Learn to Represent Queries and Videos for Ad-hoc Video Search

Xirong Li, **Chaoxi Xu**, Jianfeng Dong

Renmin University of China

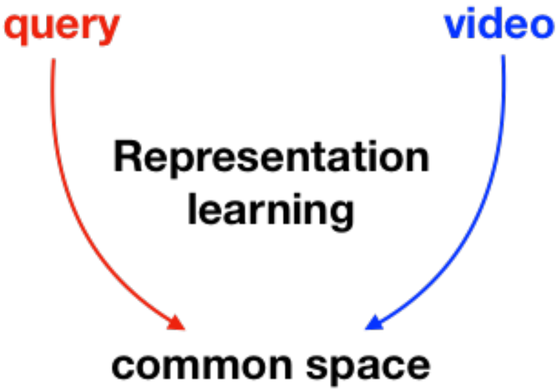
Zhejiang Gangshang University

TRECVID 2019 Workshop

2019-11-12

Key question in ad-hoc video search

How to estimate the relevance of an *unlabeled* video (clip) with respect to a specific query expressed solely in *natural-language* text?

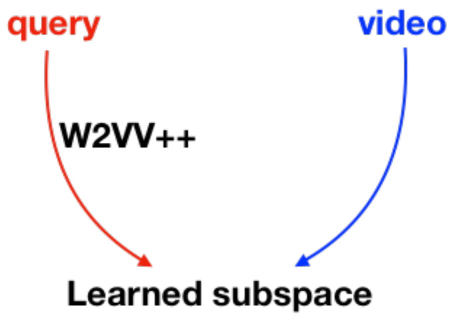


Three dimensions to explore

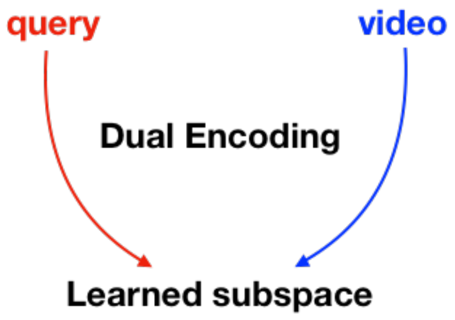
- Query representation
- Video representation
- Common space

Our approach

Based on two deep learning (and concept-free) models



W2VV++ [Li et al., ACMMM'19]
Focus on the query side

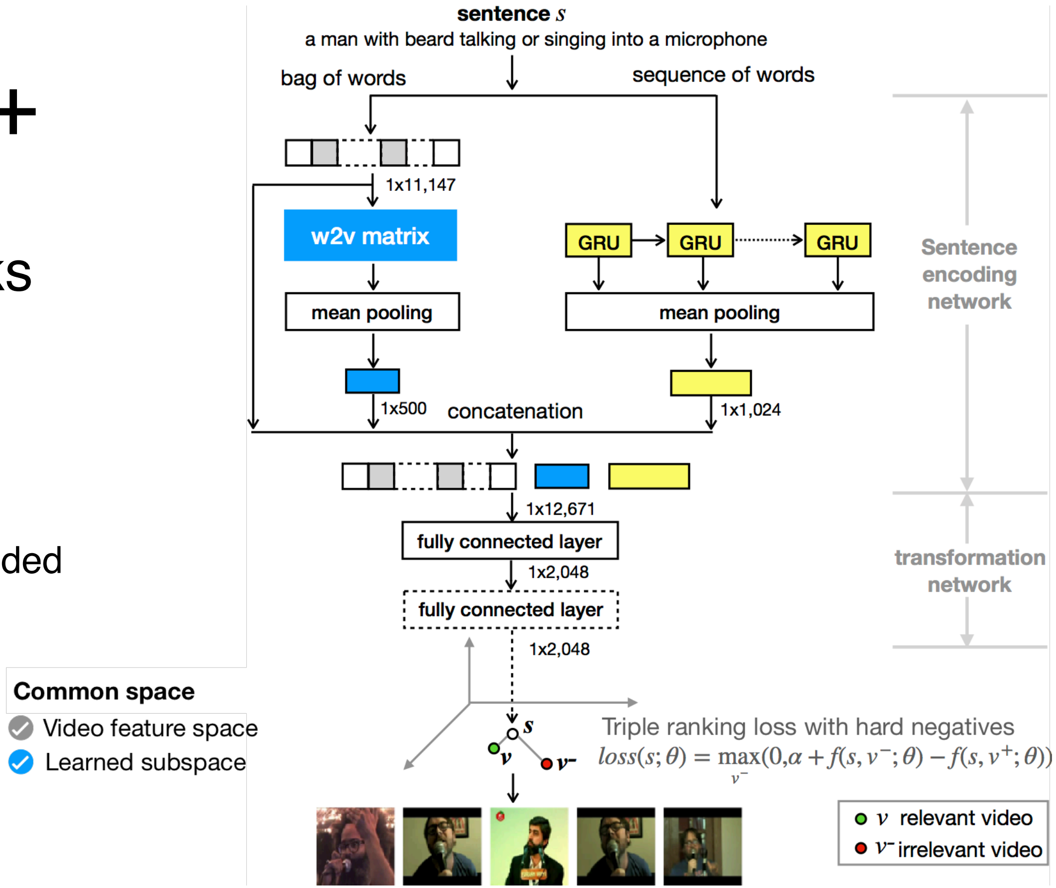


Dual Encoding [Dong et al., CVPR'19]
Focus on both query and video sides

Model 1: W2VV++

Consists of two subnetworks

- A sentence encoding network
  - Bag-of-words
  - Word2Vec + mean pooling
  - GRU + mean pooling
  - ... more text encoders can be included
- A transformation network
  - Common space learning

