Video Description

Ir. He Ming Zhang
Advisor: Prof. C.-C. Jay Kuo
Outline

• Motivation
• Problem definition
• Preliminaries
• Related works
• Conclusion
Outline

• Motivation
• Problem definition
• Preliminaries
• Related works
• Conclusion
Motivation

We have ...
• huge amount of video
  Every minute, 100 hours of video are uploaded to YouTube\(^1\).

We lack ...
• time to watch all the videos
• description of videos

We want ...
• computer to understand the visual content
• computer to describe the visual content

\(^1\)https://www.youtube.com/yt/press/statistics.html accessed on 2015-02-06.
Motivation

Applications
• Tagging

 VS

woman
dog

A woman is walking a dog
A woman is chased by a dog

• Indexing

Improving indexing and search quality for online videos.
Motivation

Applications
• Human-robot interaction
  Describing movies for the blind

As well as for the lazy people...
Outline

• Motivation

• Problem definition
  • Problem for researchers
  • Datasets
  • Evaluation

• Preliminaries

• Related works

• Conclusion
Problem Definition

Problem for researchers

• From video clip to natural language
  • Input - video clip
    Typically from several to few tens of seconds
    A specific domain or open domain (“in the wild”)
  • Output - natural language that describes the content of the input
    One or more sentence(s) in natural language (usually in English)

• Different from image description
  • Video contains more information
    more or less difficult?
### Problem definition

#### Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>multi-sentence</th>
<th>domain</th>
<th>sentence source</th>
<th>videos</th>
<th>clips</th>
<th>sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>YouCook [1]</td>
<td>x</td>
<td>cooking</td>
<td>crowd</td>
<td>88</td>
<td>-</td>
<td>2668</td>
</tr>
<tr>
<td>TACoS [2]</td>
<td>x</td>
<td>cooking</td>
<td>crowd</td>
<td>127</td>
<td>7206</td>
<td>18227</td>
</tr>
<tr>
<td>TACoS Multi-Level [3]</td>
<td>x</td>
<td>cooking</td>
<td>crowd</td>
<td>185</td>
<td>14105</td>
<td>52593</td>
</tr>
<tr>
<td>MVAD [5]</td>
<td>x</td>
<td>open</td>
<td>professional</td>
<td>92</td>
<td>48986</td>
<td>55904</td>
</tr>
<tr>
<td>MPII-MD [6]</td>
<td>x</td>
<td>open</td>
<td>professional</td>
<td>94</td>
<td>68337</td>
<td>68375</td>
</tr>
</tbody>
</table>
Problem Definition

Datasets

- **Trend - more challenging**
  - **Broader domains**
    From single domain to open domain

- **Larger datasets**
  More sentences/clips
Problem Definition

Datasets

• MSVD
  • YouTube videos
    e.g. from 0:33 to 0:46,
    http://www.youtube.com/watch?v=mv89psg6zh4

• Multi-descriptions
  - A bird in a sink keeps getting under the running water from a faucet.
  - A bird is bathing in a sink.
  - A bird is splashing around under a running faucet.
  - A bird is standing in a sink drinking water that is pouring out of the faucet.
  - ...


Problem Definition

Datasets

• MSVD
  • YouTube videos
e.g. from 0:11 to 0:14,
  http://www.youtube.com/watch?v=cSDkshD2ME0

• Multi-descriptions
  – Someone behind a rock shoots a man on horseback who slumps forward onto his horse.
  – A man shoots a man on a horse.
  – A man hiding behind a rock shoots a man on horseback with a rifle.
  – A man is shooting another man.
  – ...

Problem Definition


Problem Definition

Example results from state-of-the-art [7]

A man is pouring oil into a pot.

A dog is playing in a bowl.

Problem Definition

Evaluation

• Difficulties

Natural language is rich

Description may be partially wrong/correct

No standard metric (a few metrics are used by different researchers)
Problem Definition

Evaluation

• Methods

   Human evaluation
   Binary rating (correct/ incorrect)

   Scale rating (e.g. 1~5)
Problem Definition

Evaluation

• Methods

Automated evaluation:
  BLEU (BiLingual Evaluation Understudy)
  - one of the first metrics to achieve a high correlation with human judgements of quality
  - modified version of F-score
  - example:
    Ref: Israeli officials are responsible for airport security.
    A: Israeli officials responsibility of airport safety.
    B: Airport security Israeli officials are responsible.
    Score: A - 0%
    B - 52%
Problem Definition

Evaluation

• Methods

Automated evaluation:
  METEOR (Metric for Evaluation of Translation with Explicit ORdering)
  - higher correlation with human judgements in both corpus and sentence level
  - modified version of F-score
  - flexible matching (partial credit)

Ref: Joe goes home
A: Jim went home
B: Jim walks home
Outline

• Motivation

• Problem definition

• Preliminaries
  • Statistical Machine Translation (SMT)
  • Recurrent Neural Network (RNN)

• Related works

• Conclusion
Preliminaries

We need ...

