[ECCV 2018] Find and Focus: Retrieve and Localize Video Events with Natural Language Queries

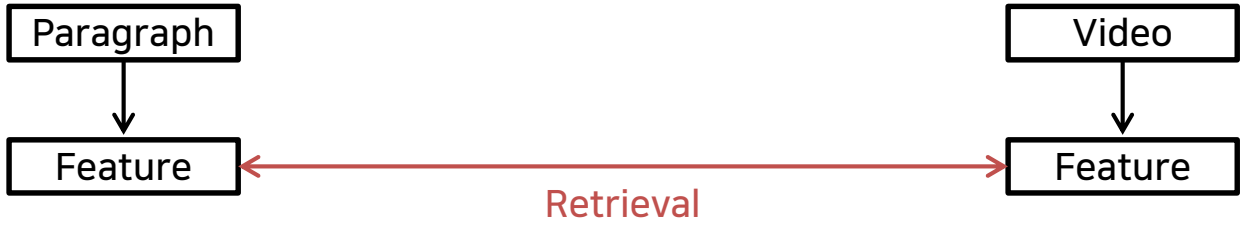
SenseTime Joint Lab

2021.10.25 Won Jo

Overview

Paragraph-to-video retrieval attempted to encode the entire video

However, as the video is getting **longer and the content diversity is growing**, the method of encoding the whole has limitations.



Overview: Find and Focus

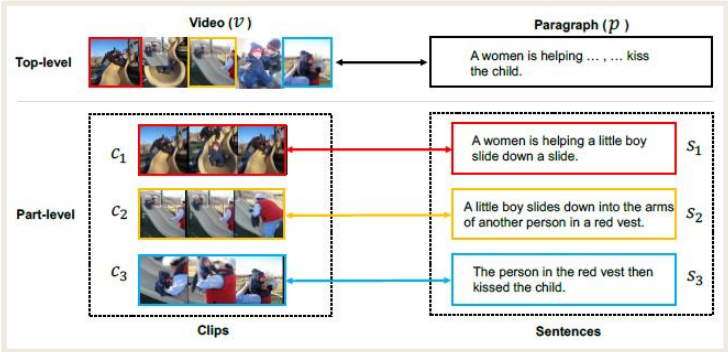
The proposed method is a method of localizing and refining
which sentence and which clip are the same in paragraph-to-video retrieval

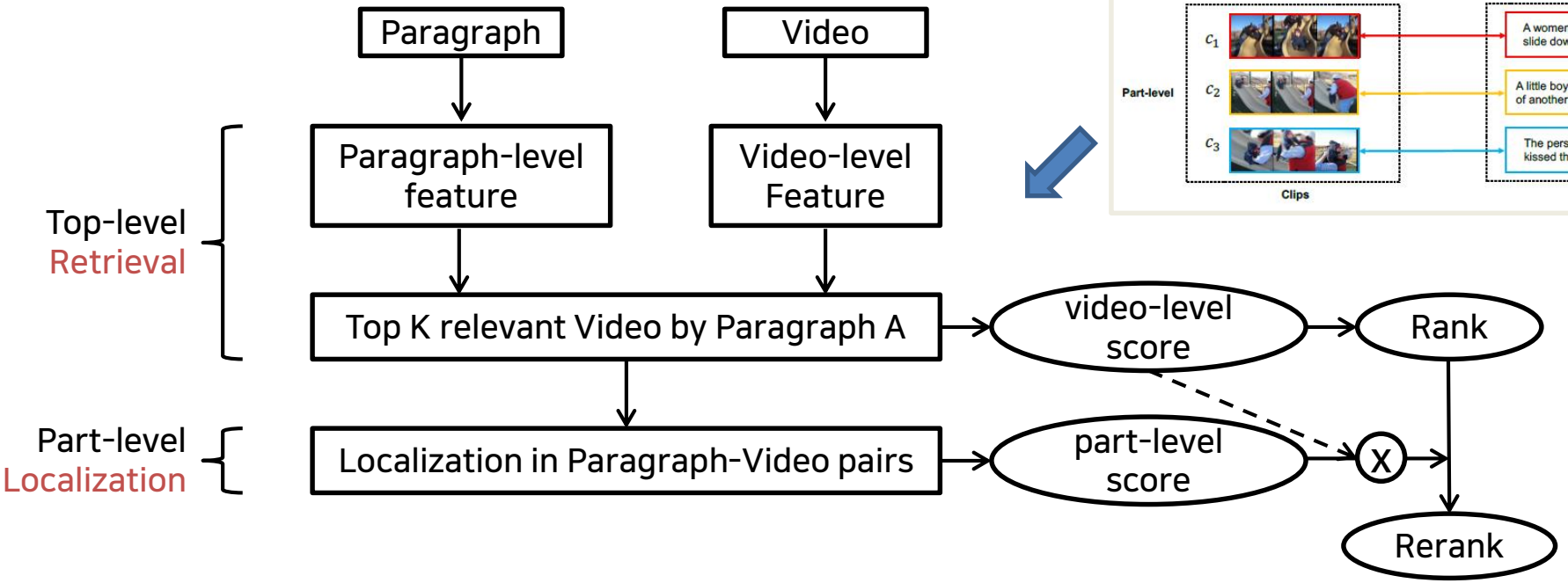
Retrieval is used to reduce the searching space, and refinement is used for re-ranking.

