Multi-Scale Word2VisualVec for Video Caption Retrieval

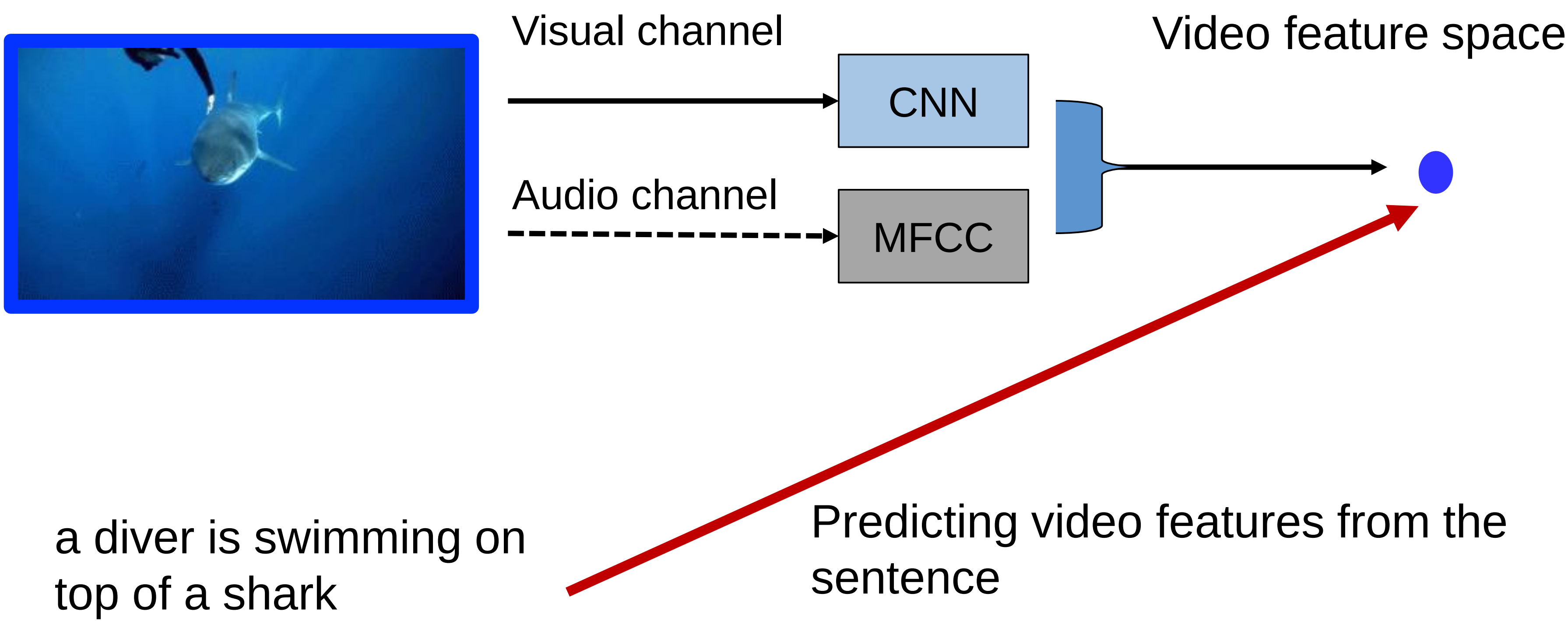
Xirong Li[*], Chaoxi Xu[*], Cees G. M. Snoek[+], Dennis Koelma[+]

Renmin University of China[*]

University of Amsterdam[+]