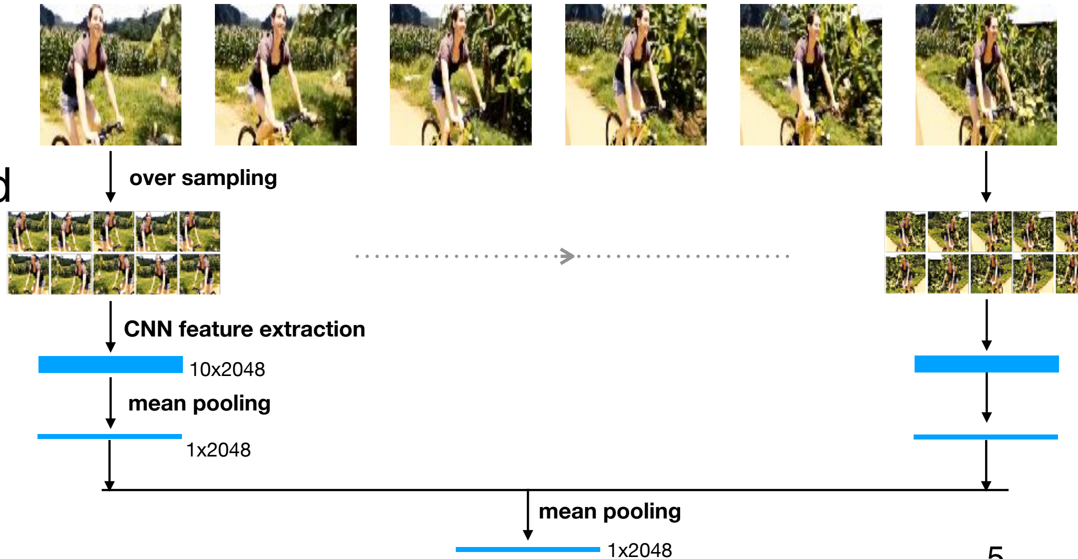


Li et al., W2VV++: Fully deep learning for ad-hoc video search, ACMMM 2019

Model 1: W2VV++

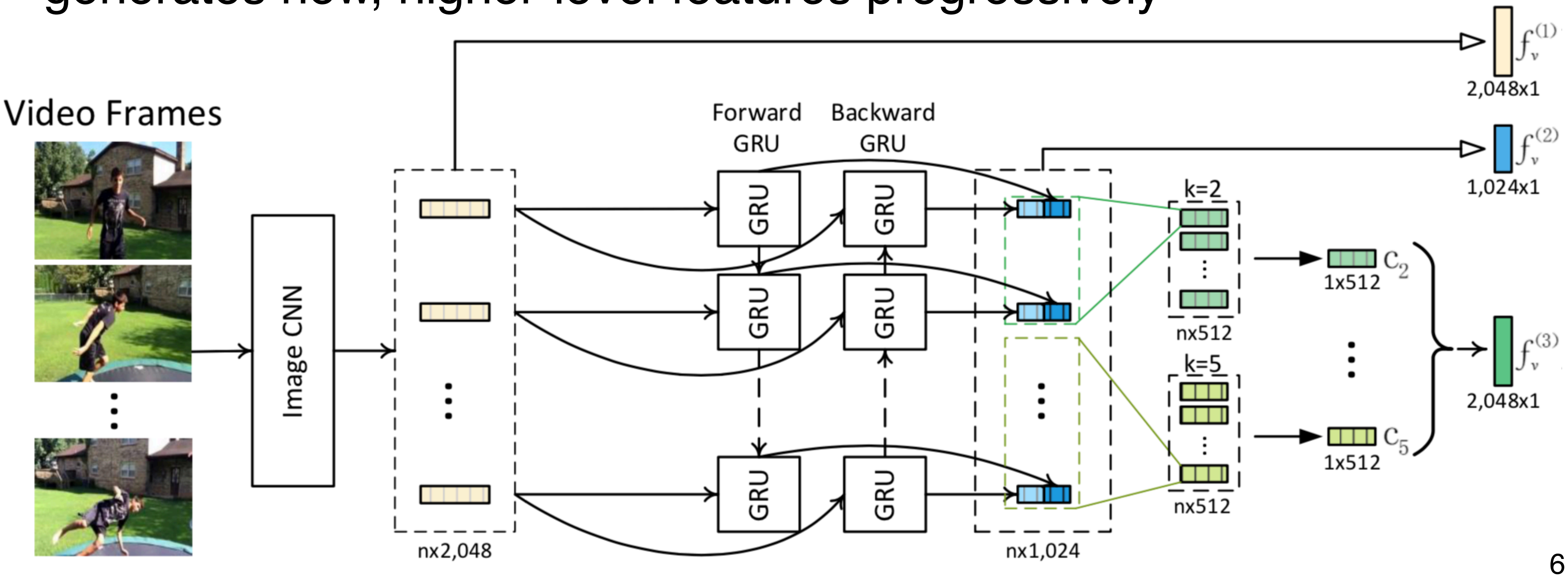
Video representation by multi-level mean pooling

- Sample frames every 0.5 second
- Extract frame-level features by
  - ResNeXt-101
  - ResNet-152
- Two cnn features concatenated
  - 4,096-dim feature per frame



