[CVPR 2020] Spatio-Temporal Graph for Video Captioning with Knowledge Distillation

Boxiao Pan[1], Haoye Cai[1], De-An Huang[1], Kuan-Hui Lee[2], Adrien Gaidon[2], Ehsan Adeli[1], Juan Carlos Niebles[1]

[1]Stanford University & [2]Toyota Research Institute

2021.11.29 Willy Fitra Hendria

Sejong ICL

Spatio-Temporal Graph for Video Captioning

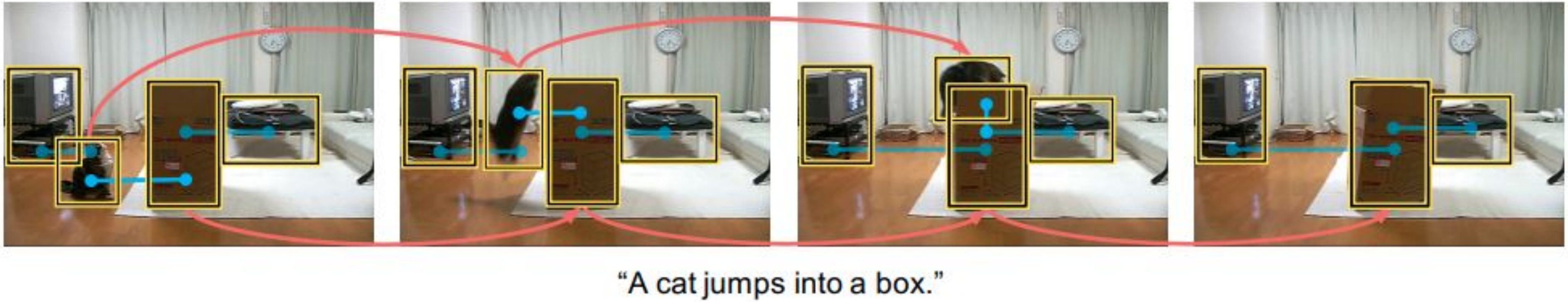
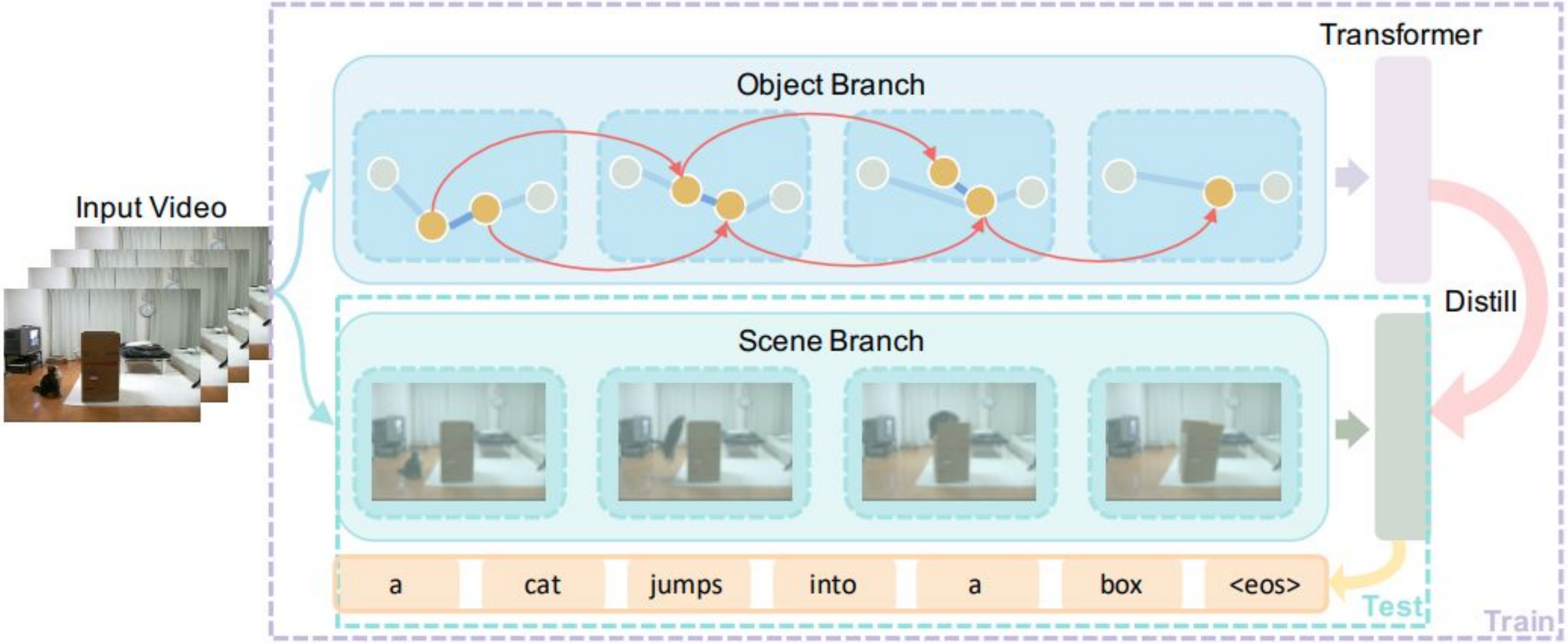