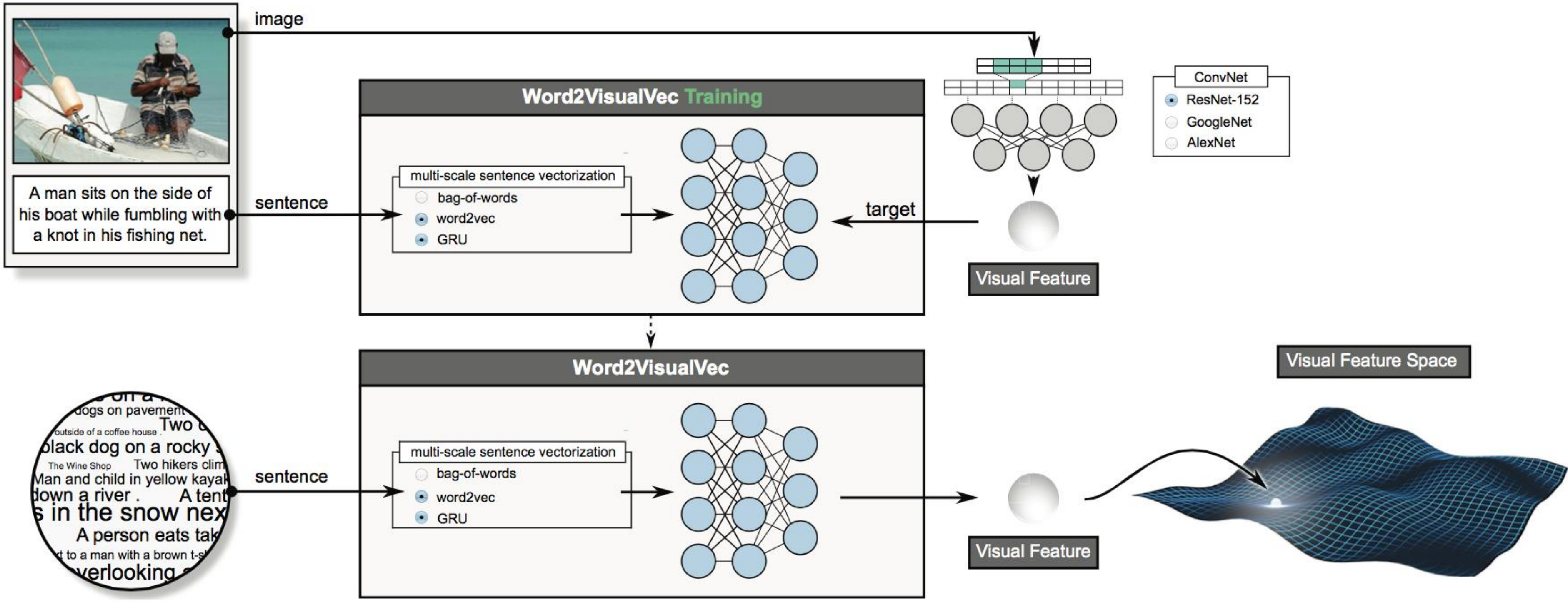
Our idea (as in TV16)

Perform video caption retrieval in a **video feature space**



Multi-Scale Word2VisualVec

Word, sentence, temporal text encoding -> MLP -> visual feature



J. Dong, X. Li, C. Snoek, **Predicting Visual Features from Text for Image and Video Caption Retrieval**, Arxiv: 1709.01362, 2017

TV17 Implementation

We improve with better sentence vectorization and better visual feature.

	TV16	TV17
training set	msrvtt10ktrain	msrvtt10k
validation set	TV16 training set	
sentence vectorization	word2vec	**multi-scale** **+ bag-of-words** **+ word2vec** **+ Gated Recurrent Unit**
visual feature	GoogleNet-shuffle (1024-dim)	**ResNext-shuffle** **(2048-dim)**
audio feature	bag of MFCC (1024-dim)	
MLP architecture	500-1000-2048	11098-2048-3072

*bag-of-words: 9,574-dim (term freq >=5), word2vec: 500-dim, GRU: 1,024-dim

TV17 Implementation cont.