Model 2: Dual Encoding

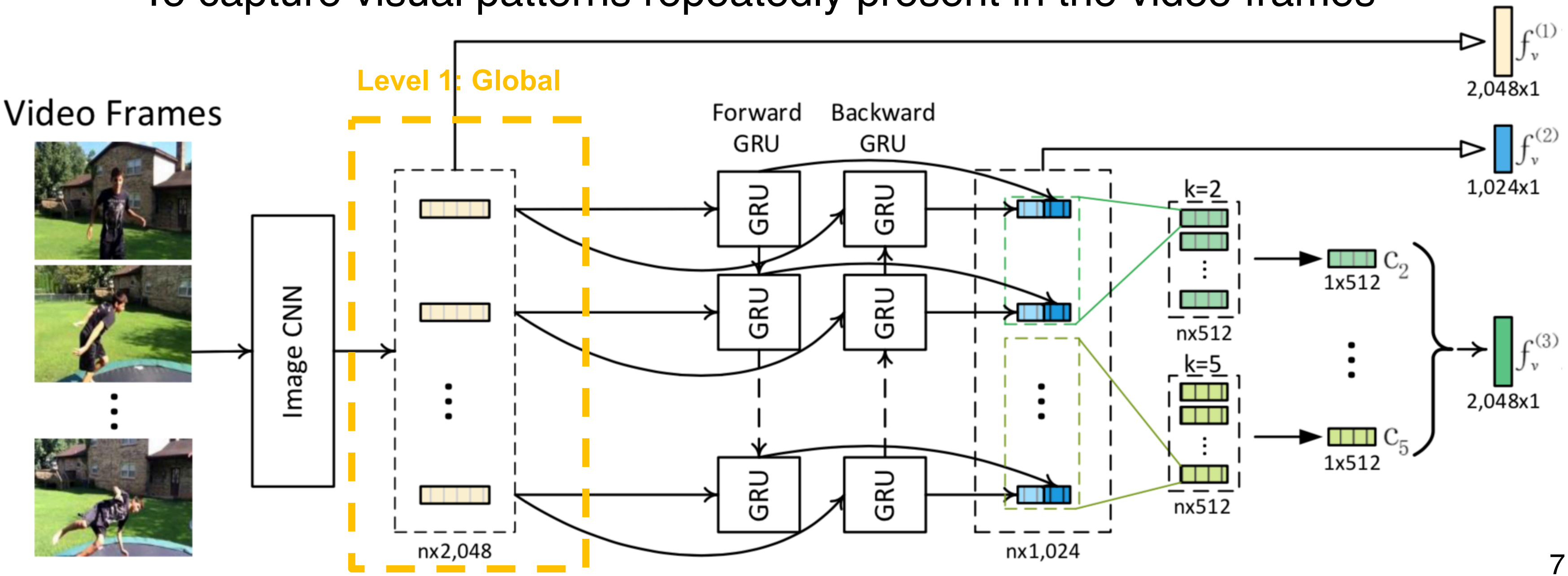
Given a sequence of frame-level CNN features, the network generates new, higher-level features progressively



Model 2: Dual Encoding

Level 1: Global encoding by mean pooling

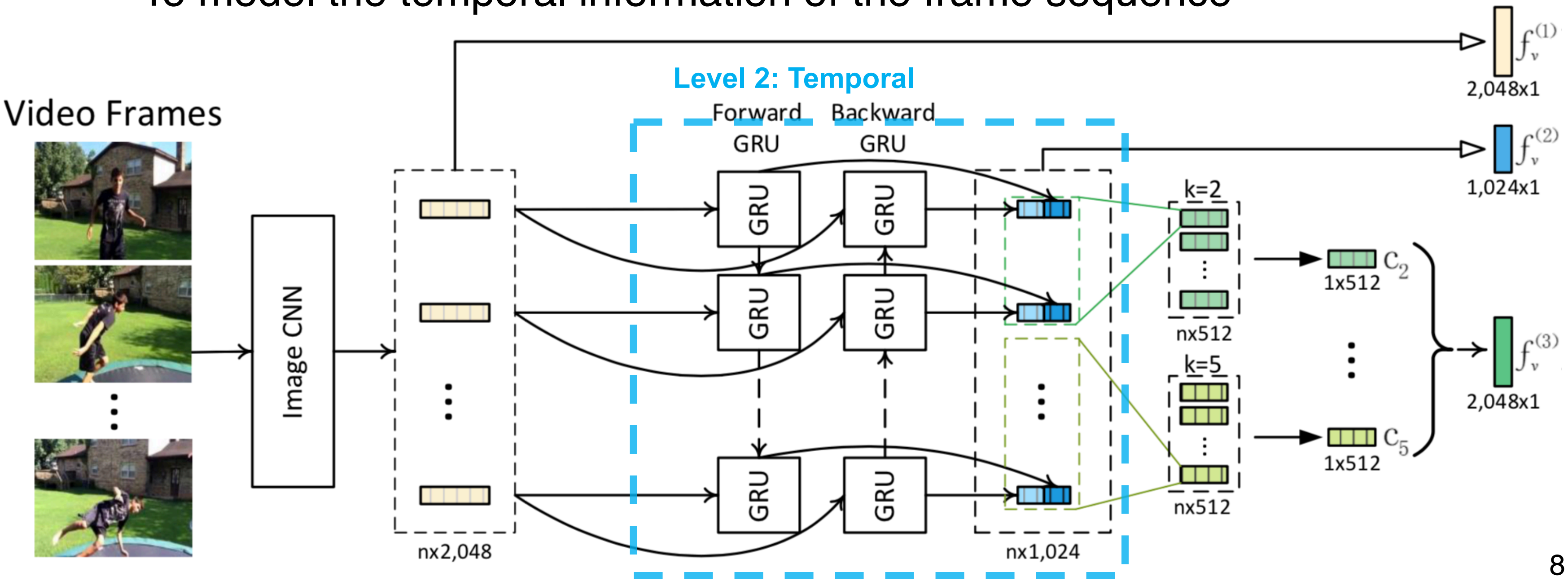
- To capture visual patterns repeatedly present in the video frames