Figure: Illustration of spatio-temporal graph for video captioning. Yellow boxes represent object proposals from object detection model. Red arrows denote directed temporal edges, while blue lines indicate undirected spatial connections.

Motivation:
Explicitly modelling objects interactions to make visually grounded predictions in interpretable manner.

Overall Framework



The object branch captures space-time object interaction information via the proposed spatio-temporal graph model, while the scene branch provides the global context absent from the object branch.

Feature Representation

Given a sequence of RGB frames $\{x_1, x_2, \ldots, x_T\}$, they extract **scene features** and **object features**.

Scene Features. 2D frame features $F_{2D} = \{f_1, f_2, \ldots, f_T\}$ are extracted using ResNet-101, and 3D clip features $F_{3D} = \{v_1, v_2, \ldots, v_L\}$ are extracted using I3D. These two features are projected to the same dimension, then concatenated along channel dimension.

Object Features. Set of object features $F_o = \{o_1^1, o_1^2, \ldots, o_t^j, \ldots, o_T^{N_T}\}$ are extracted using Faster R-CNN, where **N_t** denotes the number of objects in frame **t** and **j** is the object index within each frame. Each object has the same dimension as 2D frame features.

Spatial and Temporal Graph

Spatial Graph

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

where $\boldsymbol{G_{tij}^{space}}$ is the **(i, j)**-th element of $\mathbf{G_t^{space}} \in \mathbf{R}^{N_t \times N_t}$, which measures the spatial connectivity between the **i**-th and **j**-th objects at time step **t**. $\mathbf{N_t}$ denotes total number of objects at time step **t**. $\boldsymbol{\sigma_{tij}}$ denotes the IoU between the two objects.

Based on the observation that objects which are close to each other are more likely to be correlated.

Temporal Graph

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

where $\mathbf{G_{tij}^{time}}$ denotes the **(i, j)**-th element of $\mathbf{G_t^{time}} \in \mathbf{R}^{N_t \times N_{t+1}}$, and $\mathbf{cos(o^i , o^j)}$ measures the cosine similarity between the two feature vectors.

Spatio-Temporal Graph

Merge all spatial and temporal graphs for a video into a single spatio-temporal graph G^{st} :

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

N is the total number of objects in all time steps in the video, i.e., $N = \sum_{t=1}^{T} N_t$

Then the graph convolution is applied to this spatio-temporal graph.

Graph Convolutional Network

Pan et al. (2020) defined the propagation rule as follows:

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

where $\mathbf{W^{(l)}}$ is the weight matrix of layer **l**. **Λ** is the diagonal node degree matrix with

$$\Lambda_{ii} = \sum_j G_{ij}^{st}$$

The input $\boldsymbol{H^{(0)}}$ are the stacked object features $\boldsymbol{F_0}$ multiply with the transformation matrix $\mathbf{W_0}$:

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

Knowledge Distillation

Pan et al. performed distillation by minimizing the KL divergence between word probability distribution from the two branches:

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

$\mathbf{P_o(x)}$ be the probability distribution (pre-Softmax logits) across the vocabulary **V** from object branch and $\mathbf{P_s(x)}$ be the probability distribution from scene branch.

