Word2VisualVec for Video-To-Text Matching and Ranking

Jianfeng Dong[1], Xirong Li[2], Xiaoxu Wang[2],
Qijie Wei[2], Weiyu Lan[2], Cees G. M. Snoek[3]

Zhejiang University[1]
Renmin University of China[2]
University of Amsterdam[3]

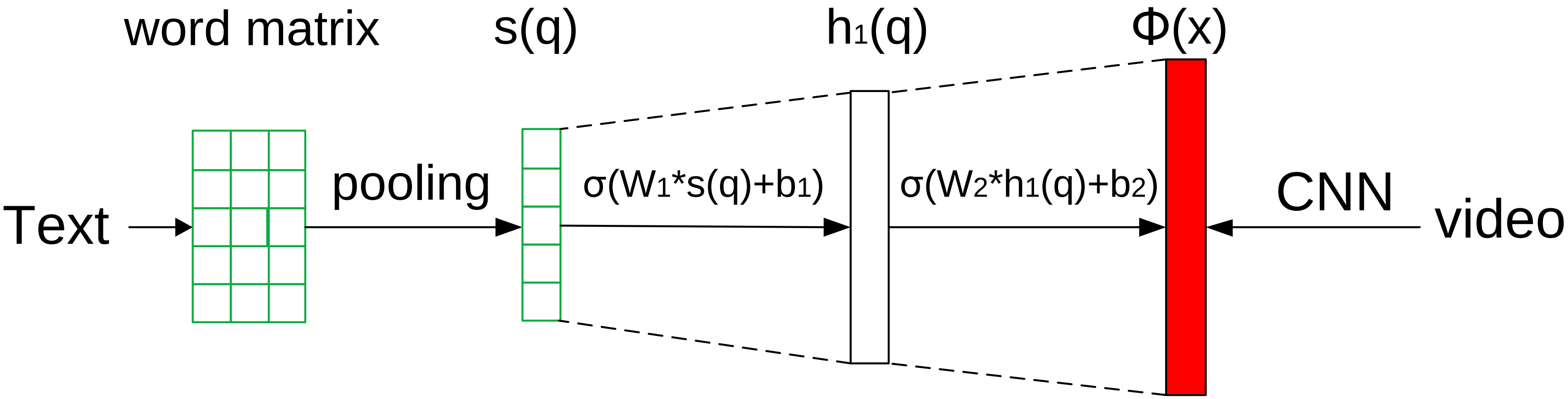
Our idea

Project sentences into a **video feature space**

Match sentences and videos in this space

Solution: Word2VisualVec

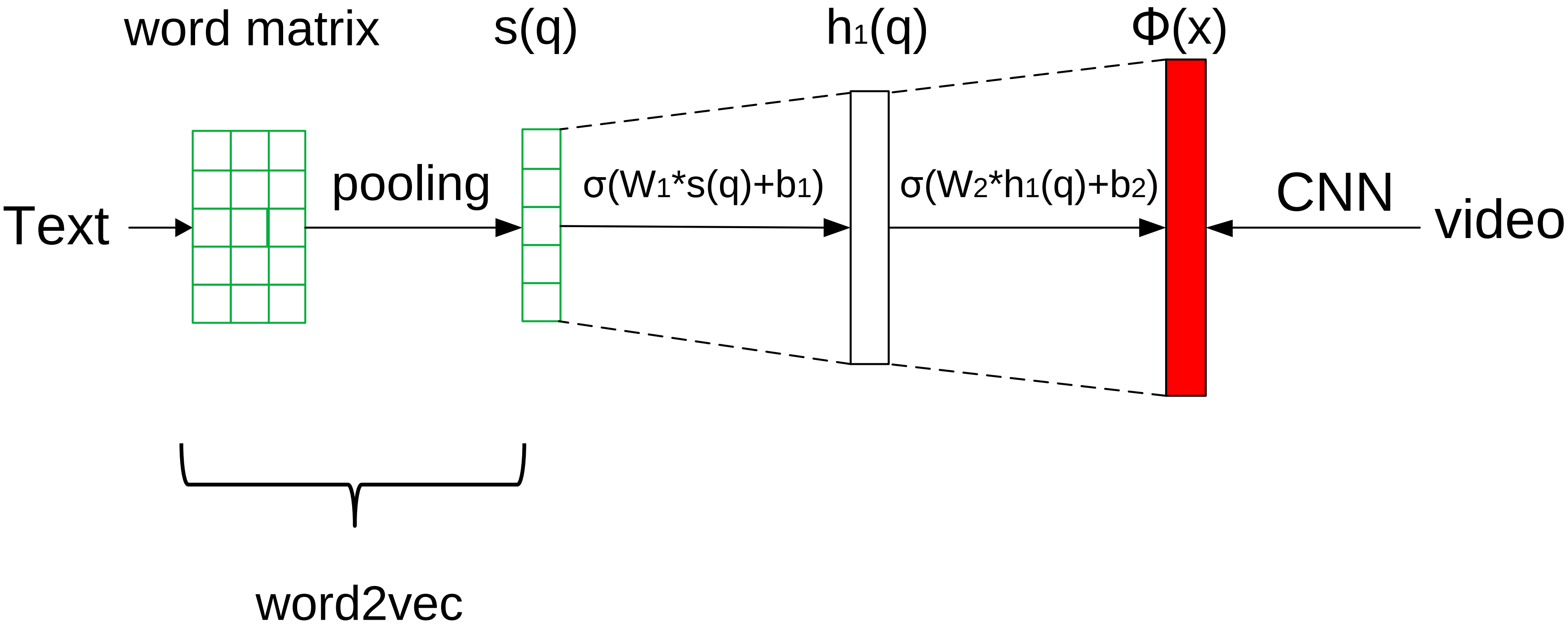
Transform text into a video feature vector