Model 2: Dual Encoding

Level 2: Temporal-aware encoding by biGRU

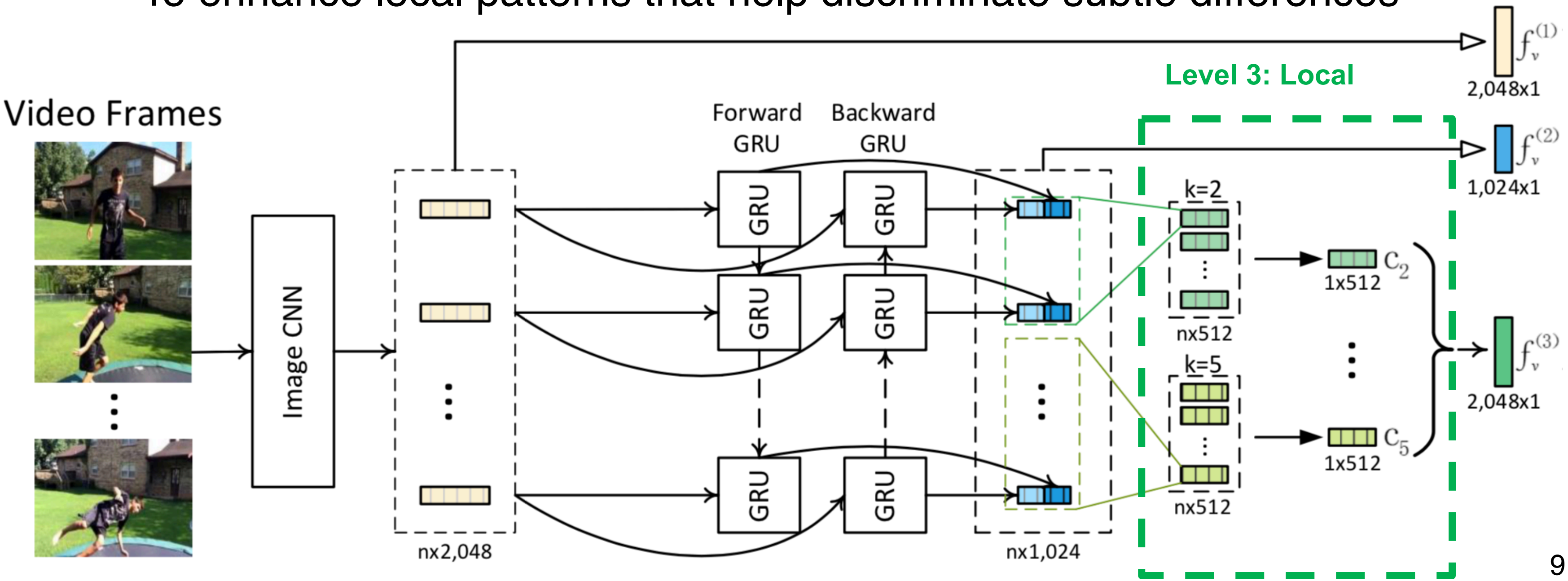
- To model the temporal information of the frame sequence



Model 2: Dual Encoding

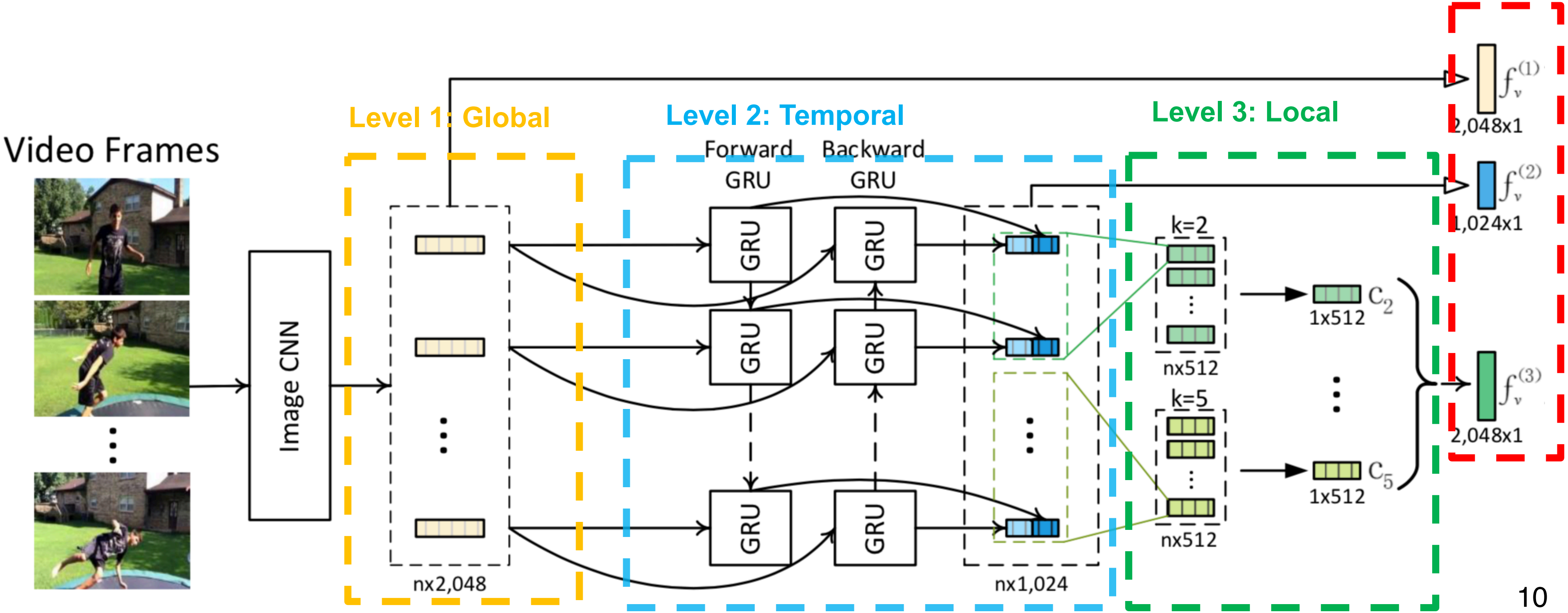
Level 3: Local-enhanced encoding by biGRU-CNN

