Sentence Encoder Assembly for Ad-hoc Video Search

Fangming Zhou, Aozhu Chen, Xirong Li

AI & Media Computing Lab, Renmin University of China

TRECVID 2020 Workshop

2020-12-8

Ad-hoc Video Search

How to retrieve unlabeled videos for ad-hoc textual queries?

query **video**

Representation Learning

Common space

Using **cross-modal representation learning**

- Sentence representation
- Video representation
- Common space

What We Focus This Year

How to fully utilize multiple text encoders?

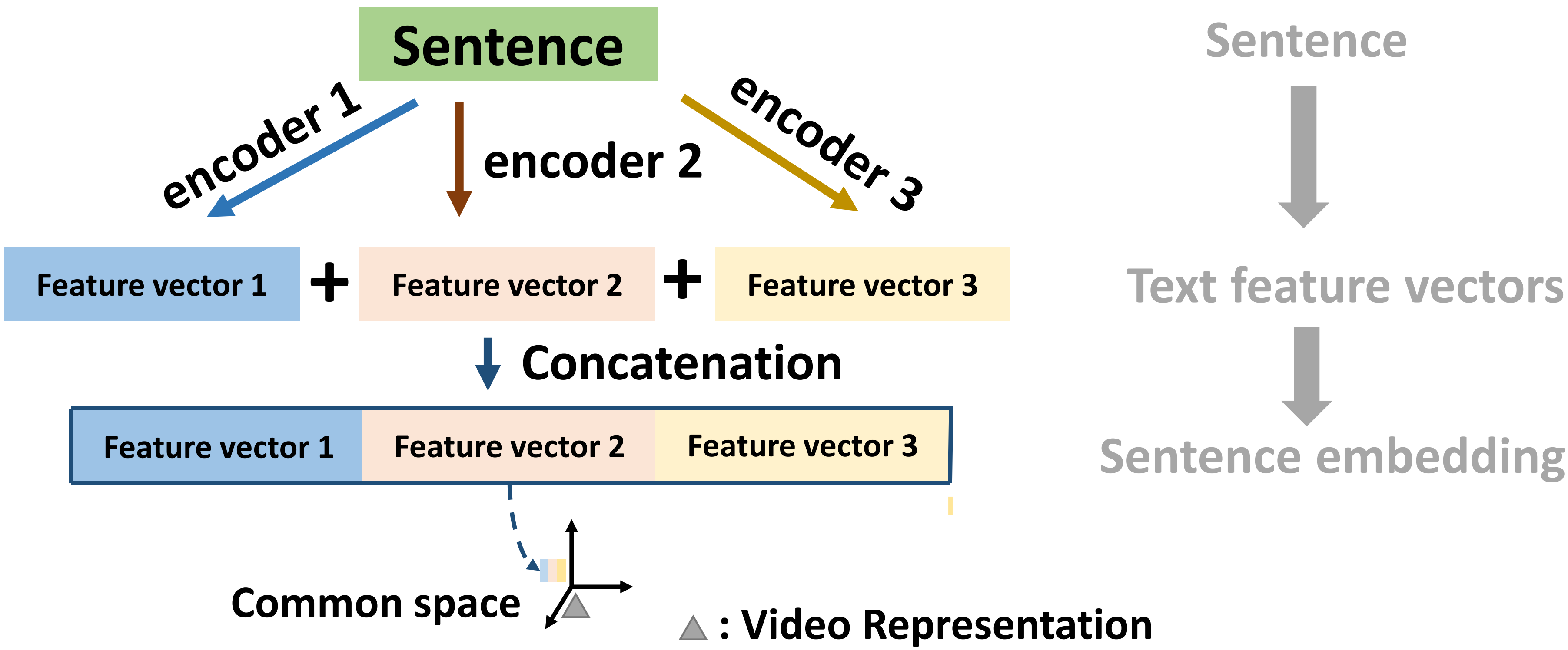
Previously some works combine multiple sentence encoders together as the sentence embedding

Works	Sentence encoders used
W2VV++ [Li et al., ACMMM'19] (AVS 2018 winner)	Bag-of-word, word2vec, GRU
Dual Encoding [Dong et al., CVPR'19]	Bag-of-word, bi-GRU, 1d-CNN
……	……