J. Dong, X. Li, C. Snoek, Word2VisualVec: Cross-Media Retrieval by Visual Feature Prediction, Arxiv:1604.06838, 2016

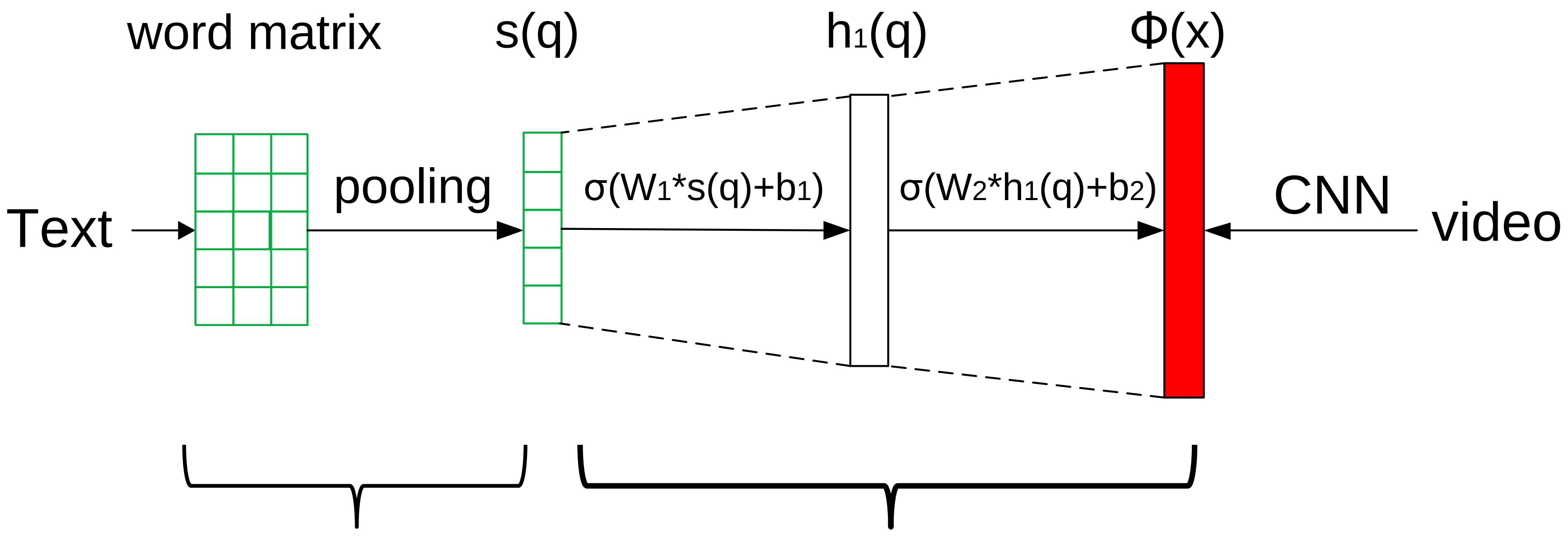
Word2VisualVec

Transform text into a video feature vector



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]