- To enhance local patterns that help discriminate subtle differences



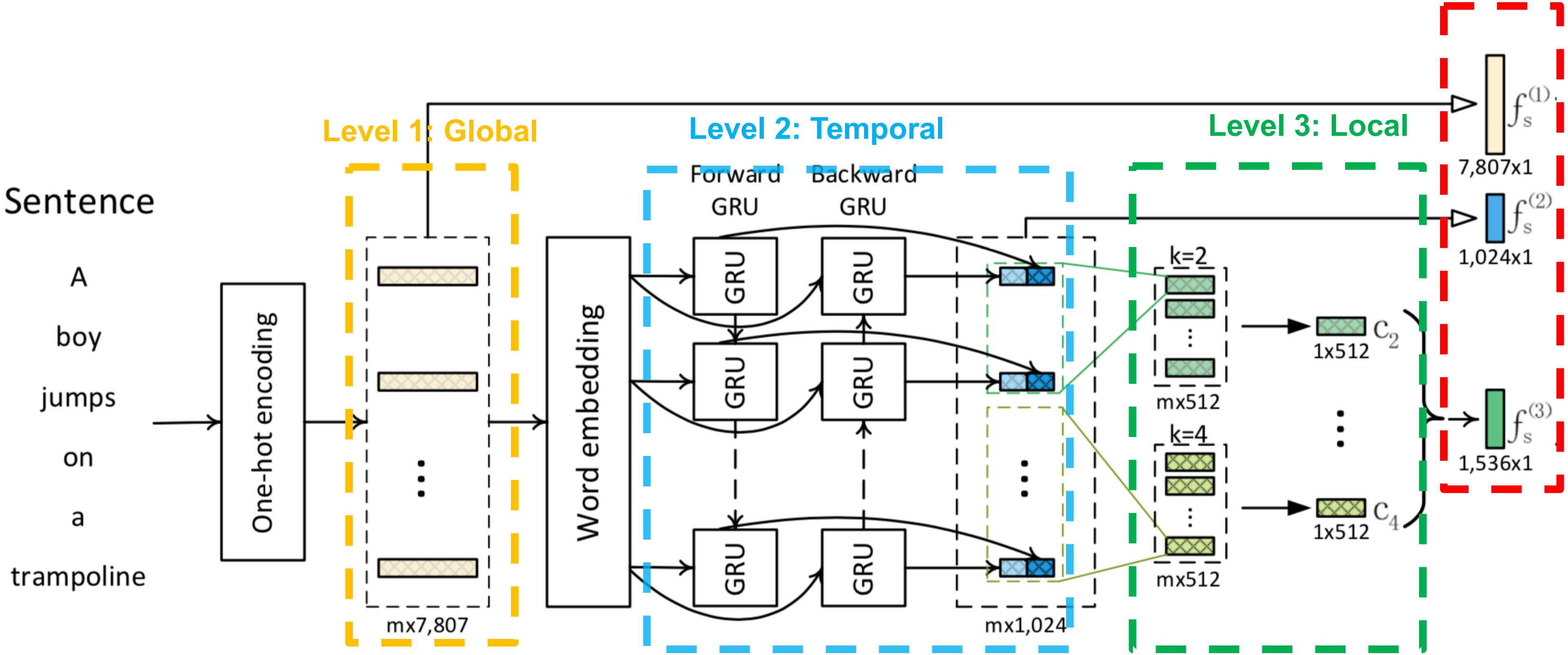
Model 2: Dual Encoding

Multi-level encoding by simple concatenation



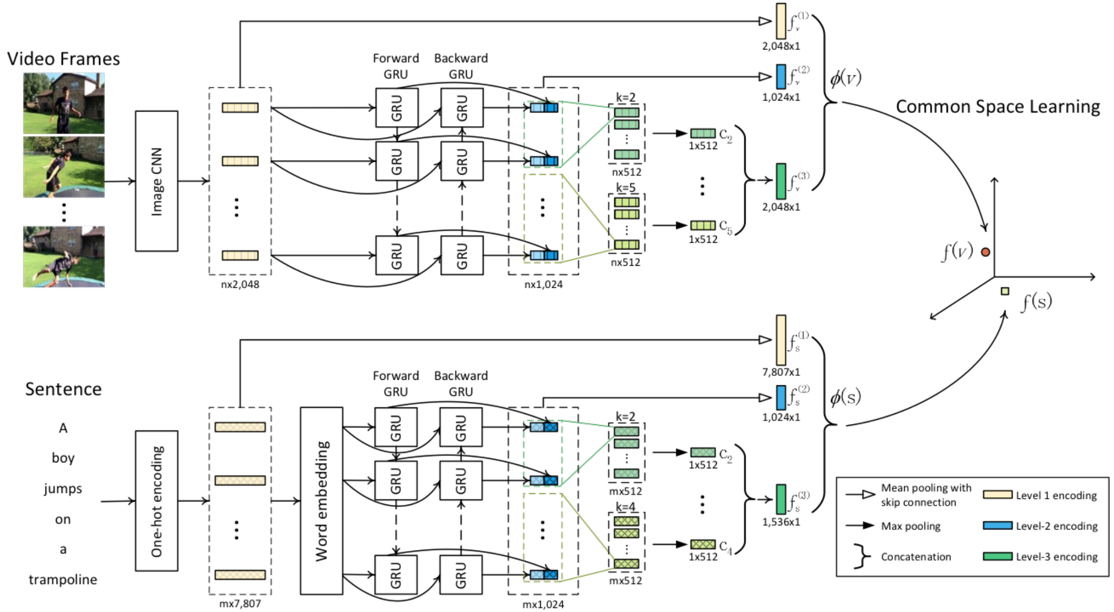
Model 2: Dual Encoding

The same network design applies on the text side



Model 2: Dual Encoding

The network encodes a given video / sentence in parallel



+ The same network design for both modalities

+ Three-level encoding for each modality

+ Separated encoding for each modality

+ Any SOTA common space learning can be used

Dong et al., Dual Encoding for Zero-Example Video Retrieval, CVPR 2019

Training / validation sets

Training

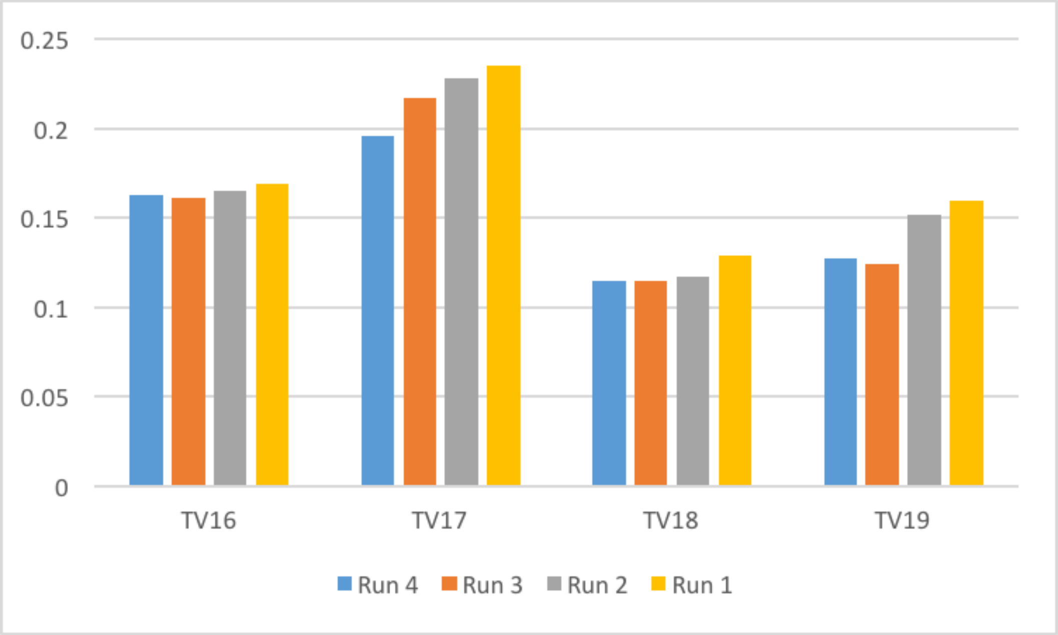
