
Paper Review:

Open-book Video Captioning with Retrieve-Copy-Generate Network

Velda Vania
Vision Language Intelligence Lab - Sejong University
31.08.2022

Zhang, Z., Qi, Z., Yuan, C., Shan, Y., Li, B., Deng, Y. and Hu, W., 2021. Open-book video captioning with retrieve-copy-generate network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9837-9846)

Contents

- Motivation
- Main Contribution
- Proposed Method
 - Overview
 - Retrieve-Copy-Generate (RCG) Network
 - Video-to-Text Retriever (VTR)
 - Copy-mechanism Caption Generator
- Results
- Conclusion

Motivation

- Most existing works have some drawbacks:
 - The caption generation process lacks appropriate guidance since the video is the only source of input which resulting in the generations of more generic sentences
 - Knowledge domain of model is fixed after training and cannot be expanded unless retraining the model

Motivation

- To address these issues, instead of performing video captioning (VC) task directly, the author propose to convert it into two stages:
 - First, perform video-text retrieval (VTR) to search sentences relevant to given video from text corpus
 - Then, the retrieval sentences are used as extra guidance for caption generation
- During inference, the generator can generate words based on video content or directly copy expressions from retrieved sentences

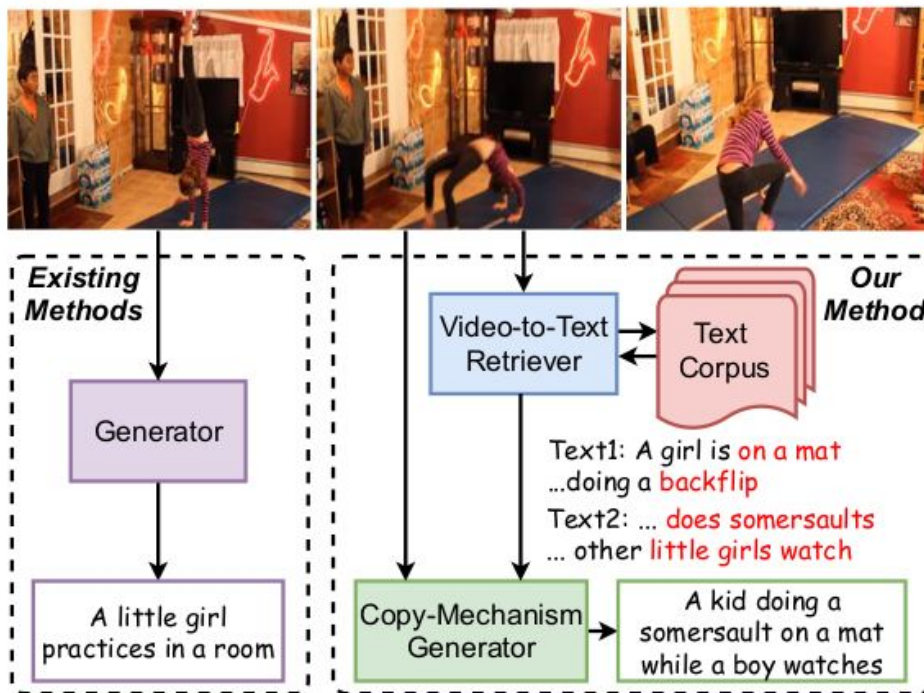


Fig 1. Comparison of existing methods and the author's method

Main Contribution

- Propose to solve the video captioning task with open-book paradigm, which generates captions under the guidance of video-content-retrieval sentences
- Introduce novel Retrieve-Copy-Generate / RCG
 - Improved cross modal retrieval is utilized to provide hints for generator
 - Copy-mechanism generator is proposed for dynamical copying and better generation
- Extensive experiments show that the proposed approach achieves state-of-the-art results in VATEX and competitive performance on MSR-VTT

Proposed Methods - Overview

Pipeline of proposed method (RCG) consist of : **VTR** that searches for video content relevant sentences from corpus containing all the sentences in the training set **Copy-mechanism Generator** produces the words by steps under the hints or guidance of retrieved sentences and the visual features