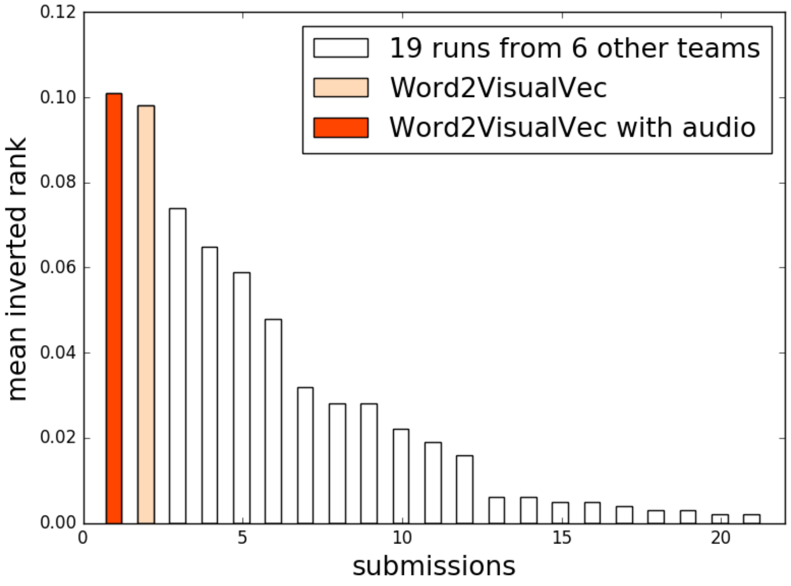
Validation set

- TRECVID 200 training videos

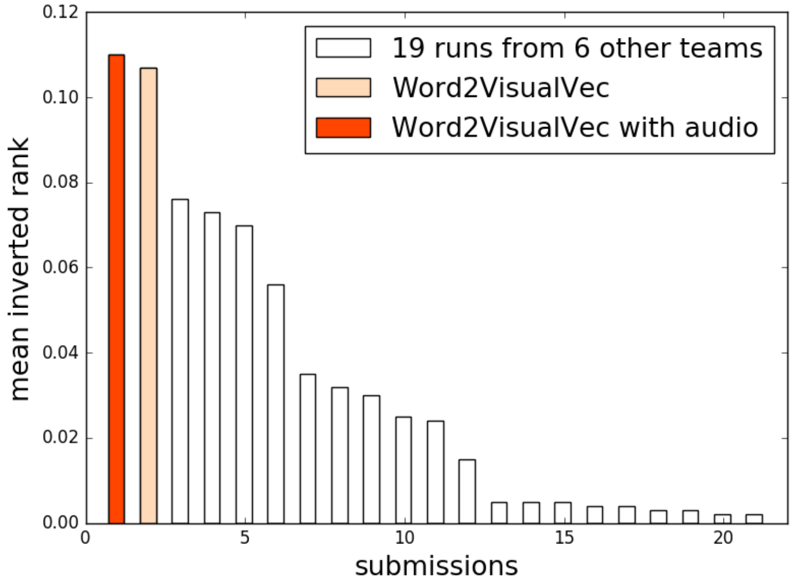
Video-to-text results

Word2VisualVec is effective

set A

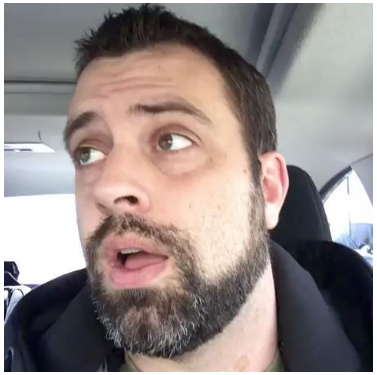


set B



Adding the audio feature provides some improvement

Video-to-text results

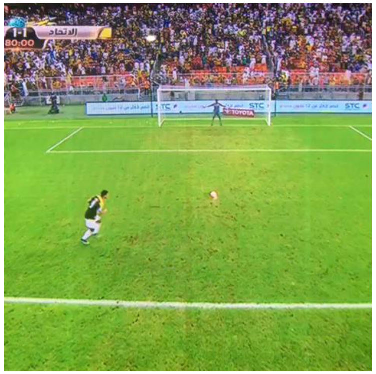


Text → Visual

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

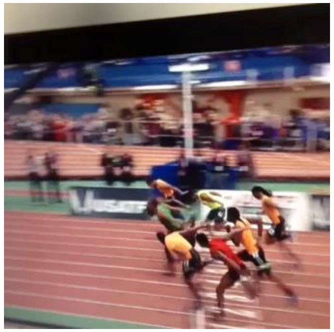
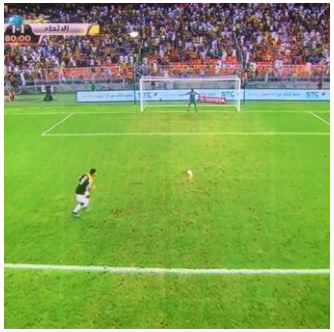
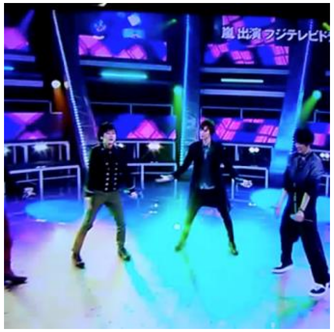
