Word2VisualVec for Video-To-Text Matching and Ranking

Jianfeng Dong\textsuperscript{1}, Xirong Li\textsuperscript{2}, Xiaoxu Wang\textsuperscript{2}, Qijie Wei\textsuperscript{2}, Weiyu Lan\textsuperscript{2}, Cees G. M. Snoek\textsuperscript{3}

Zhejiang University\textsuperscript{1}
Renmin University of China\textsuperscript{2}
University of Amsterdam\textsuperscript{3}
Our idea

Project sentences into a **video feature space**

Match sentences and videos in this space
Solution: Word2VisualVec

Transform text into a video feature vector

\[ \sigma(W_1 * s(q) + b_1) \]

\[ \sigma(W_2 * h_1(q) + b_2) \]

Text \rightarrow pooling \rightarrow \Phi(x) \rightarrow CNN \rightarrow video

Word2VisualVec

Transform text into a video feature vector

word matrix $\rightarrow$ pooling $\rightarrow$ $s(q)$ $\rightarrow$ $h_1(q)$ $\rightarrow$ $\Phi(x)$ $\rightarrow$ CNN $\rightarrow$ video

$\sigma(W_1*s(q)+b_1)$ $\sigma(W_2*h_1(q)+b_2)$
Word2VisualVec

Transform text into a video feature vector

Word2vec + Multi-layer perceptron

Minimize Mean Squared Error between text vector and video vector
Implementation

Two video features
- Visual: Mean pooling over frame-level CNN feature extracted by GoogleNet-shuffle [Mettes et al. ICMR16]
- Visual + Audio: GoogleNet-shuffle + Bag of quantized MFCC

Word2Vec
- 500-dim, trained on user tags of 30m Flickr images

Word2VisualVec architecture
- For predicting the visual feature: 500-1000-1024
- For predicting the visual + audio feature: 500-1000-2048

Training set
- MSR-VTT training set of 6,513 videos [Xu et al. CVPR16]

Validation set
- TRECVID 200 training videos
Adding the audio feature provides some improvement
Video-to-text results

Text → Visual
a man with a beard is wearing glasses

Text → Visual + Audio
man talks into the camera

Text → Visual
soccer players are blocking the ball on a soccer field

Text → Visual + Audio
a soccer player scores a goal on a soccer field

More results at http://lixirong.net/demo/vtt/tv16.html
Video Description Generation

J. Dong, X. Li, W. Lan, Y. Huo, C. Snoek,
Early embedding and late reranking for video captioning,
ACM Multimedia 2016
## 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 race track</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 soccer player is playing 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>people are dancing 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

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

Rerank sentences by matching with video tags

Maximize tag matches

models are walking in a fashion show

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.
Description generation results

Adding “where” improve the performance
Live demo

http://lixirong.net/demo/vtt

Video-to-Text

a fashion model is walking down a runway
Tags: model, runway, walking, woman

accept video file less than 10 MB
Conclusion

**Word2VisualVec** for video-to-text matching in video space

**Early embedding and late reranking** improves LSTM based video captioning

Winning results in the VTT task

Xirong Li