Overview: Find and Focus

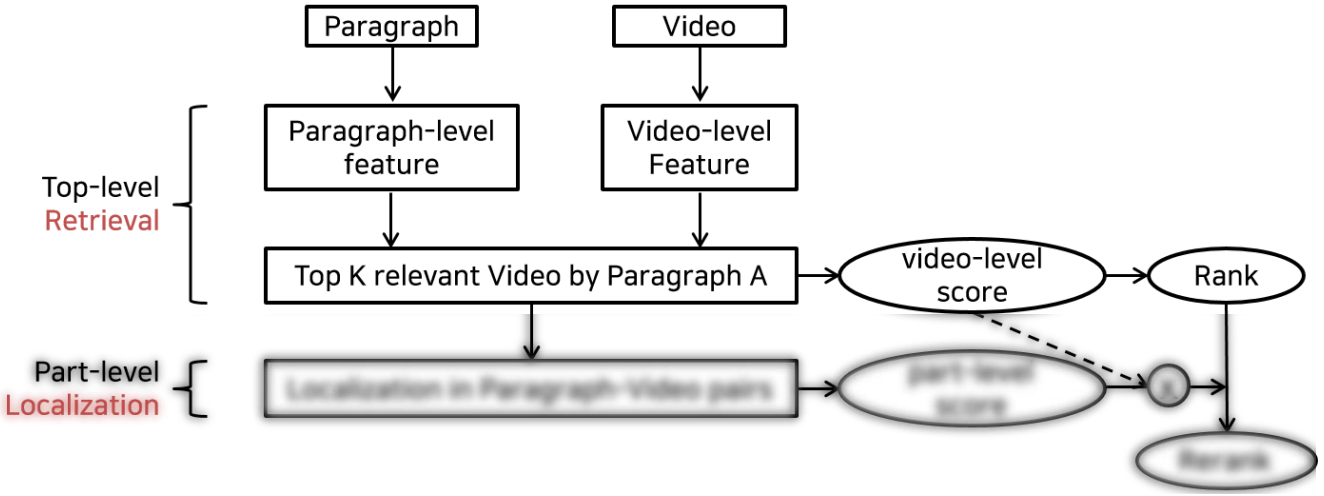
The details of Find and Focus framework

Video-level return is performed first,
and part-level refinement is performed for reranking in the generated pair to return.





