Visual and Textual Encoder Assembly for Ad-hoc Video Search

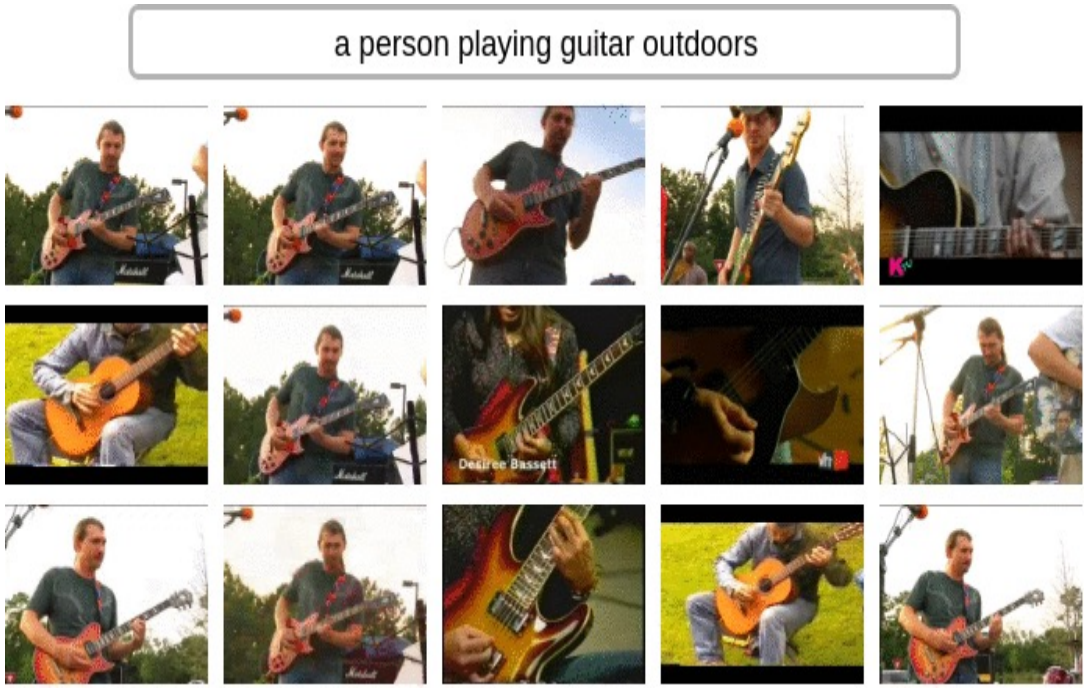
Aozhu Chen, Fan Hu, Xirong Li

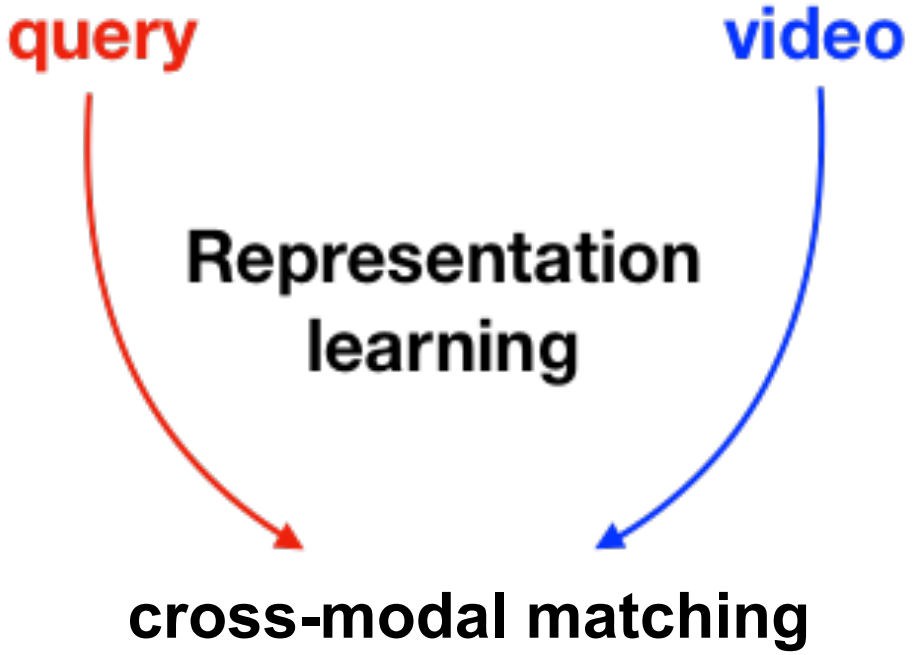
AIMC Lab, School of Information, Renmin University of China
https://ruc-aimc-lab.github.io/