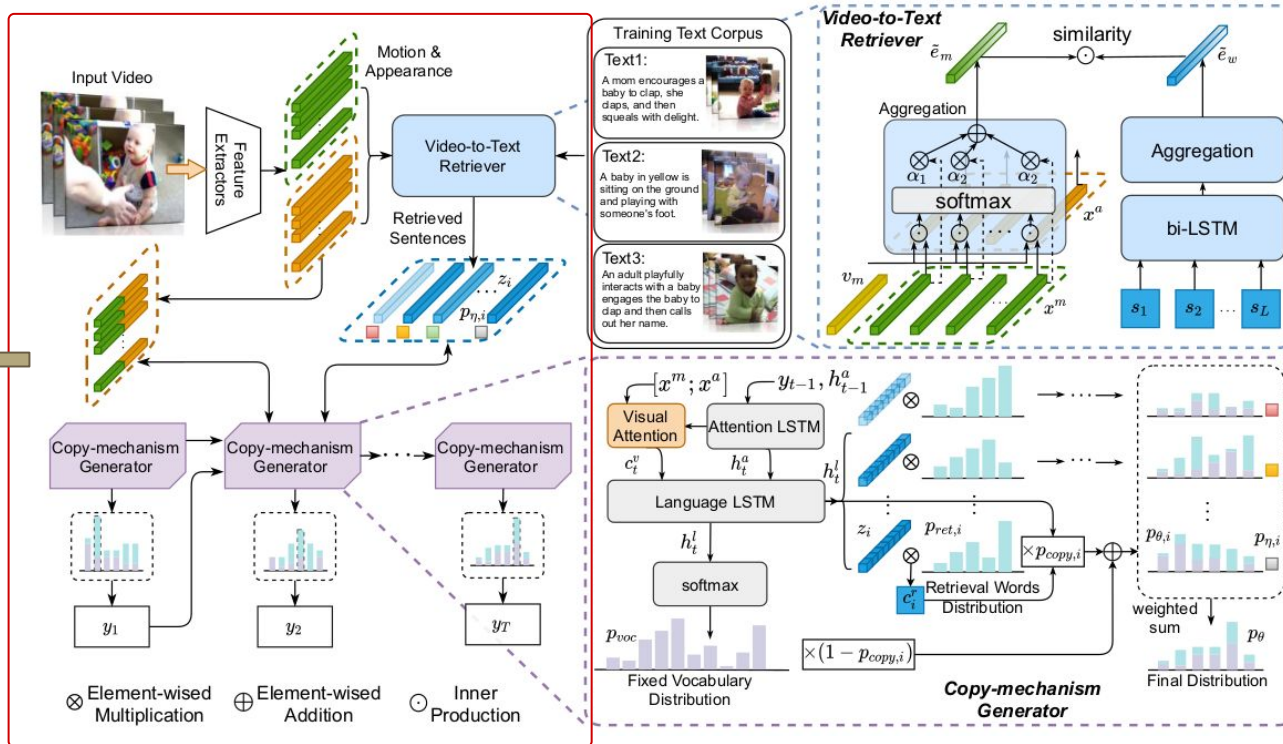
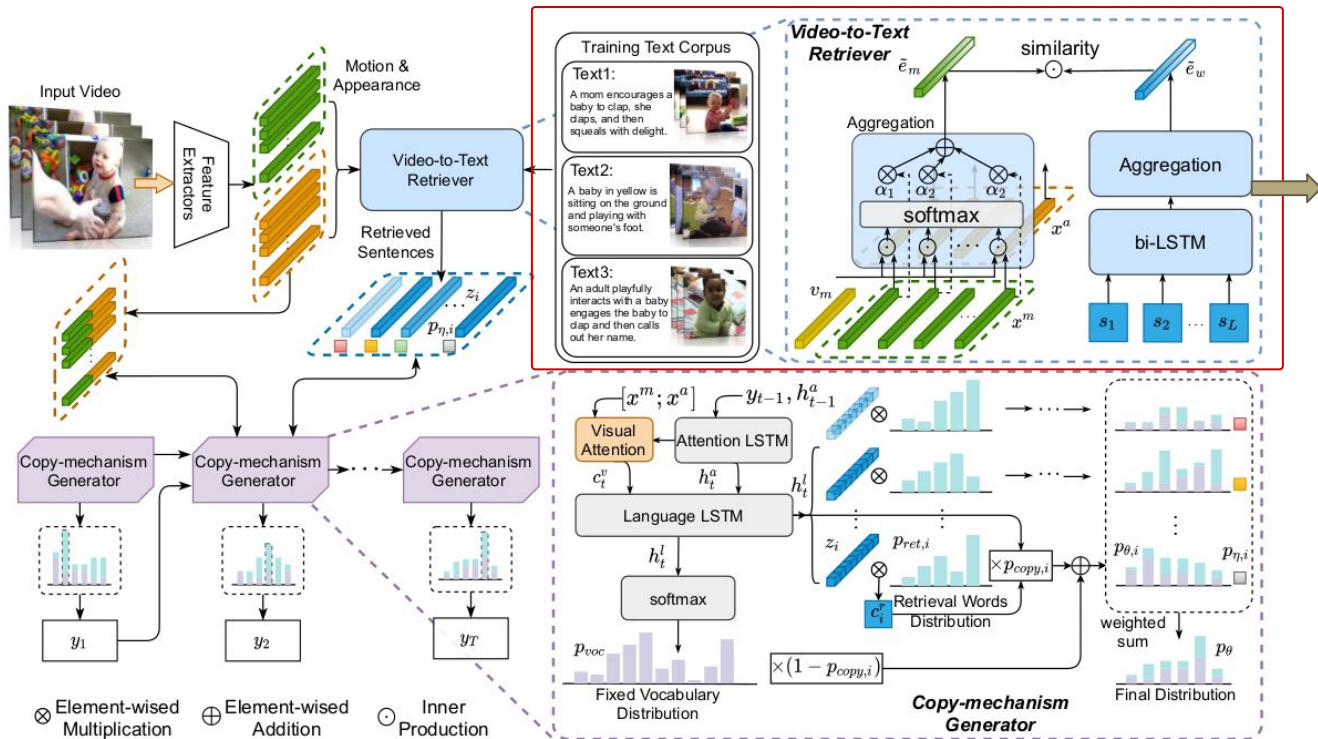


