Dense Encoding for Video-to-Text Matching

Jianfeng Dong[1], Xirong Li[2], Chaoxi Xu[2], Jing Cao[2], Xun Wang[1], Gang Yang[2]

[1]Zhejiang Gongshang University

[2]AI & Media Computing Lab, Renmin University of China

Video to Text (VTT) Task @ TRECVID 2018

Matching and Ranking Task

Task: given a query video, participants are asked to rank a list of pre-defined sentences.

Given video

Candidate sentences

a man speaks to audiences indoors

a boy jumps on a trampoline

. . .

a person skates indoors

Ranked sentences

similarity

High

a boy jumps on a trampoline

a person skates indoors

. . .

a man speaks to audiences indoors

Low

Cross-modal Similarity

Key question: how to compute cross-modal similarity?

Video

Similarity

Sentence

?

Athletics make a choreography in gym.

Common space based cross-modal retrieval

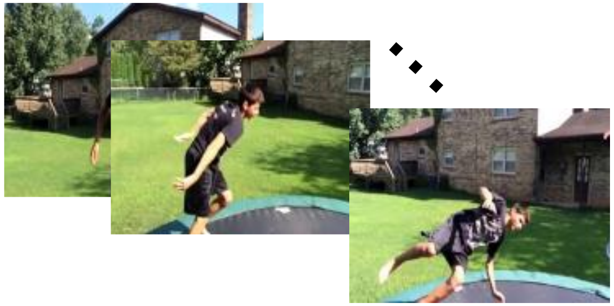
Cross-modal Retrieval

Common space based cross-modal retrieval models can be typically decomposed into two modules:

- Data encoding
- Common space learning

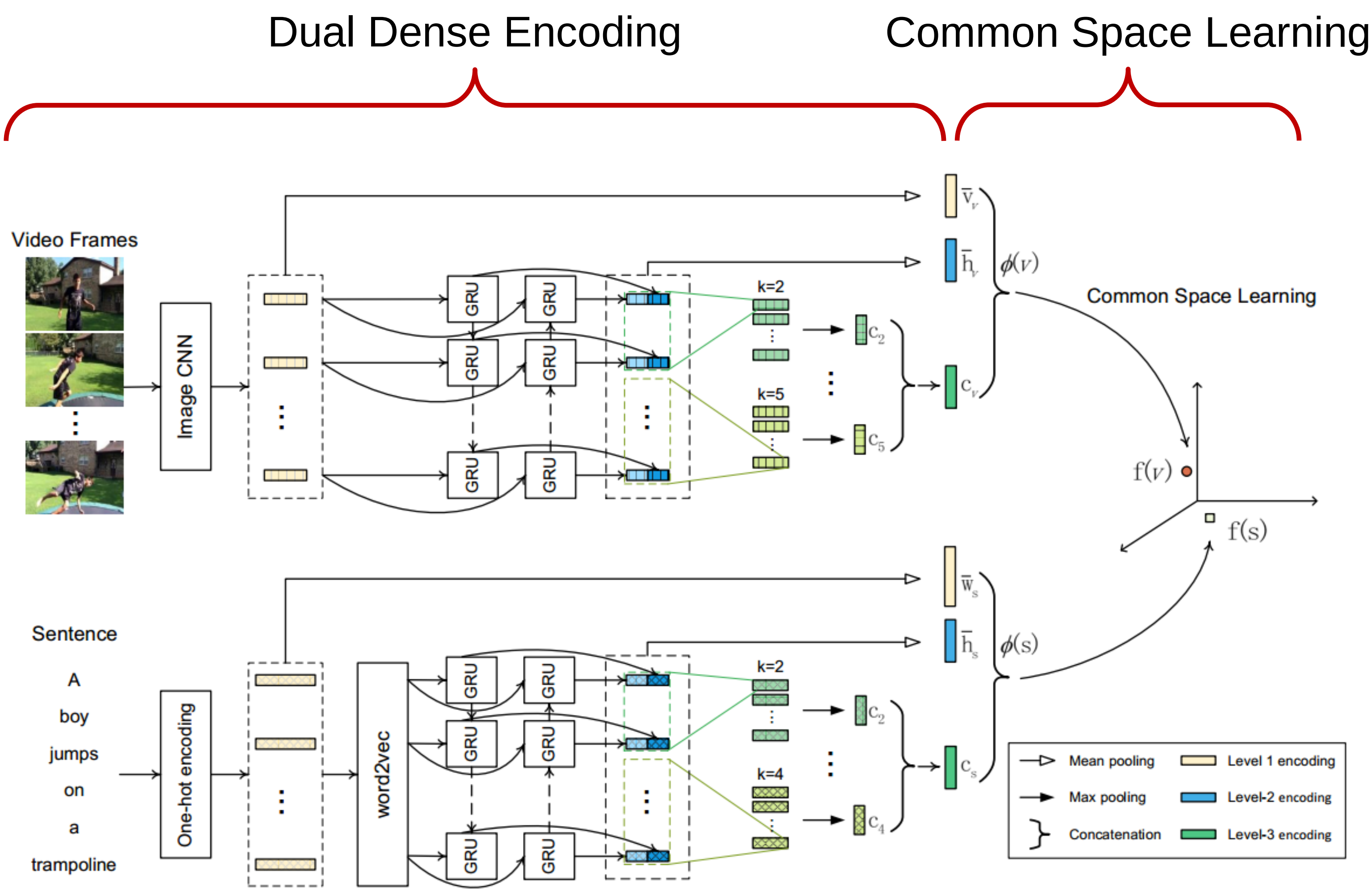
video as a sequence of frames

sentence as a sequence of words



A boy jumps on a trampoline

Our Model



Dual Dense Encoding
Common Space Learning
Video Frames
Image CNN
GRU
k=2
k=5
c_2
c_5
c_v
$\bar{v}_v$
$\bar{h}_v$
$\phi(v)$
Common Space Learning
$f(v)$
$f(s)$
Sentence
A
boy
jumps
on
a
trampoline
One-hot encoding
word2vec
k=4
c_4
c_s