7 December 2021

Ad-hoc Video Search (AVS)

- Search for **unlabeled** videos by **natural-language** queries

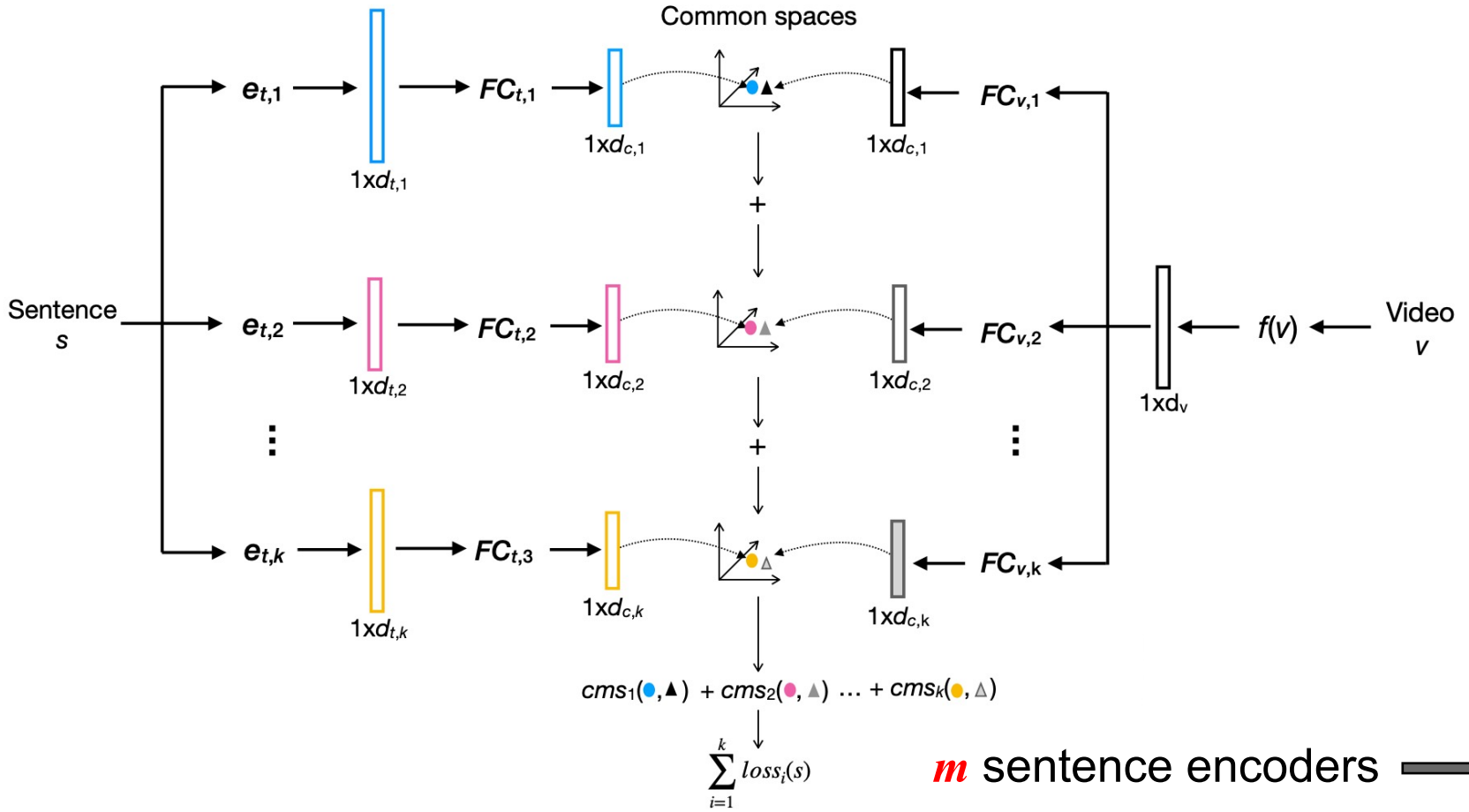




Our Solution

Sentence Encoder Assembly (SEA)

- It supports text-video matching in multiple encoder-specific common spaces.



m sentence encoders ⟹ ***m*** common spaces

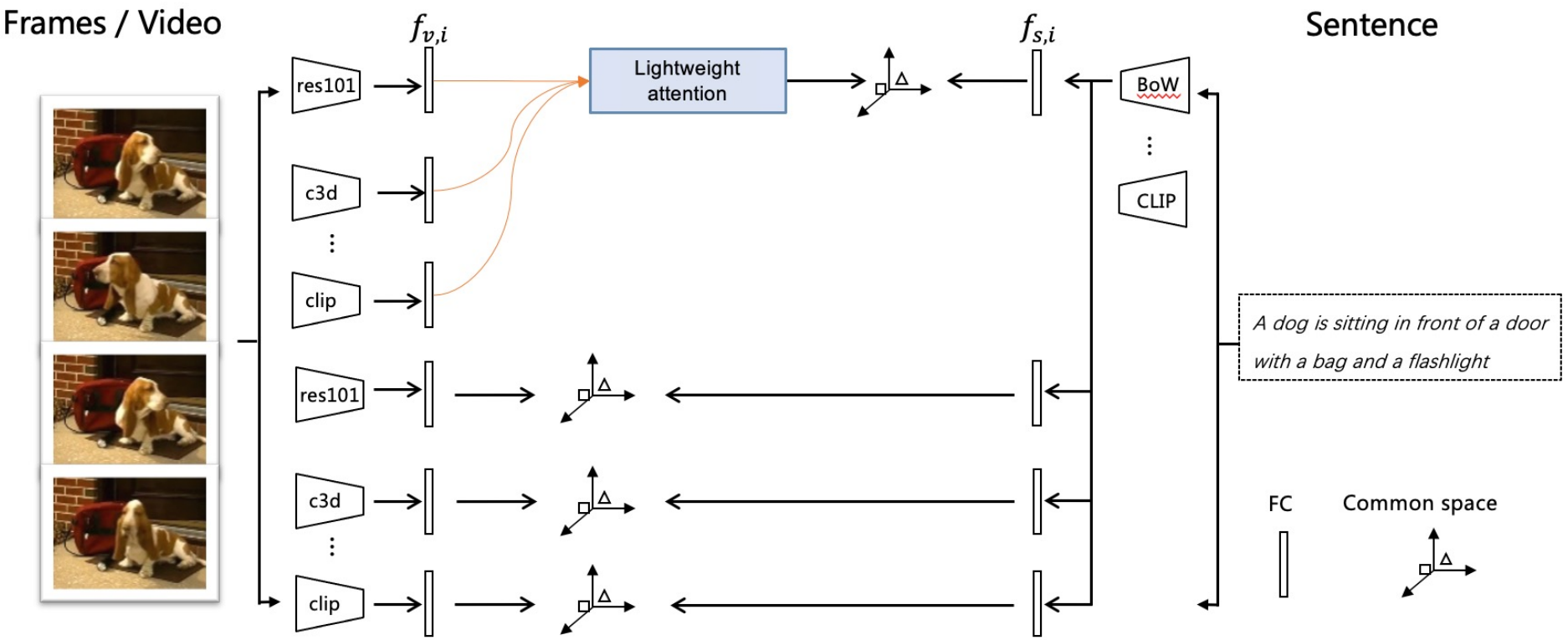
[1] Li et al., SEA: Sentence Encoder Assembly for Video Retrieval by Textual Queries, TMM 2020

[2] Li et al., Renmin University of China at TRECVID 2020: Sentence Encoder Assembly for Ad-hoc Video Search, TRECVID 2020

Our Solution

Multiple Encoder Assembly (MEA)

- It supports text-video matching in multiple feature-specific common spaces.