- recognition (CRF, CNN, etc)
  - objects
  - scene / background
  - events

- language processing (manual rules, SMT, RNN, etc)
  - word selection
  - sentence generation
Preliminaries

n-gram

• Markov model with higher order
  In a language model, the probability of a word is conditioned on some number of previous words.

• Properties and usages
  It is used in statistical natural language processing.
Preliminaries

Statistical Machine Translation (SMT)

• **Statistical model**

  It *translates* the document according to the probability distribution $p(T|S)$;

  Examples:
  
  - **Word-level**
    
    S (Dutch): *Ik* *ben* *een* *promovendus*.
    T (English): *I* *am* *a* *PhD student*.

  - **Semantic-level**
    
    S (Dutch): *Ik* *ben* *het* *er* *mee* *eens*.
    T (English): *I* *am* *it* *here with* *in agreement*.
    
    T (English): *I* agree *with* *it*.

  The system can not store all native strings and their translation, therefore the language models are approximated by n-gram models.
Preliminaries

Recurrent Neural Network (RNN)

• Internal memory
  A class of neural network where connections between units form a directed cycle;

\[
\begin{align*}
  x_t &\rightarrow c_t \\
  c_t &\rightarrow h_t
\end{align*}
\]

• Properties and usages
  It can process sequential data and be used for language modeling, handwriting recognition, etc.

Traditional RNNs are very hard to train;
Preliminaries

Recurrent Neural Network (RNN)

- LSTM (Long Short-Term Memory)

Internal memory for an arbitrary length of time;
- **Input** gate: determines when the unit should let the input flow into its memory;
- **Forget** gate: determines when the unit should forget the value in its memory;
- **Output** gate: determines when the unit should output the value in its memory.

Outline

• Motivation

• Problem definition

• Preliminaries

• Related works
  • Early works
  • Recent works
  • Summary

• Conclusion
Related works

Early works

• Youtube2text [9]

Mine (Subject, Verb, Object) triplets from the natural language descriptions of the videos.
Build a separate semantic hierarchy for each part of the triplet ($H_S$, $H_V$, and $H_O$).
Detect objects and activities using existing object and motion descriptors.

Figure 3: Small portions of the Hierarchies learned over Subjects, Verbs and Objects

Early works

- Youtube2text

Language model

- For activities that are unseen during training, they expand detected verbs with similar verbs.
  
  e.g. for (person, move, car), expand "move" with "ride" and "drive" without training videos for "ride" or "drive"

- Select the best triplet

\[
\text{score} = p(S \mid \text{video}) \times p(V_{\text{expand}} \mid \text{video}) \times \text{Similarity}(V_{\text{expand}}, V_{\text{original}}) \times p(O \mid \text{video}) \times SVO\_\text{likelihood}
\]

- Generate sentences using **manual template**
Related works

Early works
  • Youtube2text

Experimental results on MSVD

Automated evaluation

Human evaluation
  - For each test video, retrieve the 3 most similar videos according to the SVO triplet
  - Ask workers to rate, on a scale of 1 to 5, how relevant the retrieved videos are with respect to the given video.
  - Average rating obtained is 1.99

Figure 4: Zero-shot Activity Recognition
Early works

- Translating video content to natural language descriptions [10]

Encoder-decoder framework:
Video description is phrased as a translation problem from video content to natural language and used a semantic representation of the video content as intermediate step.

Encoder : Conditional Random Field
Decoder : Statistical Machine Translation

Related works

Early works

• Translating video content to natural language descriptions

Experimental results on TACoS

CRF+SMT: the person cracks the eggs
Human: the person dumps any remaining whites of the eggs from the shells into the cup with the egg whites

CRF+SMT: the person gets out a cutting board from the loaf of bread from the fridge
Human: the person gets the lime, a knife and a cutting board
Recent works


Recent works

- Long-term RNN

Figure 4: Our approaches to video description. (a) LSTM encoder & decoder with CRF max (b) LSTM decoder with CRF max (c) LSTM decoder with CRF probabilities. (For larger figure zoom or see supplemental).
Related works

Recent works
• Long-term RNN

LSTM as both encoder and decoder

Use CRF max
Related works

Recent works
• Long-term RNN

LSTM as decoder
Use CRF max
Related works

Recent works
• Long-term RNN

LSTM as decoder

Use CRF probabilities
Related works

Recent works
• Long-term RNN

Experimental results on TACoS

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Input</th>
<th>BLEU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT[9]</td>
<td>CRF max</td>
<td>24.9</td>
</tr>
<tr>
<td>LSTM (a)</td>
<td>CRF max</td>
<td>25.3</td>
</tr>
<tr>
<td>LSTM (b)</td>
<td>CRF max</td>
<td>27.4</td>
</tr>
<tr>
<td>LSTM (c)</td>
<td>CRF probabilities</td>
<td>28.8</td>
</tr>
</tbody>
</table>
Related works

