NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Video Corpus Moment Retrieval with Contrastive Learning

Hao Zhang[1,2], Aixin Sun[1], Wei Jing[3], Guoshun Nan[4],
Liangli Zhen[2], Joey Tianyi Zhou[2], Rich Siow Mong Goh[2]

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Institute of High Performance Computing, A*STAR, Singapore
[3]Institute of Infocomm Research, A*STAR, Singapore
[4]Singapore University of Technology and Design, Singapore

SIGIR 2021

Single Video Moment Retrieval (SVMR)

a.k.a., temporal sentence grounding in video

Inputs:

An untrimmed video + **a language query**

Outputs:

The target moment

Query: Rachel explains to her dad on the phone why she can't marry her fiancé.

Video:

00:44 00:54

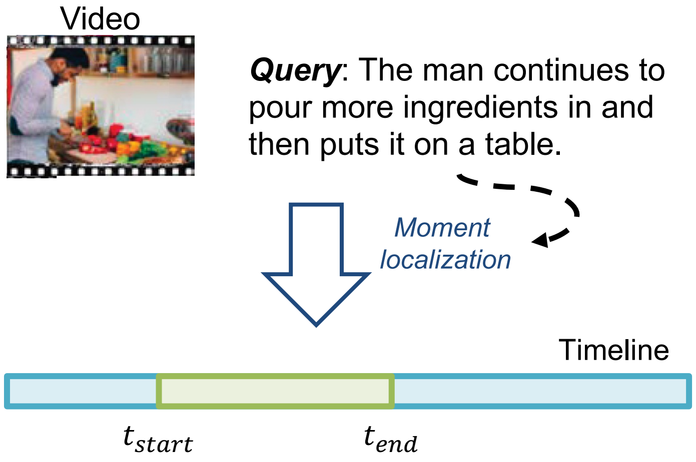
Target Moment

The example from TVRetrieval.

Video Corpus Moment Retrieval (VCMR)

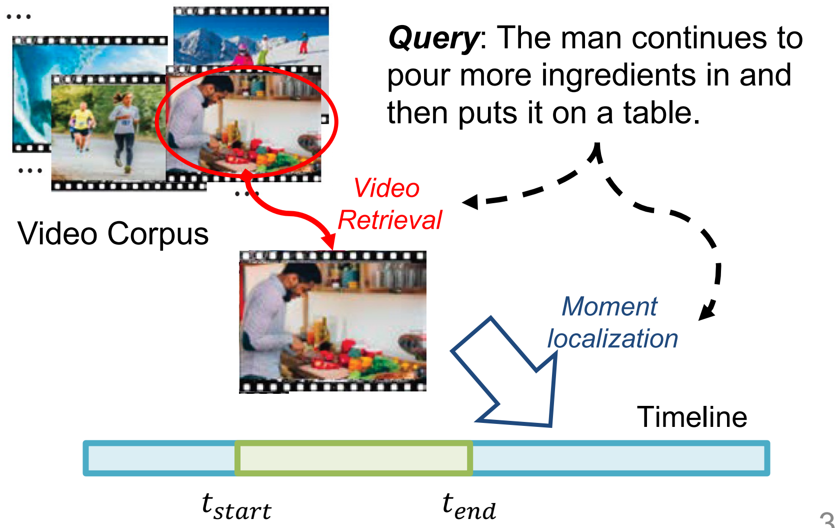
SVMR

Input: ***an untrimmed video***, language query
Output: target moment



VCMR

Input: ***video corpus with multiple videos***, language query
Output: target moment