[1] Hu et al., Lightweight Attentional Feature Fusion for Video Retrieval by Text, arXiv 2021

Choice of Visual Encoders

Seven 2D / 3D features :

2D feature (frame-level)

- rx101
- re152
- wsl
- clip

3D feature (segment-level)

- c3d
- ircsn
- tf

Rank	Feature	TV19	TV20	MEAN
1	*c3d*	0.036	0.098	0.067
2	*tf*	0.102	0.142	0.122
3	*ircsn*	0.088	0.193	0.14
4	*clip*	0.155	0.180	0.168
5	*wsl*	0.135	0.210	0.173
6	*rx101-wsl*	0.148	0.208	0.178
7	*clip-wsl*	0.148	0.208	0.178
8	*rx101-re152-wsl*	0.159	0.208	0.184
9	*rx101-re152*	0.167	0.201	0.184
10	*rx101-clip*	0.196	0.223	0.210
11	*rx101-re152-wsl-clip*	0.182	0.238	0.210
12	*rx101-wsl-clip*	0.183	0.239	0.211
13	*rx101-re152-clip*	0.199	0.233	0.216

Model: SEA(bow, w2v)

Choice of Textual Encoders

Three sentence encoders:

- Bag-of-Words (bow)
- Word2Vec (w2v)
- CLIP (ViT-B/32)[1]

Feature	TV19	TV20	MEAN
Model: SEA(bow, w2v)			
rx101-wsl-clip	0.183	0.239	0.211
rx101-re152-clip	0.199	0.233	0.216
Model: SEA(bow, w2v, clip)			
rx101-re152-clip	0.199	0.249	0.224
rx101-wsl-clip	0.204	0.262	0.233

[1] Radford et al., Learning Transferable Visual Models from Natural Language Supervision, ICML 2021

Choice of (Pre-)Training Data

Five datasets :

- MSCOCO
- GCC
- MSR-VTT
- TGIF
- VATEX

Feature	TV19	TV20	MEAN
Pre-training on MS-COCO:			
rx101-wsl-clip-c3d	0.203	0.339	0.271
rx101-wsl-clip-tf	0.205	0.313	0.259
rx101-wsl-clip-ircsn	0.196	0.347	0.272
Pre-training on GCC:			
rx101-wsl-clip-ircsn	0.203	0.335	0.269
Pre-training on GCC and MS-COCO:			
rx101-wsl-clip-ircsn	0.209	0.357	0.283

Training data: MSR-VTT，TGIF and VATEX

Submissions(fully automatic track)

We submitted the following 4 runs:

MEA, SEA, CLIP and their combinations

Run id	**Model**	**Feature**	**(Pre-)Training Data**
Run1	Late fusion of Run3, Run4, CLIP	*rx101* *wsl* *clip* *ircsn*	MSCOCO, GCC MSR-VTT TGIF VATEX
Run2	MEA-reRank		
Run3	MEA(bow, w2v,clip)		
Run4	SEA(bow, w2v,clip)		
-	CLIP（ViT/B32）	-	-

