DL-61-86 at TRECVID 2017: Video-to-Text Description

Jianfeng Dong\textsuperscript{1,2}, Shaoli Huang\textsuperscript{1}, Duanqing Xu\textsuperscript{2}, Dacheng Tao\textsuperscript{1}

1. UBTECH Sydney AI Centre, SIT, FEIT, University of Sydney
2. College of Computer Science and Technology, Zhejiang University
Matching and Ranking subtask

Query Videos:

Candidate sentences to be ranked:
- a man speaks to audiences indoors
- a person skates indoors
- **Athletics make a choreography in gym.** ✓
- a woman is holding a phone to her ear.
Cross-media Similarity

Video

Sentence

Athletics make a choreography in gym.

Key question: how to compute cross-media similarity?
Our Model

Key components:
- Spatial Enhanced Video Representation
- Multi-scale Sentence Representation
Common way of video representation

Use RNN to capture spatio-temporal information.
Spatial Enhanced Video Representation

Learn a GRU with skip-connections that allow bypassing of the spatial features.
Multi-scale Sentence Representation

It merges bag-of-words, word2vec and GRU sentence features and letting the model figure out the optimal way for combining them.

Athletics make a choreography in gym.

Objective Function

Triplet Ranking Loss:

\[
l(x, q; \theta) = \sum_{q'} [\alpha + s(x, q') - s(x, q)]_+ \]

\[
[x]_+ \equiv \max(x, 0) \]

Improved Triplet Ranking Loss: (using hardest example)

\[
l(x, q; \theta) = \max_{q'} [\alpha + s(x, q') - s(x, q)]_+ \]
Other winning components

1. Use more training data and fine-tune the model on the data provided by TRECVID

2. Use pre-trained word2vec to initialize word embedding before the LSTM/GRU

3. Use batch normalization after the FC layer

4. Fuse different models
Datasets

Table 1. Overview of datasets used in our submission.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Videos</th>
<th># Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSVD</td>
<td>1,970</td>
<td>80,863</td>
</tr>
<tr>
<td>MSR-VTT</td>
<td>10,000</td>
<td>200,000</td>
</tr>
<tr>
<td>TGIF</td>
<td>101,980</td>
<td>125,672</td>
</tr>
<tr>
<td>Validation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tv2016train</td>
<td>200</td>
<td>400</td>
</tr>
<tr>
<td>Fine-Tune</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tv2016test</td>
<td>1,915</td>
<td>3,830</td>
</tr>
</tbody>
</table>

Datasets provided by TRECVID 2016

External datasets
Other winning components

1. Use more training data and fine-tune the model on the data provided by TRECVID

2. Use pre-trained word2vec to initialize word embedding before the LSTM/GRU

3. Use batch normalization after the FC layer

4. Fuse different models
**Pre-trained word2vec**

1. Word2vec trained on the Google news documents

2. Word2vec trained on the tags of Flickr images

Word2vec with the dimensionality of 500 trained on 30 million Flickr tags.

**URL:** https://drive.google.com/open?id=0B1OT7LFjhrF_RWptMjY2TVBqLWc

Other winning components

1. use more training data and fine-tune the model on the data provided by TRECVID

2. Use pre-trained word2vec to initialize word embedding before the LSTM/GRU

3. Use batch normalization after the FC layer

4. Fuse different models
Other winning components

1. Use more training data and fine-tune the model on the data provided by TRECVID

2. Use pre-trained word2vec to initialize word embedding before the LSTM/GRU

3. Use batch normalization after the FC layer

4. Fuse different models
Fuse with Word2VisualVec

Model fusion is simple but it is effective.

Improve Word2VisualVec:
1. Use multi-scale text representation to embed sentence
2. Use the improved triplet ranking loss

Evaluation Results

Our submissions lead the evaluation with a great margin.

(a) Results on the test set 2

(d) Results on the test set 5
Take-home Messages

- Use Spatial Enhanced Video Representation to embed videos
- Use Multi-scale Sentence Vectorization to embed sentences
- Some other winning components
  1. use more training data and fine-tune the model on the data provided by TRECVID
  2. Use pre-trained word2vec to initialize word embedding before the LSTM/GRU
  3. Use batch normalization after the FC layer
  4. Fuse different models