Paper Review:
Object Relational Graph with Teacher-Recommended Learning for Video Captioning

Z. Zhang et al., "Object relational graph with teacher-recommended learning for video captioning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 13275-13285.

Putra Bahy Helmi Hartoyo
Vision Language Intelligence Lab - Sejong University
18.07.2022

Contents

- **Background**
- **Main Contribution**
- **Architecture**
- **Proposed Modules**
  - Object Relational Graph
  - Teacher Recommended Learning via ELM
- **Results**
  - Quantitative Results
  - Qualitative Results
- **Conclusion**

Background

- Information from both **vision** and **language** is important in video captioning task.
- Existing models **lack of adequate visual representation**.
    - Neglecting the explicit interactions between objects in the spatial/temporal domain.
- In the caption corpus, it is found that the **majority** of words are **function words** and **common words** e.g. “the” and “man” than the real **content-specific words**.
    - Called as a **long-tailed problem**.

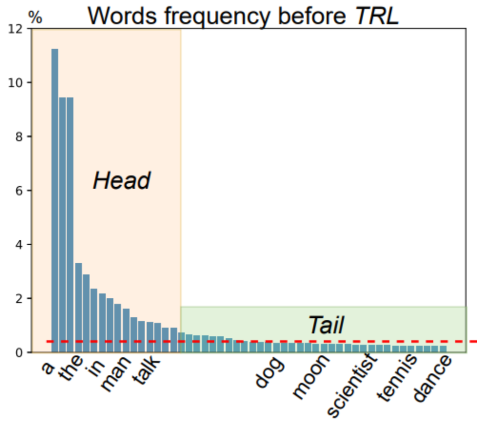


Figure 1: Long-tailed problem observed in corpus of MSR-VTT

Main Contribution

- Propose an **object relational graph (ORG)** based encoder, which **captures more detailed interaction features** between objects to enrich visual representation.
- Design a **teacher-recommended learning (TRL)** method to **make full use of** the successful **external language model (ELM)** to integrate the abundant linguistic knowledge into the caption model.

Architecture

- Consisted of 3 main modules; **Object Encoder**, **Teacher-recommended Learning,** and Description Generator.

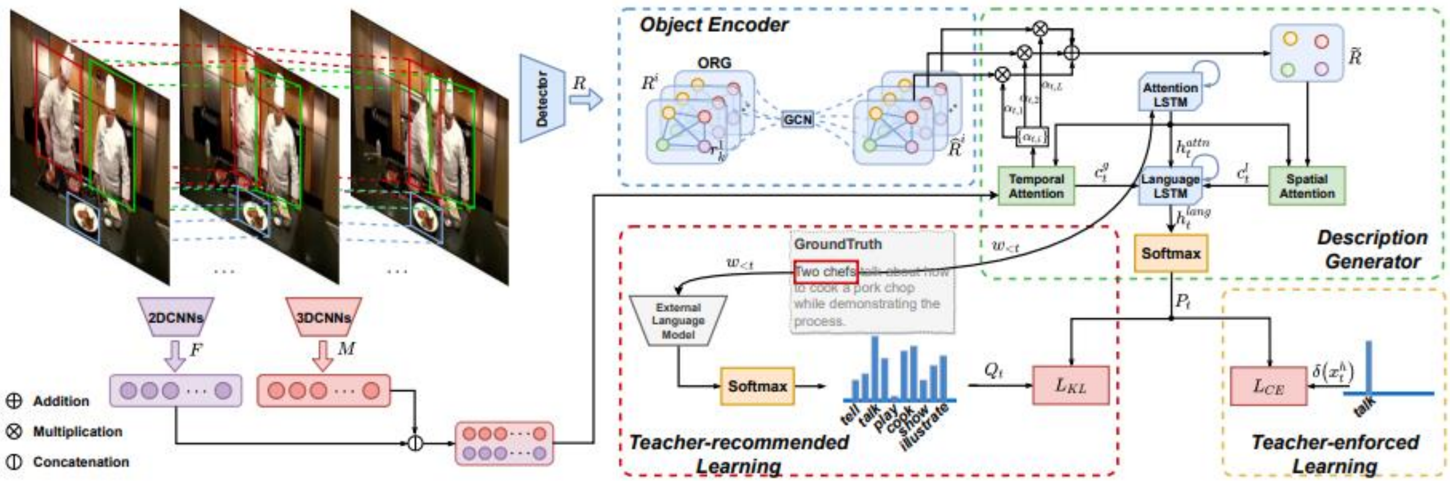


Figure 2: Architecture overview of ORG-TRL

Proposed Modules

- **Object Relational Graph**
  - A **graph-based object encoder** which can learn the interaction among different objects dynamically.
  - This paper proposed two kinds of object relational graph:
    - **Partial** object relational graph (P-ORG)
    - **Complete** object relational graph (C-ORG)
  - The difference among the two is, **P-ORG** only consider relationship between objects **in the same frame** while **C-ORG** also accounts the relationship of objects **across all frames**.