a soccer player scores a goal on a soccer field

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

Video Description Generation

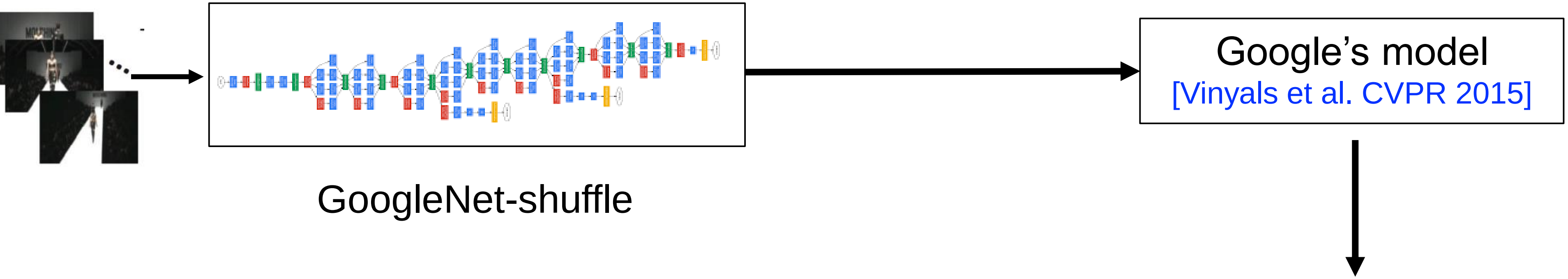
J. Dong, X. Li, W. Lan, Y. Huo, C. Snoek,
Early embedding and late reranking for video captioning,
ACM Multimedia 2016

Idea: Re-use Video Tags for Captioning

	Predicted tags	Generated caption
	track race field woman	a group of people are running in a **race track**
	soccer player game playing	a **soccer player** is **playing** a goal on a soccer field
	dance people woman dancing	**people** are **dancing** on a stage