$\bar{w}_s$
$\bar{h}_s$
$\phi(s)$
Mean pooling
Max pooling
Concatenation
Level 1 encoding
Level-2 encoding
Level-3 encoding

Dual Dense Encoding

By jointly exploiting multi-level encodings, dual dense encoding is designed to explicitly model global, local and temporal patterns in videos and sentences.

Level 1. Global Encoding by Mean Pooling

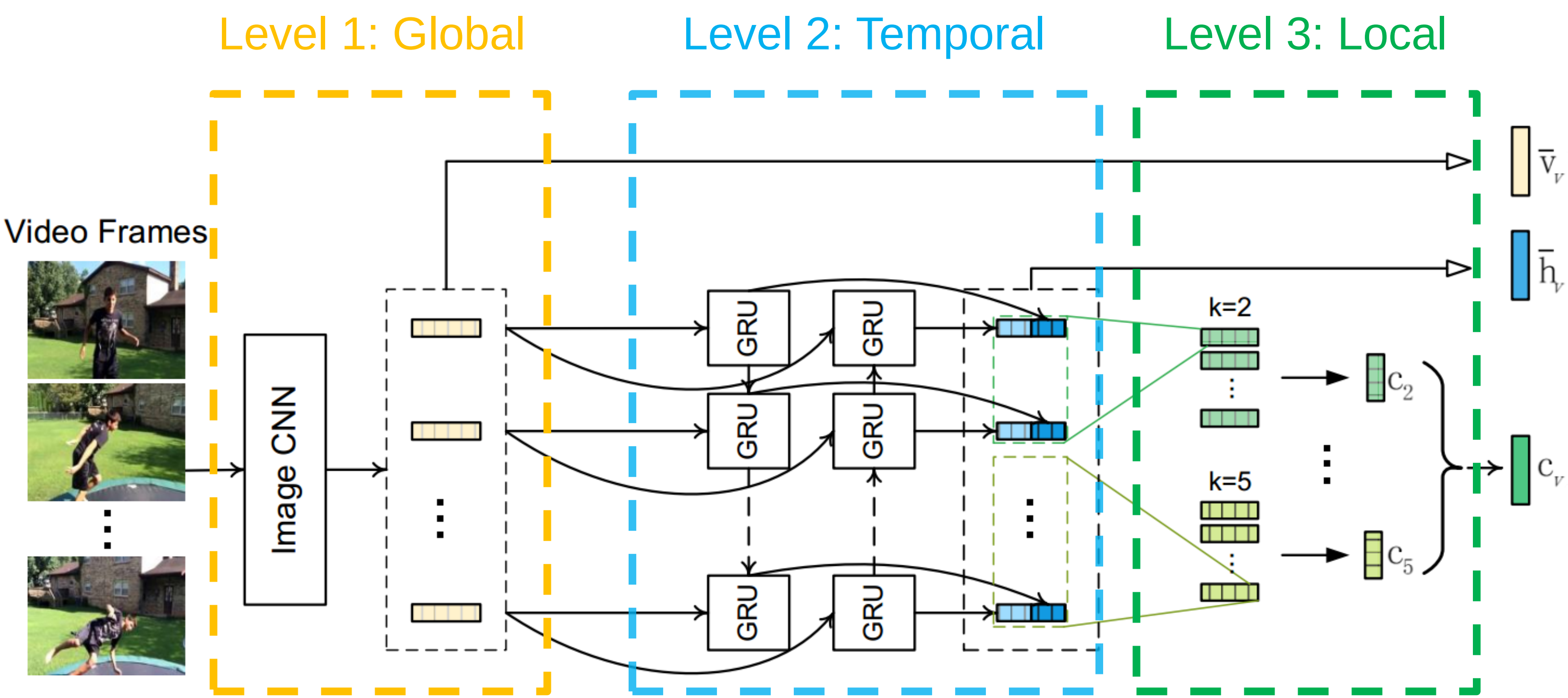
Level 2. Temporal-Aware Encoding by biGRU