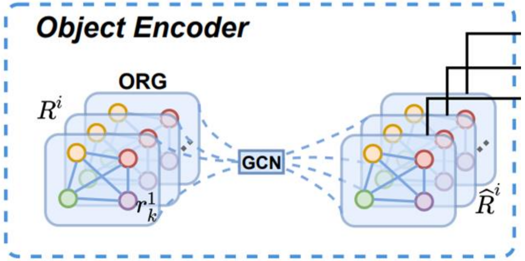
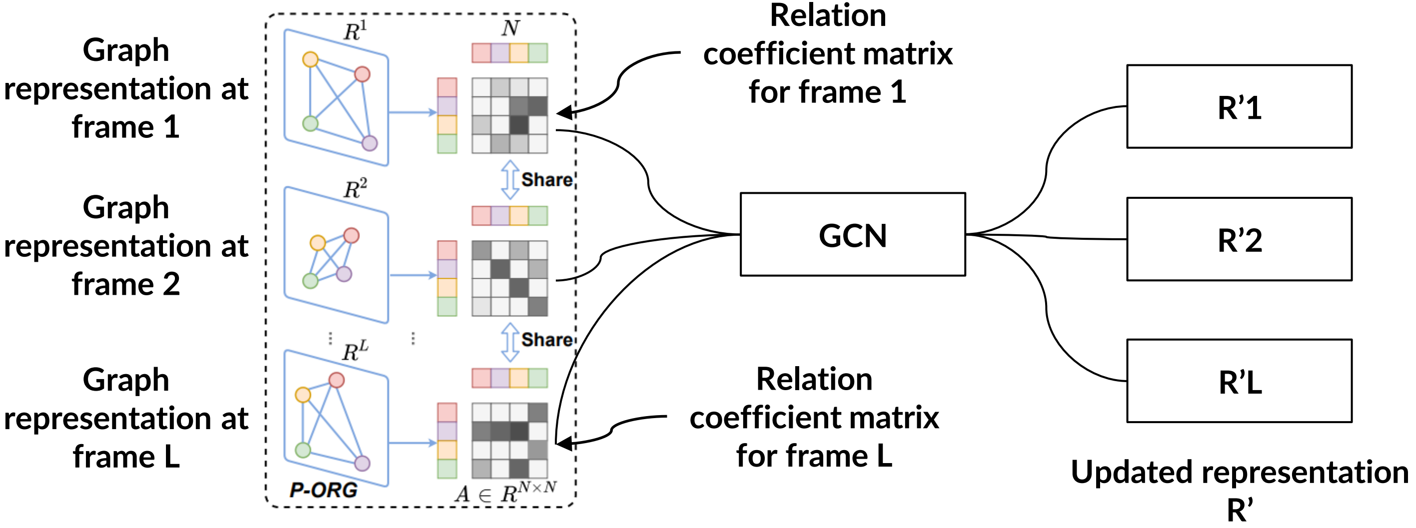


Figure 3: ORG as an object encoder, uses the help of GCN to update its objects representations (R)

Proposed Modules

- **Object Relational Graph**
  - **Partial** object relational graph will have different relation coefficient matrix (A) for each frame. It denotes the relationships between objects in each frame.