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

What We Focus This Year

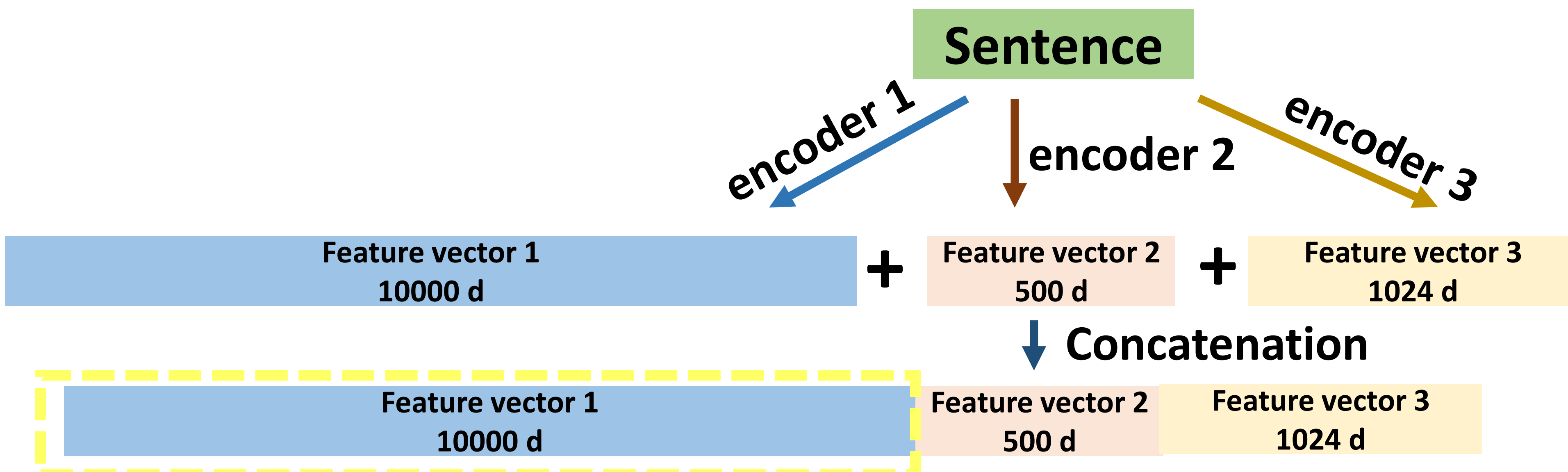
But they just concatenate text feature vectors



What We Focus This Year

Their **disadvantages**? **Longer** vector is **dominant**

- The combined feature can be dominated by a specific encoder



Longer vector dominated

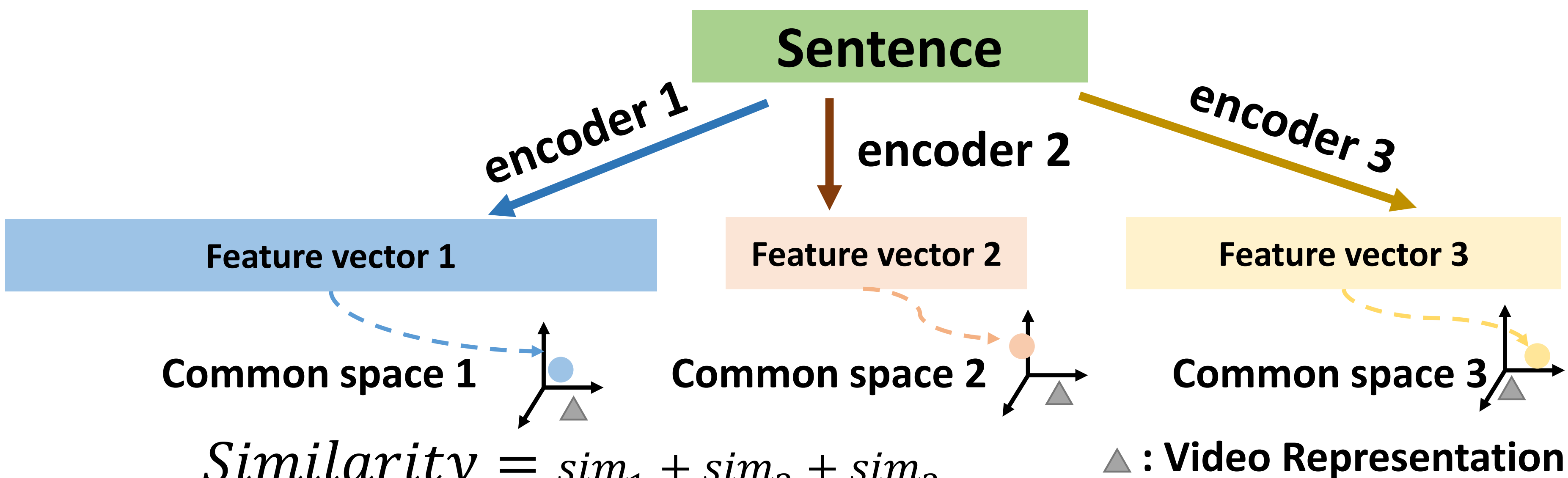
How We Do

Sentence Encoder Assembly (SEA)