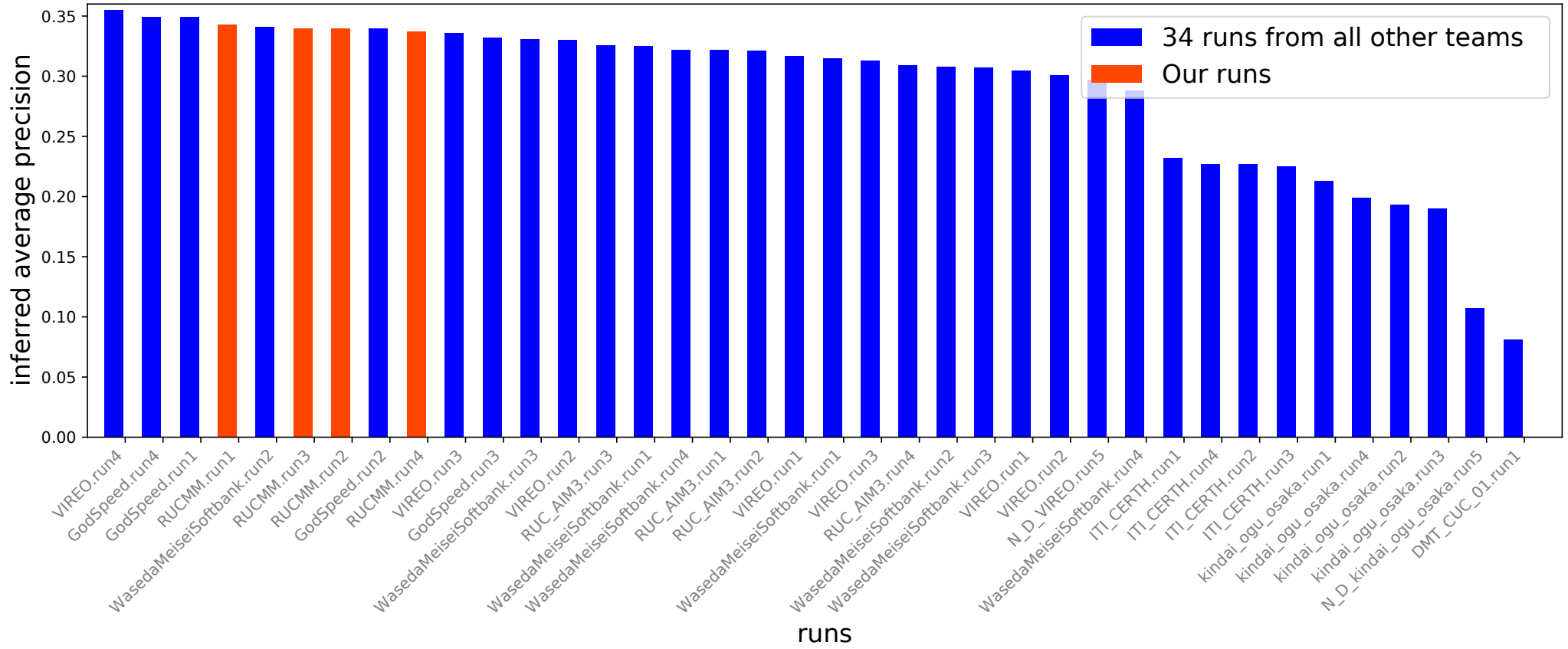
Performance on TV16-21 AVS Task

The performance of CLIP is lower, but it can bring better performance when doing late fusion.

Run id	**Model**	**TV16**	**TV17**	**TV18**	**TV19**	**TV20**	**TV21**	**MEAN**
Run1	Late fusion of Run3, Run4, CLIP	0.246	0.320	0.161	0.227	0.366	0.343	0.277
Run2	MEA-rerank	0.224	0.342	0.168	0.224	0.361	0.340	0.293
Run3	MEA	0.223	0.342	0.167	0.223	0.361	0.340	0.292
Run4	SEA(bow, w2v,clip)	0.232	0.255	0.135	0.213	0.358	0.337	0.239
-	CLIP	0.173	0.208	0.087	0.136	0.161	0.194	0.160
-	Late fusion of Run3, Run4	0.235	0.300	0.156	0.225	0.365	0.339	0.270

All fully automatic AVS submissions

Our submissions ranked the 3rd



Results of individual topics

More **blue** is better

More **red** is worse

Can be divided into 2 types:

- Easy topics: all **blue**
- Hard topics: all **red**

topic id	Run1	Run2	Run3	Run4
661	0.395	0.388	0.384	0.423
662	0.792	0.811	0.824	0.657
663	0.504	0.485	0.486	0.533
664	0.269	0.260	0.261	0.260
665	0.127	0.120	0.120	0.134
666	0.207	0.214	0.212	0.186
667	0.110	0.092	0.093	0.128
668	0.287	0.262	0.260	0.299
669	0.171	0.163	0.165	0.184
670	0.514	0.506	0.505	0.516
671	0.166	0.166	0.166	0.143
672	0.013	0.020	0.019	0.009
673	0.056	0.068	0.067	0.035
674	0.879	0.906	0.904	0.864
675	0.511	0.506	0.506	0.511
676	0.176	0.162	0.163	0.193
677	0.833	0.835	0.836	0.830
678	0.336	0.337	0.337	0.296
679	0.040	0.042	0.043	0.031
680	0.478	0.460	0.460	0.517

Per-topic analysis on TV21

topic id	Run1	Run2	Run3	Run4
661	0.395	0.388	0.384	0.423
662	0.792	0.811	0.824	0.657
663	0.504	0.485	0.486	0.533
664	0.269	0.260	0.261	0.260
665	0.127	0.120	0.120	0.134
666	0.207	0.214	0.212	0.186
667	0.110	0.092	0.093	0.128
668	0.287	0.262	0.260	0.299
669	0.171	0.163	0.165	0.184
670	0.514	0.506	0.505	0.516
671	0.166	0.166	0.166	0.143
672	0.013	0.020	0.019	0.009
673	0.056	0.068	0.067	0.035
674	0.879	0.906	0.904	0.864
675	0.511	0.506	0.506	0.511
676	0.176	0.162	0.163	0.193
677	0.833	0.835	0.836	0.830
678	0.336	0.337	0.337	0.296
679	0.040	0.042	0.043	0.031
680	0.478	0.460	0.460	0.517

Easy query

674 a **parachutist** descending towards a field on the ground **in the daytime**

Per-topic analysis on TV21

topic id	Run1	Run2	Run3	Run4
661	0.395	0.388	0.384	0.423
662	0.792	0.811	0.824	0.657
663	0.504	0.485	0.486	0.533
664	0.269	0.260	0.261	0.260
665	0.127	0.120	0.120	0.134
666	0.207	0.214	0.212	0.186
667	0.110	0.092	0.093	0.128
668	0.287	0.262	0.260	0.299
669	0.171	0.163	0.165	0.184
670	0.514	0.506	0.505	0.516
671	0.166	0.166	0.166	0.143
672	0.013	0.020	0.019	0.009
673	0.056	0.068	0.067	0.035
674	0.879	0.906	0.904	0.864
675	0.511	0.506	0.506	0.511
676	0.176	0.162	0.163	0.193
677	0.833	0.835	0.836	0.830
678	0.336	0.337	0.337	0.296
679	0.040	0.042	0.043	0.031
680	0.478	0.460	0.460	0.517

Easy query

677 two **boxers** in a **ring**

Per-topic analysis on TV21

topic id	Run1	Run2	Run3	Run4
661	0.395	0.388	0.384	0.423
662	0.792	0.811	0.824	0.657
663	0.504	0.485	0.486	0.533
664	0.269	0.260	0.261	0.260
665	0.127	0.120	0.120	0.134
666	0.207	0.214	0.212	0.186
667	0.110	0.092	0.093	0.128
668	0.287	0.262	0.260	0.299
669	0.171	0.163	0.165	0.184
670	0.514	0.506	0.505	0.516
671	0.166	0.166	0.166	0.143
672	0.013	0.020	0.019	0.009
673	0.056	0.068	0.067	0.035
674	0.879	0.906	0.904	0.864
675	0.511	0.506	0.506	0.511
676	0.176	0.162	0.163	0.193
677	0.833	0.835	0.836	0.830
678	0.336	0.337	0.337	0.296
679	0.040	0.042	0.043	0.031
680	0.478	0.460	0.460	0.517