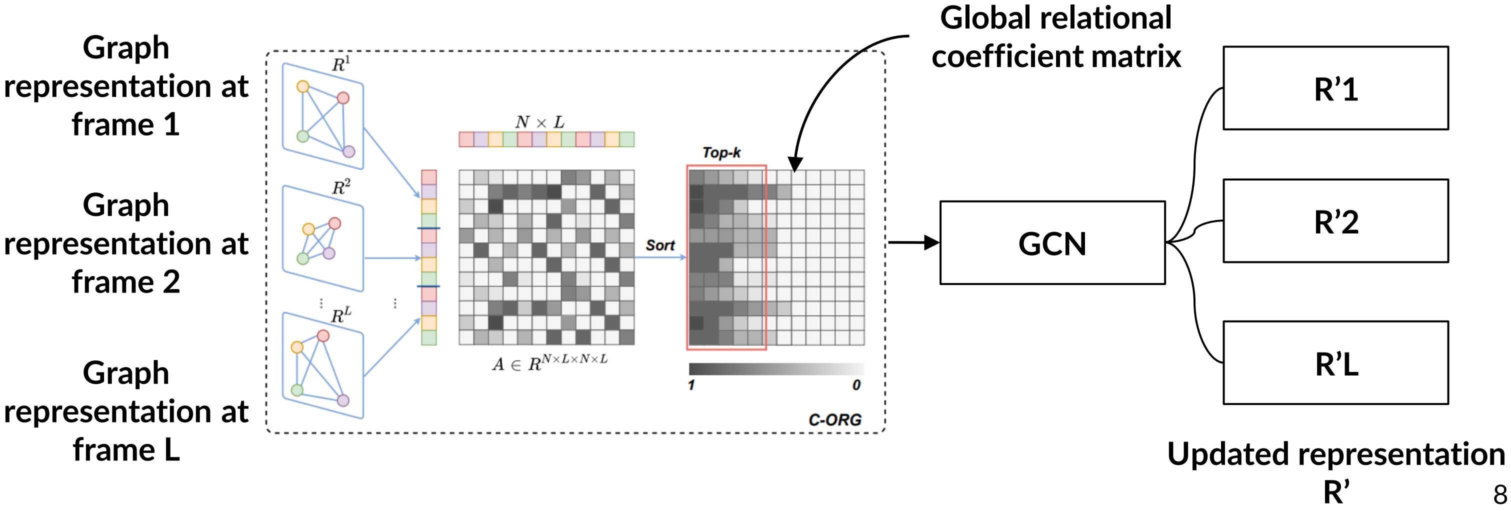
Figure 4: P-ORG, each frame will have its own relation coefficient matrix (A)



Proposed Modules

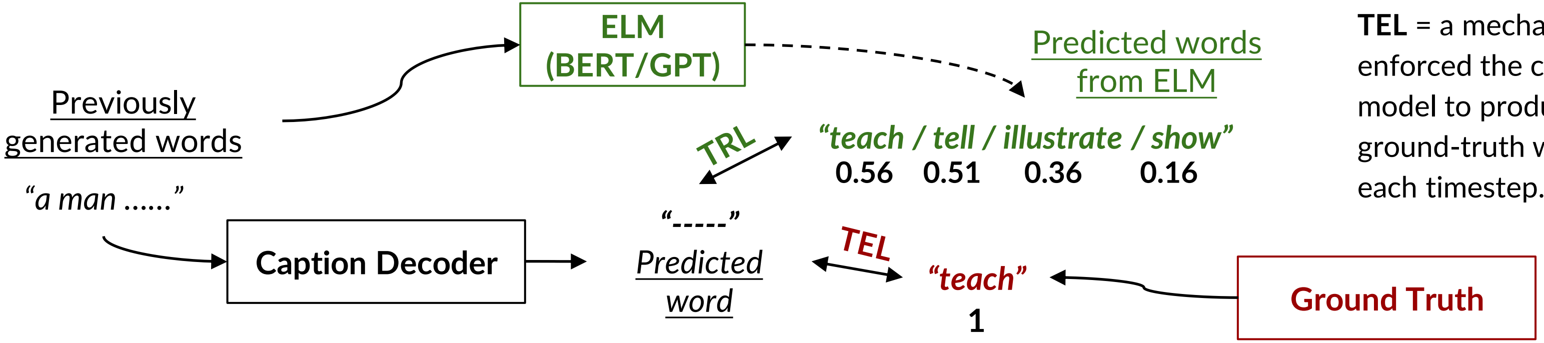
- **Object Relational Graph**
    - **Complete** object relational graph connects all objects in the video in all time frames by creating a single relational coefficient matrix (A).

Figure 4: C-ORG, will have single relation coefficient matrix (A) that stores all the objects relationships



Proposed Modules

- **Teacher Recommended Learning (TRL) via ELM**
  - A module that improves the common teacher-enforced learning (TEL) mechanism by exploiting external language model (ELM).
  - ELM provides rich choices of words and can provides more options to the captioning model by providing ***soft targets*** of all possible words.
  - There are many ready-made models that can be used as an ELM for TRL.



TEL = a mechanism that enforced the captioning model to produce ground-truth word at each timestep.

