Spatio-Temporal Graph for Video Captioning with Knowledge Distillation

Boxiao Pan, Haoye cai, De-An Huangg, Kuan-Hui Lee, Adrien Gasidon, Ehsan Adeli, Juan Carlos Niebles

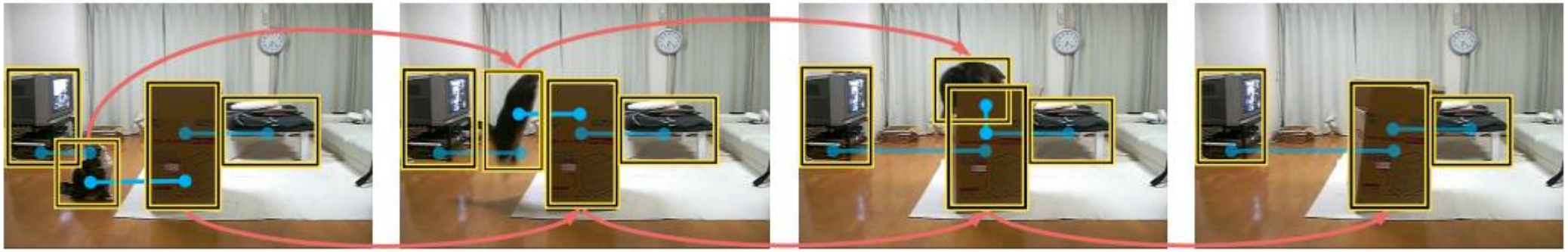
CVPR 2020

임희주

Index

Introduction

Introduction



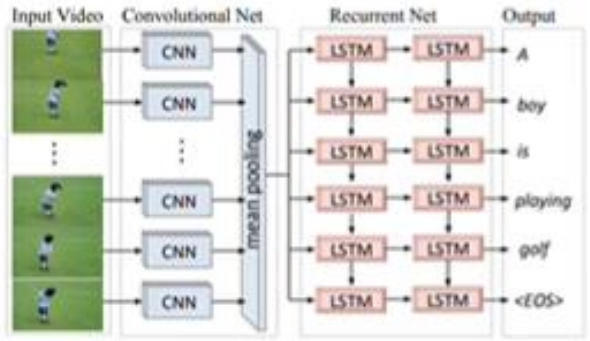
"A cat jumps into a box."

Video Captioning

비디오 입력에서 어떻게 장면을 이해하고 묘사할까?

Introduction

Video-level approaches

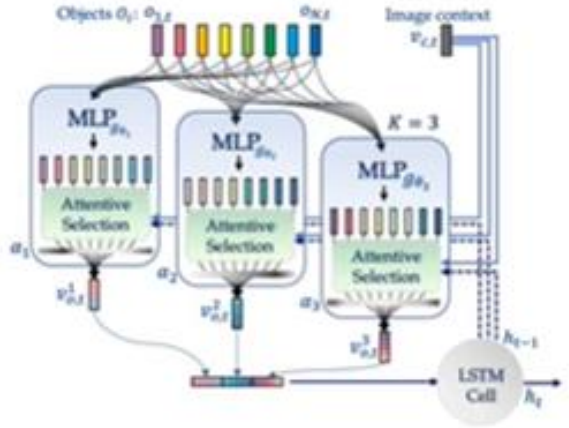


[Venugopalan et al. 2014, Venugopalan et al. 2015, Chen et al. 2018, Pei et al. 2019]

✓ Global context modeling

Ignore object-level information

Spatial object interaction modeling



[Ma et al. 2018]

✓ Spatial attentive pooling

No temporal transformation

Temporal object transformation modeling

[Zhang et al. 2019]