Hard query

672 a person wearing a **cap backwards**

Per-topic analysis on TV21

topic id	Run1	Run2	Run3	Run4
661	0.395	0.388	0.384	0.423
662	0.792	0.811	0.824	0.657
663	0.504	0.485	0.486	0.533
664	0.269	0.260	0.261	0.260
665	0.127	0.120	0.120	0.134
666	0.207	0.214	0.212	0.186
667	0.110	0.092	0.093	0.128
668	0.287	0.262	0.260	0.299
669	0.171	0.163	0.165	0.184
670	0.514	0.506	0.505	0.516
671	0.166	0.166	0.166	0.143
672	0.013	0.020	0.019	0.009
673	0.056	0.068	0.067	0.035
674	0.879	0.906	0.904	0.864
675	0.511	0.506	0.506	0.511
676	0.176	0.162	0.163	0.193
677	0.833	0.835	0.836	0.830
678	0.336	0.337	0.337	0.296
679	0.040	0.042	0.043	0.031
680	0.478	0.460	0.460	0.517

Hard query

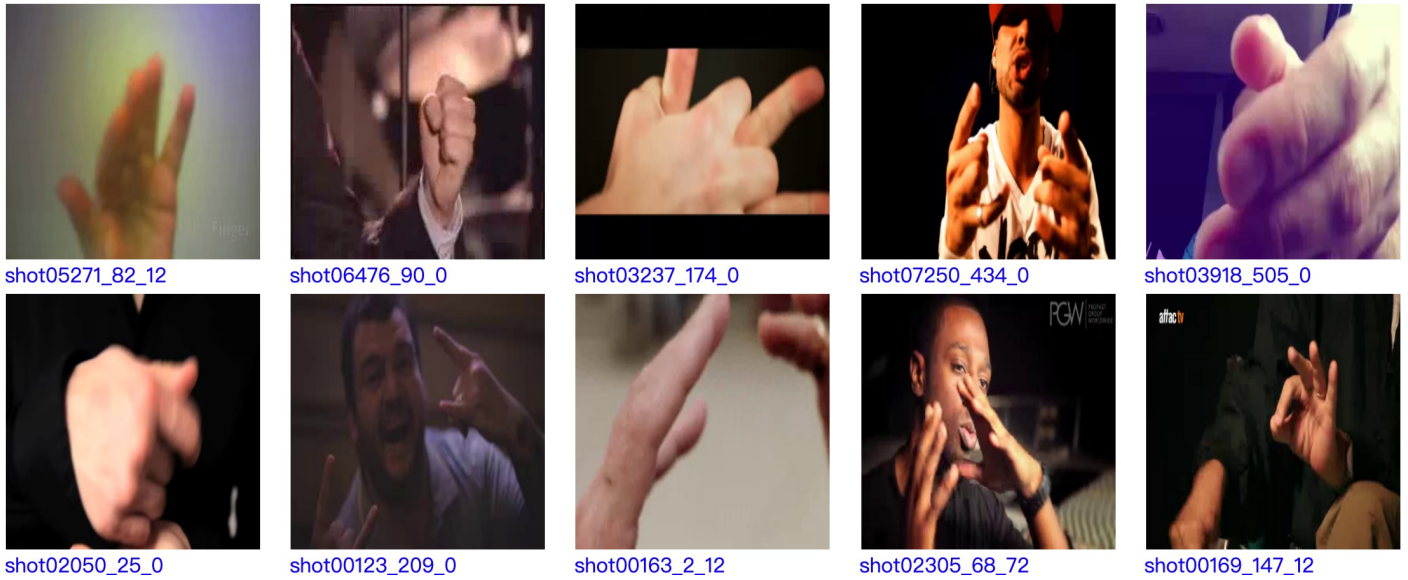
679 a ladder with **less than 6 steps**

Per-topic analysis on TV21

topic id	Run1	Run2	Run3	Run4
661	0.395	0.388	0.384	0.423
662	0.792	0.811	0.824	0.657
663	0.504	0.485	0.486	0.533
664	0.269	0.260	0.261	0.260
665	0.127	0.120	0.120	0.134
666	0.207	0.214	0.212	0.186
667	0.110	0.092	0.093	0.128
668	0.287	0.262	0.260	0.299
669	0.171	0.163	0.165	0.184
670	0.514	0.506	0.505	0.516
671	0.166	0.166	0.166	0.143
672	0.013	0.020	0.019	0.009
673	0.056	0.068	0.067	0.035
674	0.879	0.906	0.904	0.864
675	0.511	0.506	0.506	0.511
676	0.176	0.162	0.163	0.193
677	0.833	0.835	0.836	0.830
678	0.336	0.337	0.337	0.296
679	0.040	0.042	0.043	0.031
680	0.478	0.460	0.460	0.517

Hard query

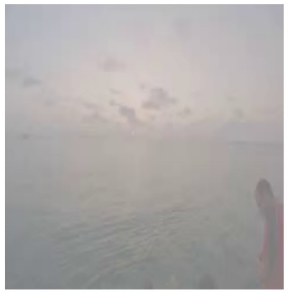
673 a man **pointing with** his finger



Are we making progress?

Per-topic analysis on TV20

655 one or more persons **standing in a body of water**

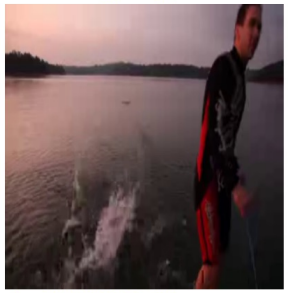