A new and general architecture

① Multi-space Learning

Long vector dominated problem is solved



$$Similarity = sim_1 + sim_2 + sim_3$$

How We Do

Loss for **multi space learning**

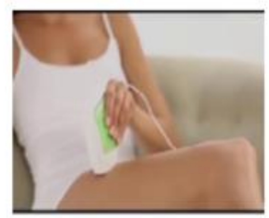
We use **the marginal ranking loss**[1] to select the **hard negative examples** during training

- Selection based on the **combined similarity** (a)
- Selections based on **individual similarities** (b-d)

More diverse hard negatives should be used

A female giving a nail art tutorial

(a) *Combined*



(b) *BoW*

(c) *w2v*

(d) *GRU*

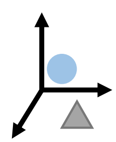
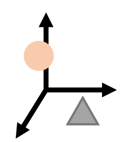
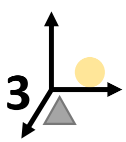
[1] Faghri et al., VSE++: Improving visual-semantic embeddings with hard negatives, BMVC 2018

How We Do

Sentence Encoder Assembly (SEA)

② Multi-loss Learning

- Learning $\boldsymbol{k}$ common spaces for $\boldsymbol{k}$ encoders
- **Combined loss** $loss = \sum_{i=1}^{k} loss_i(sentence, video)$
- **30%** extra hard negatives can be provided

Common space 1  **Common space 2**  **Common space 3** 

$loss_1$ $loss_2$ $loss_3$

$$loss = loss_1 + loss_2 + loss_3$$

Is SEA Effective?

Retrospective experiments on TV16-19 in [1] can prove:

Sentence encoders	Model	TV16	TV17	TV18	TV19	SUM
bow,w2v	W2VV++	0.144	0.218	0.111	0.143	0.616
	SEA	0.157	0.234	0.128	0.166	0.685
bow,w2v,gru	W2VV++	0.162	0.223	0.101	0.139	0.625
	SEA	0.150	0.234	0.122	0.166	0.672
bow,w2v,bigru	W2VV++	0.161	0.217	0.104	0.135	0.617
	SEA	0.164	0.228	0.125	0.167	0.684
bow, w2v,bert	W2VV++	0.151	0.225	0.102	0.128	0.606
	SEA	0.153	0.228	0.121	0.148	0.650
bow,w2v,gru,bert	W2VV++	0.143	0.193	0.093	0.101	0.530
	SEA	0.160	0.231	0.121	0.154	0.666
bow,w2v,bigru,bert	W2VV++	0.158	0.206	0.090	0.105	0.559
	SEA	0.159	0.229	0.117	0.155	0.660

- Darker color indicates higher infAP
- SEA models surpass almost all the corresponding W2VV++ models

Yes, it is

* **W2VV++** uses a single common space

[1] Li et al., SEA: Sentence encoder assembly for video retrieval by textual queries, T-MM *2021*.

Choice of Sentence Encoder

We consider 5 sentence encoders

1. Bag-of-word (BoW)
2. word2vec (w2v)
3. NetVlad
4. bi-GRU *
5. BERT *

Different combinations are designed

1. SEA (BoW, NetVlad)
2. SEA (BoW, w2v)
3. SEA (BoW, w2v, bi-GRU)
4. SEA (BoW, w2v, bi-GRU, BERT)

complexity

* Indicates **Sequence model**

Datasets and Visual Features

Datasets	Usage
msrvtt10k	training
tgif	training
TV2016 VTT training set	validation
the Google's Conceptual Captions (GCC)	pre-training

Image caption dataset pre-training is **not** suitable for models with **C3D feature**

Video Features	Dim.
ResNext-101 (frame-level)	2,048
ResNet-152 (frame-level)	2,048
C3D (video-level)	2,048

Combined Video Features	Dim.
ResNext-101 + ResNet-152	4,096
ResNext-101 + ResNet-152 + C3D	6,144