✓ Temporal object trajectory

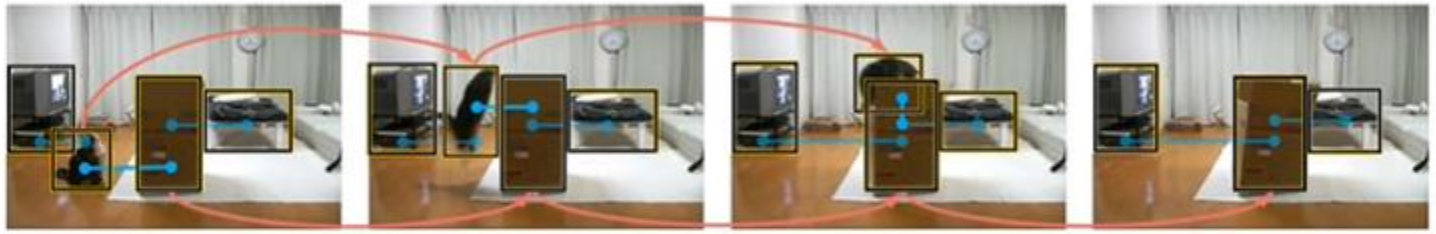
No spatial interaction

Introduction

Ignore object-level information

No temporal transformation

No spatial interaction



Object detection

Temporal transformation modeling

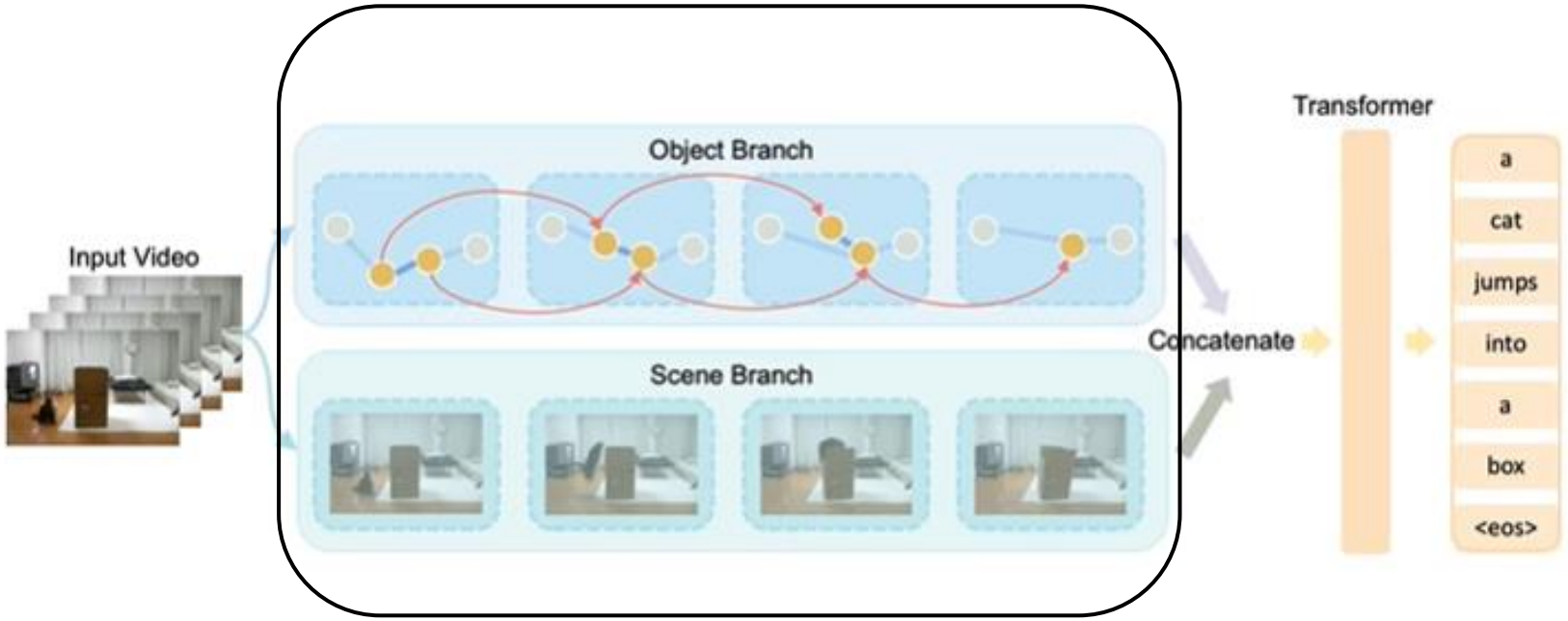
Spatial interaction modeling

Spatio-temporal 한 object의 관계 모델링

Methodology

Methodology

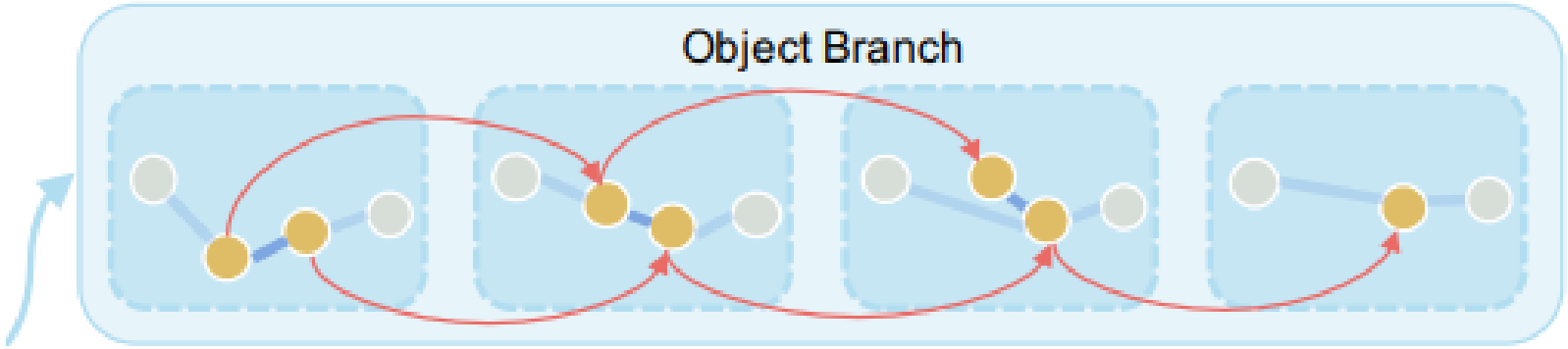
: Two branch



Two branch 구조

Methodology

: Object branch



- Object features by Faster R-CNN
- Spatial edges: spatial connectivity (IoU):

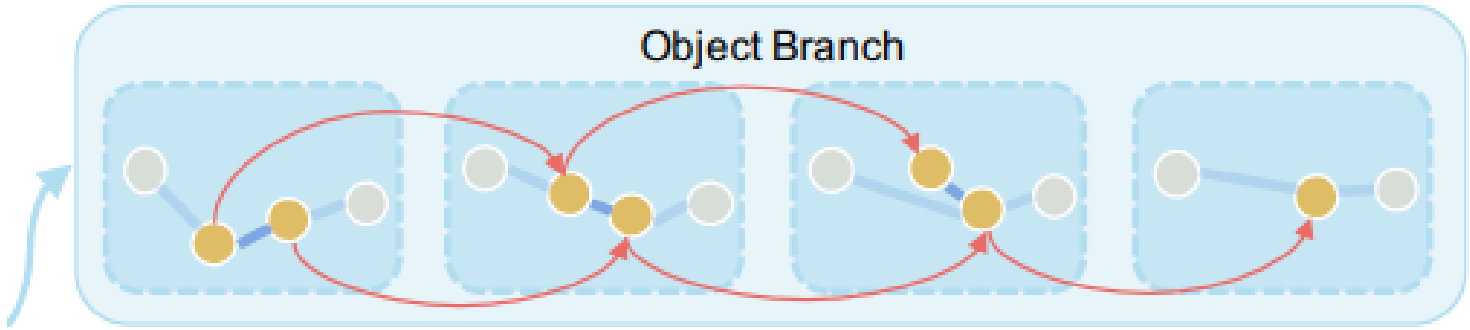
$$G_{tij}^{space} = \frac{\exp \sigma_{tij}}{\sum_{j=1}^{N_t} \exp \sigma_{tij}}$$

- Temporal edges: semantic similarity (cosine distance of object features):

$$G_{tij}^{time} = \frac{\exp \cos (o_t^i, o_{t+1}^j)}{\sum_{j=1}^{N_{t+1}} \exp \cos (o_t^i, o_{t+1}^j)}$$

Methodology

: Object branch



- Spatio-Temporal Graph

$$G^{st} = \begin{bmatrix} G_1^{space} & G_1^{time} & 0 & \dots & 0 \\ 0 & G_2^{space} & G_2^{time} & \dots & 0 \\ 0 & 0 & G_3^{space} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & G_T^{space} \end{bmatrix} \in \mathbb{R}^{N \times N}$$

