

Early Embedding and Late Reranking for Video Captioning

Jianfeng Dong¹, **Xirong Li**², 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



0:00 / 0:13

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?

...

提交

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

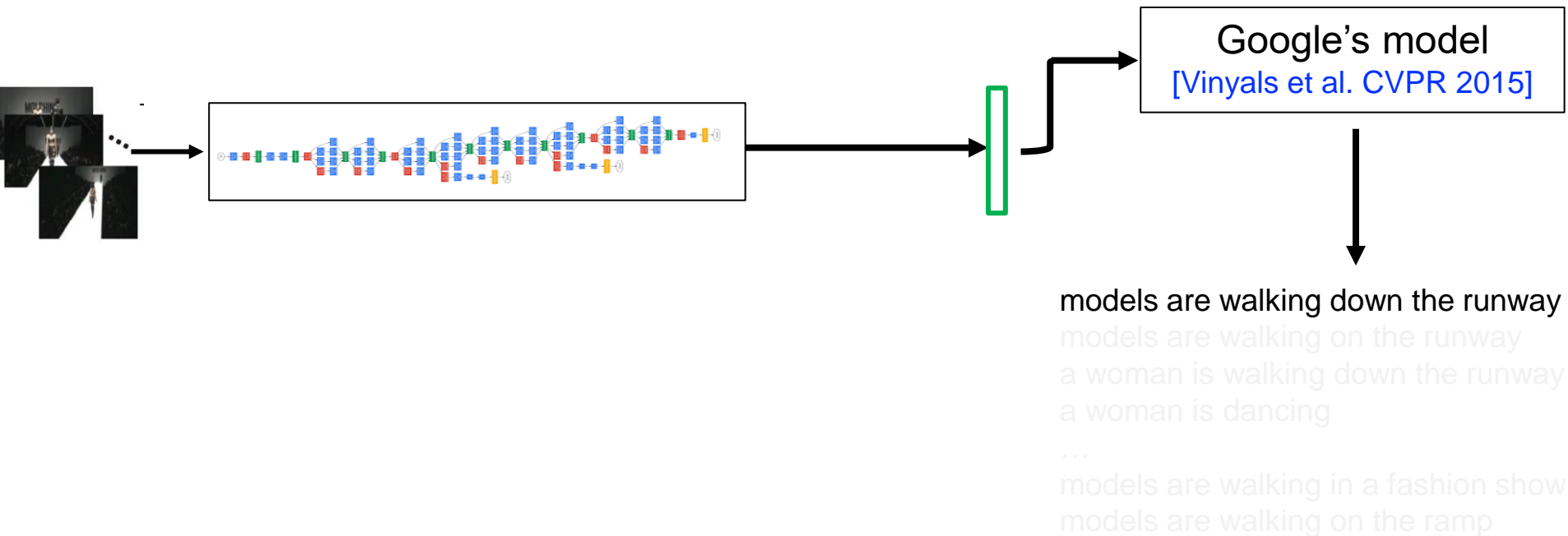


dance
people
woman
dancing

people are **dancing** on a stage

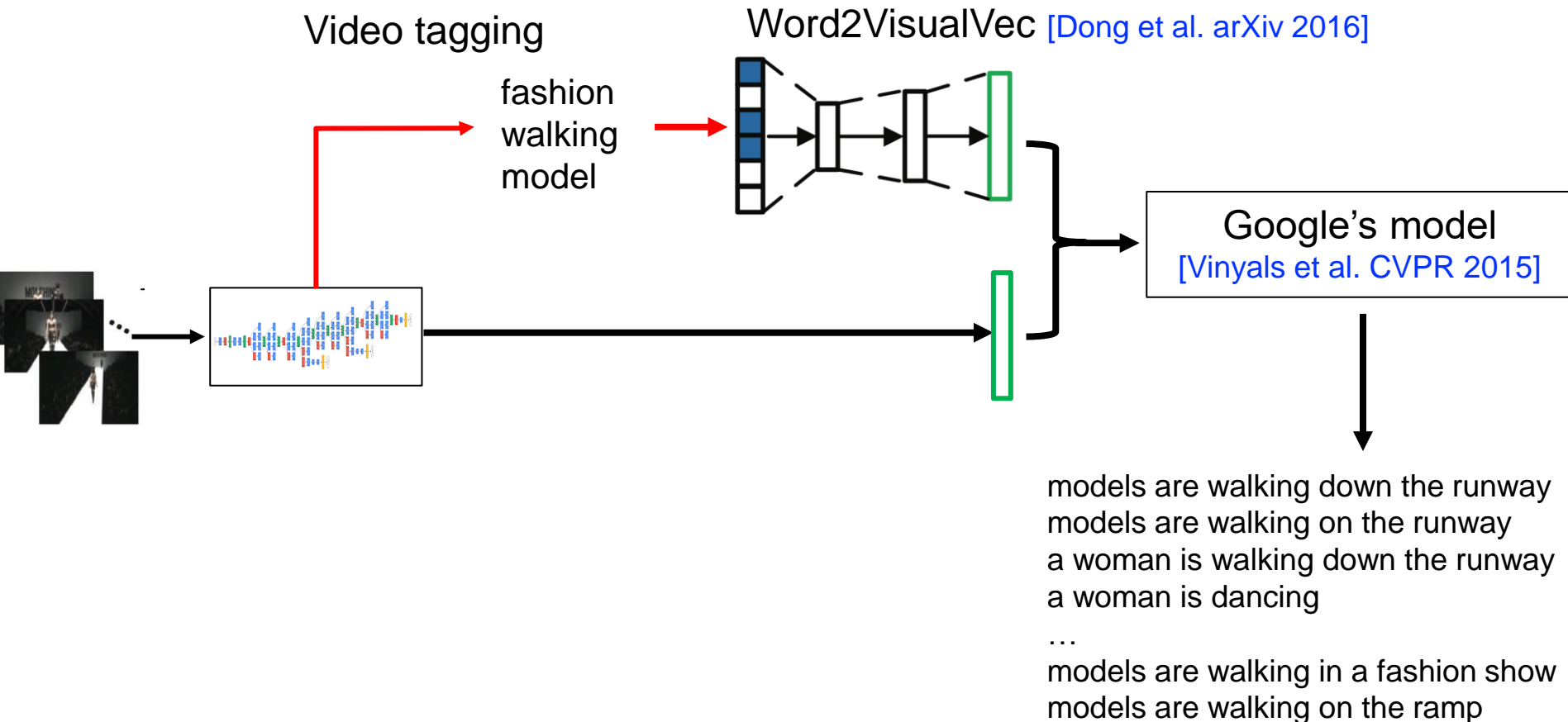
Proposed Video Captioning System

Google's model for sentence generation



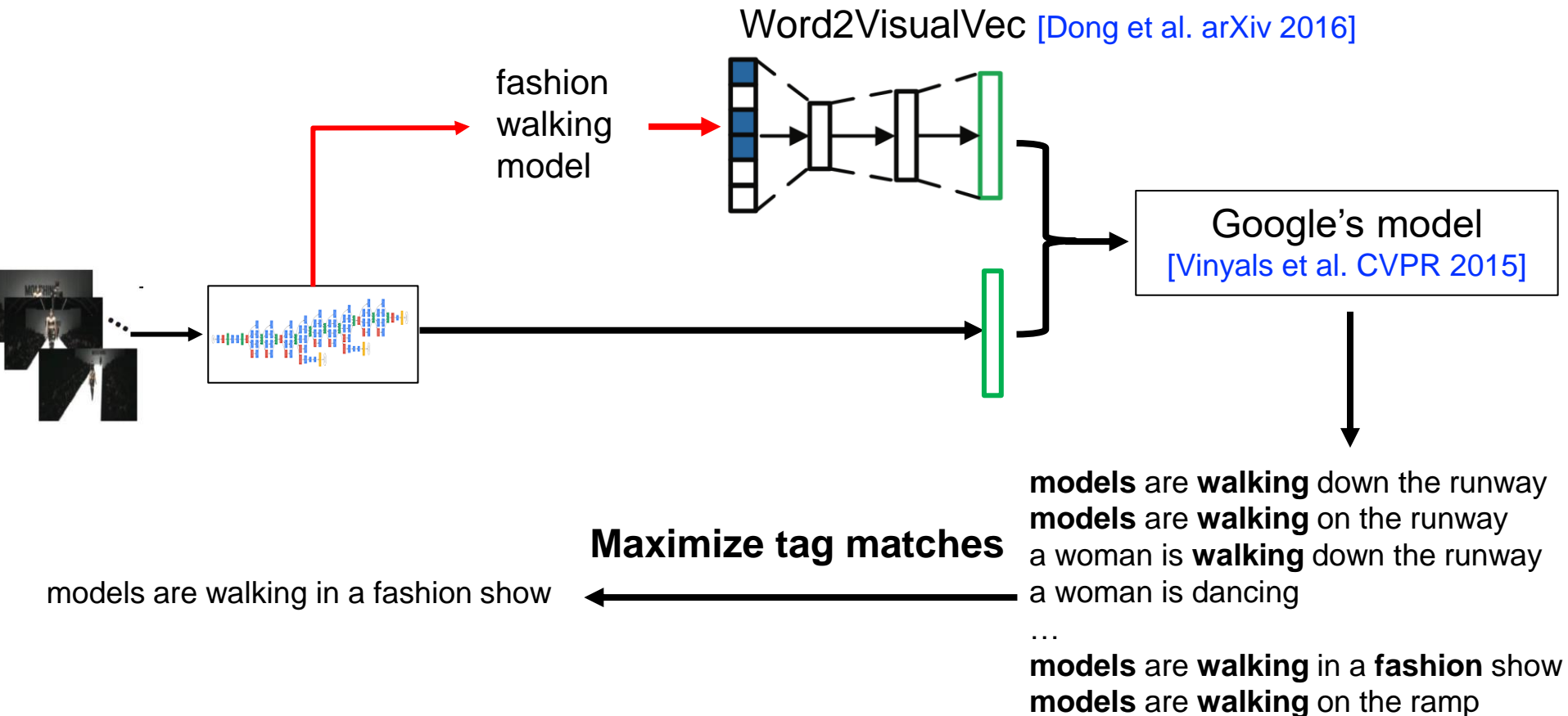