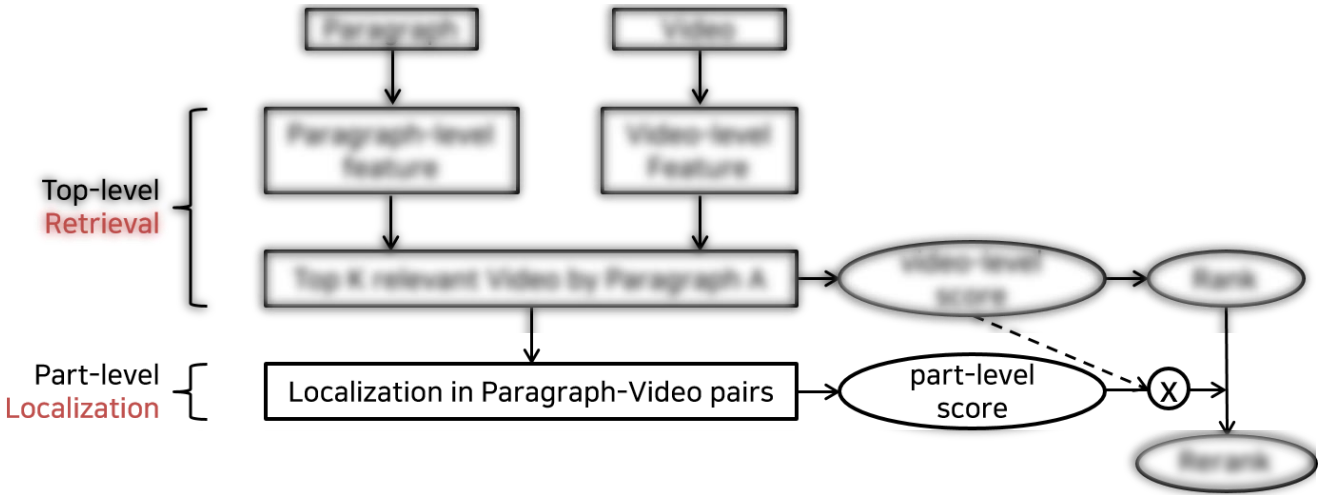
Find and Focus: Top-level



Top-level Matching (Retrieval)

- Describe top-level feature
  - ✓ Paragraph-level feature -> Bag of Word
  - ✓ Video-level feature -> VSE
- Projection to the common space
  - ✓ Not mentioned, maybe FC Layers
- Metric Learning with Margin Loss
- Calculate the top-level(video-level) score

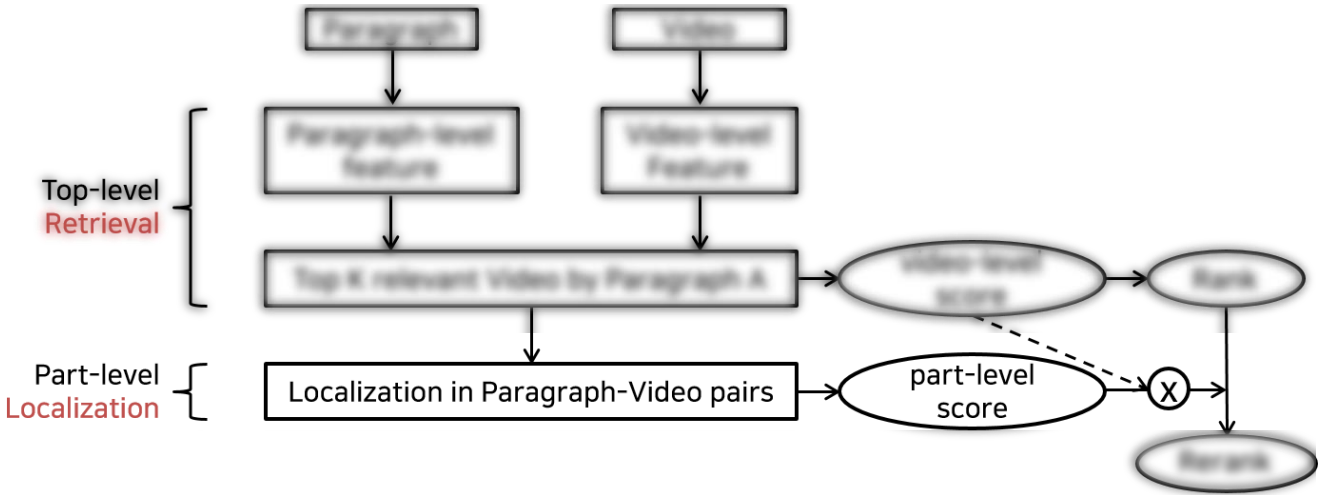
Find and Focus: Part-level



Part-level Matching (Localization)