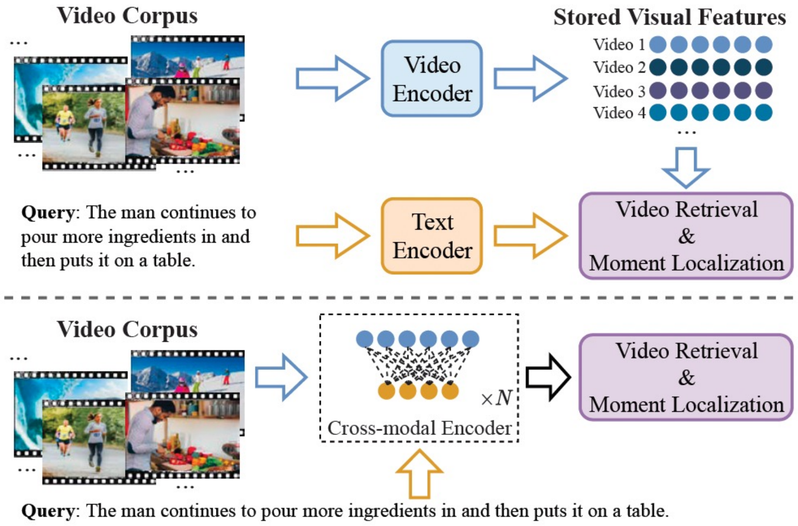


Existing VCMR Approaches

Video retrieval and **moment localization**

$$V^* = \arg\max_{V} p(V|Q) \quad \text{and} \quad \boldsymbol{m}^* \approx \arg\max_{\boldsymbol{m}\in V^*} p(\boldsymbol{m}|V^*, Q)$$

$\boldsymbol{V}^*$ denotes the target video
$\boldsymbol{m}^*$ is the target temporal moment.



NANYANG TECHNOLOGICAL UNIVERSITY | SINGAPORE

Existing VCMR Approaches

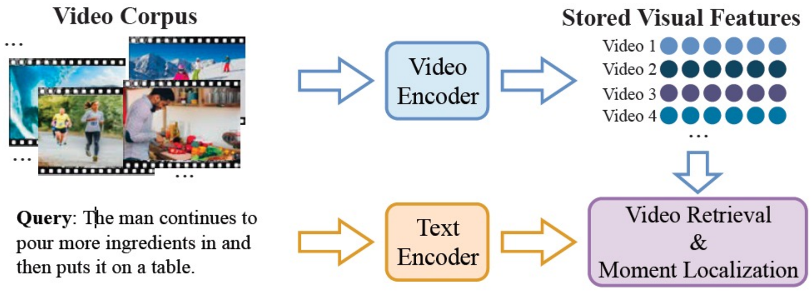
Unimodal Encoding Approach: to encode video and text **separately** and learn the matching through **late feature fusion**.

Pros:

High efficiency

Cons:

Low retrieval accuracy



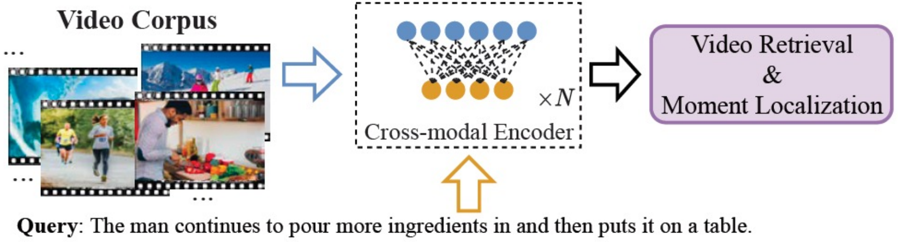
Cross-modal Encoding Approach: to **jointly** encode query words and video features by **cross-modal reasoning** at fine-grained granularity.

Pros:

High retrieval accuracy

Cons:

Low efficiency



Our Solution

Remedy the **contradiction** between *high efficiency* and *high-quality retrieval* in VCMR

To achieve the *pros of both unimodal and cross-modal encoding approaches*.

Key idea:

1. Adopt **unimodal encoding approach** to **keep** the *high efficiency*.
2. Adopt **contrastive learning** to simulate *cross-modal interaction* for *high-quality retrieval*.

Our Solution

Key idea:
1. Adopt **unimodal encoding approach** to **keep** the *high efficiency*.
2. Adopt **contrastive learning** to simulate *cross-modal interaction* for *high-quality retrieval*.