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

…

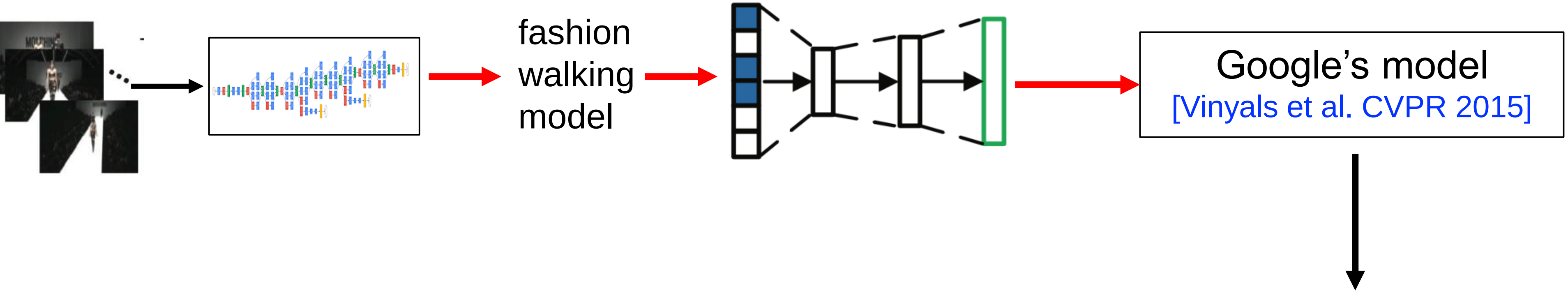
models are walking in a fashion show

models are walking on the ramp

Our solution

Better initialization by tag embedding

Re-encoding by Word2VisualVec



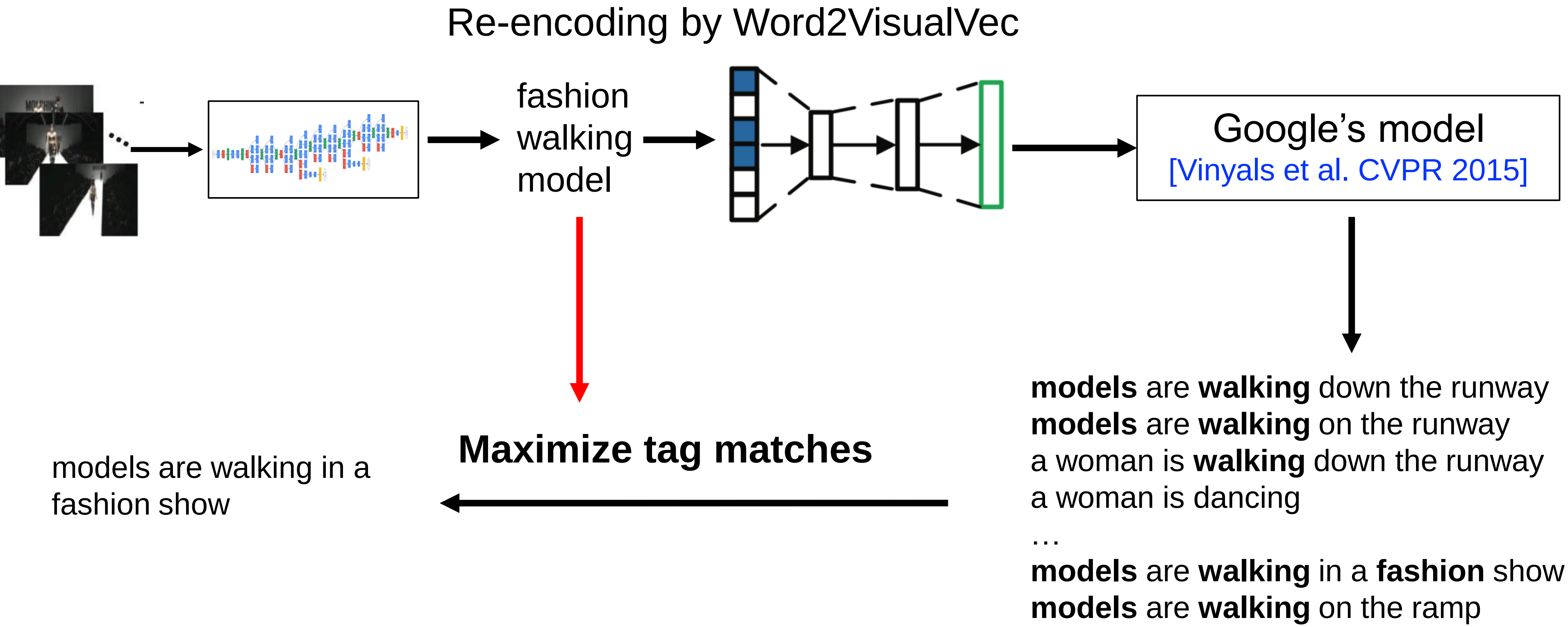
fashion
walking
model