- MSR-VTT
  - 10k web video clips and 200k sentences
- TGIF
  - 100k animated GIFs and 120k sentences
- Validation
  - 90 topics from TV16 / 17 / 18
  - IACC.3, 335k video clips

Our submissions (fully automatic track)

- Four runs based on W2VV++, Dual Encoding and their combinations

run id	description
run 4	W2VV++
run 3	W2VV++ with a BERT encoder
run 2	Dual Encoding
run 1 (primary)	Late average fusion of W2VV++ and Dual Encoding

On TV 2016 - 2019 AVS tasks

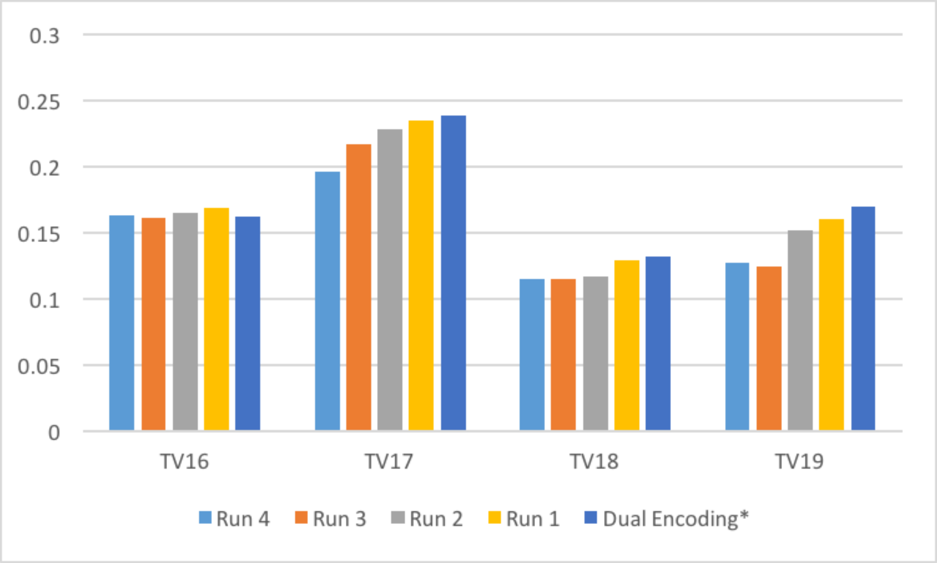


- Dual Encoding is better than W2VV++
  - Marginally on TV16 and TV18
  - Clearly on TV17 and TV19
- Including BERT not always helps
  - Helpful only for TV17
- Model ensemble is better than individual models

