Dual Encoding for Zero-Example Video Retrieval

Jianfeng Dong[1], Xirong Li[2], Chaoxi Xu[2], Shouling Ji[3], Yuan He[4], Gang Yang[2], and Xun Wang[1]

[1]Zhejiang Gangshang University [2]Renmin University of China [3]Zhejiang University [4]Alibaba

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Introduction

In **zero-example video retrieval (ZEVR),** an end user searches for unlabeled videos by ad-hoc queries described in natural language text with no visual example provided.

How to properly associate visual and linguistic information presented in temporal order?

State-of-the-Art

Two types of methods

- The majority are concept based
  - ✓ Representing both video and text by concept vectors
  - ✓ Challenges exist in concept detection, selection and representation
- Few works consider deep learning
  - ✓ Lack of multi-level encoding
  - ✓ Lack of unified encoding

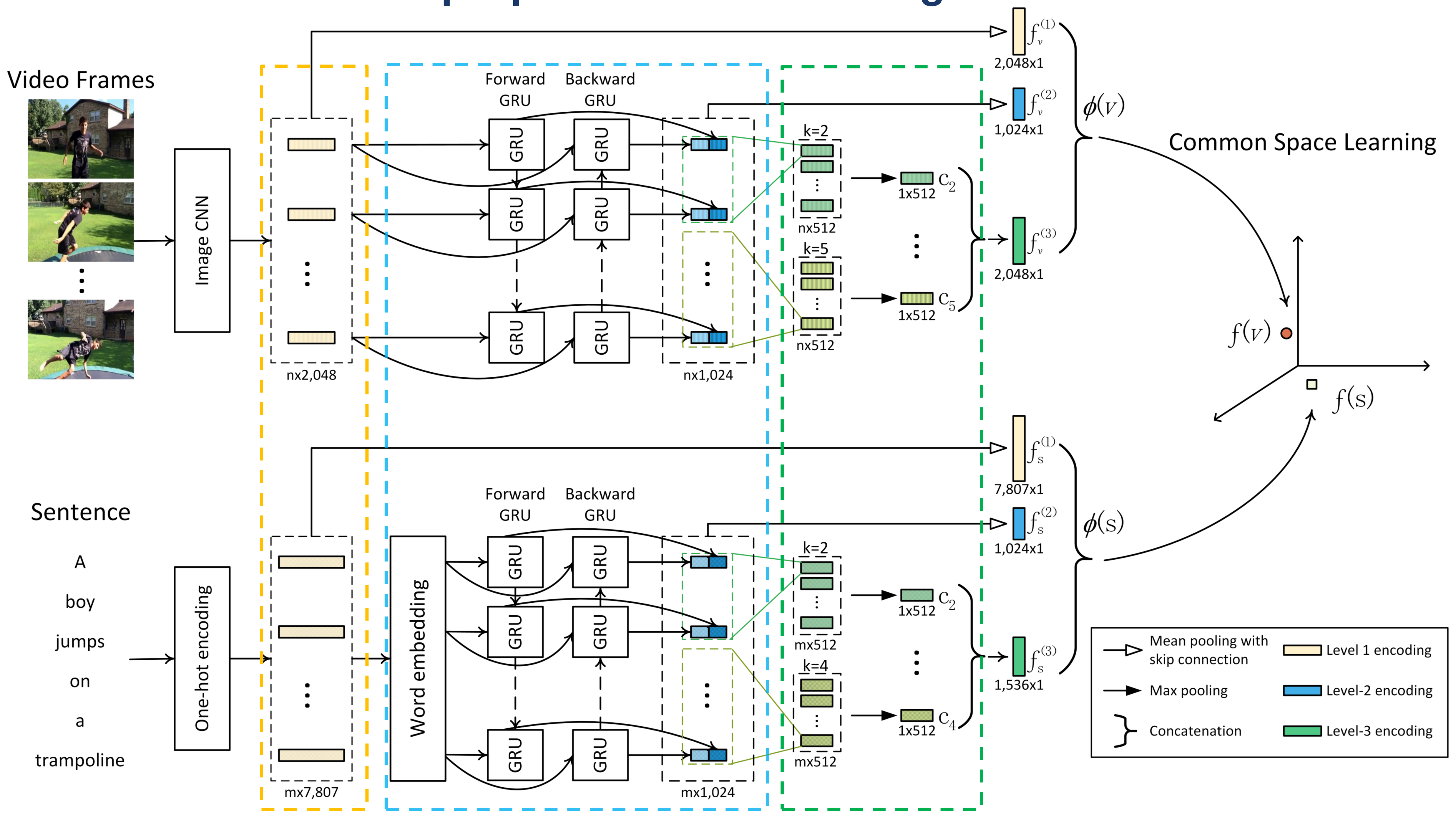
Method	Video-side encoding	Text-side encoding
Xu *et al.* AAAI15	Mean pooling	Recursive Neural Networks
Habibian *et al.* T-PAMI17	Mean pooling	Bag-of-Words (BoW)
Yu *et al.* CVPR17	LSTM	LSTM
Yu *et al.* ECCV18	CNN	bi-LSTM
Mithun *et al.* ICMR18	Mean pooling	GRU
Dong *et al.* T-MM18	Mean pooling	[BoW; Word2Vec; GRU]

Mean pooling over frame-level CNN features

https://github.com/danieljf24/dual_encoding

xirong@ruc.edu.cn dongjf24@gmail.com

Our proposal: Dual encoding network



Level 1 Global : To capture visual patterns repeatedly present in the video frames

Level 2 Temporal-aware : To model the temporal information of the frame sequence

Level 3 Local-enhanced : To enhance local patterns that help discriminate subtle differences

Experiments

- Is multi-level encoding better than single-level encoding?