Fig 2. Overview of the proposed RCG Network for Open-book Video Captioning

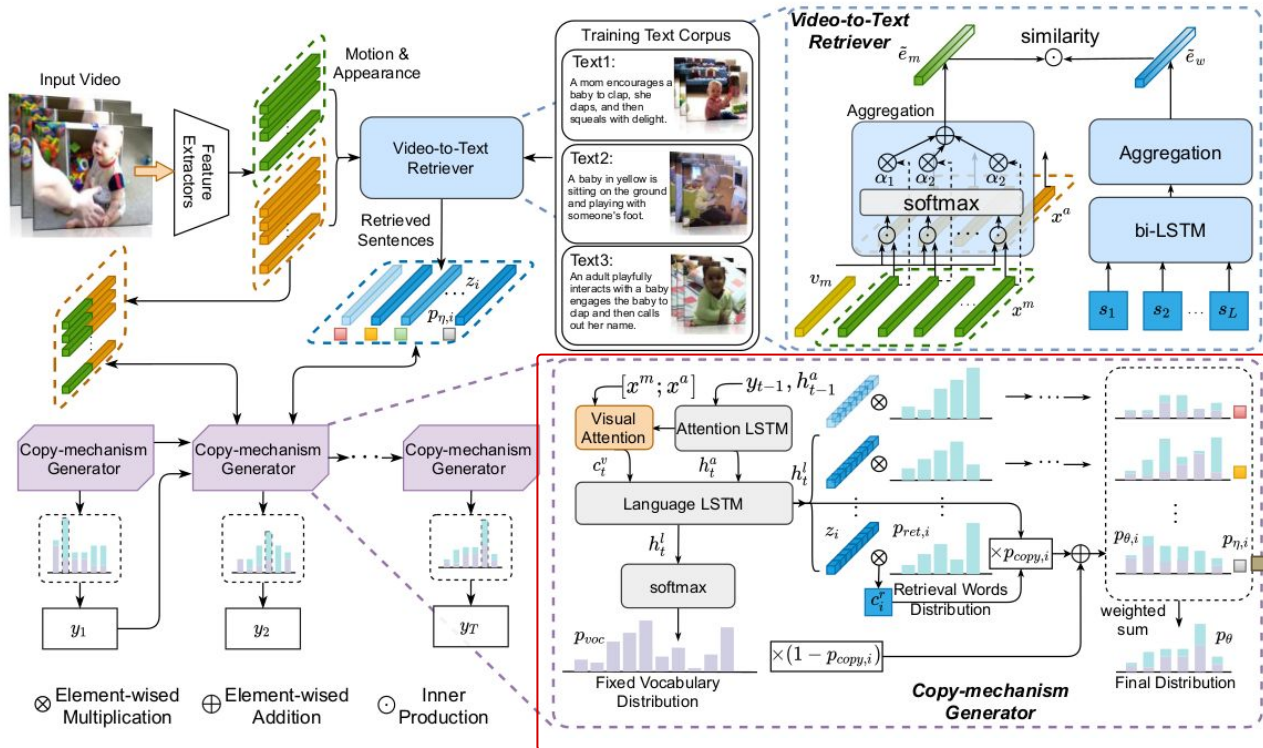
Proposed Methods - Overview



The Bi-encoders architecture is leveraged to efficiently and effectively achieve the cross-modal retrieval

Fig 2. Overview of the proposed Retrieve-Copy-Generate (RCG) Network for Open-book Video Captioning

Proposed Methods - Overview

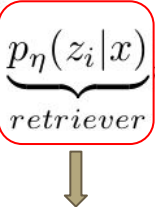


A hierarchical caption decoder is used to generate the fixed vocabulary based on the video content. Meanwhile, an improved multi-pointer module directly copies the expressions from the retrieved sentences for a better generation

Fig 2. Overview of the proposed Retrieve-Copy-Generate (RCG) Network for Open-book Video Captioning

Proposed Methods - RCG Network

- Consists of two components :
 - **Video-to-text retriever**
 - Retrieves top k semantically similar sentences z according to video x
 - **Copy-mechanism generator**
 - Conditioned on the retrieved sentence , the original visual information and previous generated tokens to generate the current target token
 - Formally, the conditional probability of producing caption are defined as:

$$p(y|x) = \prod_{t=1}^T \sum_{i=1}^{topk} \underbrace{p_{\eta}(z_i|x)}_{\text{retriever}} \underbrace{p_{\theta}(y_t|z_i, x, y_{1:t-1})}_{\text{generator}}$$


Represents the confidence of whether the generator can copy words directly from retrieval sentences or not

η : VTR's parameter
 θ : copy-mechanism's parameter
 z_i : retrieved sentence
 x : video visual information
 t : time step
 y_t : generated token at t

Proposed Methods - RCG Network - VTR

- Major function : find the top-k most similar sentences z given video x in a massive retrieval corpus