- Update (through graph conv)

$$H^{(l+1)} = \mathrm{ReLU}(H^{(l)} + \Lambda^{-\frac{1}{2}} G^{st} \Lambda^{-\frac{1}{2}} H^{(l)} W^{(l)}),$$

$$H^{(0)} = \mathrm{stack}(F_o) W_o \in \mathbb{R}^{N \times d_{model}},$$

Methodology

: Scene branch

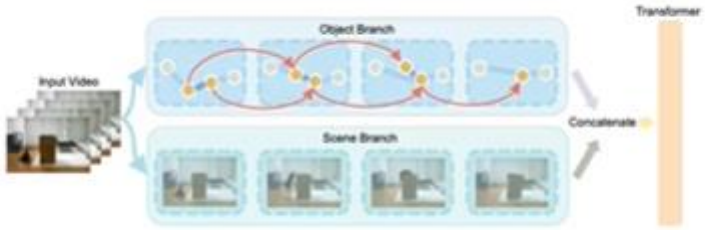
- **Object branch 에서 누락된 global context 제공**

- **2D frame features 와 3D clip features 의 concatenation**

Methodology

: Problem

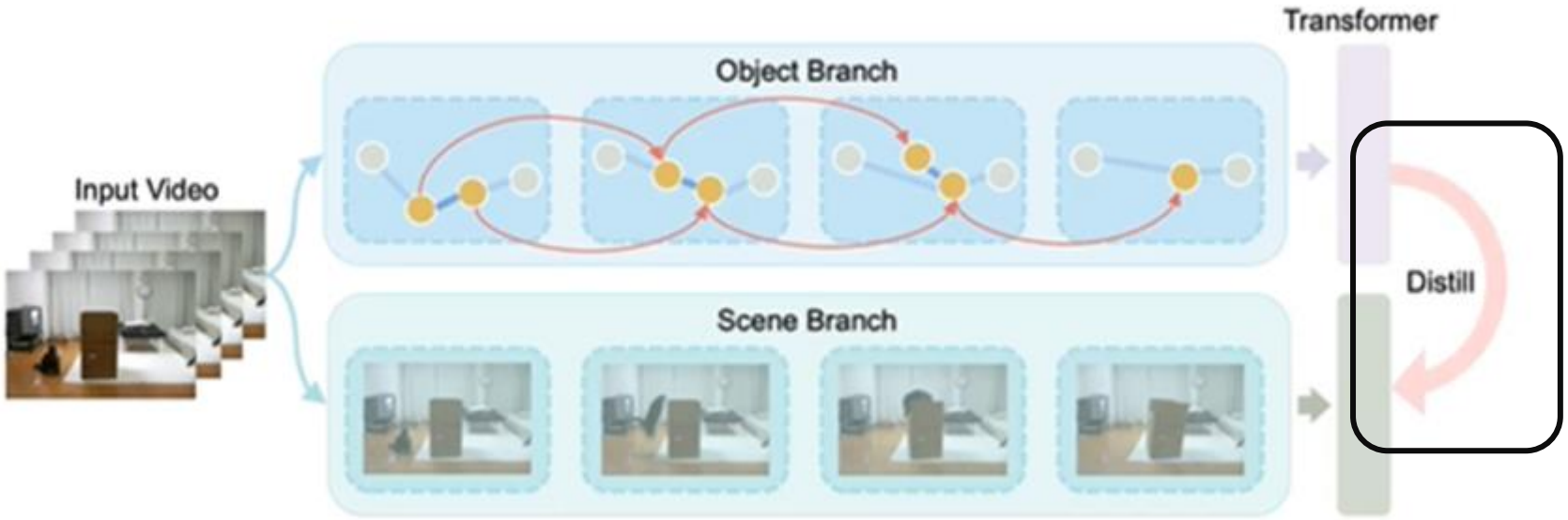
Problem : Object branch 는 안정적이지 못함



- 비디오에 a variable number of objects 포함되어 있음
- 학습된 object는 매우 noisy
- 또한 Early feature fusion 이 성능 최적화 하기엔 부적합

Methodology

: Solution (Distillation)