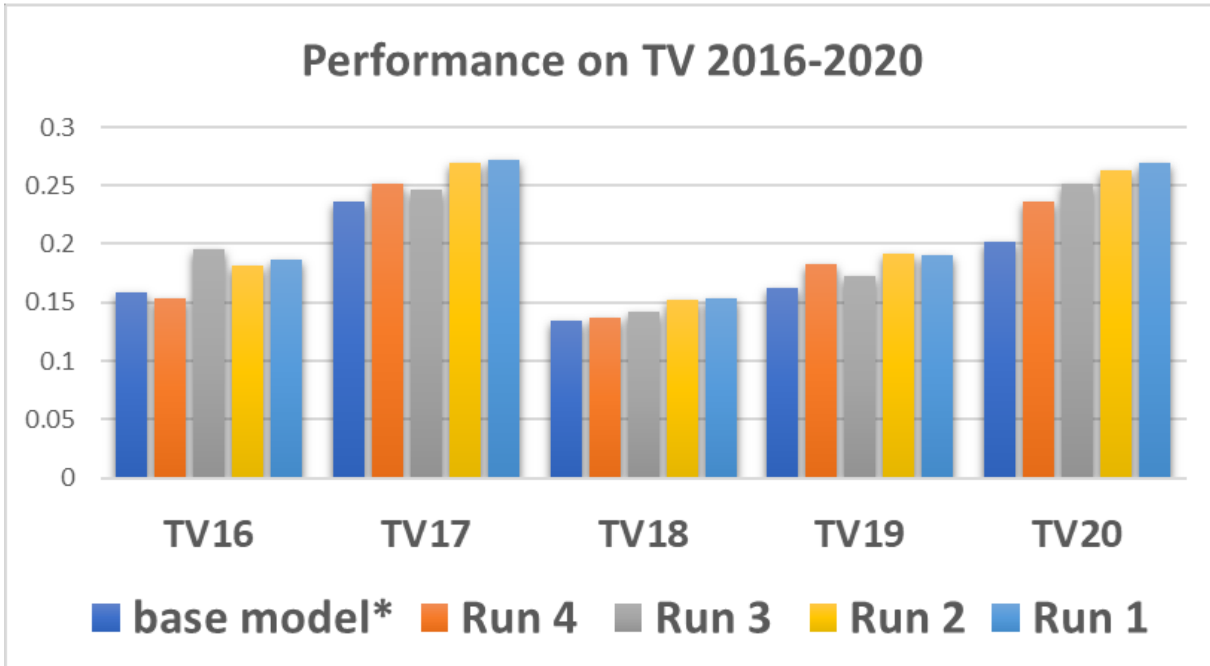
We use **combined video features** for better performance

Submissions
(fully automatic track)

We submitted the following 4 runs:
SEA model with different setups and their combinations

Run id	Description
Run 4	SEA (BoW, w2v), **ResNeXt-ResNet-C3D**
Run 3	SEA (BoW, NetVlad), ResNeXt-ResNet, and **pre-trained on GCC**
Run 2	Late average fusion of three SEA models
Run 1 (primary run)	Late average fusion of four SEA models