Training Objectives

- Cross entropy loss (TEL) + KL divergence loss (TRL)

Tradeoff parameter

$$\mathcal{L}(\theta) = \lambda \mathcal{L}_{KL}(\theta) + (1-\lambda)\mathcal{L}_{CE}(\theta)$$

Possible word *d* probability from ELM

Ordinary TEL (CE Loss)

$$\mathcal{L}_{CE}(\theta) = -\sum_{t=1}^{T} \delta(x_t^h)^T \cdot log P_t$$

Proposed TRL (KLD Loss)

$$\mathcal{L}_{KL}(\theta) = -\sum_{t=1}^{T} \sum_{d \in \boldsymbol{x}_t^s} Q_t^d \cdot log P_t^d$$

Ground truth word probability

Predicted word probability

Results

- **Qualitative results**

GT: a woman is mixing something in a bowl
Baseline: there is a woman is making a dish
ORG-TRL: a person is mixing some food in a bowl

Effects of ORG:
- Detects more detailed objects.
- Recognize explicit interactions between objects, i.e. person -> "mixing" -> some food.

Effects of TRL:
- Supply the model with words that rarely appears in captioning dataset, i.e. "*climate change*".
- Give richer choice of words.

GT: narrator talks about some people not believing in climate change

[EOS]	and	the	to	[UNK]
0.531	0.031	0.026	0.021	0.0170

change	effect	[EOS]	country	weather
0.673	0.072	0.055	0.005	0.004

Results

- **Quantitative results**
  - Achieved **competitive results** in both MSVD and MSR-VTT dataset.

Models	Year	Features			MSVD				MSR-VTT			
		Appearence	Motion	Object	B@4	M	R	C	B@4	M	R	C
SA-LSTM [38]	2018	Inception-V4	-	-	45.3	31.9	64.2	76.2	36.3	25.5	58.3	39.9
M3 [40]	2018	VGG	C3D	-	52.8	33.3	-	-	38.1	26.6	-	-
RecNet [38]	2018	Inception-V4	-	-	52.3	34.1	69.8	80.3	39.1	26.6	59.3	42.7
PickNet* [6]	2018	ResNet-152	-	-	52.3	33.3	69.6	76.5	41.3	27.7	59.8	44.1
MARN [27]	2019	ResNet-101	C3D	-	48.6	35.1	71.9	92.2	40.4	28.1	60.7	47.1
SibNet [21]	2019	GoogleNet	-	-	54.2	34.8	71.7	88.2	40.9	27.5	60.2	47.5
OA-BTG [53]	2019	ResNet-200	-	Mask-RCNN	**56.9**	36.2	-	90.6	41.4	28.2	-	46.9
GRU-EVE [1]	2019	InceptionResnetV2	C3D	YOLO	47.9	35.0	71.5	78.1	38.3	28.4	60.7	48.1
MGSA [5]	2019	InceptionResnetV2	C3D	-	53.4	35.0	-	86.7	42.4	27.6	-	47.5
POS+CG [36]	2019	InceptionResnetV2	OpticalFlow	-	52.5	34.1	71.3	88.7	42.0	28.2	61.6	48.7
POS+VCT [12]	2019	InceptionResnetV2	C3D	-	52.8	36.1	71.8	87.8	42.3	**29.7**	**62.8**	49.1
ORG-TRL	Ours	InceptionResnetV2	C3D	FasterRCNN	54.3	**36.4**	**73.9**	**95.2**	**43.6**	28.8	62.1	**50.9**

Results

- **Quantitative results (Ablation study)**
  - The **presence of ORG or TRL** or the combination of both in the architecture **shows performance increment** compared to the Baseline, in both dataset.

Methods		MSVD				MSR-VTT			
ORG	TRL	B@4	M	R	C	B@4	M	R	C
×	×	53.3	35.2	72.4	91.7	41.9	27.5	61.0	47.9
✓	×	54.0	36.0	73.2	94.1	43.3	28.4	61.5	50.1
×	✓	54.0	36.0	73.7	93.3	43.2	28.6	61.7	50.4
✓	✓	**54.3**	**36.4**	**73.9**	**95.2**	**43.6**	**28.8**	**62.1**	**50.9**

Conclusion

- This paper has proposed **a novel architecture** by **modeling** object interaction in video with a graph-based encoder, called **object relational graph (ORG)**.
- Another contribution of this paper is to proposed a **teacher-recommended learning (TRL)** which **exploit** a well-trained **external language model (ELM)** to enhance the vocabulary of the captioning model.
- The effectiveness of these two modules has successfully proven by **showing competitive results** in both MSVD and MSR-VTT dataset.