Overall Loss Function

Loss of object branch Loss of scene branch Distilation Loss

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

where $\boldsymbol{\lambda_{sl}}$ and $\boldsymbol{\lambda_d}$ are trade-off hyper-parameters.

Quantitative Results (1)

They follow the standard practice [30] to not compare to methods based on reinforcement learning (RL) [39].

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**

Table: Comparison with other methods on MSVD

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**

Table: Comparison with other methods on MSR-VTT

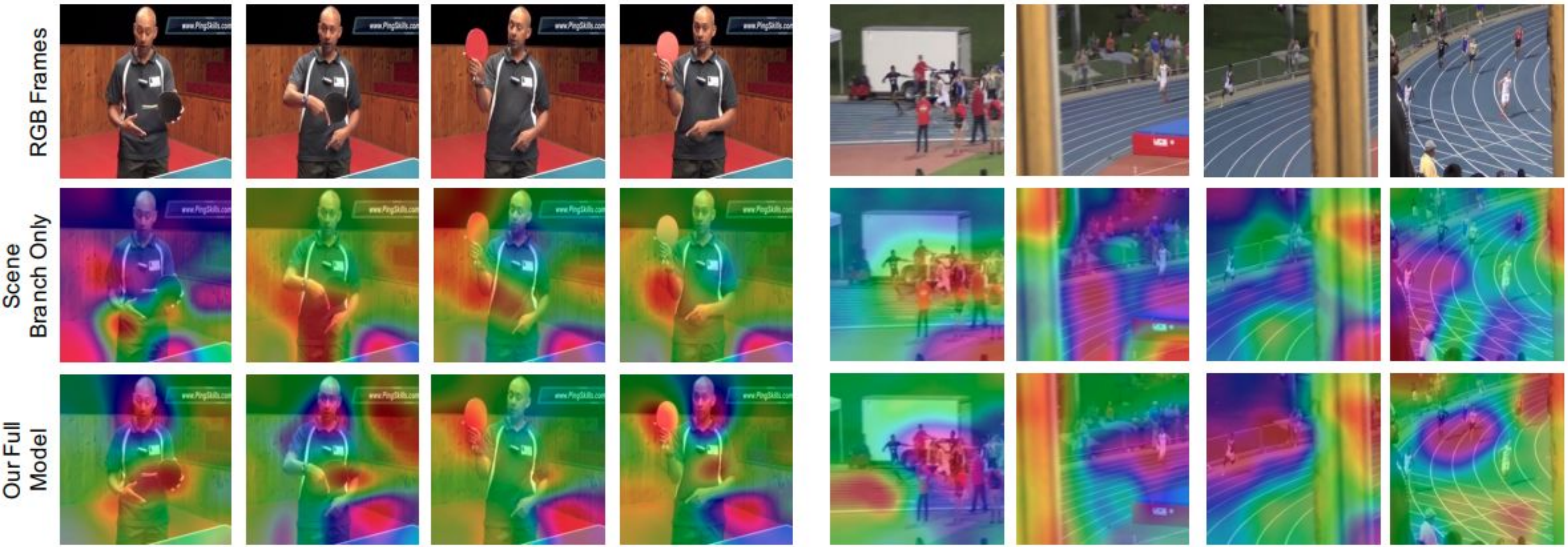
The first section (in the tables) includes methods that optimize language decoding, while the second section is for those that focus on visual encoding.