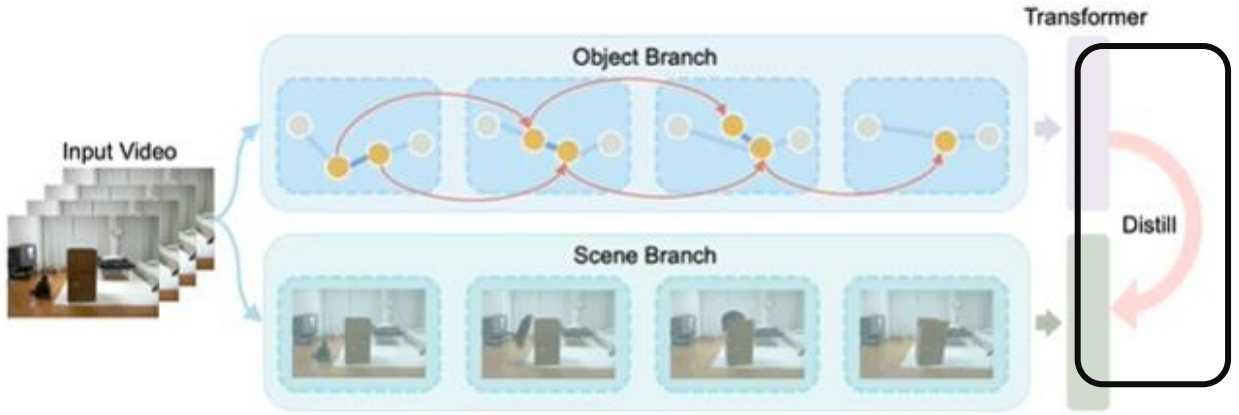


해결방안 : late fusion(with distillation)

Distillation: $q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$

Methodology

: Solution (Distillation)



- 두 branch로 부터 온 logits의 KL divergence를 최소화 하는 방안으로 Fuse함

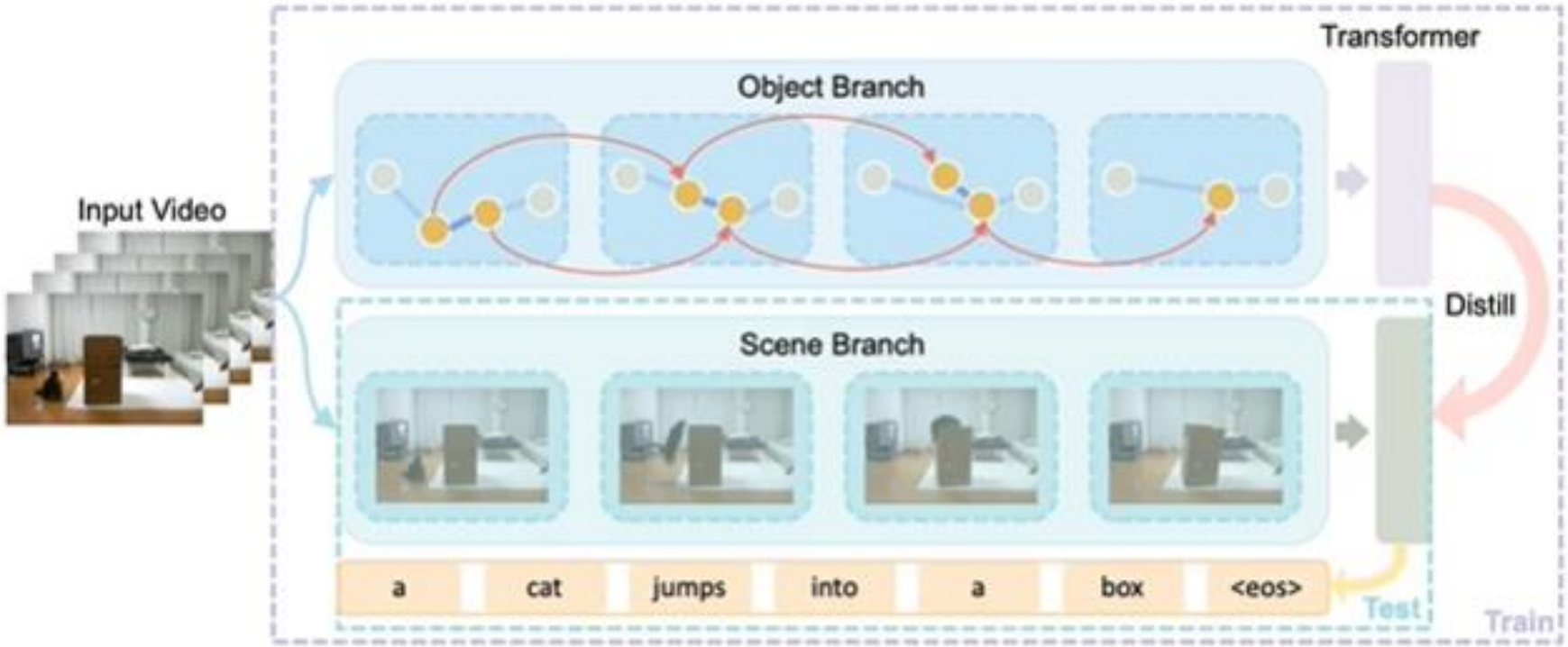
$$L_{distill} = -\sum_{x \in V} P_s(x) \log\left(\frac{P_o(x)}{P_s(x)}\right).$$

