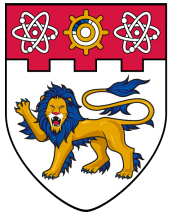


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Agency for Science, Technology and Research SINGAPORE

Parallel Attention Network with Sequence Matching for Video Grounding

Hao Zhang[1,2], Aixin Sun[1], Wei Jing[3], 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

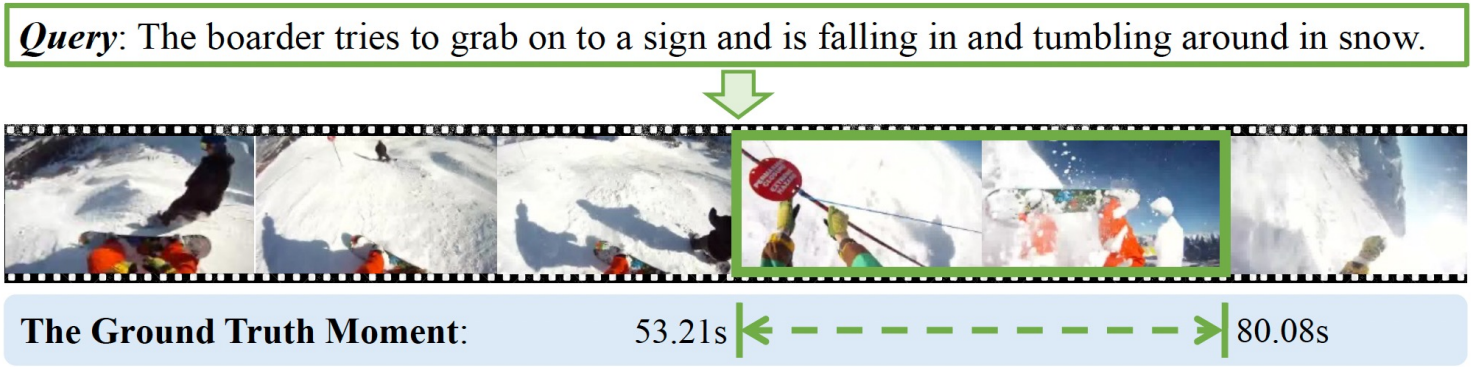
ACL 2021

Introduction

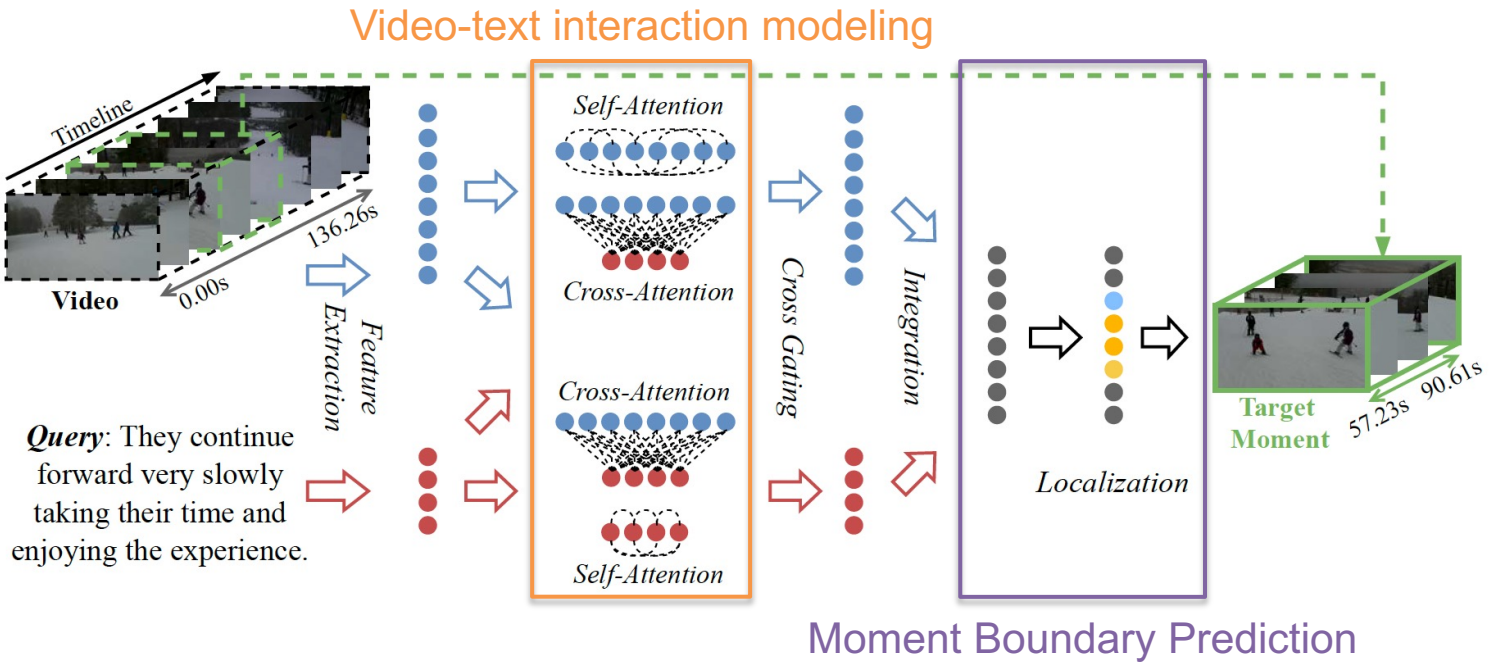
Language Query

Untrimmed Video

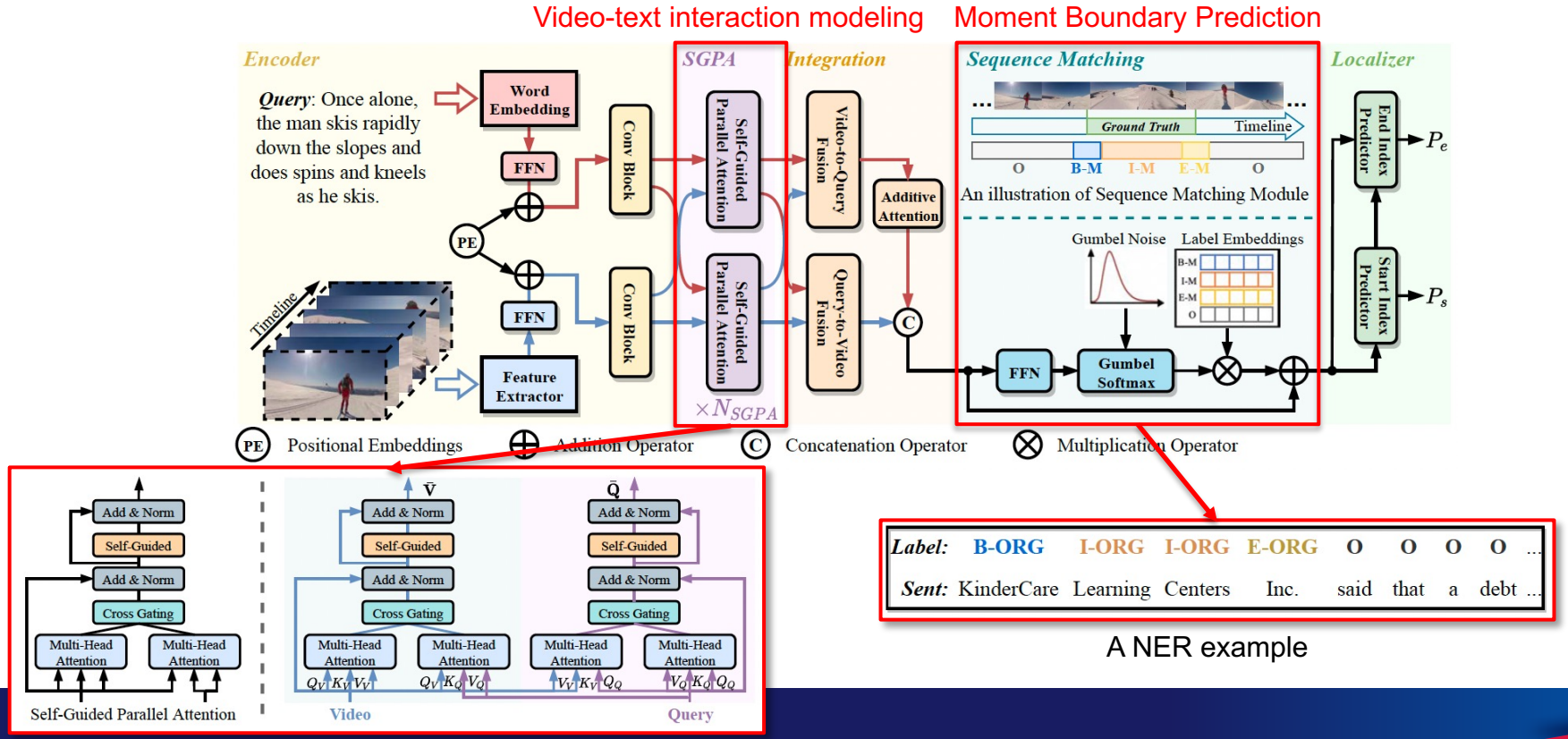
Moment Answer



Motivation


Video-text interaction modeling
Timeline
136.26s
0.00s
Video
Feature Extraction
Query: They continue forward very slowly taking their time and enjoying the experience.
Self-Attention
Cross-Attention
Cross-Attention
Self-Attention