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

Rerank sentences by matching with video tags



Re-encoding by Word2VisualVec

fashion
walking
model

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

Maximize tag matches

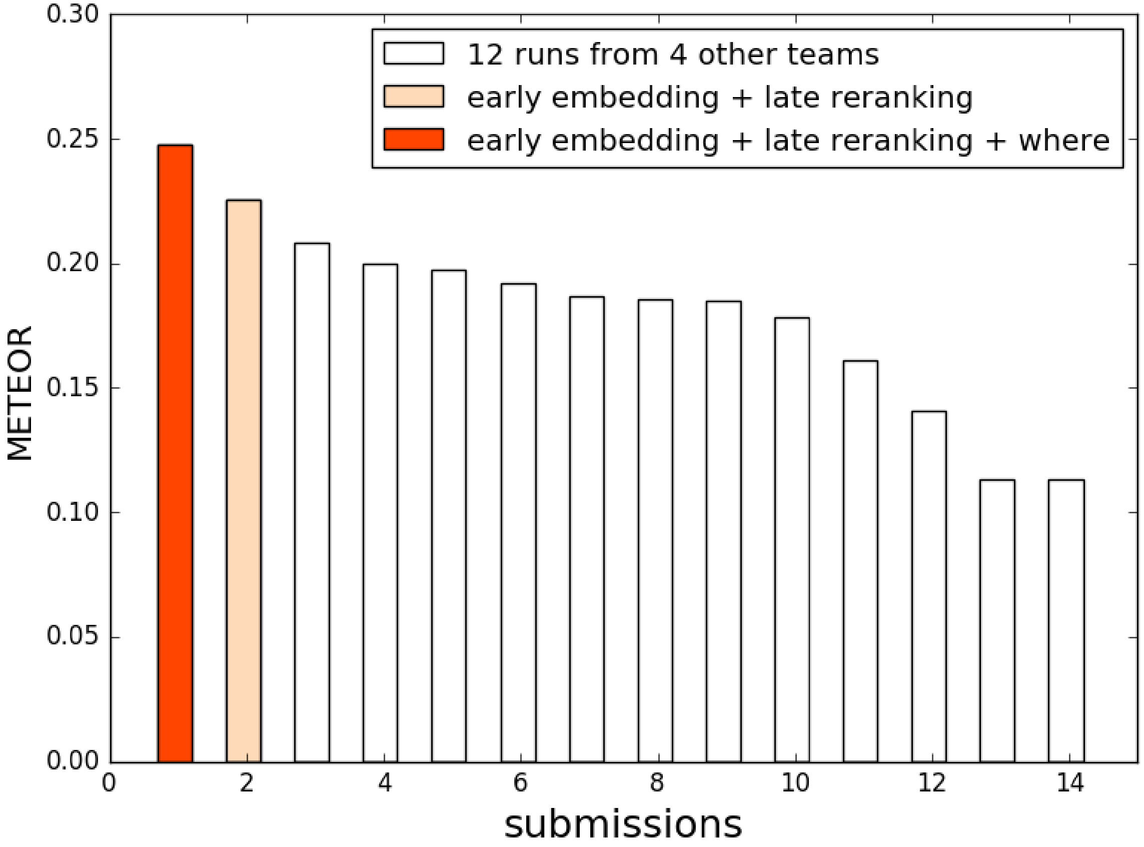
models are walking in a fashion show

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.

