

Goal: Automatically detecting the video segments semantically relevant to the language description from untrimmed video, without temporal boundary annotation.

- Drawbacks of existing methods
- (1) Negative samples from other videos
- (2) Data-independent proposal generation procedure
- Advantages of our method
- (1) Higher quality proposals
- (2) Stronger ability to distinguish confusing scenes

Method Overview								
Masked Query: Word $\widehat{W} \in \mathbb{R}^{M \times D_W}$ Mask θ_2 Reconditioned Person <mask> off. Embedding Conditioned Reconditioned Attention Attention <math>Methods Mask θ_2 Reconditioned Mask off. Mask θ_2 Reconditioned Attention Attention <math>Methods <math>Methods Reconditioned Mask off. Methods <math>Methods <math>Methods Reconditioned Mask off. <math>Methods $$</math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></mask>	structed query Target: Person takes their Shoes off. Loss							
(a) Overall Framework	(b) Mask Conditioned Attention							

Weakly Supervised Video Moment Localization with Contrastive Negative Sample Mining

Minghang Zheng¹, Yanjie Huang², Qingchao Chen³, Yang Liu^{1,4}

¹Wangxuan Institute of Computer Technology, Peking University, ²School of Integrated Circuits and Electronics, Beijing Institute of Technology, ³National Institute of Health Data Science, Peking University ⁴Beijing Institute for General Artificial Intelligence

Mask Generator

Aim: Generate high-quality content-based proposals Feature Extraction

Word embedding: Glove

Video encoder: CLIP or I3D

Mask Generation

(1) Obtain fused feature *H* (2) Predict Gaussian center c and width w through h_N in HThe positive Gaussian mask m^p :

$$m_i^p = \exp(-\frac{\alpha(i/N-c)^2}{w^2}), i = 1, ..., N$$

Negative Sample Mining Aim: Enable our model to distinguish highly confusing scenes (1) Easy negatives: Frames suppressed by m^p (2) Hard negatives: The entire video

Mask Conditioned Reconstructor

Aim: Reconstruct query conditioned on arbitrary sample masks Mask Conditioned Attention Aggregated context information:

 $E_m(V,m) = \mathsf{Softmax}(A \otimes m) V_a \in \mathbb{R}^{N \times D_H}$

Mask Conditioned Semantic Completion

Aim: Segments highlighted by positive mask reconstruct the query better (1) Mask words in query

(2) Reconstruct the original query

(3) Calculate the difference between probability and real distribution with cro entropy loss

(4)The final reconstruction loss \mathcal{L}_{rec} :

$$\mathcal{L}_{rec} = \mathcal{L}_{ce}^p + \mathcal{L}_{ce}^h$$

where \mathcal{L}_{ce}^{p} and \mathcal{L}_{ce}^{h} means the cross-entropy loss mentioned above conditioned m^p and m^h .

Intra-Video Contrastive

Intra-Video Contrastive loss L_{IVC} :

 $\mathcal{L}_{IVC} = \max(\mathcal{L}_{ce}^p - \mathcal{L}_{ce}^h + \beta_1, 0) + \max(\mathcal{L}_{ce}^p - \mathcal{L}_{ce}^e + \beta_2, 0)$

Experiments

Table 1. ActivityNet Captions

Table 2. Charades-STA

Method	loU=0.1	Recall IoU=0.3	loU=0.5	•	Method	loU=0.3	Recall IoU=0.5	IoU=0.7
Random	38.23	18.64	7.63	-	Random	20.12	8.61	3.39
WS-DEC	62.71	41.98	23.34		TGA	32.14	19.94	8.84
EC-SL	68.48	44.29	24.16		WSTG	39.8	27.3	12.9
MARN	-	47.01	29.95		SCN	42.96	23.58	9.97
SCN	71.48	47.23	29.22		WSTAN	43.39	29.35	12.28
RTBPN	73.73	49.77	29.63		VLANet	45.24	31.83	14.17
WSLLN	75.4	42.8	22.7		LoGAN	48.04	31.74	13.71
LCNet	78.58	48.49	26.33		MARN	48.55	31.94	14.81
WSTAN	79.78	52.45	30.01		CRM	53.66	34.76	<u>16.37</u>
CRM	81.61	55.26	32.19		LCNet	59.60	39.19	18.87
	70.40				RTBPN	<u>60.04</u>	32.36	13.24
CINIM (ours)	/8.13	55.68	33.33		CNM (ours)	60.39	35.43	1545
						00.07	JJ. T J	тр.т.



