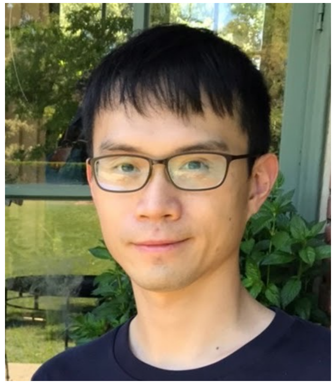
NEURAL INFORMATION PROCESSING SYSTEMS

End-to-end Multi-modal Video Temporal Grounding

NeurIPS 2021



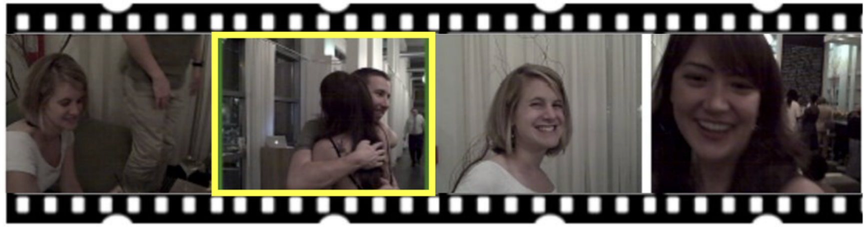
Yi-Wen Chen
UC Merced

Yi-Hsuan Tsai
Phiar

Ming-Hsuan Yang
UC Merced,
Google Research

Text-Guided Video Temporal Grounding

- Goal
  - Identify the starting and ending time of an event based on a natural language description
- Applications
  - Video retrieval, video editing, human-computer interaction
- Challenges
  - Joint understanding of scenes, objects, activities and sentences in videos

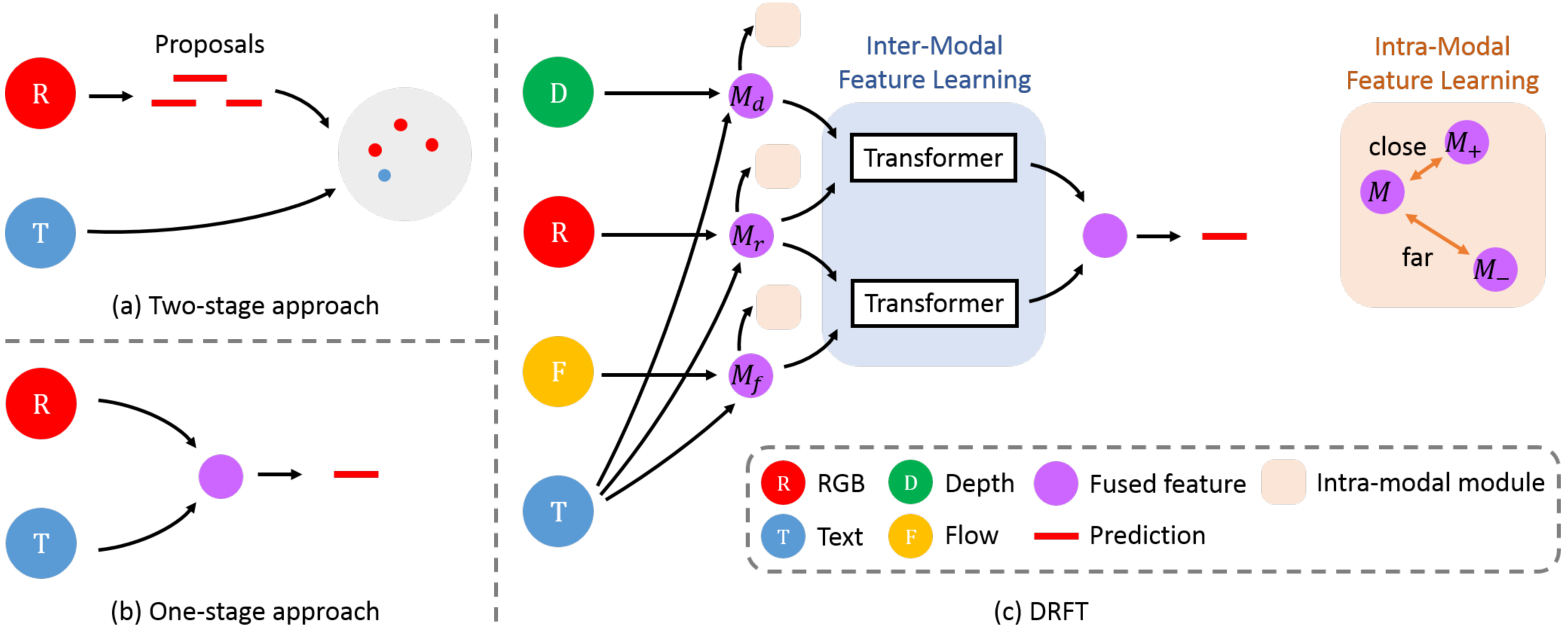


Query: “A girl and a guy hug each other”