Performance on TV2016-2020 AVS Task

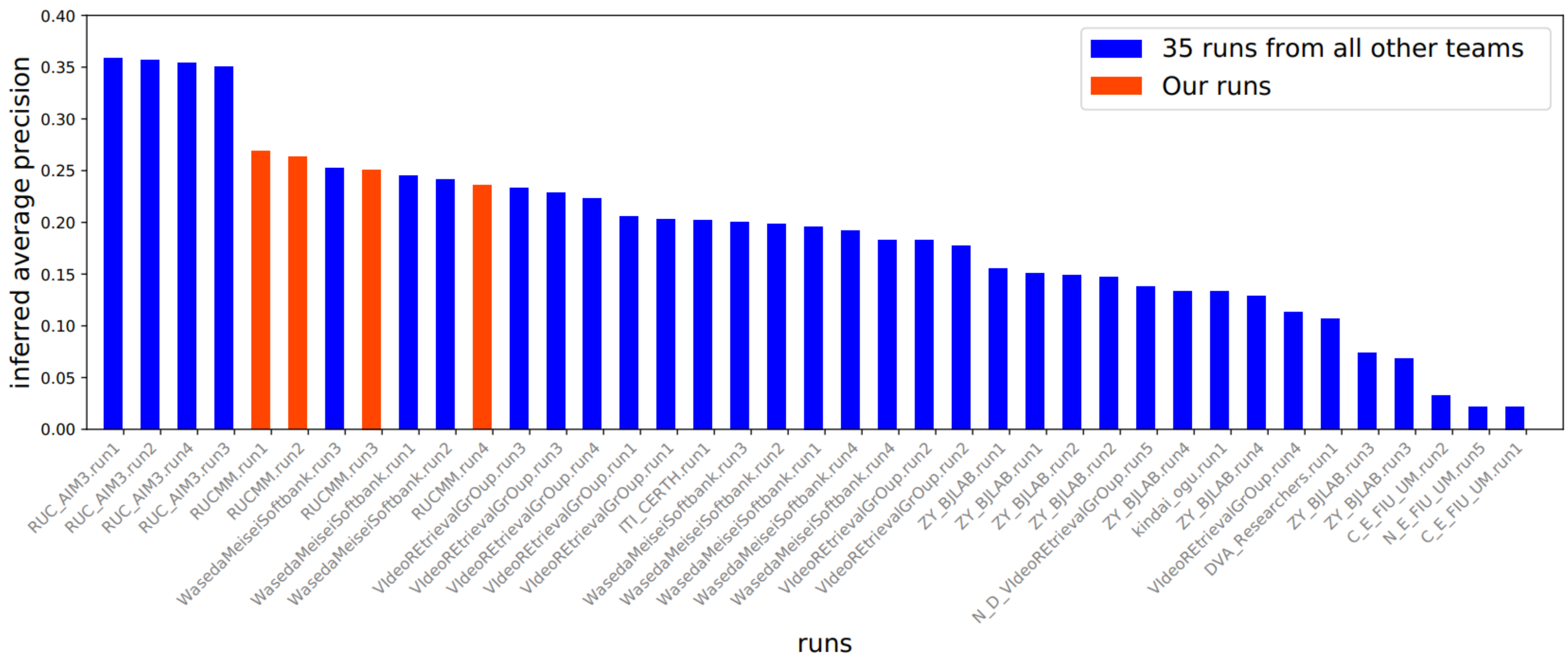


- Appending extra C3D feature helps (Run 4)
- Pre-training strategy helps (Run 3)
- Model ensemble helps (Run 2, Run 1)

* **base model**: SEA(BoW, NetVlad), no C3D feature included, no pre-training

All fully automatic AVS submissions

Our submissions ranked the 2nd



Results of individual topics

topic id	runs1	runs2	runs3	runs4
641	0.281	0.279	0.232	0.231
642	0.522	0.541	0.454	0.516
643	0.092	0.097	0.053	0.088
644	0.662	0.446	0.795	0.038
645	0.166	0.168	0.153	0.154
646	0.155	0.158	0.132	0.136
647	0.436	0.436	0.400	0.398
648	0.099	0.092	0.105	0.078
649	0.469	0.500	0.546	0.477
650	0.079	0.083	0.057	0.103
651	0.072	0.080	0.001	0.213
652	0.130	0.132	0.123	0.125
653	0.231	0.286	0.045	0.313
654	0.341	0.337	0.320	0.302
655	0.053	0.047	0.013	0.087
656	0.629	0.627	0.618	0.559
657	0.084	0.077	0.068	0.031
658	0.055	0.050	0.080	0.046
659	0.342	0.355	0.323	0.371
660	0.476	0.477	0.512	0.455

More **blue** is better

More **red** is worse

Can be divided into 3 types:

- Easy topics: all blue
- Not easy topics: not all blue
- Hard topics: all red

Case Study

Easy query

- All models perform well

topic id	runs1	runs2	runs3	runs4
641	0.281	0.279	0.232	0.231
642	0.522	0.541	0.454	0.516
643	0.092	0.097	0.053	0.088
644	0.662	0.446	0.795	0.038
645	0.166	0.168	0.153	0.154
646	0.155	0.158	0.132	0.136
647	0.436	0.436	0.400	0.398
648	0.099	0.092	0.105	0.078
649	0.469	0.500	0.546	0.477
650	0.079	0.083	0.057	0.103
651	0.072	0.080	0.001	0.213
652	0.130	0.132	0.123	0.125
653	0.231	0.286	0.045	0.31[illegible]
654	0.341	0.337	0.320	0.302
655	0.053	0.047	0.013	0.087
656	0.629	0.627	0.618	0.559
657	0.084	0.077	0.068	0.031
658	0.055	0.050	0.080	0.046
659	0.342	0.355	0.323	0.371
660	0.476	0.477	0.512	0.455

656 a long haired man

Case Study

Easy query

- All models perform well

topic id	runs1	runs2	runs3	runs4
641	0.281	0.279	0.232	0.231
642	0.522	0.541	0.454	0.516
643	0.092	0.097	0.053	0.088
644	0.662	0.446	0.795	0.03[illegible]
645	0.166	0.168	0.153	0.154
646	0.155	0.158	0.132	0.136
647	0.436	0.436	0.400	0.398
648	0.099	0.092	0.105	0.078
649	0.469	0.500	0.546	0.477
650	0.079	0.083	0.057	0.103
651	0.072	0.080	0.001	0.213
652	0.130	0.132	0.123	0.125
653	0.231	0.286	0.045	0.313
654	0.341	0.337	0.320	0.302
655	0.053	0.047	0.013	0.087
656	0.629	0.627	0.618	0.559
657	0.084	0.077	0.068	0.031
658	0.055	0.050	0.080	0.046
659	0.342	0.355	0.323	0.371
660	0.476	0.477	0.512	0.455