Recent works
• Mean pooling [12]

Basic encoder-decoder framework

Encoder: pre-trained CNN
  for each frame separately
  mean-pooling on all frames

Decoder: LSTM

Related works

Recent works
• Mean pooling [12]

Experimental results using METEOR (%)

<table>
<thead>
<tr>
<th>Methods</th>
<th>MSVD</th>
<th>MVAD</th>
<th>MPII-MD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean pool - AlexNet</td>
<td>26.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean pool - VGG</td>
<td>27.7</td>
<td>6.1</td>
<td>6.7</td>
</tr>
<tr>
<td>Mean pool - AlexNet COCO pre-trained</td>
<td>29.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean pool - GoogleNet</td>
<td>28.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Related works

Recent works
• Temporal attention [13]

Basic encoder-decoder framework

Encoder: pre-trained CNN on ImageNet
used for each frame separately
+ temporal information

Decoder: LSTM

Related works

Recent works

- Temporal attention

Exploiting temporal structure

Local: 3D-CNN
three 3D convolutional layer
temporal features obtained by
max-pooling

Global: temporal attention
mechanism
Related works

Recent works
• Temporal attention

Experimental results using METEOR (%)

<table>
<thead>
<tr>
<th>Methods</th>
<th>MSVD</th>
<th>MVAD</th>
<th>MPII-MD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean pool - GoogleNet</td>
<td>28.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temporal attention - GoogleNet</td>
<td>29.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temporal attention - GoogleNet + 3D-CNN</td>
<td>29.6</td>
<td>4.3</td>
<td></td>
</tr>
</tbody>
</table>
Recent works

- S2VT [14]

Recent works

- **S2VT [17]**

No separate encoder-decoder

Use the same LSTM for both encoder and decoder
Related works

Recent works
- S2VT [17]

Experimental results using METEOR (%)

<table>
<thead>
<tr>
<th>Methods</th>
<th>MSVD</th>
<th>MVAD</th>
<th>MPII-MD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean pool - AlexNet</td>
<td>26.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean pool - VGG</td>
<td>27.7</td>
<td>6.1</td>
<td>6.7</td>
</tr>
<tr>
<td>Mean pool - GoogleNet</td>
<td>28.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temporal attention - GoogleNet</td>
<td>29.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temporal attention - GoogleNet + 3D-CNN</td>
<td>29.6</td>
<td>4.3</td>
<td></td>
</tr>
<tr>
<td>S2VT (Flow) - AlexNet</td>
<td>24.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2VT (RGB) - AlexNet</td>
<td>27.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2VT (RGB) - VGG</td>
<td>29.2</td>
<td>6.7</td>
<td>7.1</td>
</tr>
<tr>
<td>S2VT (RGB + Flow) - VGG for RGB, AlexNet for Flow</td>
<td>29.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 2. Our hierarchical RNN for video captioning. **Green** denotes the input to the framework, **blue** denotes the output, and **red** denotes the recurrent components. The **orange** arrow represents the reinitialization of the sentence generator with the current paragraph state. For simplicity, we only draw a single video feature pool in the figure. In fact, both appearance and action features go through a similar attention process before they are fed into the multimodal layer.
Recent works

• hRNN

Two language generators: sentence generator and paragraph generator

Multimodal layer after the recurrent layer to combine video content features

2D CNN for frame feature extraction, 3D CNN for video feature extraction
Related works

Recent works

- hRNN

Experimental results using METEOR (%)

<table>
<thead>
<tr>
<th>Methods</th>
<th>MSVD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean pool - VGG</td>
<td>27.7</td>
</tr>
<tr>
<td>Temporal attention - GoogleNet + 3D-CNN</td>
<td>29.6</td>
</tr>
<tr>
<td>S2VT (RGB) - VGG</td>
<td>29.2</td>
</tr>
<tr>
<td>S2VT (RGB + Flow) - VGG for RGB, AlexNet for Flow</td>
<td>29.8</td>
</tr>
<tr>
<td>hRNN - VGG</td>
<td>31.1</td>
</tr>
<tr>
<td>hRNN- C3D</td>
<td>30.3</td>
</tr>
<tr>
<td>hRNN - VGG + C3D</td>
<td>32.6</td>
</tr>
</tbody>
</table>
Related works

Summary

Keyword-sentence frameworks

Encoder-decoder frameworks

Encoder

Decoder

CRF

2D CNN

2D CNN

3D CNN

RNN

SMT

RNN
Related works

Future

• Encoder-decoder framework
  - encoder: +scene classification
  - encoder to decoder: better structure
  - decoder

• Other framework
Conclusion

Video description is ...

• **important**
  - tagging
  - indexing
  - human-robot interaction

• **difficult**
  - implementation
  - evaluation

• **under development**
  - datasets
  - evaluation methods
  - algorithms