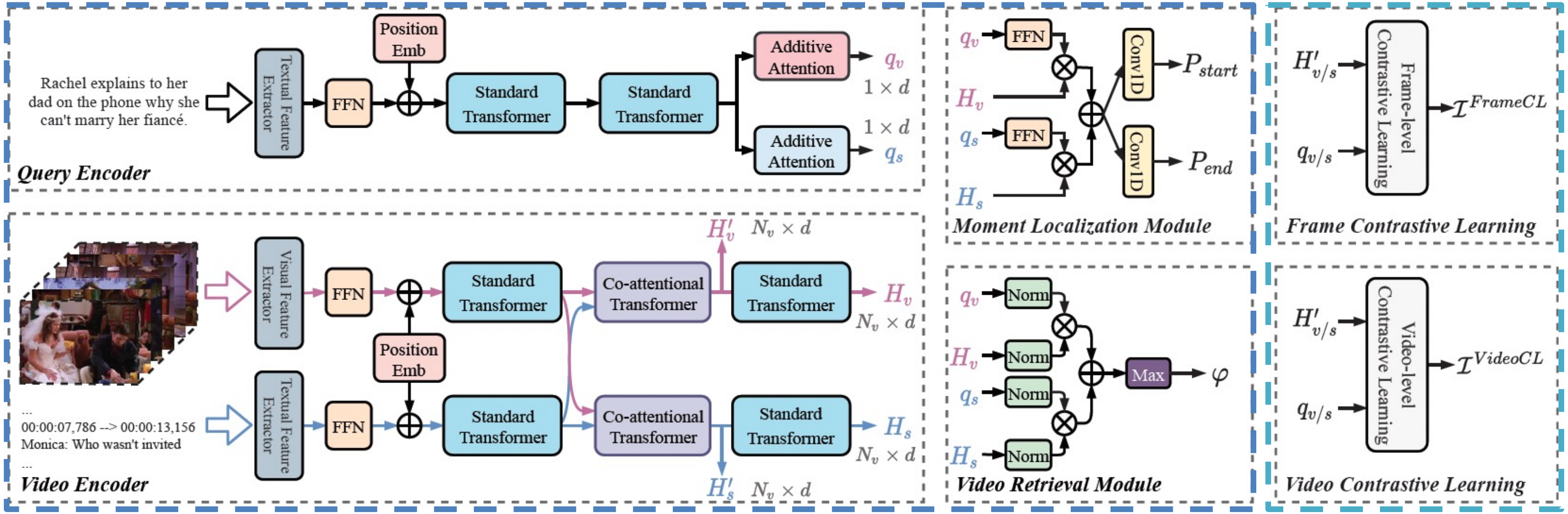
Cross-modal Interaction
It is to *highlight* the relevant and important information from both modalities through **co-attention mechanisms**.

Contrastive Learning
It is to *maximize* the mutual information (MI) of positive pairs and to *minimize* the MI of negative pairs.

A pair of matching video and query is a positive pair, and a non-matching pair is a negative pair in training.

The ***cross-modal interaction learning*** and ***contrastive learning*** share a similar objective of *emphasizing the relevant information of input pairs*.

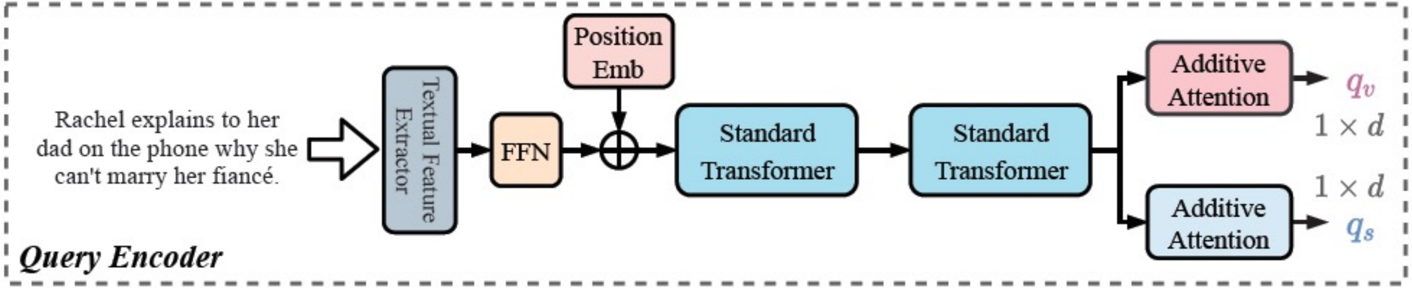
Retrieval and Localization Network with Contrastive Learning (ReLoCLNet)