642 a person paddling **kayak** in the water

shot06958_261_75 shot06951_88_101 shot02352_87_145 shot06958_233_37 shot01357_97_11
shot06958_222_150 shot02312_93_383 shot02312_94_29 shot06806_159_130 shot02862_131_45

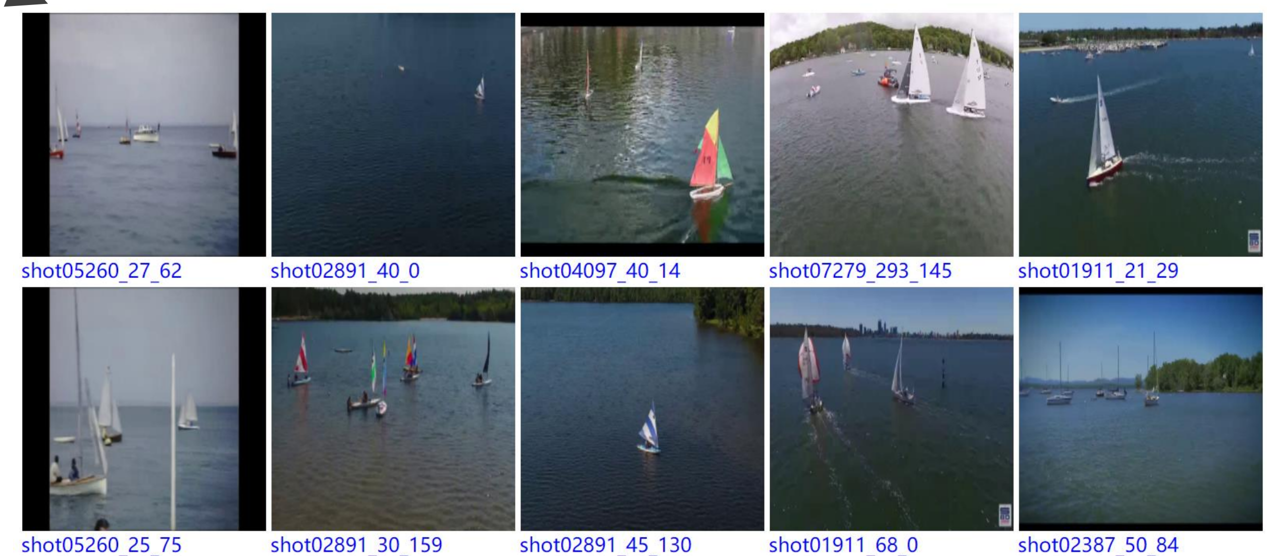
topic id	runs1	runs2	runs3	runs4
641	0.281	0.279	0.232	0.231
642	0.522	0.541	0.454	0.516
643	0.092	0.097	0.053	0.088
644	0.662	0.446	0.795	0.038
645	0.166	0.168	0.153	0.154
646	0.155	0.158	0.132	[illegible]
647	0.436	0.436	0.400	0.398
648	0.099	0.092	0.105	0.078
649	0.469	0.500	0.546	0.477
650	0.079	0.083	0.057	0.103
651	0.072	0.080	0.001	0.213
652	0.130	0.132	0.123	0.125
653	0.231	0.286	0.045	0.313
654	0.341	0.337	0.320	0.302
655	0.053	0.047	0.013	0.087
656	0.629	0.627	0.618	0.559
657	0.084	0.077	0.068	0.031
658	0.055	0.050	0.080	0.046
659	0.342	0.355	0.323	0.371
660	0.476	0.477	0.512	0.455