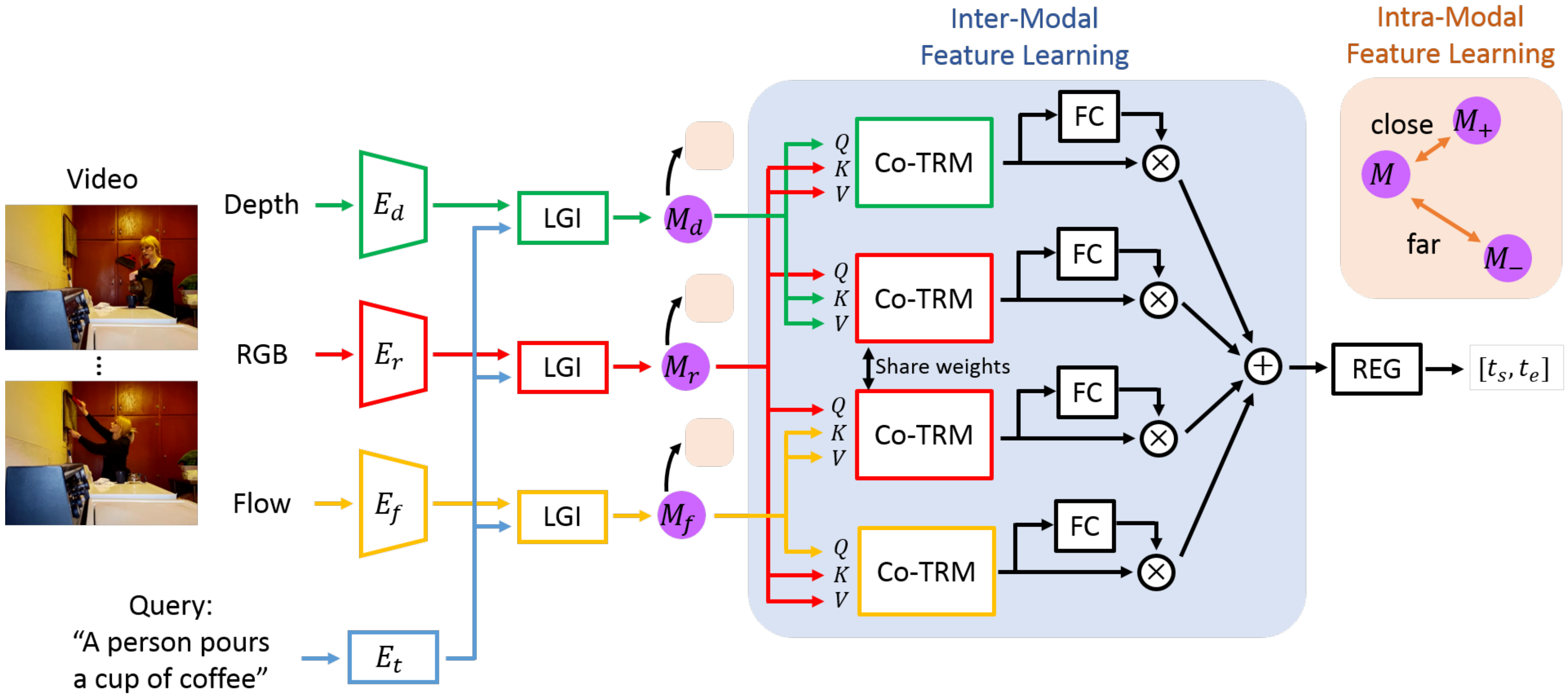
Query: “Train begins to move”

Motivation

- RGB features are affected by background clutters
- Optical flow can identify actions with large motion, e.g., “closing a door”, “throwing a pillow”
- Depth helps recognize actions involving objects with distinct shapes, e.g., “sitting in a bed”, “working at a table”