Unimodal encoding baseline: ReLoNet

ReLoCLNet: ReLoNet + CL objectives

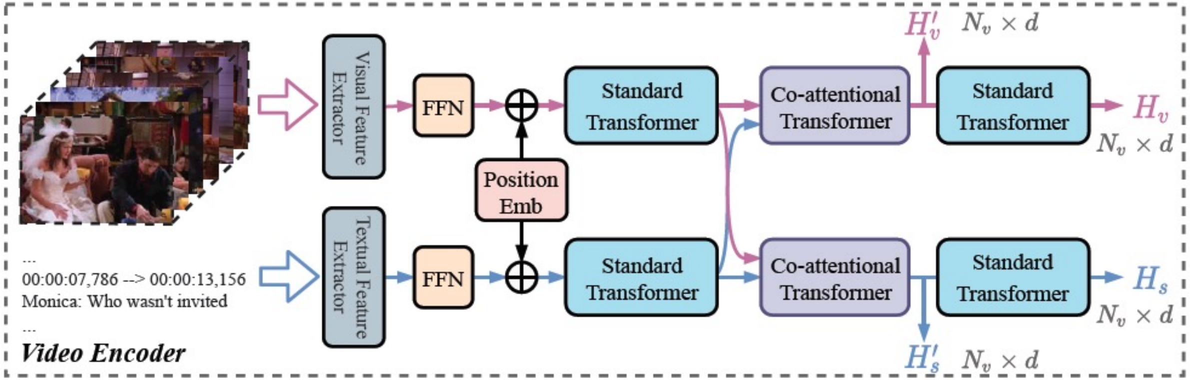
ReLoCLNet: Query Encoder



- The textural feature extractor can be pre-trained word embeddings, *e.g.*, GloVe, or language model, *e.g.*, RoBERTa.
- Two standard transformers is used to encode the contextual representations of query: $\widetilde{\boldsymbol{Q}} = \{\widetilde{\boldsymbol{q}}_i\}_{i=0}^{n_q-1} \in \mathbb{R}^{n_q \times d}, m \in \{v, s\}$.
- The additive attention strategy is applied to aggregate the information of $\widetilde{\boldsymbol{Q}}$ to generate modularized query vectors $\boldsymbol{q}_m \in \mathbb{R}^d$.

$$\boldsymbol{\alpha}^q = \mathtt{Softmax}(\boldsymbol{W}_{m,\alpha} \cdot \widetilde{\boldsymbol{Q}})) \in \mathbb{R}^{n_q}, \quad \boldsymbol{q}_m = \sum_{i=0}^{n_q-1} \alpha_i^q \times \boldsymbol{q}_i \in \mathbb{R}^d$$

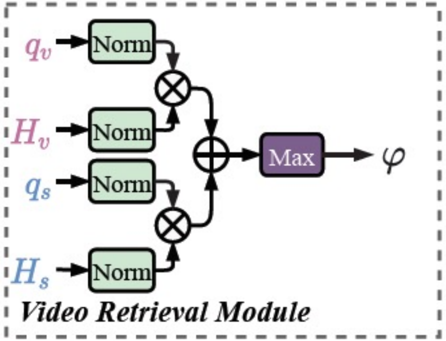
ReLoCLNet: Video Encoder



- Visual extractor can be pre-trained C3D or I3D model.
- Textual extractor is same as the query encoder.
- The visual and subtitle features are encoded parallelly.

- The co-attentional transformer encodes their cross-modal representations as $\boldsymbol{H}'_m = \{\boldsymbol{h}'_{m,i}\}_{i=0}^{n_v-1} \in \mathbb{R}^{n_v \times d}, m \in \{v, s\}$.
- The encoded cross-modal representations are refined by another transformer block: $\boldsymbol{H}_m = \{\boldsymbol{h}_{m,i}\}_{i=0}^{n_v-1} \in \mathbb{R}^{n_v \times d}$.

ReLoCLNet: Video Retrieval Module



- The video retrieval score is generated by computing the cosine similarities between $\boldsymbol{H}_m$ and $\boldsymbol{q}_m$.

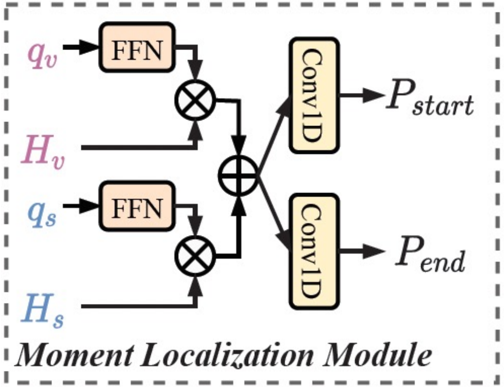
$$\boldsymbol{\varphi}_m = \mathrm{norm}(\boldsymbol{H}_m^\top) \cdot \mathrm{norm}(\boldsymbol{q}_m)$$

$$\varphi_m = \max(\boldsymbol{\varphi}_m) = \max([\varphi_m^0, \varphi_m^1, \dots, \varphi_m^{n_v-1}])$$

- The video retrieval is trained with ranking loss (hinge loss) as:

$$\mathcal{L}^{VR} = \max(0, \Delta + \frac{1}{N}\sum \boldsymbol{\varphi}' - \varphi) + \max(0, \Delta + \frac{1}{N}\sum \boldsymbol{\varphi}'' - \varphi)$$