Cross Gating
Integration
Localization
Target Moment
57.23s
90.61s
Moment Boundary Prediction

Method


Video-text interaction modeling
Moment Boundary Prediction
Encoder
Query: Once alone, the man skis rapidly down the slopes and does spins and kneels as he skis.
Word Embedding
FFN
PE
FFN
Feature Extractor
Timeline
Conv Block
Conv Block
SGPA
Self-Guided Parallel Attention
Self-Guided Parallel Attention
×N_SGPA
Integration
Video-to-Query Fusion
Query-to-Video Fusion
Additive Attention
C
Sequence Matching
Ground Truth
Timeline
O
B-M
I-M
E-M
O
An illustration of Sequence Matching Module
Gumbel Noise
Label Embeddings
B-M
I-M
E-M
O
FFN
Gumbel Softmax
Localizer
End Index Predictor
Start Index Predictor
P_e
P_s
PE Positional Embeddings
Addition Operator
C Concatenation Operator
Multiplication Operator
Add & Norm
Self-Guided
Add & Norm
Cross Gating
Multi-Head Attention
Multi-Head Attention
Self-Guided Parallel Attention
$\bar{V}$
$\bar{Q}$
Video
Query
Label: B-ORG I-ORG I-ORG E-ORG O O O O ...
Sent: KinderCare Learning Centers Inc. said that a debt ...
A NER example

Results

Methods	R@1, IoU = μ			mIoU
	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.7$	
DEBUG	54.95	37.39	17.69	36.34
ExCL	61.50	44.10	22.40	-
MAN	-	46.53	22.72	-
SCDM	-	54.44	33.43	-
CBP	-	36.80	18.87	-
GDP	54.54	39.47	18.49	-
2D-TAN	-	39.81	23.31	-
TSP-PRL	-	45.30	24.73	40.93
TMLGA	67.53	52.02	33.74	-
VSLNet	70.46	54.19	35.22	*50.02*
DRN	-	53.09	31.75	-
LGI	*72.96*	*59.46*	*35.48*	-
SeqPAN	**73.84**	**60.86**	**41.34**	**53.92**