shot07119_264_0


shot03670_19_0

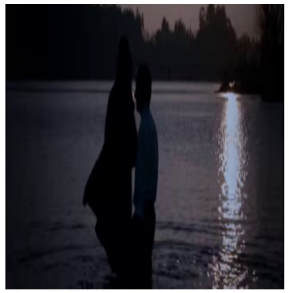

shot04386_80_0


shot05625_54_0

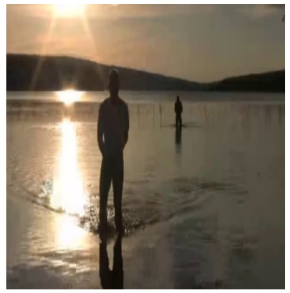

shot04342_137_12


shot02623_6_15

shot06853_69_0


shot04155_393_12


shot00119_72_0


shot00823_192_25

Topic id	RUCMM@TV20	RUCMM@TV21
641	0.281	0.316
642	0.522	0.640
643	0.092	0.228
644	0.662	0.861
645	0.166	0.089
646	0.155	0.254
647	0.436	0.364
648	0.099	0.063
649	0.469	0.835
650	0.079	0.085
651	0.072	0.610
652	0.130	0.228
653	0.231	0.416
654	0.341	0.365
655	0.053	0.064
656	0.629	0.730
657	0.084	0.106
658	0.055	0.062
659	0.342	0.426
660	0.476	0.574

Easy queries become easier, hard queries remain hard.

Are we making progress?

Per-topic analysis on TV20

658 two or more people **under a tree**


shot03193_249_37


shot03193_238_175


shot03193_242_275


shot06805_52_25


shot02499_169_25


shot06677_100_60


shot03484_8_29


shot03193_243_62


shot04413_19_72


shot05693_160_0

Topic id	RUCMM@TV20	RUCMM@TV21
641	0.281	0.316
642	0.522	0.640
643	0.092	0.228
644	0.662	0.861
645	0.166	0.089
646	0.155	0.254
647	0.436	0.364
648	0.099	0.063
649	0.469	0.835
650	0.079	0.085
651	0.072	0.610
652	0.130	0.228
653	0.231	0.416
654	0.341	0.365
655	0.053	0.064
656	0.629	0.730
657	0.084	0.106
658	0.055	0.062
659	0.342	0.426
660	0.476	0.574

Easy queries become easier, hard queries remain hard.

Conclusions

To boost AVS performance

- Cross modal Matching Model
- Video/text feature
- (Pre-)Training data

Understanding **fine-grained** queries is still hard

- Attributes: number
- Positions: in the water, under a tree

caz@ruc.edu.cn