ReLoCLNet: Moment Localization Module



- The video-query similarity scores in moment localization module is computed as:

$$\mathcal{S}_{mq} = H_m^{\top} \cdot \boldsymbol{q}'_m \in \mathbb{R}^{n_v}, \quad \text{where} \quad m \in \{v, s\} \ \boldsymbol{q}'_m = \boldsymbol{W}_m \cdot \boldsymbol{q}_m + \boldsymbol{b}_m \in \mathbb{R}^d$$

$$\mathcal{S} = \frac{1}{2}(\mathcal{S}_{vq} + \mathcal{S}_{sq})$$

- The video-query similarity scores $\mathcal{S} \in \mathbb{R}^{n_v}$. Then, the 1d convolutional layers are applied to compute the start and end boundaries scores.

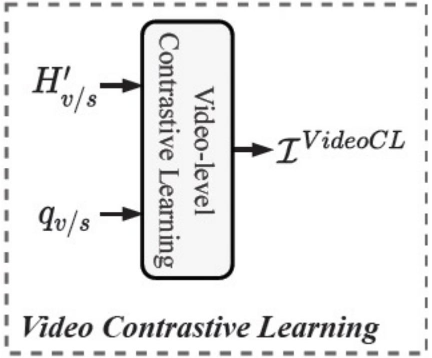
$$\mathcal{S}_{\text{start}} = \text{Conv1D}_{\text{start}}(\mathcal{S}), \quad \mathcal{S}_{\text{end}} = \text{Conv1D}_{\text{end}}(\mathcal{S})$$

- The moment localization is trained with cross-entropy loss:

$$P_{\text{start}} = \text{Softmax}(\mathcal{S}_{\text{start}}), \quad P_{\text{end}} = \text{Softmax}(\mathcal{S}_{\text{end}})$$

$$\mathcal{L}^{ML} = \frac{1}{2} \times \left(f_{\text{XE}}(P_{\text{start}}, Y_{\text{start}}) + f_{\text{XE}}(P_{\text{end}}, Y_{\text{end}}) \right)$$

ReLoCLNet: Video Contrastive Learning (VideoCL) Module



VideoCL aims to learn a joint feature space:

1. **semantically related videos and queries are close to each other**
2. **far away otherwise**.

- Compute the modularized representation of latent representations H'_m as:

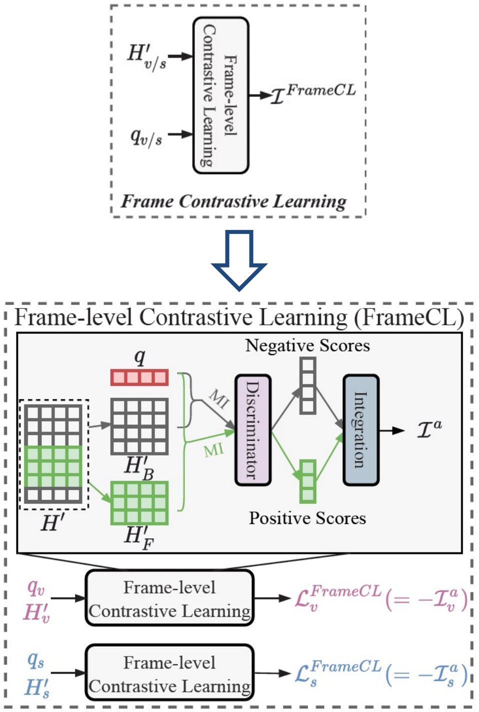
$$\boldsymbol{\alpha}^m = \texttt{Softmax}(\boldsymbol{W}_{m,\alpha} \cdot \boldsymbol{H}'_m) \in \mathbb{R}^{n_v}, \quad \boldsymbol{c}_m = \sum_{i=0}^{n_v-1} \alpha_i^m \times \boldsymbol{h}'_{m,i}$$

- Let $\mathcal{P} = \{(\boldsymbol{c}_m, \boldsymbol{q}_m)\}$ as positive pairs and $\mathcal{N} = \{(\boldsymbol{c}'_m, \boldsymbol{q}'_m)\}$ as negatives

$$\mathcal{I}_m^e = \log\left(\frac{\sum_{(\boldsymbol{c}_m,\boldsymbol{q}_m)\in\mathcal{P}} e^{f(\boldsymbol{c}_m)^\top \cdot g(\boldsymbol{q}_m)}}{\sum_{(\boldsymbol{c}_m,\boldsymbol{q}_m)\in\mathcal{P}} e^{f(\boldsymbol{c}_m)^\top \cdot g(\boldsymbol{q}_m)} + \sum_{(\boldsymbol{c}'_m,\boldsymbol{q}'_m)\sim\mathcal{N}} e^{f(\boldsymbol{c}'_m)^\top \cdot g(\boldsymbol{q}'_m)}}\right)$$

$$\mathcal{I}^e = \frac{1}{2}(\mathcal{I}_v^e + \mathcal{I}_s^e) \qquad \mathcal{L}^{VideoCL} = -\mathcal{I}^e$$

ReLoCLNet: Frame Contrastive Learning (FrameCL) Module



- FrameCL focuses on moment localization within a given video-query pair.
- The video features that reside **within boundaries of target moment as positive samples**, and **the rest as negatives**.