Quantitative Results (2)

Method	BLEU@4	METEOR	ROUGE-L	CIDEr
Scene Branch Only	45.8	34.3	71.0	86.0
Two Branch + Concat	45.5	34.1	70.7	79.3
Two Branch + L2	46.1	33.7	70.6	80.3
Spatial Graph Only	50.8	36.1	72.9	91.8
Temporal Graph Only	50.7	36.1	73.1	92.1
Dense Graph	51.4	35.9	72.8	91.3
Our Full Model	**52.2**	**36.9**	**73.9**	**93.0**

Table: Ablation study on MSVD

Qualitative Results (1)

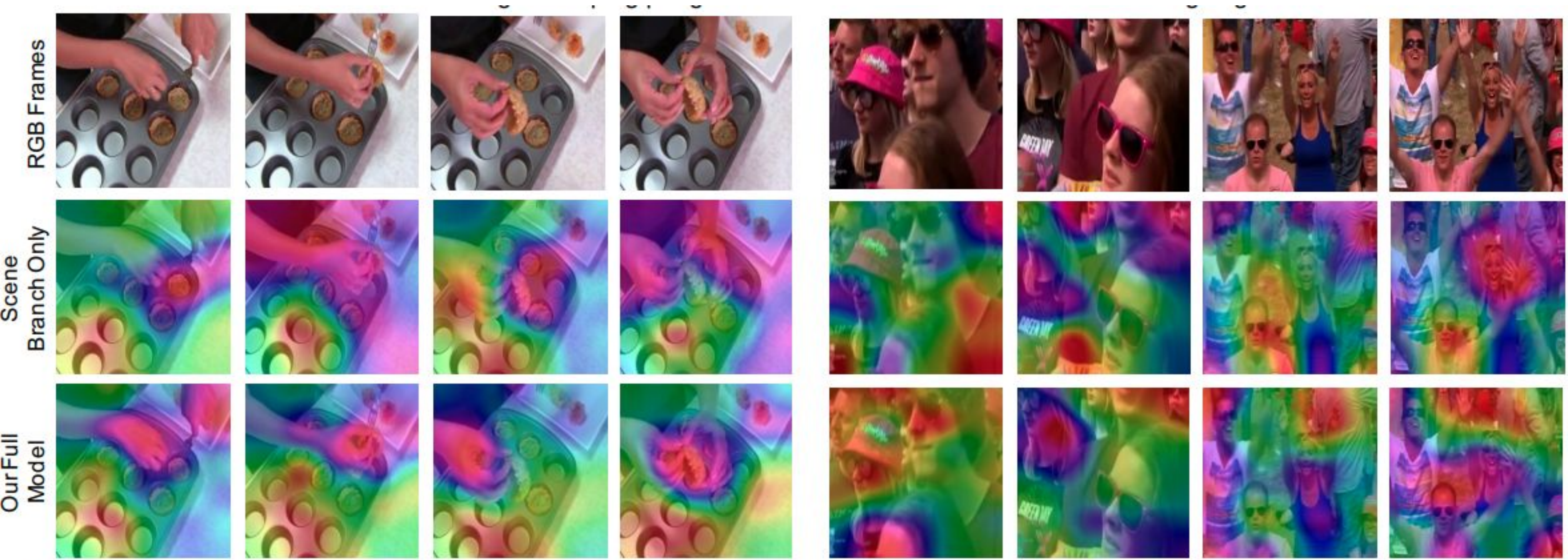


GT: a man in a black shirt demonstrates how to play ping pong
Wang *et al*. [39]: there is a man is talking about table tennis
Ours: a man in a **black shirt** is talking about ping pong

GT: a group of men are running down a race track
Wang *et al*. [39]: there is a man running on the track
Ours: a **race** is going on the track

Red color indicates high attention scores, while blue means the opposite.

Qualitative Results (2)



GT: a woman is showing how to make little baskets from potatoes
Wang *et al.* [39]: a person is preparing a recipe
Ours: a woman is showing how to make a **potato** salad

GT: people are dancing and singing
Wang *et al.* [39]: a man is singing
Ours: **a group of people** are singing and dancing

Red color indicates high attention scores, while blue means the opposite.

Main Contributions

1. Design a **novel spatio-temporal graph network** to perform video captioning by exploiting object interactions.
2. Propose an **object-aware knowledge distillation mechanism** to solve the problem of noisy feature learning that exists in the spatio-temporal graph models.

Thank you