ETRI Electronics and Telecommunications Research Institute

UST UNIVERSITY OF SCIENCE & TECHNOLOGY

BMRN: Boundary Matching and Refinement Network for Temporal Moment Localization

Muah Seol[1,2] Jonghee Kim[1] Jinyoung Moon[1,2]

[1]ETRI, [2]UST

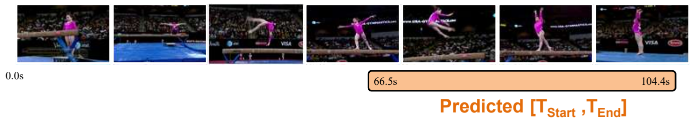
O-DRUM @ JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Background

1. Task: Temporal Moment Localization with Natural Language (**TML**)
(also called **Temporal Sentence Grounding** (**TSG**)

Query: She jumps and flips herself around and ends by jumping down with her arms up.

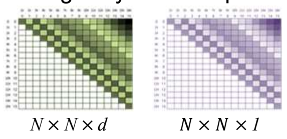
Input Video



Predicted [T_{Start} ,T_{End}]

2. 2D Map Representations for Proposal Features and Scores

- Originally from temporal adjacent maps in 2D-TAN [1]



$[m\cdot\tau,\ (n+1)\cdot\tau],\ 0 \le m \le n \le N-1$

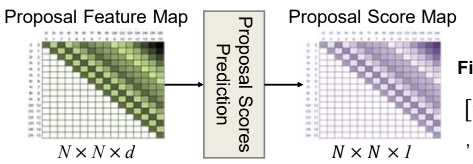
top-K moment proposals with the highest proposal scores, which are not highly intersected between them through **NMS**

Motivation and Contribution

1. Boundary Matchingg and Refinement for

- **Variable Boundary** Proposals in 2D map based approaches

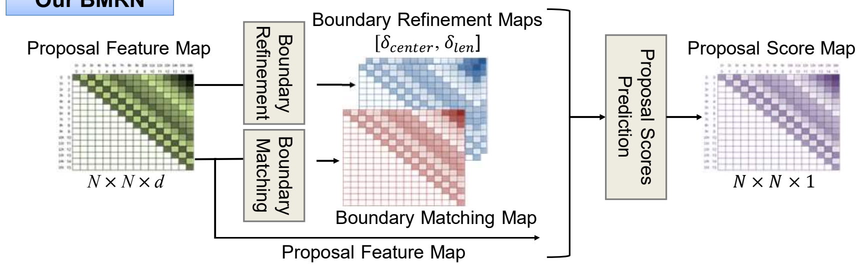
2D Map-based (2D-TAN [1])



Fixed Boundary Proposals

$[m\cdot\tau,\ (n+1)\cdot\tau]$ $,\ 0 \le m \le n \le N-1$

Our BMRN

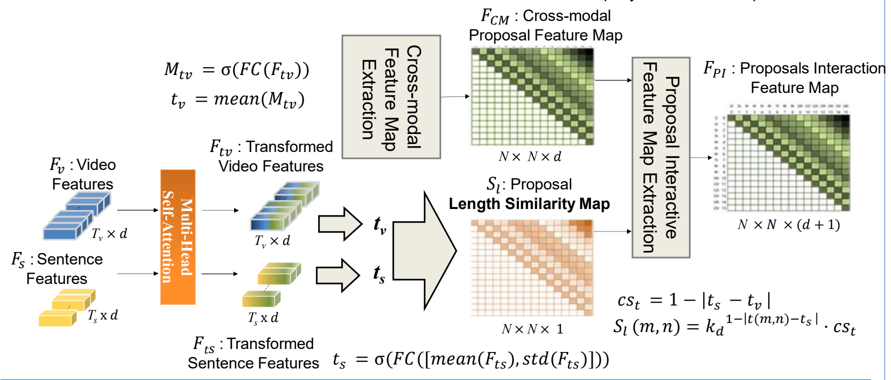


Refined Variable Boundary Proposals

$$[m\cdot\tau + \delta_{center}^{(m,n)} - \frac{\delta_{len}^{(m,n)}}{2},\ (n+1)\cdot\tau + \delta_{center}^{(m,n)} + \frac{\delta_{len}^{(m,n)}}{2}]\ ,\ 0 \le m \le n \le N-1$$