 - The corpus contains the whole sentences of training set
 - During training and testing, the sentences are retrieved only from the corpus of training set
 - In addition, sentences belonging to own video are excluded
- VTR applies bi-encoder architecture:
 - Text encoder: maps all sentences in corpus into vectors
 - Visual encoder: maps video x into vectors
- VTR is pre-trained using contrastive learning
 - Each positive video-retrieved sentence pair (x^+, z^+) should be closer than any other negative video-retrieved sentence pairs (x^+, z^-) and (x^-, z^+) in a mini batch

Proposed Methods - RCG Network - VTR

- Thus, given video x , the probability of retrieved sentence z_i is estimated as:

$$p_{\eta}(z_i|x) = \text{softmax}(\text{sim}(x, z_i)), z_i \in \{z_1, \dots, z_{\text{topk}}\}.$$

$$e_x = \frac{\text{Agg}(x_m) + \text{Agg}(x_a)}{2}$$

$$\text{sim}(x, z_i) = \frac{e_{z_i} \cdot e_x}{||e_{z_i}|| ||e_x||}$$

z_i : retrieved sentence

x : video

topk : number of most relevant sentences of x

x_m : motion features of video x

x_a : appearance features of video x

$\text{Agg}(\cdot)$: aggregation function

e_{z_i} : embedding of z_i

Proposed Methods - RCG Network - Copy-mechanism Caption Generator

- In generation process, the words of multiple sentences are copied adaptively where the proportion of copy word or generate the word is adjusted according to the context
- Consists of two components :
 - Hierarchical Caption Decoder
 - The attention LSTM focus on different visual features according to the current hidden state to achieve the visual context
 - The language LSTM aggregate the current state and visual context to generate the probability distribution of the fixed vocabulary p_{voc} at each time step

