Text-Visual Prompting for Efficient 2D Temporal Video Grounding

Yimeng Zhang[1,2], Xin Chen[2], Jinghan Jia[1], Sijia Liu[1], Ke Ding[2]
[1]OPTML lab, Michigan State University, [2]Applied ML, Intel

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

- **What is temporal video grounding (TVG) ?**

TVG is to predict the **starting/ending time points** of moments described by a text sentence within a long untrimmed video.

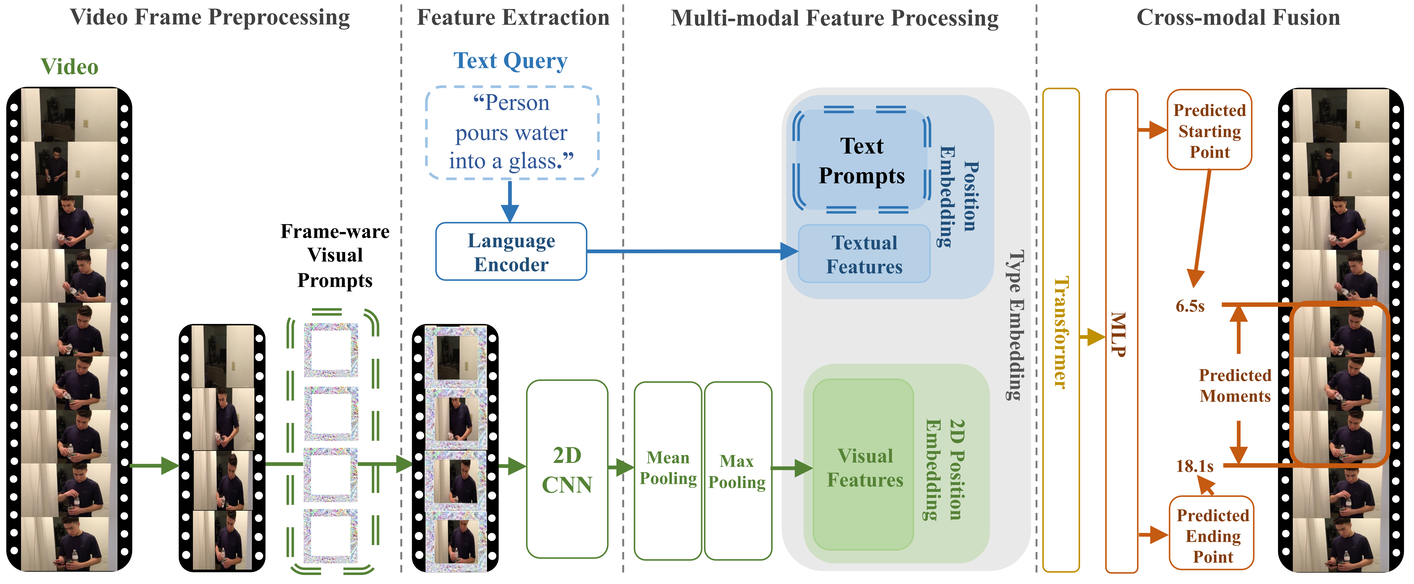
- **Motivation**

High complexity of 3D CNNs makes extracting dense 3D visual features time-consuming, which calls for intensive memory and computing resources.

- **Challenges**

How to advance 2D TVG methods so as to achieve comparable results to 3D TVG methods?

Text-Visual Prompting (TVP) Framework for TVG

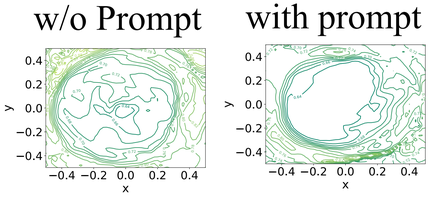


Loss Function Design

$$\mathcal{L} = \mathcal{L}_{\text{tIoU}} + \beta_1 \mathcal{L}_{\text{dis}} + \beta_2 \mathcal{L}_{\text{dur}}$$

Temporal IoU Loss $\mathcal{L}_{tIoU}$: maximize overlapping between the predicted time interval and its ground truth.
Distance Loss $\mathcal{L}_{dis}$: minimize the normalized central time point distance.
Duration Loss $\mathcal{L}_{dur}$: minimize the duration differences.