2. Query Length-aware Proposal Feature Map

- **Query length**, which is inspired by the time span in STCM-Net [2]
- **Length Similarity Map** between the estimated Len_{query} and $Len_{Proposal}$


F_v : Video Features
$T_v \times d$
F_s : Sentence Features
T_s x d
Multi-Head Self-Attention
F_{tv} : Transformed Video Features
$T_v \times d$
T_s x d
F_{ts} : Transformed Sentence Features
t_v
t_s
Cross-modal Feature Map Extraction
F_{CM} : Cross-modal Proposal Feature Map
$N\times N\times d$
S_l: Proposal Length Similarity Map
$N\times N\times 1$
Proposal Interactive Feature Map Extraction
F_{PI} : Proposals Interaction Feature Map
$N \times N \times (d+1)$

$M_{tv} = \sigma(FC(F_{tv}))$

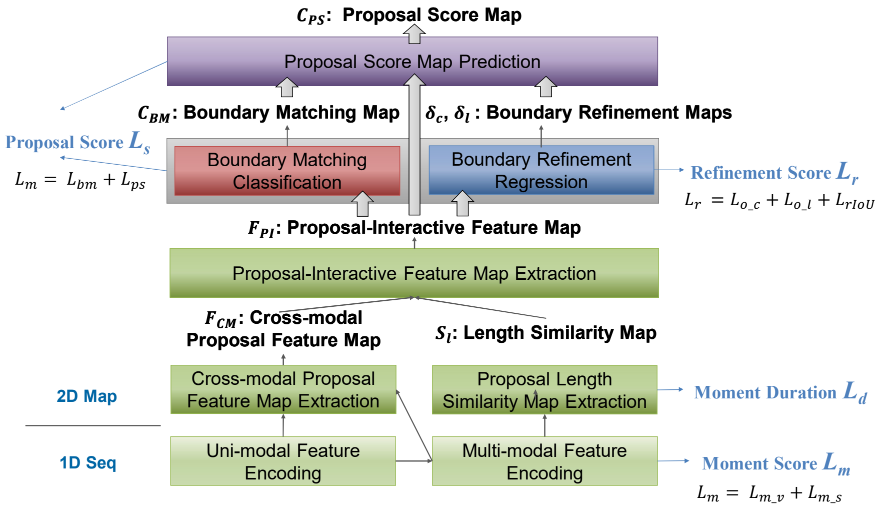
$t_v = mean(M_{tv})$

$cs_t = 1 - |t_s - t_v|$

$S_l\,(m,n) = k_d{}^{1-|t(m,n)-t_s|} \cdot cs_t$

$t_s = \sigma(FC([mean(F_{ts}), std(F_{ts})]))$

Method


C_{PS}: Proposal Score Map
Proposal Score Map Prediction
C_{BM}: Boundary Matching Map
δ_c, δ_l : Boundary Refinement Maps
Proposal Score L_s
Boundary Matching Classification
Boundary Refinement Regression
Refinement Score L_r
F_{PI}: Proposal-Interactive Feature Map
Proposal-Interactive Feature Map Extraction
F_{CM}: Cross-modal Proposal Feature Map
S_l: Length Similarity Map
2D Map
1D Seq
Cross-modal Proposal Feature Map Extraction
Proposal Length Similarity Map Extraction
Moment Duration L_d
Uni-modal Feature Encoding
Multi-modal Feature Encoding
Moment Score L_m