- The contrastive loss is computed by measuring MI between the **query** and the **positive**/**negative** visual features:

Positive: $\boldsymbol{H}'_{m,F} = \{\boldsymbol{h}'_{m,i} | i = i^s, \ldots, i^e\} \in \mathbb{R}^{d \times n_t}$

Negative: $\boldsymbol{H}'_{m,B} = \{\boldsymbol{h}'_{m,i} | i = 0, \ldots, i^s - 1, i^e + 1, \ldots, n_v - 1\} \in \mathbb{R}^{d \times (n_v - n_t)}$

$$\mathcal{I}^a_m = \mathbb{E}_{\boldsymbol{H}'_{m,F}}\Big[-\mathrm{sp}\big(-C_\theta(\boldsymbol{q}, \boldsymbol{H}'_{m,F})\big)\Big] - \mathbb{E}_{\boldsymbol{H}'_{m,B}}\Big[\mathrm{sp}\big(C_\theta(\boldsymbol{q}, \boldsymbol{H}'_{m,B})\big)\Big]$$

$$\mathcal{I}^a = \frac{1}{2}(\mathcal{I}^a_v + \mathcal{I}^a_s) \qquad \mathcal{L}^{FrameCL} = -\mathcal{I}^a$$

Experiments

Datasets:

- TVR dataset
- ActivityNet Captions (ANetCaps) dataset

Metric

- Recall@k, where $k \in \{1, 5, 10, 100\}$.
- Recall@k, IoU=μ, where $k \in \{1, 10, 100\}$ and $\mu \in \{0.5, 0.7\}$.

The definition of a **correct** prediction by VCMR model is that:

i. The predicted video matches the ground truth (GT) video;
ii. The predicted moment within the video has **high overlap** with the GT moment.

(The overlap is measured by temporal *Intersection over Union*, *IoU*)

Experiments

Comparison of the VCMR results on
TVR and ANetCaps datasets

Dataset	Method	Recall@k, IoU = 0.5			Recall@k, IoU = 0.7		
		R1	R10	R100	R1	R10	R100
TVR	XML [37]	-	-	-	2.62	9.05	22.47
	HERO [38]	-	-	-	2.98	10.65	18.25
	FLAT [78]	*8.45*	21.14	30.75	*4.61*	11.29	16.24
	HAMMER [78]	**9.19**	*21.28*	31.25	**5.13**	*11.38*	16.71
	ReLoNet	5.46	16.65	*35.08*	2.71	9.37	*22.87*
	ReLoCLNet	8.03	**21.37**	**44.10**	4.15	**14.06**	**32.42**
ANetCaps	MCN [30]	0.02	0.18	1.26	0.01	0.09	0.70
	CAL [16]	0.21	1.32	6.82	0.12	0.89	4.79
	FLAT [78]	2.57	*13.07*	*30.66*	1.51	*7.69*	17.67
	HAMMER [78]	*2.94*	**14.49**	**32.49**	*1.74*	**8.75**	**19.08**
	ReLoNet	2.16	9.96	24.54	1.26	5.64	17.43
	ReLoCLNet	**3.09**	11.28	25.95	**1.82**	6.91	*18.33*

XML, HERO: unimodal encoding approaches.
FLAT, HAMMER: cross-modal encoding approaches.
MCN, CAL: ranking-based approaches.

ReLoNet is comparable to the unimodal encoding approaches, XML and HERO.

ReLoCLNet surpasses the unimodal encoding approaches significantly, while achieves comparable performance to the cross-modal encoding methods, HAMMER.

Experiments

Comparison of Retrieval efficiency on TVR dataset

Method	Retrieval Efficiency	
	Total Time	Average Per Query
XML [37]	39.34 seconds	3.61 milliseconds
HAMMER [78]	2, 378.67 seconds	218.33 milliseconds
ReLoNet ReLoCLNet	42.07 seconds	3.86 milliseconds

The time spent on data pre-processing and feature extraction by pre-trained extractor are not counted since the same process applies to all methods.