- 결과 학습된 Feature는 더 강건해짐

- Test 시에는 distilled된 Scene branch 만 사용

Methodology

: Final Framework



$$L = L_{o\_lang} + \lambda_{sl} L_{s\_lang} + \lambda_d L_{distill},$$

Experiment

Experiment

Dataset

MSVD

A baseball player hits a baseball.

MSR-VTT

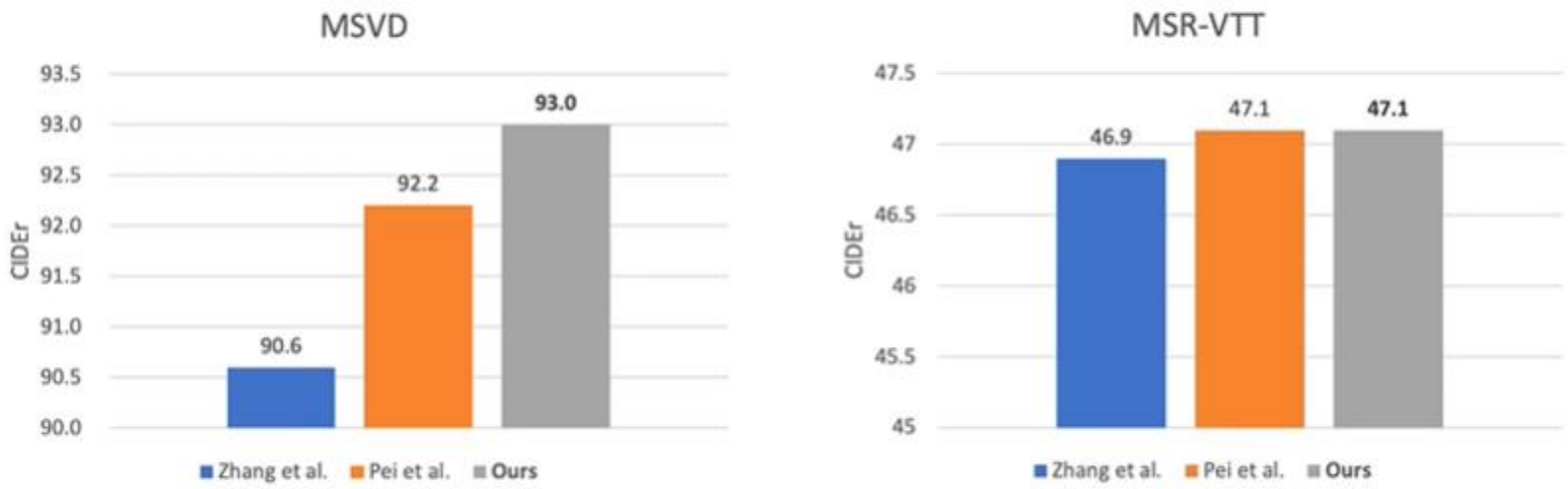
A cat is playing with a baby.

	No. of Videos	No. of Sentences / Video
MSVD	1970	~40
MSR-VTT	10000	20

Experiment

Results

Quantitative results – comparison



Our approach achieves state-of-the-art performance on MSVD and competitive results on MSR-VTT

Experiment

Results

MSVD

Method	BLEU@4	METEOR	ROUGE-L	CIDEr
Wang *et al.* [39]	52.5	34.1	71.3	88.7
Hou *et al.* [19]	52.8	36.1	71.8	87.8
RecNet [40]	52.3	34.1	69.8	80.3
PickNet [6]	52.3	33.3	69.6	76.5
OA-BTG [49]	**56.9**	36.2	-	90.6
MARN [30]	48.6	35.1	71.9	92.2
Ours	52.2	**36.9**	**73.9**	**93.0**