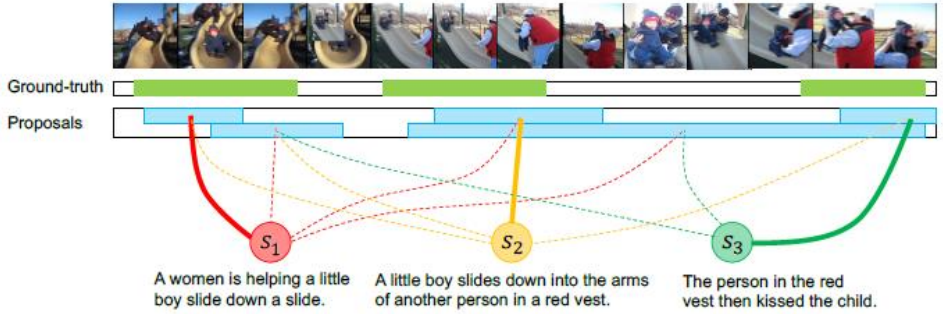
- Describe feature
  - ✓ $Paragraph = \{Sentence_i\}, Video = \{Clip_i\}$
  - ✓ Sentence-level feature -> Bag of Words
  - ✓ Clip-level feature -> TSN
- Projection to the common space
  - ✓ Not mentioned, maybe FC Layers
- Metric Learning with Margin Loss

Find and Focus: Part-level

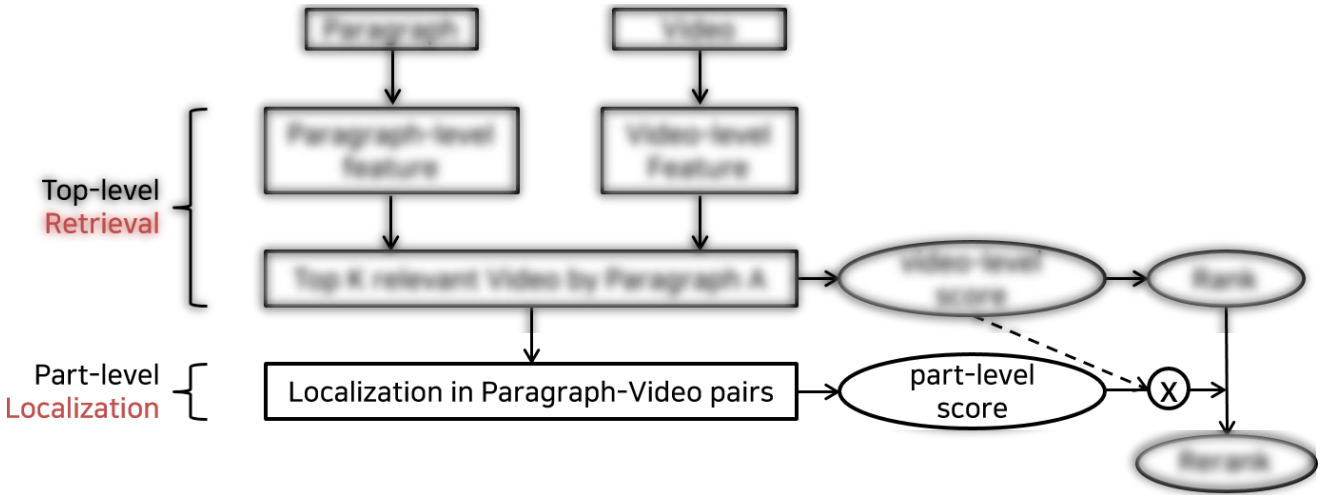


Part-level Matching (Localization)

- Clip Proposal
  - ✓ Match between sentence feature and clip feature in the paragraph and video that become pairs
  - ✓ Continuous clips matched to the same sentence are grouped to generate proposal