Ì



		Ablation	Studies				
(1) Effect of	components						
Та	ble 3. Mask Generate	or	Table 4. Int	ra-Video	Negative	e Sample	e M
Method	Recall		Hard Facy		Reca	all	
Full Model	10U=0.1 IoU=0.3 IoU	1=0.5 mloU 3 33 37 14		loU=0.1	IoU=0.3	IoU=0.5	5 m
w/o. Mask	79.35 47.71 26	5.98 34.73	\mathbf{x}	78.13	55.68 55.67	33.33 31.40	37 36
			√ × × ×	80.99	55.19 40.26	30.94 24.93	36
	-			02.27	-10.20	24.73	20
(2) Effect of	Iraining Strategy						
		Table 5. Train	ing Strategy				
	Mathad		Re	call			
		loU=	0.1 IoU=0.	3 IoU=(0.5 mlc	U	
	$\min_{\theta_1} \mathcal{L}_{IVC} + \mathbf{m}$	$\operatorname{in}_{\theta_2} \mathcal{L}_{rec} \mid 78.$	13 55.68	33.3	3 37.	14	
	$\lim_{\theta_1,\theta_2} (\mathcal{L}_{IVC})$	$+ \mathcal{L}_{rec}) \mid $ 63.	J7 43.8U	24.5	0 28.9	70	
	Qualitative E			(Nlat			
	Gualitative E	xamples o	MACUVIL	ynet (Japtic	ons	
	Query: She laughs	s and continues	to brush her	teeth.			
		15.18s G	70.59s			_	
	2 12¢	SCN	63 88s				
	2.12s	SCN 16.59s O	63.88s urs 70.59s				
	2.12s Ouerv: He blows i	SCN 16.59s O (a) into the harmon	63.88s urs 70.59s nica and star	ts to nla	v it.		
	2.12s Query: He blows i	SCN 16.59s O (a) into the harmon	63.88s urs 70.59s nica and star	ts to pla	y it.		
	2.12s Query: He blows i	SCN 16.59s O (a) into the harmon 16.59s O (a) (a) 37.37s	63.88s urs 70.59s nica and star	ts to play	y it. 162.22s	5	
	2.12s Query: He blows i Second	SCN 16.59s O (a) into the harmon 37.37s	63.88s urs 70.59s nica and star GT GT SCN	ts to play	y it. 162.22s 160.60s 162.22s		
	2.12s Query: He blows f Is.66s	SCN 16.59s O (a) into the harmon 37.37s (b)	63.88s urs 70.59s nica and star GT GT SCN Ours	ts to play	y it. 162.22s 160.60s 162.22s		
	2.12s Query: He blows i Is.66s Query: Once com	SCN 16.59s O (a) into the harmon 37.37s (b) aplete, she jum ful as the crow	63.88s urs 70.59s nica and star GT GT SCN Ours ps up and do	ts to play	y it. 162.22s 160.60s 162.22s py that l	her	
	2.12s Query: He blows f isofo 18.66s Query: Once com jump was success	SCN 16.59s O (a) into the harmon 37.37s (b) plete, she jum ful as the crow	63.88s urs 70.59s nica and star GT SCN Ours ps up and do d begins to c	ts to play	y it. 162.22s 160.60s 162.22s py that l her.	her	
	2.12s Query: He blows i Source I B.66s Query: Once cons jump was success	SCN 16.59s O (a) into the harmon 37.37s (b) aplete, she jum ful as the crow	63.88s urs 70.59s nica and star GT GT SCN Ours ps up and do d begins to c	ts to play	y it. 162.22s 160.60s 162.22s py that l her. GT 16.9	her 93s	
	2.12s Query: He blows f Source of 18.66s Query: Once cons jump was success	SCN 16.59s O (a) into the harmon into the harmon $into the harmon into the harmoninto the harmon into the harmon into the harmoninto the harmon into the harmon into the harmoninto the harmon into the harmoninto the harmoninto the harmon into the harmoninto the harm$	63.88s urs 70.59s nica and star	ts to play	y it. 162.22s 160.60s 162.22s py that l her. GT 16.9 16.9	her 93s	

- Generate Gaussian mask as the positive sample
- Mine the hard and easy negative samples within the same video
- Experiments and ablation studies demonstrate our advantages

Acknowledgement

Supported by: State Key Laboratory of Media Convergence Production Technology and Systems, National Engineering Laboratory of big data analysis and application technology