Post processing

Refine the top rankings by matching with tags predicted by

- ResNext-ImageNet13k
- ResNext-Places2
- ResNext-FCVID
- Neighbor Tag Voting using msrvtt10k

Late fusion of two W2VV models:
ResNext -ImageNet13k and ResNext-Places2

- Rank based fusion
- Score based fusion

Video tagging results

State-of-the-art is still not good enough




places	igloo ✖	hospital_room ✖	martial_arts_gym	basketball_court_indoor
ImageNet13k	feather_boa	earphone	parasol ✖	professional_basketball
FCVID	brushing_teeth ✖	trumpet_performance ✖	group_dance ✖	basketball_professional
NeighborVot.	woman	man	game	basketball

vague

Ranking Performance on TV16test

Video feature	w2vv	Set A	Set B
GoogleNet + mfcc	single-scale	0.096	0.106
	multi-scale	0.114	0.127
ResNext + mfcc	single-scale	0.158	0.174
	multi-scale	**0.169**	**0.188**

- Multi-scale sentence vectorization improves Word2VisualVec
- Bigger improvement comes from better video feature

Predict ResNext + mfcc from text using multi-scale w2vv

Ranking Performance on TV17test

run	Set 2-A	Set 2-B	MEAN
multi-scale w2vv	0.223	0.226	0.225
+ rank-fusion	0.218	0.225	0.222
+ score-fusion	0.225	0.227	0.226
+ score-fusion + refine	**0.229**	**0.229**	**0.229**

run	Set 3-A	Set 3-B	Set 3-C	MEAN
multi-scale w2vv	0.303	0.306	0.304	0.304
+ rank-fusion	0.303	0.306	0.307	0.305
+ score-fusion	0.309	0.308	0.306	0.308
+ score-fusion + refine	**0.316**	**0.312**	**0.310**	**0.313**

score-fusion + refine performs the best on both Set 2 and Set 3

Ranking Performance on TV17test

run	Set 4-A	Set 4-B	Set 4-C	Set 4-D	MEAN
multi-scale w2vv	0.401	0.387	0.398	0.395	0.395
+ rank-fusion	0.407	0.384	0.416	0.398	0.401
+ score-fusion	0.406	**0.392**	0.417	0.400	0.404
+ score-fusion + refine	**0.407**	0.388	**0.421**	**0.404**	**0.405**

run	Set 5-A	Set 5-B	Set 5-C	Set 5-D	Set 5-E	MEAN
multi-scale w2vv	0.517	0.548	0.514	0.514	0.531	0.539
+ rank-fusion	0.523	0.557	0.576	**0.528**	0.532	0.543
+ score-fusion	**0.532**	**0.561**	**0.585**	0.513	0.547	**0.548**
+ score-fusion + refine	0.528	0.555	**0.585**	0.513	**0.548**	0.546

score-fusion + refine improves over the baseline but not always the best on Set 4 and Set 5.

Post-evaluation experiments

To study the influence of training data on w2vv

Training data	Set 2-A	Set 2-B	MEAN
msrvtt10k	0.223	0.226	0.225
tgif-train (78,800 gifs)[Li et al. CVPR16]	0.282	0.260	0.271
tgif (100,857 gifs)	**0.290**	0.271	**0.281**
msrvtt10k + tgif	0.286	**0.274**	0.280

*Use ResNext feature alone without mfcc, as gifs have no audio channel.

- tgif as training data contributes a lot
- How to combine msrvtt10k and tgif needs attention