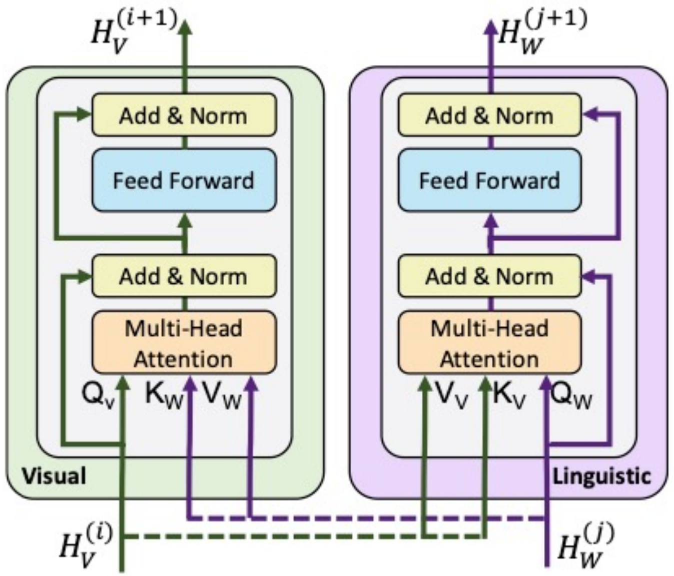


Proposed Framework



LGI: Local-Global Video-Text Interactions
Co-TRM: Co-Attentional Transformer
FC: Fully-Connected Layer
REG: Regression

Inter-Modal Feature Learning



- Co-Attentional Feature Fusion
  - Co-attentional transformer layer [1] with multi-head attention blocks
  - Each block takes a pair of features as input
  - Query (Q) from one modality, key (K) and value (V) from the other modality

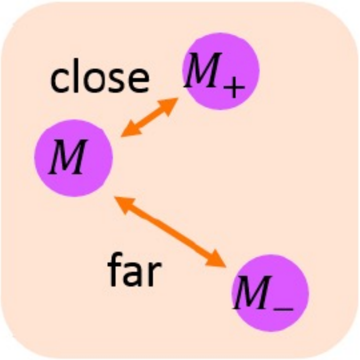
- Dynamic Feature Fusion
  - Importance of each modality depends on the input
  - Dynamically learn the weights for each multi-modal feature and linearly combine the features using the weights

[1] Lu et al. “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks” In NeurIPS, 2019

Intra-Modal Feature Learning

- Anchor V: input video
- Positive samples V_+: videos with same action category
- Negative samples V_-: videos with different action categories
- Contrastive learning on the multi-modal features M_d, M_r and M_f

$$L_{cl} = -\log \frac{\sum_{M_+ \in Q_+} e^{h(M)^\top h(M_+)/\tau}}{\sum_{M_+ \in Q_+} e^{h(M)^\top h(M_+)/\tau} + \sum_{M_- \in Q_-} e^{h(M)^\top h(M_-)/\tau}}$$



Experimental Settings

- Language encoder: Bi-LSTM
- Visual encoder: I3D/C3D
- Optical flow: RAFT [1]
- Depth: MiDaS [2]
- Datasets: Charades-STA, ActivityNet Captions
- Evaluation metric
  - Recall at various thresholds of temporal IoU (R@0.3, R@0.5, R@0.7)
  - mean temporal IoU (mIoU)

[1] Teed et al. “RAFT: Recurrent All-Pairs Field Transforms for Optical Flow” In ECCV, 2020
[2] Ranftl et al. “Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer” In PAMI, 2020