Proposed Video Captioning System

Better initialization by early bmbdding



Proposed Video Captioning System

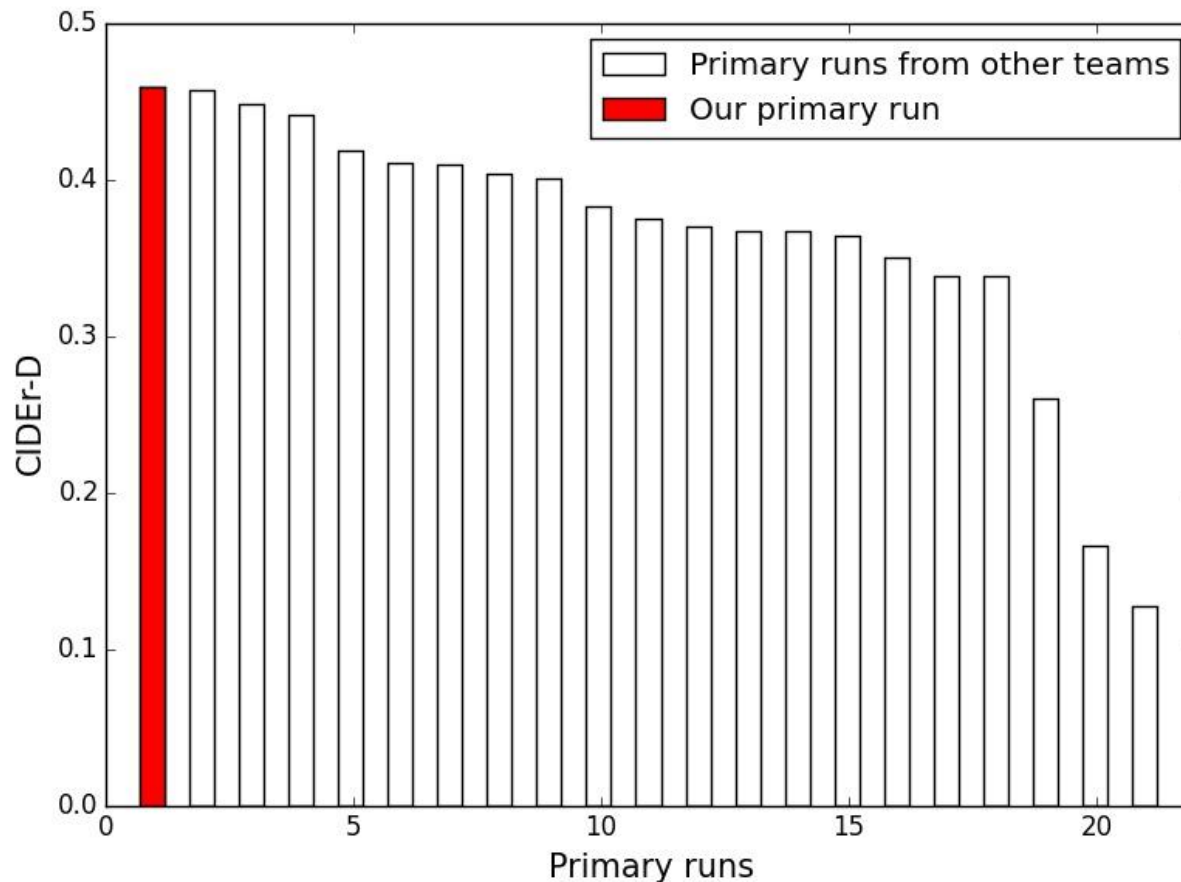
Rerank sentences by matching with video tags



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