Early Embedding and Late Reranking for Video Captioning

Jianfeng Dong¹, Xirong Li², Weiyu Lan², Yujia Huo², Cees G. M. Snoek³

Zhejiang University¹
Renmin University of China²
University of Amsterdam³
Live demo

http://lixirong.net/demo/vtt

Video-to-Text

Upload Video

How is the generated sentence?
- [ ] good
- [ ] just so so
- [ ] bad

How would you describe this video?

[Submit]

a fashion model is walking down a runway
Tags: model, runway, walking, woman
Re-use Video Tags for Captioning

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

Google’s model for sentence generation

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
Proposed Video Captioning System

Better initialization by early embedding

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
Proposed Video Captioning System

Rerank sentences by matching with video tags

models are walking in a fashion show

Maximize tag matches

models are walking down 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 runway

Word2VisualVec [Dong et al. arXiv 2016]

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

Best CIDEr-D
measuring human-likeness of generated captions
Conclusion

Early embedding and Late reranking improves LSTM based video captioning

Word2VisualVec plus our winning TRECVID Video-to-Text results highlighted in Rising Star Symposium