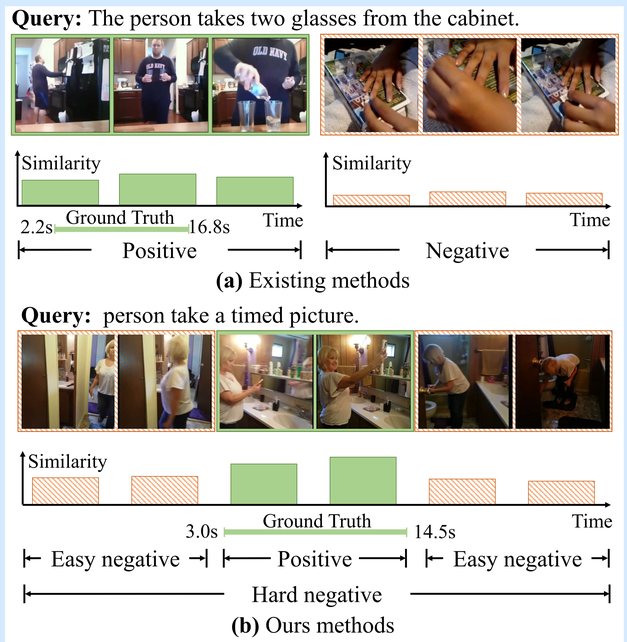
Weakly Supervised Video Moment Localization with Contrastive Negative Sample Mining

Minghang Zheng[1], **Yanjie Huang**[2], **Qingchao Chen**[3], **Yang Liu**[1,4]

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

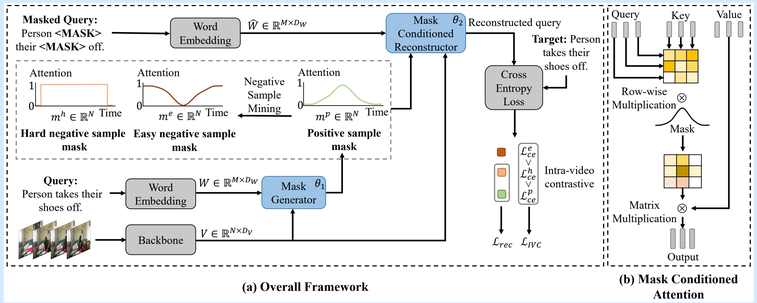
Introduction



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



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 H
The 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:

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

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 cross-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 on 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

Method	Recall IoU=0.1	IoU=0.3	IoU=0.5
Random	38.23	18.64	7.63
WS-DEC	62.71	41.98	23.34
EC-SL	68.48	44.29	24.16
MARN	-	47.01	29.95
SCN	71.48	47.23	29.22
RTBPN	73.73	49.77	29.63
WSLLN	75.4	42.8	22.7
LCNet	78.58	48.49	26.33
WSTAN	79.78	52.45	30.01
CRM	**81.61**	55.26	32.19
CNM (ours)	78.13	**55.68**	**33.33**

Table 2. Charades-STA

Method	Recall IoU=0.3	IoU=0.5	IoU=0.7
Random	20.12	8.61	3.39
TGA	32.14	19.94	8.84
WSTG	39.8	27.3	12.9
SCN	42.96	23.58	9.97
WSTAN	43.39	29.35	12.28
VLANet	45.24	31.83	14.17
LoGAN	48.04	31.74	13.71
MARN	48.55	31.94	14.81
CRM	53.66	34.76	16.37
LCNet	59.60	**39.19**	**18.87**
RTBPN	60.04	32.36	13.24
CNM (ours)	**60.39**	35.43	15.45

Ablation Studies

(1) Effect of components

Table 3. Mask Generator

Method	Recall IoU=0.1	IoU=0.3	IoU=0.5	mIoU
Full Model	78.13	**55.68**	**33.33**	**37.14**
w/o. Mask	**79.35**	47.71	26.98	34.73

Table 4. Intra-Video Negative Sample Mining

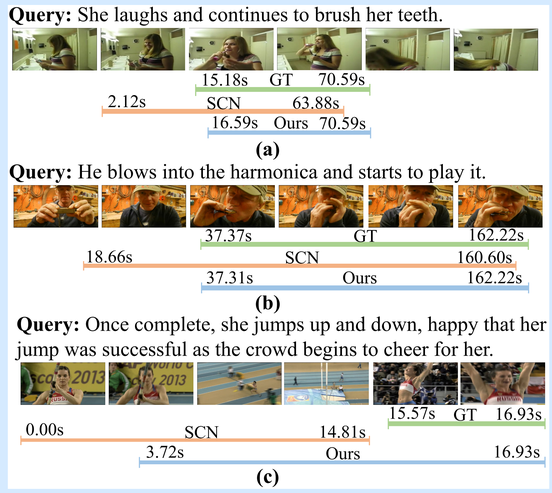
Hard	Easy	Recall IoU=0.1	IoU=0.3	IoU=0.5	mIoU
✓	✓	78.13	**55.68**	**33.33**	**37.14**
✗	✓	80.60	55.67	31.40	36.79
✓	✗	**80.99**	55.19	30.94	36.95
✗	✗	62.27	40.26	24.93	28.55

(2) Effect of Training Strategy

Table 5. Training Strategy

Method	Recall IoU=0.1	IoU=0.3	IoU=0.5	mIoU
$\min_{\theta_1} \mathcal{L}_{IVC} + \min_{\theta_2} \mathcal{L}_{rec}$	**78.13**	**55.68**	**33.33**	**37.14**
$\min_{\theta_1,\theta_2}(\mathcal{L}_{IVC} + \mathcal{L}_{rec})$	63.59	43.80	24.50	28.96

Qualitative Examples on ActivityNet Captions



Conclusion

- 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