Case Study

No Easy query

- Not all models perform well

644 sailboats in the water



Run3 works well

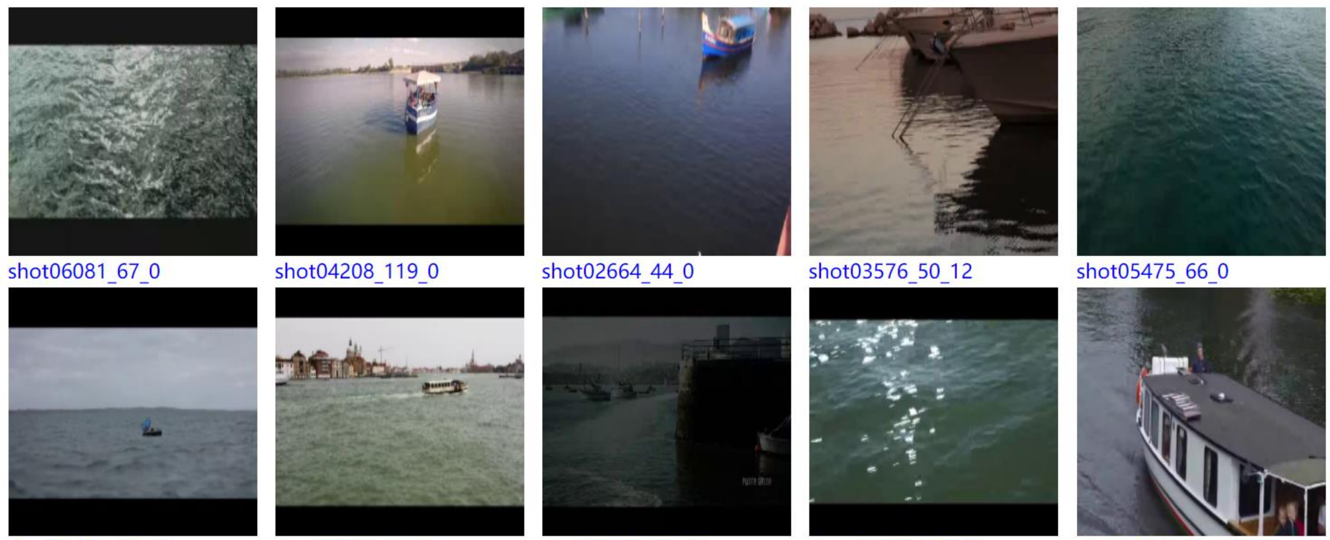
Case Study

No Easy query

- Not all models perform well

topic id	runs1	runs2	runs3	runs4
641	0.281	0.279	0.232	0.231
642	0.522	0.541	0.454	0.516
643	0.092	0.097	0.053	[illegible]
644	0.662	0.446	0.795	0.038
645	0.166	0.168	0.153	[illegible]
646	0.155	0.158	0.132	0.136
647	0.436	0.436	0.400	0.398
648	0.099	0.092	0.105	0.078
649	0.469	0.500	0.546	0.477
650	0.079	0.083	0.057	0.103
651	0.072	0.080	0.001	0.213
652	0.130	0.132	0.123	0.125
653	0.231	0.286	0.045	0.313
654	0.341	0.337	0.320	0.302
655	0.053	0.047	0.013	0.087
656	0.629	0.627	0.618	0.559
657	0.084	0.077	0.068	0.031
658	0.055	0.050	0.080	0.046
659	0.342	0.355	0.323	0.371
660	0.476	0.477	0.512	0.455

644 sailboats in the water



shot06081_67_0 shot04208_119_0 shot02664_44_0 shot03576_50_12 shot05475_66_0

shot01120 24 62 shot01679 65 0 shot00628 21 11 shot06470 133 12 shot04326 68 475

Run4 not works well, only shows **water and boat**

Case Study

No Easy query

- Not all models perform well

topic id	runs1	runs2	runs3	runs4
641	0.281	0.279	0.232	0.231
642	0.522	0.541	0.454	0.516
643	0.092	0.097	[illegible]	[illegible]
644	0.662	0.446	0.795	0.038
645	0.166	0.168	[illegible]	[illegible]
646	0.155	0.158	0.132	0.136
647	0.436	0.436	0.400	0.398
648	0.099	0.092	0.105	0.078
649	0.469	0.500	0.546	0.477
650	0.079	0.083	0.057	0.103
651	0.072	0.080	0.001	0.213
652	0.130	0.132	0.123	0.125
653	0.231	0.286	0.045	0.313
654	0.341	0.337	0.320	0.302
655	0.053	0.047	0.013	0.087
656	0.629	0.627	0.618	0.559
657	0.084	0.077	0.068	0.031
658	0.055	0.050	0.080	0.046
659	0.342	0.355	0.323	0.371
660	0.476	0.477	0.512	0.455

644 **sailboats** in the water

Datasets	Usage	
msrvtt10k + tgif	training	7 captions have '**sailboat**'
GCC	pre-training	853 captions have '**sailboat**'