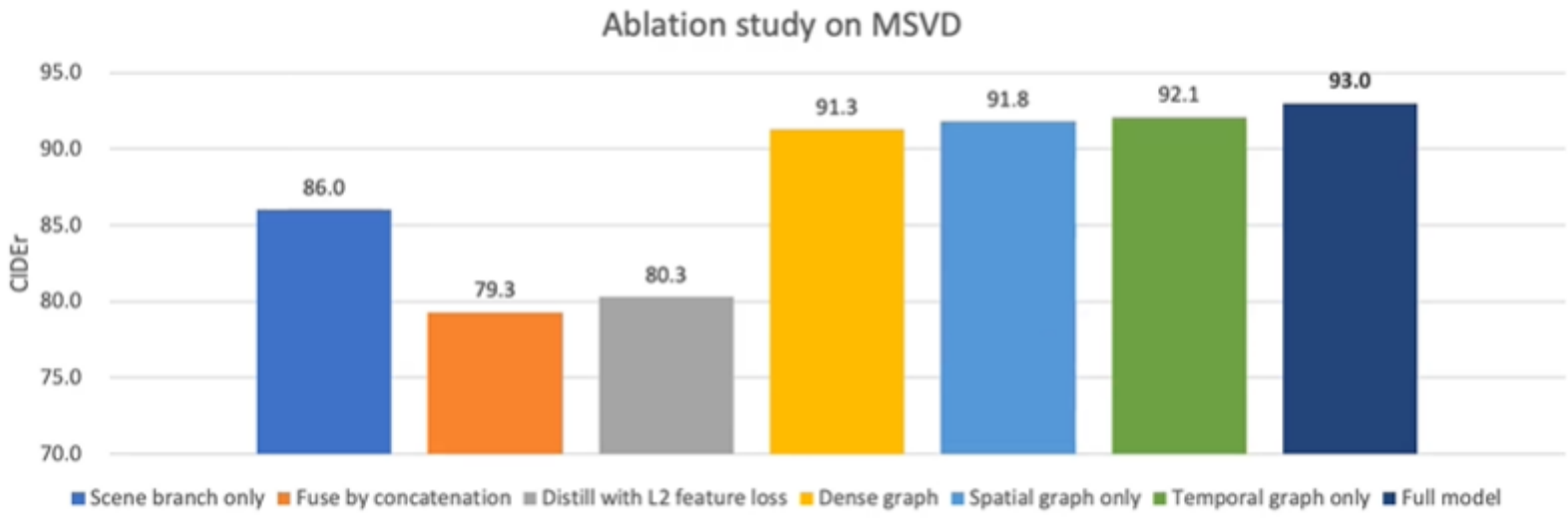
MSR-VTT

Method	BLEU@4	METEOR	ROUGE-L	CIDEr
Wang *et al.* [39]	42.0	28.2	61.6	48.7
Hou *et al.* [19]	**42.3**	**29.7**	**62.8**	**49.1**
RecNet [40]	39.1	26.6	59.3	42.7
PickNet [6]	41.3	27.7	59.8	44.1
OA-BTG [49]	**41.4**	28.2	-	46.9
MARN [30]	40.4	28.1	60.7	**47.1**
Ours (Scene only)	37.2	27.3	59.1	44.6
Ours	40.5	**28.3**	**60.9**	**47.1**

Experiment

Results

Quantitative results – ablation study



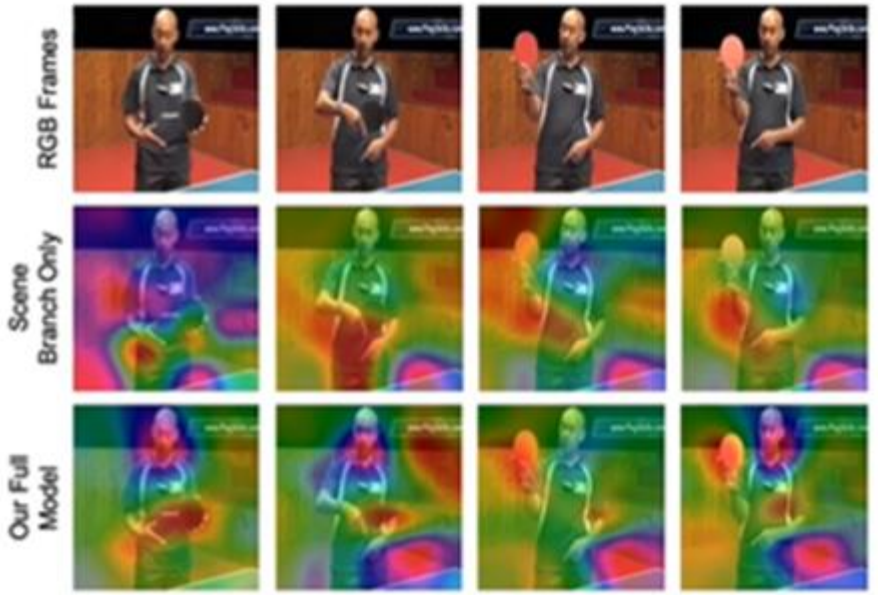
Both sub-graphs capture important and distinct information