Encoding strategy	Text-to-Video Retrieval					Video-to-Text Retrieval					Sum of Recalls
	R@1	R@5	R@10	Med r	mAP	R@1	R@5	R@10	Med r	mAP	
Level 1 (Mean pooling)	6.4	18.8	27.3	47	0.132	11.5	27.7	38.2	22	0.054	129.9
Level 2 (biGRU)	6.3	19.4	28.5	38	0.136	10.1	26.8	37.7	20	0.057	128.8
Level 3 (biGRU-CNN)	7.3	21.5	31.2	32	0.150	10.6	27.3	38.5	20	0.061	136.4
Level 1 + 2	6.9	20.4	29.1	41	0.142	11.6	29.6	40.7	18	0.058	138.3
Level 1 + 3	7.5	21.6	31.2	33	0.151	11.9	30.5	41.7	16	0.062	144.4
Level 2 + 3	7.6	**22.4**	**32.2**	**31**	**0.155**	11.9	**30.9**	42.7	16	**0.066**	147.7
Level 1 + 2 + 3	**7.7**	22.0	31.8	32	**0.155**	**13.0**	30.8	**43.3**	**15**	0.065	**148.6**

- Is dual encoding better than single-side encoding?

Video-side	Text-side	Sum of Recalls
Mean pooling	Multi-level encoding	137.1
Multi-level encoding	Bag-of-words	143.6
Dual encoding		**148.6**

- Comparison to SOTA on TRECVID'16 / '17 Ad-hoc Video Search

Method	infAP
Top-3 TRECVID finalists:	
Snoek *et al.* [28]	0.206
Ueki *et al.* [30]	0.159
Nguyen *et al.* [25]	0.120
Literature methods:	
Habibian *et al.* [10]	0.150
W2VV$_{imrl}$	0.165
Dual encoding	**0.208**

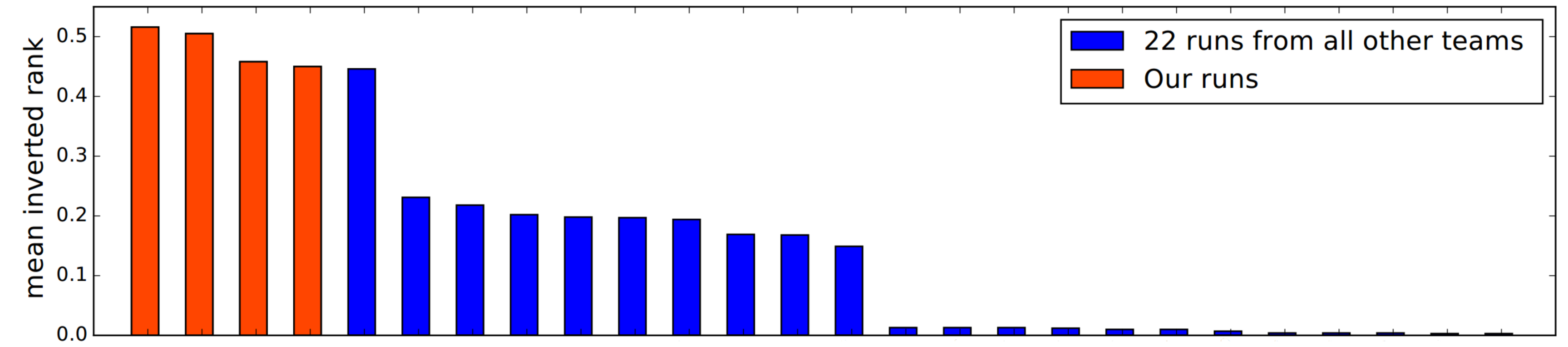
Method	infAP
Top-3 TRECVID finalists:	
Le *et al.* [15]	0.054
Markatopoulou *et al.* [22]	0.051
Liang *et al.* [18]	0.040
Literature methods:	
Habibian *et al.* [10]	0.087
Markatopoulou *et al.* [21]	0.064
W2VV$_{imrl}$	0.132
Dual encoding	**0.159**

Concept-based methods, mostly

- Comparison to SOTA on TRECVID'18 Video-to-Text Matching

Query video

→ Retrieved text *"Two dogs are playing on beach in a cloudy day"*



Take-home Messages

- One dual network to encode the video and text modalities
- Multi-level encoding plus common space learning is effective for sequence-to-sequence cross-modal matching
- Video-side multi-level encoding is more beneficial when compared with its text-side counterpart