Pre-training makes the difference

topic id	runs1	runs2	runs3	runs4
641	0.281	0.279	0.232	0.231
642	0.522	0.541	0.454	0.516
643	0.092	0.097	0.053	0.088
644	0.662	0.446	0.795	0.038
645	0.166	0.168	0.153	0.154
646	0.155	0.158	0.132	0.136
647	0.436	0.436	0.400	0.398
648	0.099	0.092	0.105	0.078
649	0.469	0.500	0.546	0.477
650	0.079	0.083	0.057	0.103
651	0.072	0.080	0.001	0.213
652	0.130	0.132	0.123	0.125
653	0.231	0.286	0.045	0.313
654	0.341	0.337	0.320	0.302
655	0.053	0.047	0.013	0.087
656	0.629	0.627	0.618	0.559
657	0.084	0.077	0.068	0.031
658	0.055	0.050	0.080	0.046
659	0.342	0.355	0.323	0.371
660	0.476	0.477	0.512	0.455

Case Study

Hard query

- All models perform bad

657 a woman with **short hair** indoors

The length of hair is not well modeled

Case Study

Hard query

- All models perform bad

topic id	runs1	runs2	runs3	runs4
641	0.281	0.279	0.232	0.231
642	0.522	0.541	0.454	0.516
643	0.092	0.097	0.053	0.088
644	0.662	0.446	0.795	0.038
645	0.166	0.168	0.153	0.154
646	0.155	0.158	0.132	0.136
647	0.436	0.436	0.400	0.398
648	0.099	0.092	0.105	0.078
649	0.469	0.500	0.546	0.477
650	0.079	0.083	0.057	0.103
651	0.072	0.080	0.001	0.213
652	0.130	0.132	0.123	0.125
653	0.231	0.286	0.045	0.313
654	0.341	0.337	0.320	0.302
655	0.053	0.047	0.013	0.087
656	0.629	0.627	0.618	0.559
657	0.084	0.077	0.068	0.031
658	0.055	0.050	0.080	0.046
659	0.342	0.355	0.323	0.371
660	0.476	0.477	0.512	0.455

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

Persons **not in** the water in some videos

Case Study

Hard query

- All models perform bad

topic id	runs1	runs2	runs3	runs4
641	0.281	0.279	0.232	0.231
642	0.522	0.541	0.454	0.516
643	0.092	0.097	0.053	0.088
644	0.662	0.446	0.795	0.038
645	0.166	0.168	0.153	0.154
646	0.155	0.158	0.132	0.136
647	0.436	0.436	0.400	0.398
648	0.099	0.092	0.105	0.078
649	0.469	0.500	0.546	0.477
650	0.079	0.083	0.057	0.103
651	0.072	0.080	0.001	0.213
652	0.130	0.132	0.123	0.125
653	0.231	0.286	0.045	0.313
654	0.341	0.337	0.320	0.302
655	0.053	0.047	0.013	0.087
656	0.629	0.627	0.618	0.559
657	0.084	0.077	0.060	0.031
658	0.055	0.050	0.080	0.046
659	0.342	0.355	0.323	0.371
660	0.476	0.477	0.512	0.455

658 two or more people under a tree

- Often **only one person** or **without people**
- Some **in** the tree instead of **under** the tree

Case Study

Hard query

- All models perform bad

topic id	runs1	runs2	runs3	runs4
641	0.281	0.279	0.232	0.231
642	0.522	0.541	0.454	0.516
643	0.092	0.097	0.053	0.088
644	0.662	0.446	0.795	0.038
645	0.166	0.168	0.153	0.154
646	0.155	0.158	0.132	0.136
647	0.436	0.436	0.400	0.398
648	0.099	0.092	0.105	0.078
649	0.469	0.500	0.546	0.477
650	0.079	0.083	0.057	0.103
651	0.072	0.080	0.001	0.213
652	0.130	0.132	0.123	0.125
653	0.231	0.286	0.045	0.313
654	0.341	0.337	0.320	0.302
655	0.053	0.047	0.013	0.087
656	0.629	0.627	0.618	0.559
657	0.084	0.077	0.068	0.031
658	0.055	0.050	0.080	0.046
659	0.342	0.355	0.323	0.371
660	0.476	0.477	0.512	0.455

643 people **dancing or singing** while wearing costumes outdoors

people **standing** or **walking**

Conclusions

To boost AVS performance

- Multi-space multi-loss Learning
- Appending extra C3D feature
- Pre-training on image caption dataset
- Late average fusion

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

- Attributes: number of persons, length of women's hair, etc.
- Actions: dancing, singing
- Positions: in the water, under a tree

Reproducibility & Reference

Code & Resources: https://github.com/li-xirong/sea (in preparation)

Papers:

- SEA: Sentence encoder assembly for video retrieval by textual queries. *IEEE Trans. Multimedia 2021*. https://arxiv.org/abs/2011.12091
- Renmin University of China at TRECVID 2020: Sentence Encoder Assembly for Ad-hoc Video Search, *TRECVID 2020 Workshop*

Contact: *xirong@ruc.edu.cn*

Thanks!