The retrieval efficiency of ReLoNet and ReLoCLNet are **comparable** to XML, *i.e.*, unimodal encoding approach.

ReLoNet and ReLoCLNet are **far more efficient** than HAMMER, *i.e.*, cross-modal encoding approach.

Experiments

Results of ***VR* subtask** on TVR and ANetCaps

Dataset	Method	Recall@k			
		$k = 1$	$k = 5$	$k = 10$	$k = 100$
TVR	MCN [30]	0.05	0.38	0.66	3.59
	CAL [16]	0.28	1.02	1.68	8.55
	MEE [48]	7.56	20.78	29.88	73.07
	XML [37]	16.54	38.11	50.41	88.22
	ReLoNet	*16.96*	*39.28*	*51.34*	*88.46*
	ReLoCLNet	**22.13**	**45.85**	**57.25**	**90.21**
ANetCaps	FLAT [78]	5.37	-	29.14	71.64
	HAMMER [78]	5.89	-	30.98	73.38
	ReLoNet	*7.02*	*24.42*	*35.24*	*78.08*
	ReLoCLNet	**9.64**	**28.02**	**40.26**	**79.13**

Results of ***SVMR* subtask** on TVR and ANetCaps

Dataset	Method	Recall@1, IoU = μ		
		$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.7$
TVR	MCN [30]	-	13.08	5.06
	CAL [16]	-	12.07	4.68
	ExCL [20]	-	*31.34*	*14.19*
	XML [37]	-	30.75	13.41
	ReLoNet	*48.14*	29.49	13.13
	ReLoCLNet	**49.87**	**31.88**	**15.04**
ANetCaps	FLAT [78]	*57.58*	*39.60*	*22.59*
	HAMMER [78]	**59.18**	**41.45**	**24.27**
	ReLoNet	39.27	23.67	14.55
	ReLoCLNet	42.65	28.54	17.76

Experiments

The effects of different objectives on TVR dataset (VR=Video Retrieval, ML=Moment Localization, VideoCL=Video Contrastive Learning, and FrameCL=Frame Contrastive Learning)

Objective				VCMR						VR			SVMR					
				Recall@k, IoU=0.5			Recall@k, IoU=0.7			Recall@k			Recall@k, IoU=0.5			Recall@k, IoU=0.7		
VR	ML	VideoCL	FrameCL	1	10	100	1	10	100	1	10	100	1	10	100	1	10	100
✔	✘	✘	✘	-	-	-	-	-	-	16.23	49.33	87.38	-	-	-	-	-	-
✘	✔	✘	✘	-	-	-	-	-	-	-	-	-	30.21	59.81	83.43	13.91	41.55	68.51
✔	✔	✘	✘	5.46	16.65	35.08	2.71	9.37	22.87	16.96	51.34	88.46	29.49	54.06	75.89	13.13	35.46	58.84
✔	✔	✔	✘	6.63	18.16	39.69	3.24	11.78	27.69	*20.69*	*55.70*	*89.71*	29.52	57.32	78.65	13.76	38.26	64.27
✔	✔	✘	✔	*7.21*	*20.04*	*42.45*	*3.75*	*12.77*	*30.32*	19.81	54.38	88.96	*31.75*	*62.20*	*85.99*	*14.73*	*44.60*	*71.44*
✔	✔	✔	✔	**8.03**	**21.37**	**44.10**	**4.15**	**14.06**	**32.42**	**22.13**	**57.25**	**90.21**	**31.88**	**63.89**	**86.67**	**15.04**	**45.24**	**72.12**

VideoCL contributes to performance improvements on both VCMR and VR, while it achieves marginal improvements on SVMR. VideoCL is in line with video retrieval objective.

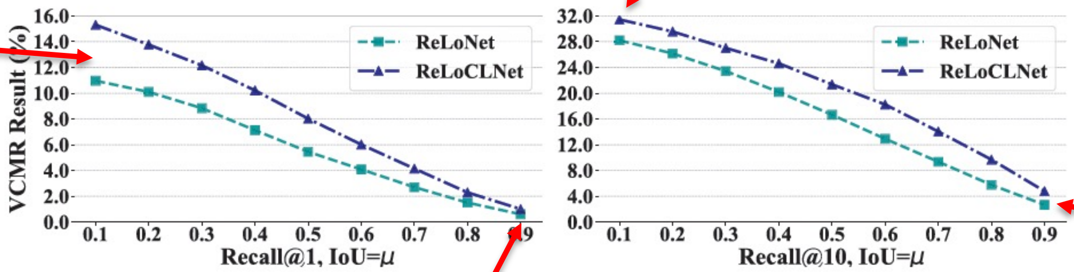
FrameCL contributes to all three tasks. FrameCL guides the model to search for boundaries of target moment for precise moment localization.