Retrospective experiment

Dual Encoding*: Combine only Dual Encoding models

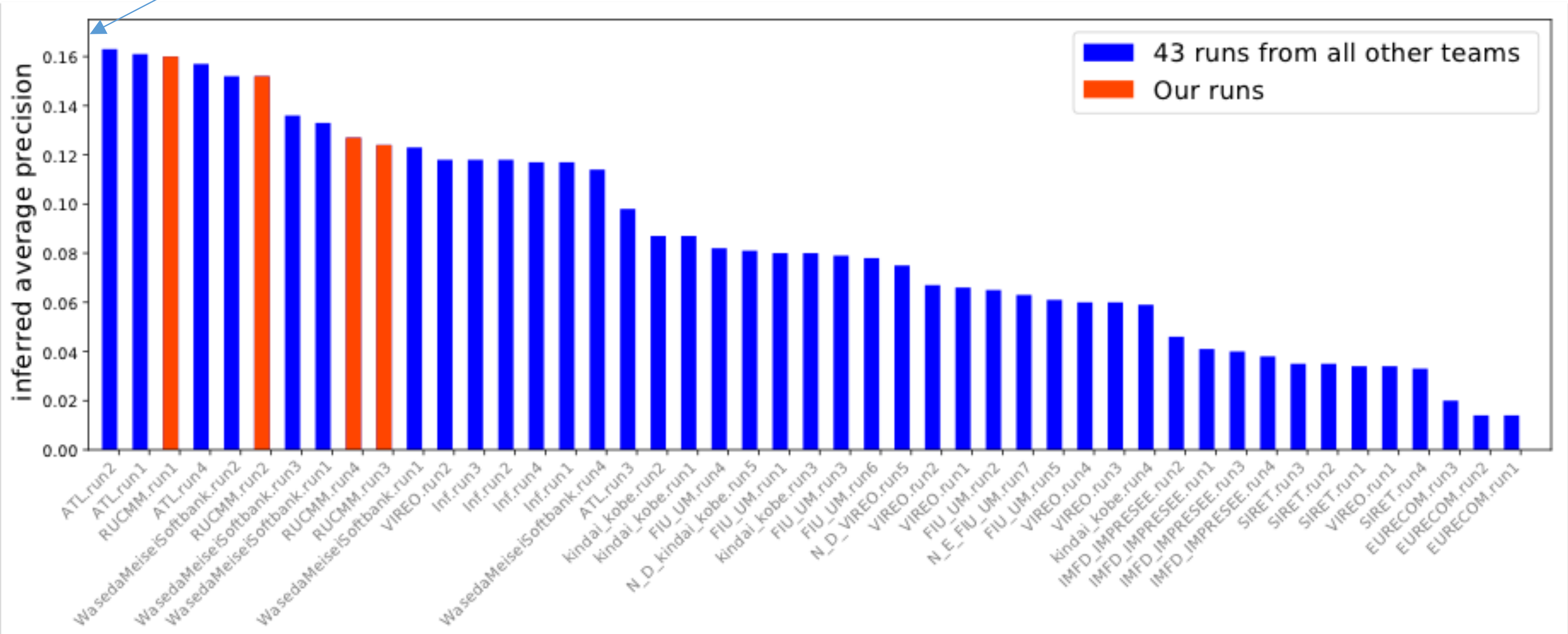
- infAP improved from 0.160 to 0.170



- Dual Encoding is clearly better than W2VV++ on TV19
- Late average fusion is safe, but suboptimal for model ensemble

All fully automatic AVS submissions

Dual Encoding* (infAP: 0.170)



#topicid	run1	run2	run3	run4
611	0.330	0.287	0.309	0.309
612	0.108	0.095	0.094	0.085
613	0.018	0.025	0.025	0.017
614	0.024	0.028	0.011	0.011
615	0.194	0.206	0.145	0.170
616	0.052	0.087	0.048	0.058
617	0.014	0.007	0.013	0.008
618	0.325	0.275	0.213	0.189
619	0.067	0.044	0.064	0.046
620	0.334	0.388	0.302	0.323
621	0.473	0.469	0.485	0.494
622	0.083	0.122	0.068	0.056
623	0.287	0.310	0.194	0.226
624	0.022	0.073	0.020	0.019
625	0.288	0.193	0.166	0.264
626	0.303	0.200	0.262	0.257
627	0.049	0.031	0.043	0.047
628	0.106	0.112	0.044	0.041
629	0.088	0.080	0.034	0.043
630	0.136	0.131	0.062	0.029
631	0.122	0.077	0.031	0.007
632	0.021	0.026	0.021	0.018
633	0.272	0.258	0.258	0.235
634	0.077	0.095	0.025	0.011
635	0.423	0.325	0.309	0.394
636	0.139	0.202	0.068	0.021
637	0.206	0.213	0.254	0.246
638	0.067	0.045	0.049	0.048
639	0.001	0.001	0.005	0.004
640	0.177	0.165	0.097	0.126

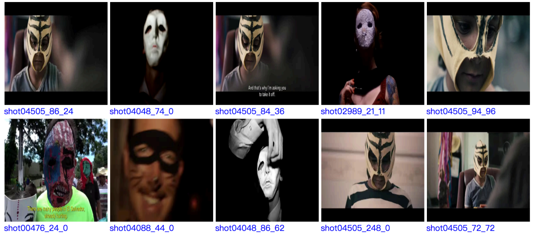
Easy query

- All models perform well

621: person in front of a graffiti painted on a wall (W2VV++, infAP: 0.4939)

635: a bald man (W2VV++: 0.3942)

620: a person with a painted face or mask (W2VV++: 0.3230)



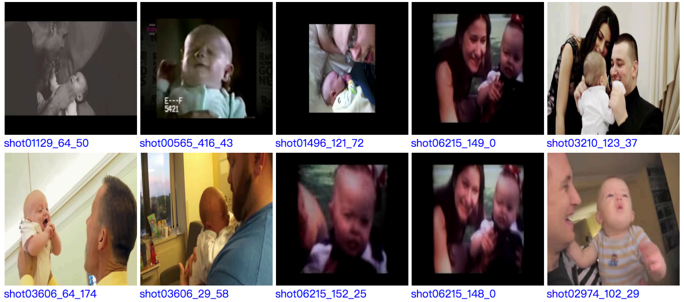
#topicid	run1	run2	run3	run4
611	0.330	0.287	0.309	0.309
612	0.108	0.095	0.094	0.085
613	0.018	0.025	0.025	0.017
614	0.024	0.028	0.011	0.011
615	0.194	0.206	0.145	0.170
616	0.052	0.087	0.048	0.058
617	0.014	0.007	0.013	0.008
618	0.325	0.275	0.213	0.189
619	0.067	0.044	0.064	0.046
620	0.334	0.388	0.302	0.323
621	0.473	0.469	0.485	0.494
622	0.083	0.122	0.068	0.056
623	0.287	0.310	0.194	0.226
624	0.022	0.073	0.020	0.019
625	0.288	0.193	0.166	0.264
626	0.303	0.200	0.262	0.257
627	0.049	0.031	0.043	0.047
628	0.106	0.112	0.044	0.041
629	0.088	0.080	0.034	0.043
630	0.136	0.131	0.062	0.029
631	0.122	0.077	0.031	0.007
632	0.021	0.026	0.021	0.018
633	0.272	0.258	0.258	0.235
634	0.077	0.095	0.025	0.011
635	0.423	0.325	0.309	0.394
636	0.139	0.202	0.068	0.021
637	0.206	0.213	0.254	0.246
638	0.067	0.045	0.049	0.048
639	0.001	0.001	0.005	0.004
640	0.177	0.165	0.097	0.126