Charades-STA

Methods	R@1, IoU = μ			mIoU
	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.7$	
DEBUG	55.91	39.72	-	39.51
ExCL	*63.00*	43.60	24.10	-
SCDM	54.80	36.75	19.86	-
CBP	54.30	35.76	17.80	-
GDP	56.17	39.27	-	39.80
2D-TAN	59.45	44.51	*27.38*	-
TSP-PRL	56.08	38.76	-	39.21
TMLGA	51.28	33.04	19.26	-
VSLNet	**63.16**	43.22	26.16	*43.19*
DRN	-	*45.45*	24.36	-
LGI	58.52	41.51	23.07	-
SeqPAN	61.65	**45.50**	**28.37**	**45.11**

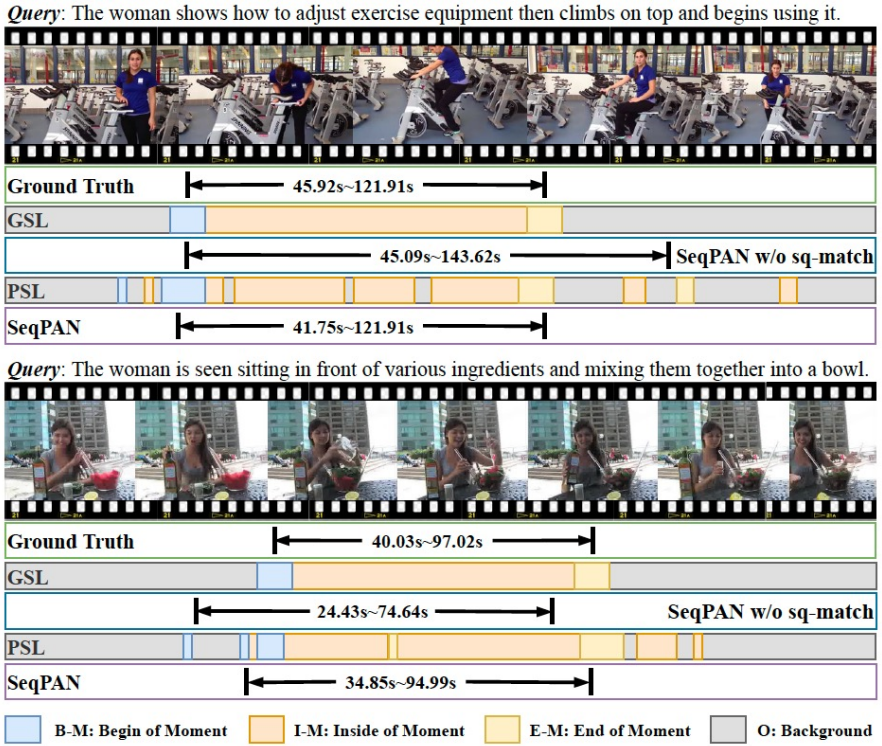
ActivityNet Captions

Methods	R@1, IoU = μ			mIoU
	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.7$	
TGN	21.77	18.90	-	-
ACL	24.17	20.01	-	-
DEBUG	23.45	11.72	-	16.03
SCDM	26.11	21.17	-	-
CBP	27.31	*24.79*	19.10	21.59
GDP	24.14	13.90	-	16.18
TMLGA	24.54	21.65	16.46	-
VSLNet	*29.61*	24.27	*20.03*	*24.11*
DRN	-	23.17	-	-
SeqPAN	**31.72**	**27.19**	**21.65**	**25.86**
2D-TAN	37.29	25.32	-	-
SeqPAN	**48.64**	**39.64**	**28.07**	**37.17**

TACoS

Visualization

- SeqPAN is superior to SeqPAN w/o sq-match module.
- The predicted moment is roughly constrained in the begin and end regions.



Thank You!