Experiments

31.44 vs. 28.21

Recall@1 and Recall@10 of VCMR on TVR over different IoU thresholds.

15.31 vs. 10.97

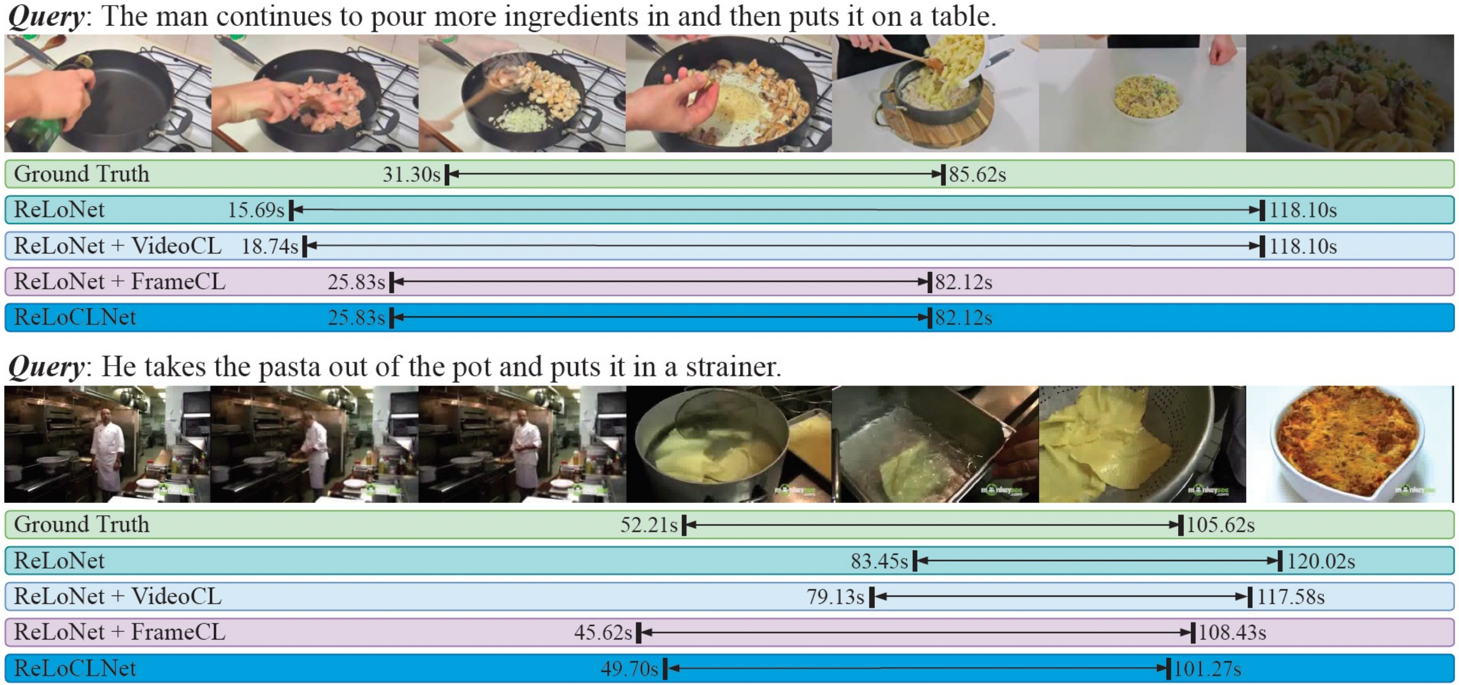


4.80 vs. 2.68

1.01 vs. 0.61

Experiments

Visualization



Conclusion

- We analyze two common approaches for VCMR task and study their pros and cons.
- We propose a Retrieval and Localization Network with Contrastive Learning (ReLoCLNet) for video corpus moment retrieval (VCMR) task.
- We introduce two contrastive learning objectives (VideoCL and FrameCL) on top of a unimodal encoding approach to address the contradiction between retrieval efficiency and retrieval quality.
- Extensive experimental studies show that ReLoCLNet addresses VCMR with high efficiency, and its retrieval accuracy is comparable with state-of-the-art methods which are much costly in terms of computation.

Thank You!