mean inverted rank is the average of the inverted ranks of results for a sample of queries.
Matching & Ranking results by run

- MediaMill
- Vireo
- Etter
- DCU
- INF(ormedia)
- NII
- Sheffield

Submitted runs

Mean Inverted Rank

<table>
<thead>
<tr>
<th>Run Name</th>
<th>Graph Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>mediamill_task1_set8.run2.txt</td>
<td>red</td>
</tr>
<tr>
<td>mediamill_task1_set8.run3.txt</td>
<td>red</td>
</tr>
<tr>
<td>mediamill_task1_set8.run4.txt</td>
<td>red</td>
</tr>
<tr>
<td>vireo_fusing_conceptB.txt</td>
<td>yellow</td>
</tr>
<tr>
<td>vireo_fusing_conceptA.txt</td>
<td>yellow</td>
</tr>
<tr>
<td>vireo_fusing_conceptA.txt</td>
<td>yellow</td>
</tr>
<tr>
<td>Etter_mandrA.1</td>
<td>green</td>
</tr>
<tr>
<td>Etter_mandrA.2</td>
<td>green</td>
</tr>
<tr>
<td>INF.ranked_listA swapped.txt</td>
<td>light blue</td>
</tr>
<tr>
<td>INF.ranked_listA new.txt</td>
<td>light blue</td>
</tr>
<tr>
<td>NII.run1.txt</td>
<td>purple</td>
</tr>
<tr>
<td>NII.run2.txt</td>
<td>purple</td>
</tr>
<tr>
<td>NII.run3.txt</td>
<td>purple</td>
</tr>
<tr>
<td>NII.run4.txt</td>
<td>purple</td>
</tr>
<tr>
<td>DCU_adapt_fusionA.txt</td>
<td>blue</td>
</tr>
<tr>
<td>INF.epoch-138.B.txt</td>
<td>green</td>
</tr>
<tr>
<td>Sheffield_UETLahore.ranklistB.testingSet.B_ranklistB.testingSet.txt</td>
<td>red</td>
</tr>
</tbody>
</table>
Samples of top 3 results (set A)

#1271
a woman and a man are kissing each other

#1387
a dog imitating a baby by crawling on the floor in a living room

#324
a dog is licking its nose
Samples of bottom 3 results (set B)

#144
A man touches his chin in a tv show

#1060
A man piggybacking another man outdoors

#414
A woman is following a man walking on the street at daytime trying to talk with him
Participant: DCU

Preprocess 10 frames/video to detect 1,000 objects (ImageNet), 94 crowd behaviour concepts (WWW dataset), locations (Places2)

4 runs, baseline BM25, Word2vec, and fusion
Participant: VIREO (CUHK)

Adopted their zero-example MED system in reverse

Used a concept bank of 2000 concepts trained on MSR-VTT, Flickr30k, MS-COCO and TGIF datasets

Runs testing concept-based approach vs attention-based deep models, finding deep models perform better, motion features dominate performance
Participant: Etter Solutions

Focused on concepts for Who, What, When, Where

Used a subset of ImageNet plus scene categories from the Places database

Applied concepts to 1 fps with sliding window, mapped this to document vector, and calculated similarity score
PARTICIPANT: MEDIAMILL

Key idea

Project sentences into a video feature space

Match sentences and videos in this space
Training time

Predict the visual representation of text

Best settings

- Use MSR-VTT for training, minimize mean squared error between text and video Vine videos for architecture hyperparameter validation
- ImageNet-Shuffle with GoogleNet mean pooling over video
- Word2vec 500-dim, trained on user tags of 30m Flickr images
Test time

Predict the visual representation of (unseen) text

For an unseen video predict its video feature, return closest sentence.
Predicting audio also, small increase
Results

Word2VisualVec: Image and Video to Sentence Matching by Visual Feature Prediction

Jianfeng Dong, Xirong Li, Cees G. M. Snoek
https://arxiv.org/abs/1604.06838
Task 2

GENERATING TEXT FROM VIDEO
Task 2: Description Generation

Given a video

Generate a textual description

Who?  What?  Where?  When?

"a dog is licking its nose"
Approximate human judgement of text quality at corpus level

Measures fraction of common N-grams in source and target.

N-gram matches for a high N rarely occur at sentence-level, so poor performance of BLEU@N when comparing only individual sentences, better comparing paragraphs or higher.

Often we see B@1, B@2, B@3, B@4 ... we do B@4.
Computes unigram precision and recall, extending exact word matches to include similar words based on WordNet synonyms and stemmed tokens.