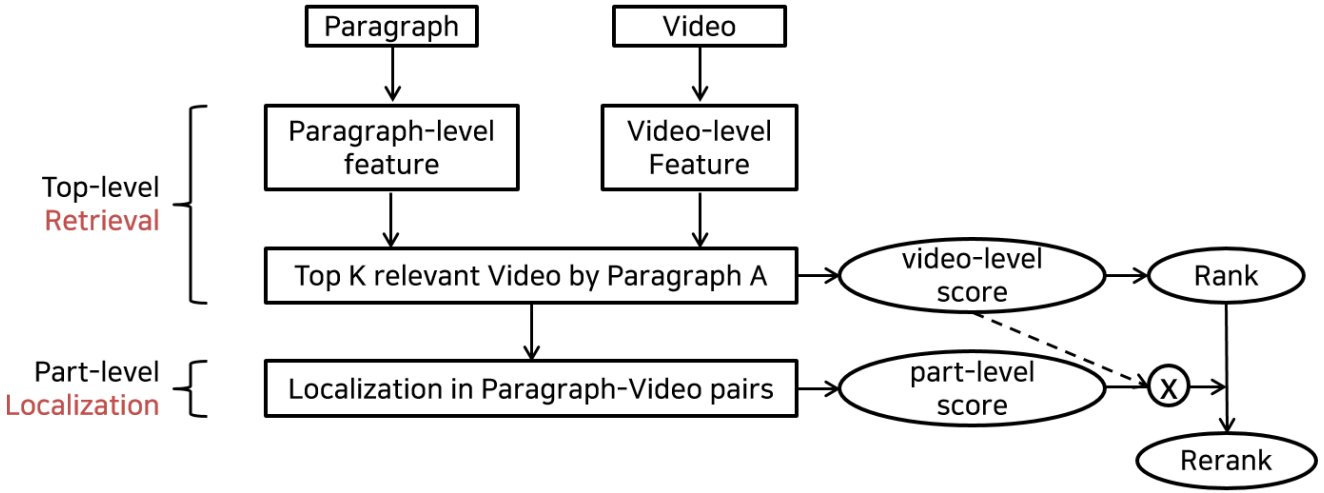
Find and Focus: Part-level



Part-level Matching (Localization)

- Cross-domainMatching
  - ✓ Using weighted cosine similarity to robust the similarity between proposal-sentences to noise.
  - ✓ The weight is calculated by cosine similarity between G, which averaged all clip features belonging to one proposal, and sensence features.
  - ✓ The sum of all proposal-sentences is selected as the part-level association score.

Find and Focus: Reranking



Reranking (Retrieval)

- Final relevance score based reranking
  - ✓ The final relevance score is composed of the product between the top-level score and the part-level score in the pair.

Experiments: Whole Video Retrieval

Find stage itself had higher performance than conventional methodologies,
and higher performance when applying focus stage refinement.

Table 1. Results for whole video retrieval on ActivityNet Captions.

	R@1	R@5	R@10	R@50	MedR
Random	0.02	0.10	0.20	1.02	2458
LSTM-YT [33]	0	4	-	24	102
S2VT [32]	5	14	-	32	78
Krishna *et al* [17]	14	32	-	65	34
VSE (Find)	11.69	34.66	50.03	85.66	10
Ours (Find + refine in Top 20)	**14.11**	37.12	52.13	-	10
Ours (Find + refine in Top 100)	14.05	**37.40**	**52.94**	**86.72**	**9**

Table 2. Results for whole video retrieval on modified LSMDC dataset.

	R@1	R@5	R@10	R@50	MedR
Random	0.20	1.02	2.04	10.22	244
VSE (Find)	2.66	10.63	16.36	52.97	45
Ours (Find + refine in Top 20)	**3.89**	**13.70**	20.04	-	45
Ours (Find + refine in Top 70)	**3.89**	13.50	**20.25**	**56.65**	**40**

Metric

- R@K
  - ✓ The number of GTs in K compared to the total number of GTs.
- MedR(Median Rank)
  - ✓ The median value of the rank to which GT is assigned.

Experiments: Proposal Generation and Clip Localization

The localization performance of the Focus stage itself is also higher than that of the previous methodology in this paper.