Non-easy query

- Not all models perform well

636: a man and a baby both visible

Dual Encoding
infAP: 0.2022



W2VV++
infAP: 0.0214

#topicid	run1	run2	run3	run4
611	0.330	0.287	0.309	0.309
612	0.108	0.095	0.094	0.085
613	0.018	0.025	0.025	0.017
614	0.024	0.028	0.011	0.011
615	0.194	0.206	0.145	0.170
616	0.052	0.087	0.048	0.058
617	0.014	0.007	0.013	0.008
618	0.325	0.275	0.213	0.189
619	0.067	0.044	0.064	0.046
620	0.334	0.388	0.302	0.323
621	0.473	0.469	0.485	0.494
622	0.083	0.122	0.068	0.056
623	0.287	0.310	0.194	0.226
624	0.022	0.073	0.020	0.019
625	0.288	0.193	0.166	0.264
626	0.303	0.200	0.262	0.257
627	0.049	0.031	0.043	0.047
628	0.106	0.112	0.044	0.041
629	0.088	0.080	0.034	0.043
630	0.136	0.131	0.062	0.029
631	0.122	0.077	0.031	0.007
632	0.021	0.026	0.021	0.018
633	0.272	0.258	0.258	0.235
634	0.077	0.095	0.025	0.011
635	0.423	0.325	0.309	0.394
636	0.139	0.202	0.068	0.021
637	0.206	0.213	0.254	0.246
638	0.067	0.045	0.049	0.048
639	0.001	0.001	0.005	0.004
640	0.177	0.165	0.097	0.126

Hard query

- All models perform bad

639: **inside view** of a small airplane flying (W2VV++, infAP 0.0036)

specific viewpoint

617:one or more **picnic tables** outdoors (Dual encoding, infAP 0.0065)

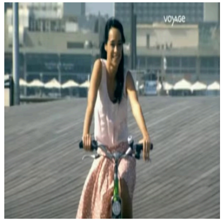
fine-grained concepts

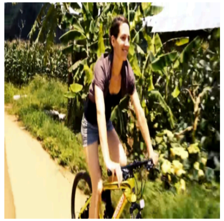
Hard query?

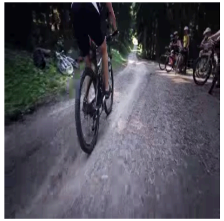
614: a woman riding or holding a bike outdoors

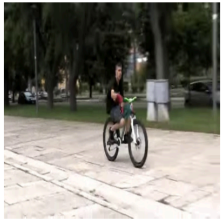
- Dual encoding, infAP 0.0276


shot04627_56_43


shot00938_55_0


shot06463_85_125

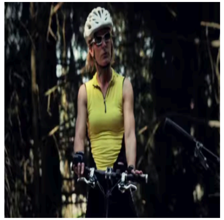

shot00763_65_12


shot02818_262_62


shot06542_30_0


shot02218_257_29


shot00938_56_0


shot00763_81_0


shot04819_71_0

Ground truth seems incomplete

Reproducibility

https://github.com/li-xirong/w2vvpp

- Test a trained W2VV++ on TV 16/17/18 AVS in few minutes

```
./do_test.sh iacc.3
~/VisualSearch/w2vvpp/w2vvpp_resnext101_resnet152_subspace_v190916.pth.tar
w2vvpp_resnext101_resnet152_subspace_v190916 tv16.avs.txt,tv17.avs.txt,tv18.avs.txt
```

```
[12 Nov 14:41:16 - util.py:line 19] /data/home/xirong/VisualSearch/iacc.3/SimilarityIndex/tv16.avs.txt/w2vvpp_resne
xt101_resnet152_subspace_v190916/id.sent.score.txt exists. overwrite
[12 Nov 14:41:16 - predictor.py:line 93] Encoding videos
335944/335944 [==============================] - 56s 167us/step
encode_vis execution time: 56.217 seconds

[12 Nov 14:42:12 - predictor.py:line 101] Encoding tv16.avs.txt captions
30/30 [==============================] - 1s 30ms/step
encode_txt execution time: 1.060 seconds

cosine_sim execution time: 53.042 seconds

writing result into file time: 18.874 seconds

[12 Nov 14:43:27 - util.py:line 19] /data/home/xirong/VisualSearch/iacc.3/SimilarityIndex/tv17.avs.txt/w2vvpp_resne
xt101_resnet152_subspace_v190916/id.sent.score.txt exists. overwrite
[12 Nov 14:43:27 - predictor.py:line 101] Encoding tv17.avs.txt captions
30/30 [==============================] - 1s 17ms/step
encode_txt execution time: 0.717 seconds

cosine_sim execution time: 75.587 seconds

writing result into file time: 18.852 seconds

[12 Nov 14:45:03 - util.py:line 19] /data/home/xirong/VisualSearch/iacc.3/SimilarityIndex/tv18.avs.txt/w2vvpp_resne
xt101_resnet152_subspace_v190916/id.sent.score.txt exists. overwrite
[12 Nov 14:45:03 - predictor.py:line 101] Encoding tv18.avs.txt captions
30/30 [==============================] - 1s 19ms/step
encode_txt execution time: 0.793 seconds
```

Conclusions

- Learn to represent query / video is effective
- Late average fusion is safe, yet suboptimal, to boost performance
- Queries with fine-grained concepts in specific viewpoints remain hard

https://github.com/li-xirong/video-retrieval

Li et al., W2VV++: Fully Deep Learning for Ad-hoc Video Search, ACMMM 2019
Dong et al., Dual Encoding for Zero-Example Video Retrieval, CVPR 2019