Proposed Methods - RCG Network - Copy-mechanism Caption Generator

- Dynamic Multi-pointers Module
 - An improved version of Pointer-Networks where the module output word probability distribution p_{ret} corresponding on each retrieved sentence
 - Since, not all the words in retrieved sentence are valid, the model decide to copy or generate the word dynamically based on probability.
 - The probability of copying words p_{copy} from each retrieved sentence are determined by the semantic context of retrieved sentence and the decoder state

Proposed Methods - RCG Network - Copy-mechanism Caption Generator

- In summary, the generation probability distribution p_{θ} is :

$$p_{\theta} = (1 - p_{copy})p_{voc} + p_{copy}p_{ret}$$

p_{copy} : copy probability

p_{ret} : dynamic words in retrieved sentences

p_{voc} : probability distribution of fixed vocabulary

- Thus, the training goal is to minimize the negative log-likelihood of each target word y_t :

$$\mathcal{L}_{gen} = - \sum_{t=1}^T \log \sum_{i=1}^{topk} p_{\eta}(z_i) p_{\theta,i}(y_t)$$

p_{θ} : generation probability distribution

p_{η} : probability similarities of the retrieved sentences

z_i : retrieved sentences index i

t : timestep

y_t : target word

Results

- Quantitative results on MSR-VTT and VATEX dataset

Dataset	Method	Ref.	CIDEr	BLEU-4	Rouge-L	Meteor
MSR-VTT	POS-CG[28]	ICCV19	48.7	42.0	61.6	28.2
	POS-VCT[13]	ICCV19	49.1	42.3	62.8	29.7
	SAAT[37]	CVPR20	49.1	40.5	60.9	28.2
	+RL		51.0	39.9	61.2	27.7
	STG-KD[22]	CVPR20	47.1	40.5	60.9	28.3
	PMI-CAP[3]	ECCV20	49.4	42.1	-	28.7
	+Audio		50.6	43.9	-	29.5
	ORG-TRL[36]	CVPR20	50.9	43.6	62.1	28.8
	Baseline	Ours	49.8	42.2	61.2	28.2
	+FixRet	Ours	52.3	43.1	61.9	29.0
	+TrainRet	Ours	52.9	42.8	61.7	29.3
VATEX	VATEX[29]	ICCV19	45.6	28.7	47.2	21.9
	ORG-TRL[36]	CVPR20	49.7	32.1	48.9	22.2
	NSA[8]	CVPR20	57.1	31.0	49.0	22.7
	Baseline	Ours	49.2	31.3	48.5	21.9
	+FixRet	Ours	56.8	33.4	50.1	23.6
	+TrainRet	Ours	57.5	33.9	50.2	23.7

Results

- Qualitative results on VATEX dataset



Ground-Truth	1-A man is scuba diving in the ocean while exploring sea creatures 2-A woman is scuba diving deep in the red sea with someone else 3-A woman is under water with her breathing apparatus on
Retrieved Sentences	1- Two scuba divers swim beneath the surface of ocean water 0.627 2- Scuba divers are visible swimming under the water of the ocean 0.616 3- Scuba divers are shown at the bottom of the water and breathing out of their masks 0.613
Baseline	a person is scuba diving in the ocean with a lot of fish
RCG	scuba divers are swimming at the base of the water and breathing out of their masks



Ground-Truth	1-A man is shown rolling a ball at another person, the person kicks the ball back 2-A man throws a ball to a boy, who kicks it back to the man, who catches it 3-a man wearing blue throws a red ball to a boy
Retrieved Sentences	1- A man and a boy are outside playing catch while another child looks on 0.536 2- A group of people are playing catch with a ball in a park 0.532 3- A family together in the park as they toss each other a baseball and catch it having fun 0.521
Baseline	a group of people are playing a game of frisbee in a park
RCG	a man standing on the grass playing catch with a ball and a child looks him

- The attention weights and **similarities** between the video of top-3 retrieved sentences are shown.
- The baseline model can generate some **correct words** and **wrong words**. Compared with it, RCG can generate more diverse captions by copying the **expressions** from retrieved sentences.

Conclusion

- This paper introduce new paradigm in video captioning domain
 - Model generates caption under the prompts of video-content relevant sentences which is not limited to the video itself
- Extensive experiments on several datasets show the effectiveness and promising of the proposed paradigm in video captioning task
- In addition, the authors' result and ablation study show that it is practical to copy knowledge not limited to retrieved sentences, e.g video subtitles for better generation

Thank You