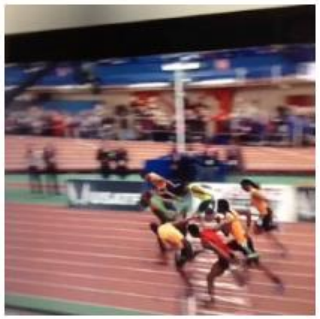
Video Description Generation

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

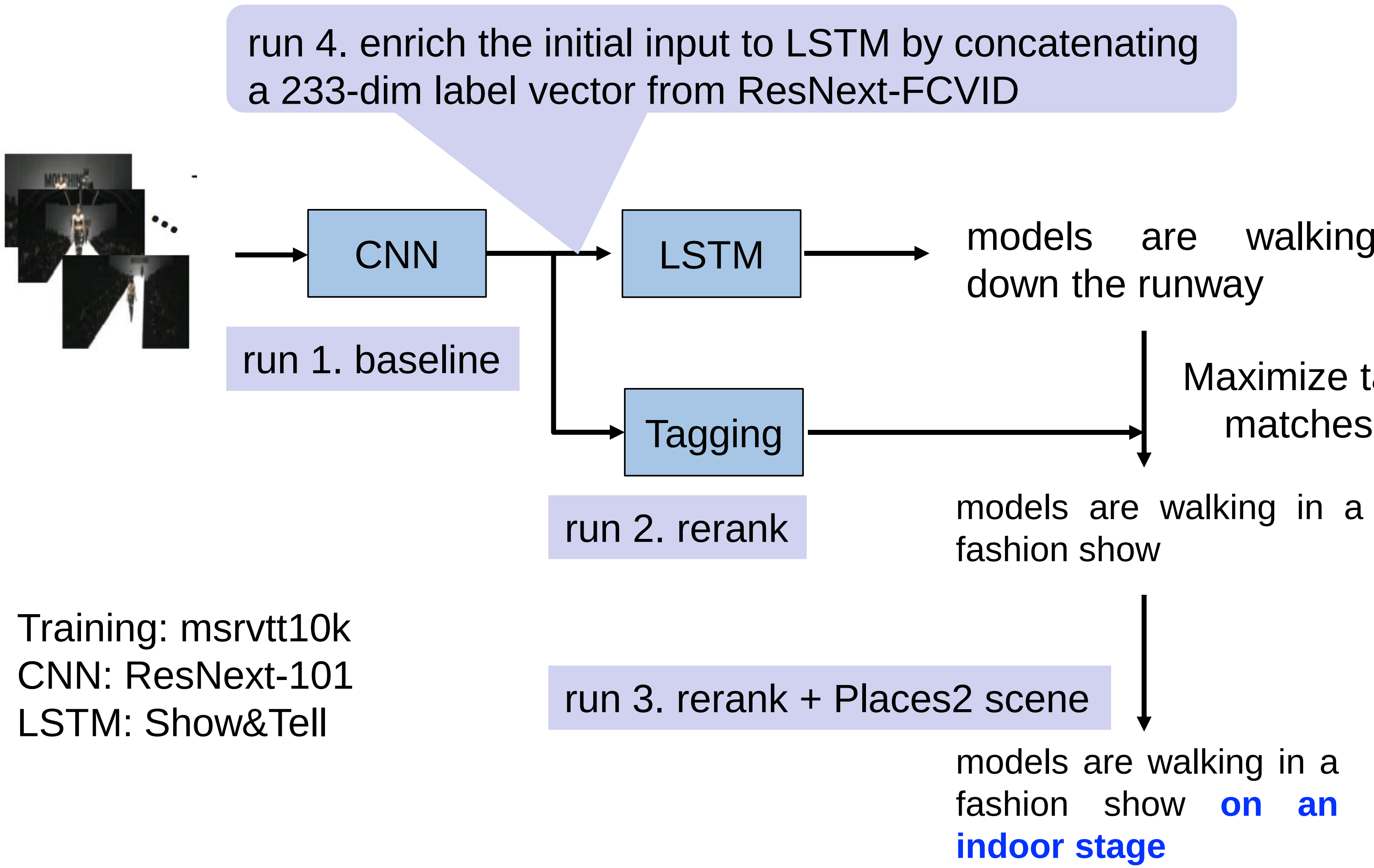
W. Lan, X. Li, J. Dong,
Fluency-guided cross-lingual image captioning,
ACM Multimedia 2017

https://github.com/weiyuk/fluent-cap

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 submissions



Training: msrvtt10k
CNN: ResNext-101
LSTM: Show&Tell

Generation Performance on TV17

run	cider	BLEU	METEOR	sts	SUM
run 1. baseline	0.291	0.013	0.152	0.418	0.875
run 2. rerank	**0.355**	**0.028**	0.181	**0.424**	**0.988**
run 3. rerank + scene	0.328	0.020	**0.196**	0.401	0.945
run 4. rerank + scene + semantic input	0.328	0.024	0.194	0.402	0.947

*Report averaged score if there are multiple references

Sentence reranking by predicted tags gives better results under all metrics.

Other tricks (scene, semantic input) do not really help.

Conclusions

Multi-scale Word2VisualVec that predicts ResNext features from text permits effective video caption retrieval

Tag-based sentence reranking improves LSTM based video captioning, in terms of all metrics

xirong@ruc.edu.cn