Loss Landscape Analysis



Performance

- **Metric**

The percentage accuracy of predicted moments whose tIoU (temporal IoU) with the ground-truth moment is larger than threshold m.

- **Charades-STA**

Type	Method	Visual Feature	Accuracy with Temporal IoU threshold m m=0.3	m=0.5	m=0.7
3D TVG	CTRL [3]	C3D	-	23.63	8.89
	ABLR [12]	C3D	-	24.36	9.01
	BPNet [10]	C3D	55.46	38.25	20.51
	LPNet [9]	C3D	59.14	40.94	21.13
	QSPN [11]	C3D	54.70	35.60	15.80
	TSP-PRL [8]	C3D	-	45.45	24.75
	TripNet [5]	C3D	54.64	38.29	16.07
	DRN [13]	C3D	-	45.40	26.40
	CPNet [6]	C3D	-	40.32	22.47
	DEBUG [7]	C3D	54.95	37.39	17.92
	ExCL [4]	I3D	61.50	44.1	22.40
	VSLNet [15]	I3D	64.30	**47.31**	**30.19**
	MAN [14]	I3D	-	46.53	22.72
2D TVG	MCN [1]	VGG	-	17.46	8.01
	SAP [2]	VGG	-	27.42	13.36
	Ours				
TVP-Based 2D TVG	Base	ResNet	61.29	40.43	19.89
	+ Visual Prompts		65.38	44.31	20.22
	+ Text Prompts		65.81	43.44	20.65
	+ Both Prompts		**65.92**	44.39	21.51

- **ActivityNet Captions**

Type	Method	Visual Feature	Accuracy with Temporal IoU threshold m m=0.3	m=0.5	m=0.7
3D TVG	CTRL [3]	C3D	28.70	14.00	-
	BPNet [10]	C3D	59.98	42.07	24.69
	LPNet [9]	C3D	**64.29**	**45.92**	25.39
	QSPN [11]	C3D	45.30	27.70	13.60
	TSP-PRL [8]	C3D	56.02	38.83	-
	TripNet [5]	C3D	48.42	32.19	13.93
	DRN [13]	C3D	-	45.45	24.36
	CPNet [6]	C3D	-	40.56	21.63
	ABLR [12]	C3D	55.67	36.79	-
	DEBUG [7]	C3D	55.91	39.72	-
	ExCL [4]	C3D	63.00	43.60	24.10
	VSLNet [15]	C3D	63.16	43.22	**26.16**
	Ours				
TVP-Based 2D TVG	Base	ResNet	57.20	40.16	19.14
	+ Visual Prompts		60.12	43.39	23.71
	+ Text Prompts		60.48	42.58	24.39
	+ Both Prompts		60.71	43.44	25.03

Ref
[1] Hendricks. et al, Localizing moments in video with natural language. (2017). [2] Chen. et al. Semantic proposal for activity localization in videos via sentence query. (2019). [3] Gao. et al. Tall: Temporal activity localization via language query (2017). [4] Ghosh. et al. Excl: Extractive clip localization using natural language descriptions. (2019). [5] Hahn. et al. Tripping through time: Efficient localization of activities in videos. (2019). [6] Li. et al. Proposal-free video grounding with contextual pyramid network. (2021). [7] Lu. et al. Debug: A dense bottom-up grounding approach for natural language video localization. (2019). [8] Wu. et al. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. (2020). [9] Xiao. et al. Natural language video localization with learnable moment proposals. (2021). [10] Xiao. et al. Boundary proposal network for two-stage natural language video localization. (2021). [11] Xu. et al. Multilevel language and vision integration for text-to-clip retrieval. (2019). [12] Yuan. et al. To find where you talk: Temporal sentence localization in video with attention based location regression. (2019). [13] Zeng. et al. Dense regression network for video grounding. (2020). [14] Zhang. et al. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. (2019). [15] Zhang. et al. Span-based localizing network for natural language video localization. (2020)