Experimental Results

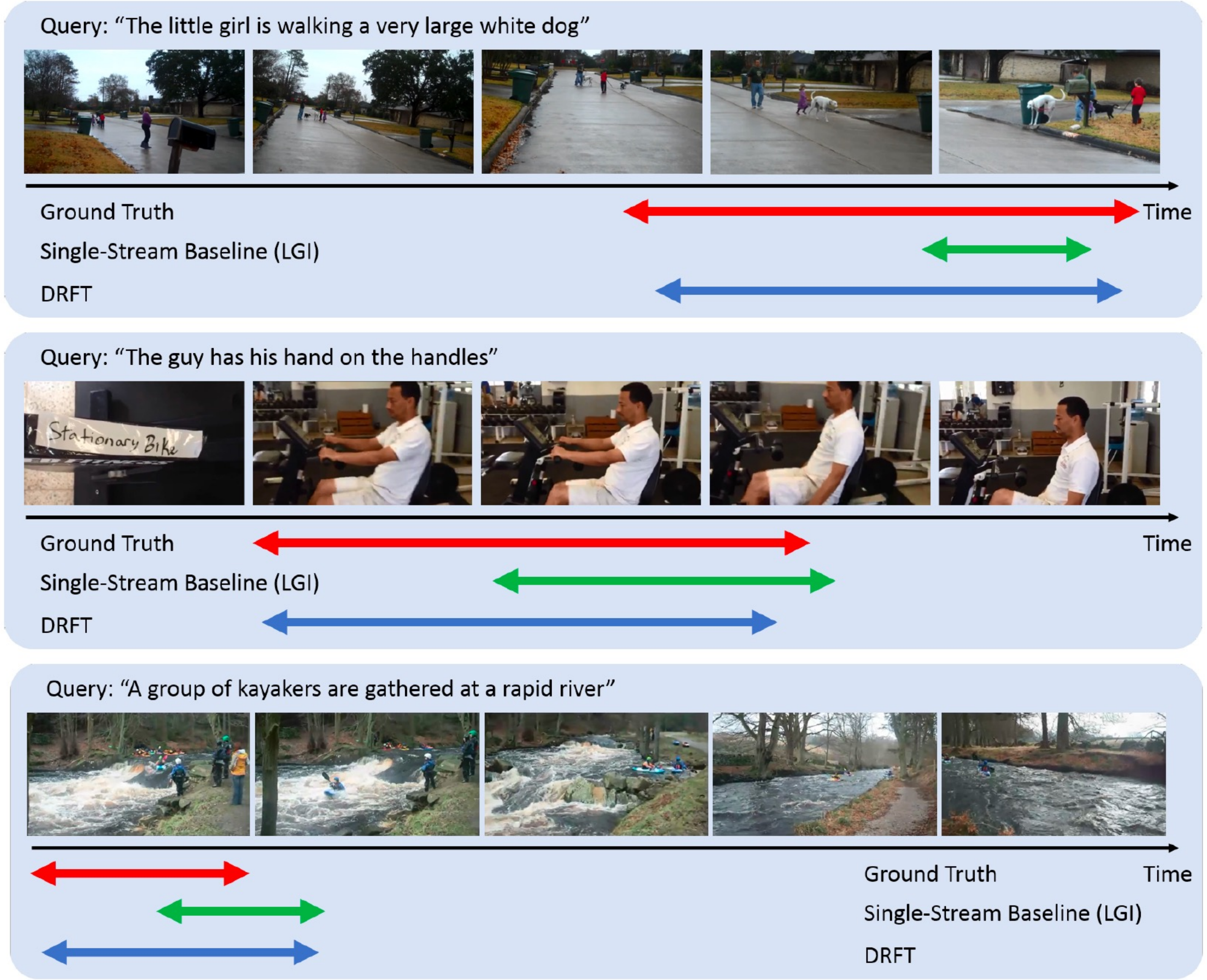
		Charades-STA				ActivityNet Captions			
	Method	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
	CTRL [Gao, ICCV'17]	-	21.42	7.15	-	28.70	14.00	20.54	-
	RWM [He, AAAI'19]	-	36.70	-	-	-	36.90	-	-
	MAN [Zhang, CVPR'19]	-	46.53	22.72	-	-	-	-	-
	TripNet [Hahn, BMVC'20]	51.33	36.61	14.50	-	45.42	32.19	13.93	-
	PfTML-GA [Rodriguez, WACV'20]	67.53	52.02	33.74	-	51.28	33.04	19.26	37.78
	DRN [Zeng, CVPR'20]	-	53.09	31.75	-	-	42.49/45.45	22.25/24.36	-
	LGI [Mun, CVPR'20]	72.96	59.46	35.48	51.38	58.52	41.51	23.07	41.13
RGB	Single-stream DRFT	73.85	60.79	36.72	52.64	60.25	42.37	25.23	43.18
RGB, flow	Two-stream DRFT	74.26	61.93	38.69	53.92	61.80	43.71	26.43	44.82
RGB, flow, depth	Three-stream DRFT	**76.68**	**63.03**	**40.15**	**54.89**	**62.91**	**45.72**	**27.79**	**45.86**

Ablation Study

Method	Charades-STA				ActivityNet Captions			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
Three-stream baseline	71.13	57.39	34.69	48.21	56.45	38.63	22.05	39.86
DRFT w/o contrastive loss	75.41	61.87	39.02	53.65	61.59	44.50	26.48	44.61
DRFT w/o learnable weight	75.03	61.65	38.78	53.11	61.47	44.42	26.31	44.39
DRFT w/o transformer	74.72	61.05	38.26	52.74	61.04	43.83	25.74	43.90
Three-stream DRFT	76.68	63.03	40.15	54.89	62.91	45.72	27.79	45.86

Qualitative Results on Charades-STA Dataset

Qualitative Results on ActivityNet Captions Dataset



Conclusions

Code Available at: https://github.com/wenz116/DRFT

- Propose a multi-modal framework for text-guided video temporal grounding by extracting complementary information from RGB, optical flow and depth features
- Design a dynamic fusion mechanism across modalities via co-attentional transformer to effectively learn inter-modal features
- Apply contrastive learning across videos for each modality to enhance intra-modal feature representations that are invariant to distracted factors with respect to actions