Level 3. Local-Enhanced Encoding by biGRU-CNN

Dong, J., Li, X., Xu, C., Ji, S., & Wang, X. (2018). Dual Dense Encoding for Zero-Example Video Retrieval. *arXiv preprint arXiv:1809.06181*.

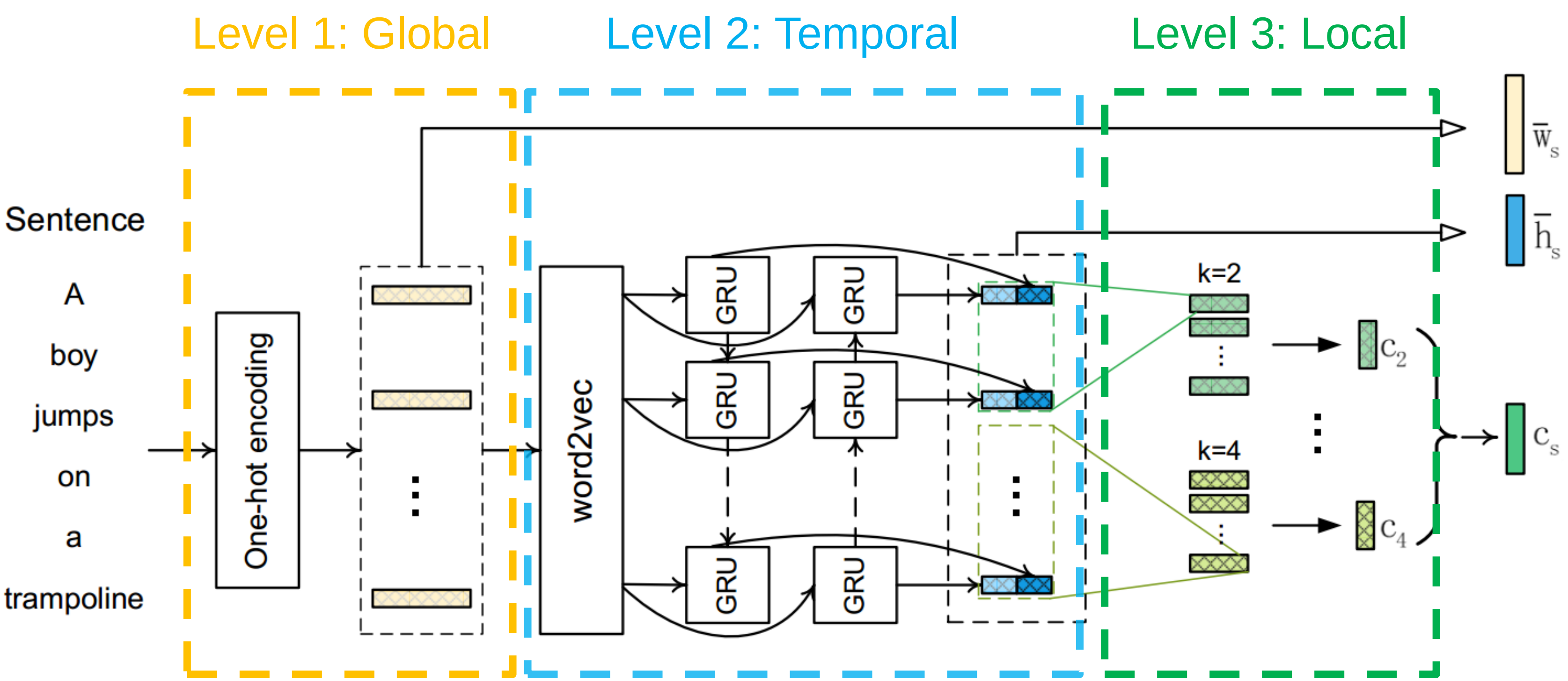
Video Encoding

Dense encoding generates new, higher-level features progressively.



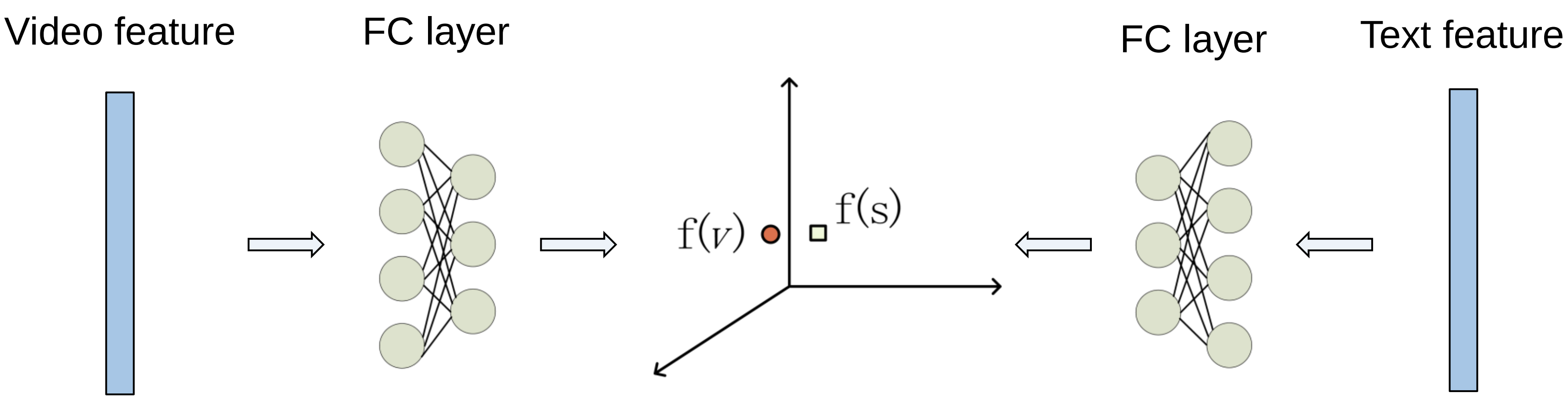
Sentence Encoding

Dense encoding for sentences is very similar to the dense encoding for videos.



Common Space Learning

We choose VSE++ as the common space learning model. Note the dual dense encoding can be flexibly applied to other common space learning models.



Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. VSE++: Improved visual-semantic embeddings. In BMVC, 2018.

Loss Function

Triplet Ranking Loss:

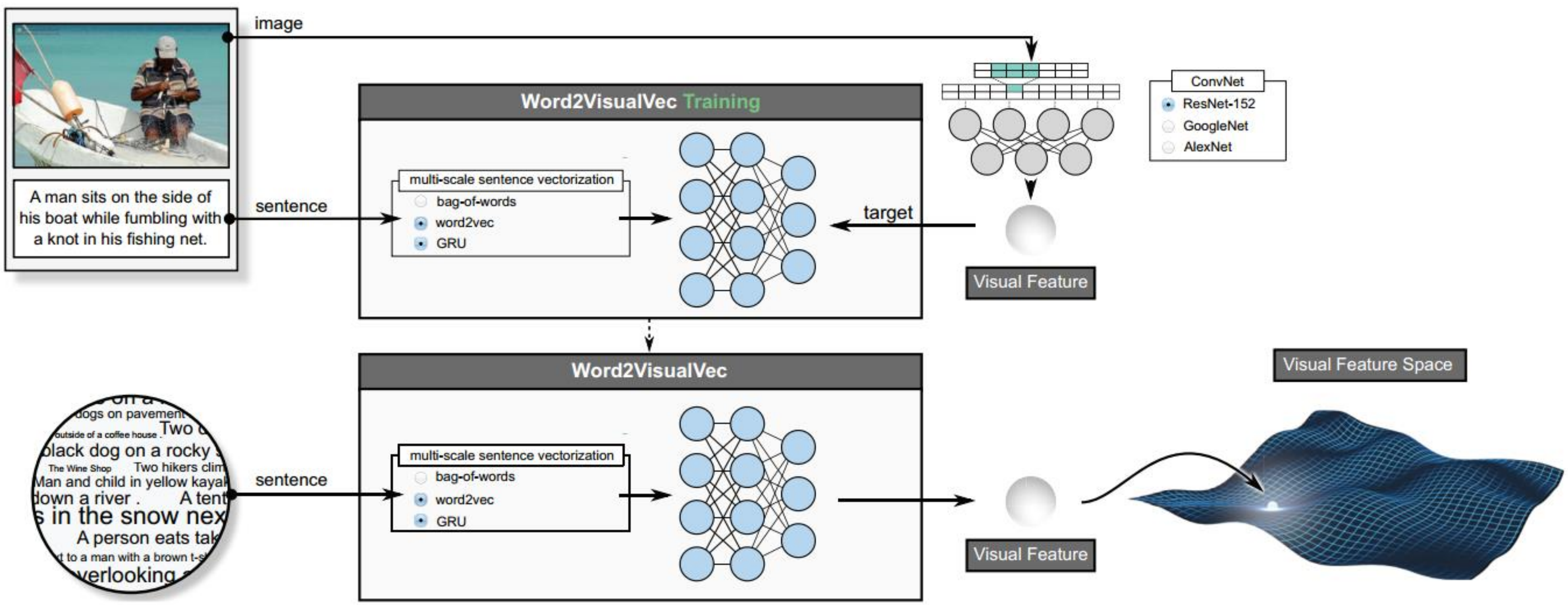
$$\mathcal{L}(v, s; \theta) = max(0, \alpha + S_\theta(v, s^-) - S_\theta(v, s)) \\ + max(0, \alpha + S_\theta(v^-, s) - S_\theta(v, s)),$$

How to select negative samples s^- and v^-:

- Randomly selected samples
- Select the most similar yet negative samples

Word2VisaulVec++

- Represent sentences into a visual feature space
- Use the improved triplet ranking loss instead of MSE



Dong, J.; Li, X.; and Snoek, C. G. Predicting visual features from text for image and video caption retrieval. IEEE Trans. Multimedia 2018.

Datasets

	Dataset	#Videos	#Sentences
Train	MSVD	1,970	80,863
	MSR-VTT	10,000	200,000
	TGIF	100,855	124,534
Validation	tv2016train	200	200

Visual Features

Video frames are extracted uniformly with an interval of 0.5 second.

CNN features:

- ResNext-101: 2,048 dim
- ResNet-152: 2,048 dim

The extracted features are available at:
https://github.com/li-xirong/avs

Ablation Study

Dense encoding exploiting all the three levels is the best.

On MSR-VTT dataset

Encoding strategy	Text-to-Video Retrieval				Video-to-Text Retrieval				Sum of Recalls
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r	
Level 1 (Mean pooling)	6.4	18.8	27.3	47	11.5	27.7	38.2	22	124.4
Level 2 (biGRU)	6.3	19.4	28.5	38	10.1	26.8	37.7	20	124.8
Level 3 (biGRU-CNN)	7.3	21.5	31.2	32	10.6	27.3	38.5	20	136.4
Level 1 + 2	6.9	20.4	29.1	41	11.6	29.6	40.7	18	138.3
Level 1 + 3	7.5	21.6	31.2	33	11.9	30.5	41.7	16	144.7
Level 2 + 3	7.6	**22.4**	**32.2**	**31**	11.9	**30.9**	42.7	16	147.6
Level 1 + 2 + 3	**7.7**	22.0	31.8	32	**13.0**	30.8	**43.3**	**15**	**148.6**