$L_m = L_{bm} + L_{ps}$

$L_r = L_{o\_c} + L_{o\_l} + L_{rIoU}$

$L_m = L_{m\_v} + L_{m\_s}$

For training, $L = \lambda_1 L_m + \lambda_2 L_d + \lambda_3 L_s + \lambda_4 L_r$

For inference, $[m\cdot\tau + \delta_{center}^{(m,n)} - \frac{\delta_{len}^{(m,n)}}{2},\ (n+1)\cdot\tau + \delta_{center}^{(m,n)} + \frac{\delta_{len}^{(m,n)}}{2}]$

Experimental Results

- Effectiveness of
  1. BM and BR maps
  2. Length Sim map

Model	Rank1@		Rank5@	
	0.5 (Δ)	0.7 (Δ)	0.5 (Δ)	0.7 (Δ)
Full BMRN	63.09	42.46	92.62	67.65
w/o **BM** and **BR** maps	60.83 (-2.26)	40.54 (-1.92)	89.95 (-2.67)	67.89 (0.24)
w/o **Length Sim** map	41.10 (-0.86)	23.25 (-0.27)	81.53 (-2.08)	48.55 (-2.11)

- Comparisons with SOTA methods on TML benchmark datasets

Method		Rank1@		Rank5@	
		0.5	0.7	0.5	0.7
C3D video features					
LPNet	EMNLP'21	40.94	21.13	-	-
DRN	CVPR'21	45.40	26.40	88.01	55.38
MS-2D-TAN	TPAMI'22	41.10	23.25	81.53	48.55
Ours		**45.93**	**28.37**	**89.12**	**57.19**
I3D video features					
LPNet	EMNLP'21	54.33	34.03	-	-
LGI	CVPR'21	59.46	35.48	-	-
CPN	CVPR'21	59.77	36.67	-	-
DTG	TCSVT'22	60.19	39.38	87.53	66.91
HiSA	TIP'22	61.10	39.70	-	-
TACI	CVIU'22	60.27	38.74	-	-
MS-2D-TAN	TPAMI'22	60.08	37.39	89.06	59.17
Ours		**63.09**	**42.46**	**92.62**	**67.65**

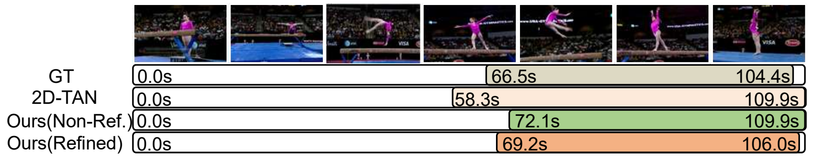
< On Charades-STA >

Method		Rank1@		Rank5@	
		0.5	0.7	0.5	0.7
C3D video features					
CMIN	SIGIR'19	43.40	23.88	67.95	50.73
2D-TAN	AAAI'20	44.51	26.54	77.13	61.96
LGI	CVPR'21	41.51	23.07	-	-
DRN	CVPR'21	45.45	24.36	77.97	50.30
CPN	CVPR'21	45.10	28.10	-	-
MSA	CVPR'21	48.02	**31.78**	78.02	63.18
LPNet	EMNLP'21	45.92	25.39	-	-
HiSA	TIP'22	45.36	27.68	-	-
TACI	CVIU'22	45.50	27.23	-	-
MS-2D-TAN	TPAMI'22	46.16	29.21	78.80	60.85
STCM-Net	Neuro.'22	46.23	29.04	78.43	63.46
Ours		**48.47**	31.15	**81.37**	**64.44**

< On ActivityNet Captions >

- Qualitative Results

Sentence Query: She jumps and flips herself around and ends by jumping down with her arms up.



References

[1] S. Zhang et al., Learning 2d temporal adjacent networks for moment localization with natural language. In AAAI, pages 12870–12877, 2020.

[2] Z. Jia et al., STCM-Net: A symmetrical one-stage network for temporal language localization in videos. Neurocomputing, 471:194–207, 2022.