Table 3. Comparison of clip localization performance for different proposal methods.

ActivityNet, clip localization Recall@tIoU			
	Recall@0.3	Recall@0.5	Recall@0.7
SSN [43]	15.85	7.33	3.20
SSN [43]+shot [1]	16.71	8.74	4.30
Ours (VSS)	**28.52**	**13.46**	**5.21**

Metric

- Recall@tIoU
  - ✓ Calculate with TP when localized proposal overlaps GT and tIoU or more.

Experiments: Ablation & Qualitative Results

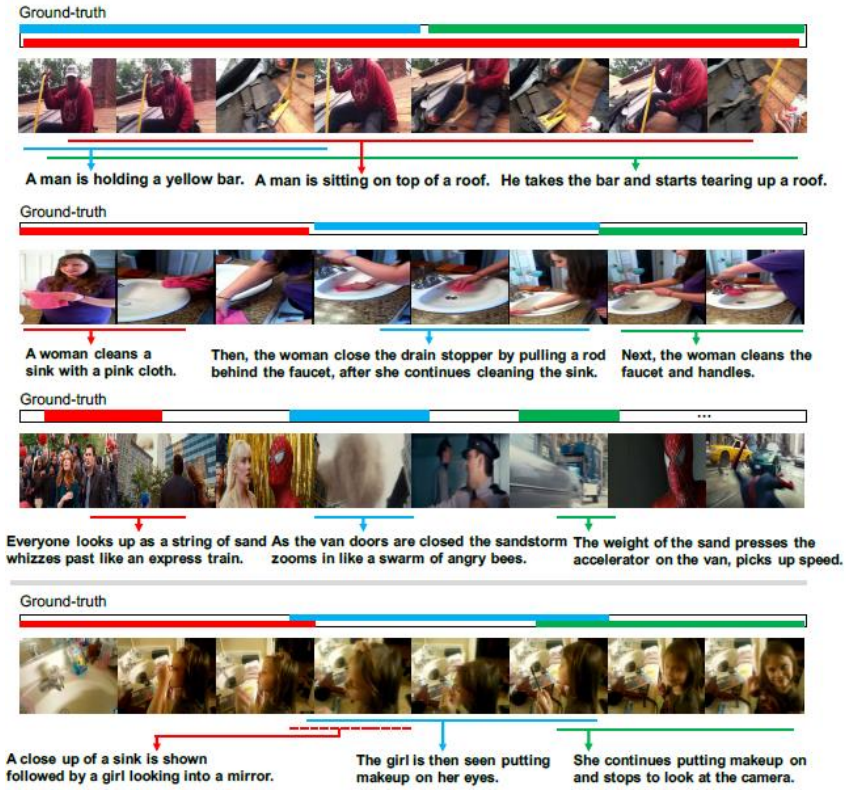
Ablation 1: Word Representation

		R@1	R@5	R@10	R@50	MedR
BoW with tf-idf	(Find)	11.69	34.66	50.03	85.66	10
	(Find + refine in Top 100)	14.05	37.40	52.94	86.72	9
BoW without tf-idf	(Find)	11.57	33.03	49.89	85.66	11
	(Find + refine in Top 100)	13.46	36.67	52.09	86.26	9
word2vec	(Find)	9.05	27.96	42.95	81.55	14
	(Find + refine in Top 100)	10.92	32.38	46.55	82.06	12
word2vec + Fisher Vec	(Find)	11.80	34.35	50.07	85.93	10
	(Find + refine in Top 100)	13.75	37.93	53.41	86.30	9

Ablation 2: K

	Recall@1	Recall@5	Recall@10	Recall@15	Recall@20	Recall@50
No Refinement	11.69	34.66	50.03	59.90	67.34	85.66
$K = 10$	13.93	36.65	-	-	-	-
$K = 20$	**14.11**	37.12	52.13	61.62	-	-
$K = 50$	14.05	37.40	52.90	**63.29**	70.53	-
$K = 100$	14.05	37.40	52.94	63.27	**70.75**	**86.72**
$K = 1000$	14.01	**37.44**	**53.06**	63.11	70.34	86.62

Qualitative Results



QnA