Our Runs

Run 0: dual dense encoding model (single)

Run 1: equally combines eight dual dense encoding models with their last FC layer and visual feature varies

Run 2: equally combines eight Word2VisaulVec++ models with sentence encoding and visual feature varies

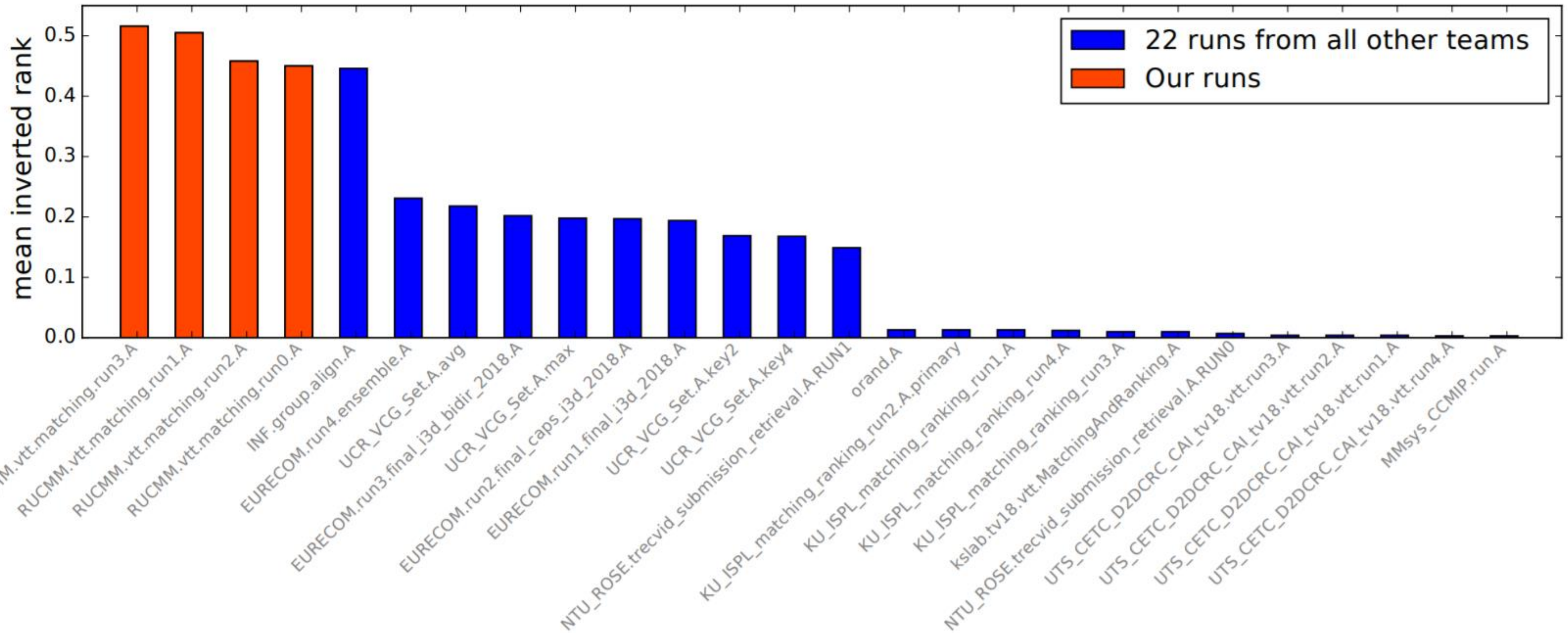
Run 3: combines run 1, run 2 and eight VSE++ models with sentence encoding and visual feature varies

Evaluation Results

	Model	Fusion	Set A	Set B	Set C	Set D	Set E
Run 0	Dense	×	0.450	0.448	0.430	0.433	0.448
Run 1	Dense	√	0.505	0.502	0.495	0.494	0.500
Run 2	W2VV++	√	0.458	0.453	0.448	0.436	0.455
Run 3	Dense W2VV++ VSE++	√	0.516	0.505	0.492	0.491	0.509

Leaderboard

Our runs lead the evaluation on five test sets.



Set A

Take-home Messages

- Dual dense encoding explicitly modeling global, local and temporal patterns is effective to encode videos and sentence
- Late fusion of multiple models is an important trick

The extracted features are available at:
https://github.com/li-xirong/avs

Thanks!