Based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision.
CIDEr (Consensus-Based Image Description Evaluation) is another recent metric

It weights each n-gram with TF-IDF

The CIDEr score for n-grams of length n:
   Computed using the average cosine similarity between the candidate and reference sentences

Vedantam et al., CVPR 2015
BLEU results

Overall system scores

INF(ormedia)  Sheffield  NII  MediaMill  DCU

BLEU score

INF.jiac-msvd.txt  Sheffield_UETLahore_discrip...  mediamill_UETLahore_discrip...  mediamill_UETLahore_discrip...  INF.hrme_24.txt  mediamill_task2_run4.txt  NII.description.run_2_resnet.txt  mediamill_task2_run2.txt  NII.description.run_2_resnet...  NII.description.run_1_resnet...  NII.description.run_3_c3d.txt
METEOR results

Overall system score

METEOR score

- INF(ormedia)
- Sheffield
- NII
- MediaMill
- DCU
An example from run submissions – 7 unique examples

1. a girl is playing with a baby
2. a little girl is playing with a dog
3. a man is playing with a woman in a room
4. a woman is playing with a baby
5. a man is playing a video game and singing
6. a man is talking to a car
7. A toddler and a dog
Participant: DCU

Train on MS-COCO using NeuralTalk2, a RNN

One caption per keyframe, captions then fused
Participant: Informedia

Focus on generalization ability of caption models

Trained 4 caption models on 3 datasets (MS-COCO, MS-VD, MSR-VTT), achieving sota on those models based on VGGNet concepts and Hierarchical Recurrent Neural Encoder for temporal aspects

Results explore transfer models to TRECVID-VTT
Participant: Sheffield / Lahore

Identified a variety of high level concepts for frames

Detect and recognize faces, age and gender, emotion, objects, (human) actions

Varied the frequency of frames for each type of recognition

Runs based on combinations of feature types
PARTICIPANT: MEDIAMILL

**Idea:** Re-use Video Tags for Captioning

<table>
<thead>
<tr>
<th>Predicted tags</th>
<th>Generated caption</th>
</tr>
</thead>
<tbody>
<tr>
<td>track</td>
<td>a group of people are running in a <strong>race track</strong></td>
</tr>
<tr>
<td>race</td>
<td></td>
</tr>
<tr>
<td>field</td>
<td></td>
</tr>
<tr>
<td>woman</td>
<td></td>
</tr>
<tr>
<td>soccer</td>
<td>a <strong>soccer player</strong> is <strong>playing</strong> a goal on a soccer field</td>
</tr>
<tr>
<td>player</td>
<td></td>
</tr>
<tr>
<td>game</td>
<td></td>
</tr>
<tr>
<td>playing</td>
<td></td>
</tr>
<tr>
<td>dance</td>
<td><strong>people</strong> are <strong>dancing</strong> on a stage</td>
</tr>
<tr>
<td>people</td>
<td></td>
</tr>
<tr>
<td>woman</td>
<td></td>
</tr>
<tr>
<td>dancing</td>
<td></td>
</tr>
</tbody>
</table>
Our solution

Google’s model for sentence generation

GoogleNet-shuffle

Google’s model [Vinyals et al. CVPR 2015]

models are walking down the runway
models are walking on the runway
a woman is walking down the runway
a woman is dancing
...
models are walking in a fashion show
models are walking on the ramp
Our solution

Better initialization by tag embedding

Re-encoding by Word2VisualVec

models are walking down the runway
models are walking on the runway
a woman is walking down the runway
a woman is dancing
...
models are walking in a fashion show
models are walking on the ramp
Our solution

Rerank sentences by matching with video tags

models are walking in a fashion show

Maximize tag matches

Re-encoding by Word2VisualVec

Google’s model [Vinyals et al. CVPR 2015]

models are walking down the runway
models are walking on the runway
a woman is walking down the runway
a woman is dancing
...
models are walking in a fashion show
models are walking on the ramp
Heuristics to add ‘where’

Two simple rules to append ‘where’ description to the end of the generated sentences:

1. Add “on a $sport\_name$ field” if $sport$ appear in the sentence, such as basketball, baseball, and football.

2. Add “on a stage” if “sing” or “dance” appear in the sentence.
Adding “where” improves performance slightly.
Conclusions on video to text task

Good participation, good finishing

METEOR scores are higher than BLEU, should have used CIDEr also

Lots of available training sets, some overlap ... MSR-VTT, MS-COCO, Places2, ImageNet, YouTube2Text, MS-VD ..

Only few applied the 4-facet description in their submissions

Task continues in 2017

www.ceessnoek.info