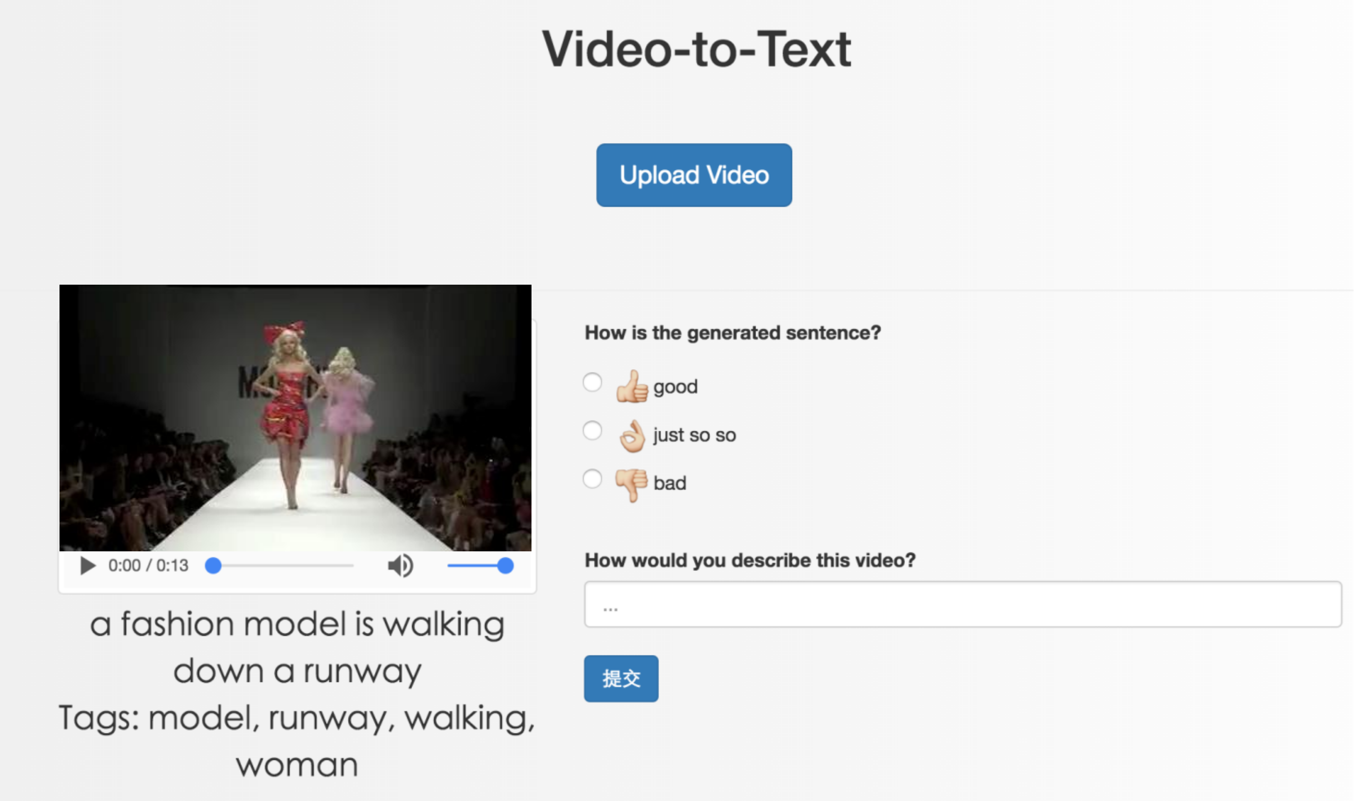
Description generation results



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

How is the generated sentence?

○ 👍 good
○ 👌 just so so
○ 👎 bad

How would you describe this video?

...

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