Experiment

Results

Scene Branch Only 모델 보다 Full model 이 key regions 을 더 잘 찾아냄

논문의 모델이 GT에 더 가까운 결과를 보임

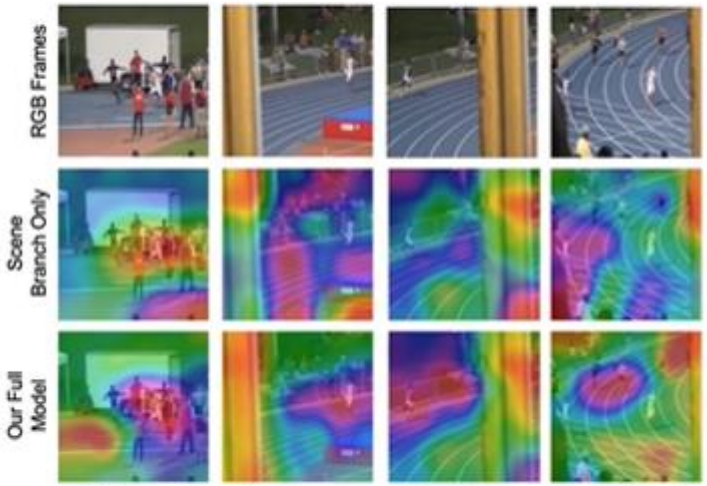


GT: a man in a black shirt demonstrates how to play ping pong
Ours: a man in a **black shirt** is talking about ping pong
[Wang et al. 2019]: there is a man is talking about table tennis

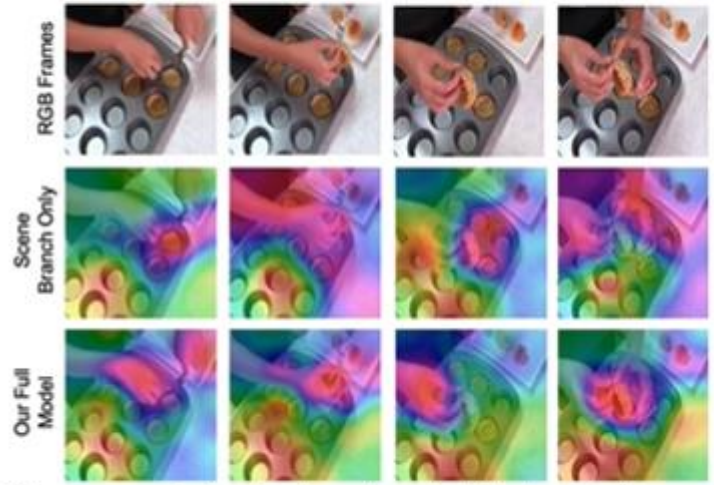
Experiment

Results

More results



GT: a group of men are running down a race track
Ours: a **race** is going on the track
[Wang et al. 2019]: there is a man running on the track



GT: a woman is showing how to make little baskets from potatoes
Ours: a woman is showing how to make a **potato** salad
[Wang et al. 2019]: a person is preparing a recipe

Conclusion & Discussion

Conclusion

Late fusion 및 Distill 기법을 포함한 Two branch 구조의 framework 제안
-> spatio temporal object interaction 을 잘 활용
-> video의 Early fusion 의 문제점 해결

Thanks

Appendix

Appendix
references

- **Paper**

- https://openaccess.thecvf.com/content_CVPR_2020/papers/Pan_Spatio-Temporal_Graph_for_Video_Captioning_With_Knowledge_Distillation_CVPR_2020_paper.pdf

- **Etc.**

- https://www.youtube.com/watch?v=QxHttaZF_Xc

- https://github.com/HYU-AILAB/ai-seminar/blob/master/season_13/07.%20A%20Graph%20Convolutional%20Neural%20Network%20for%20Emotion%20Recognition%20in%20Conversation/200831_DialogueGCN_Yuri.pdf