Dual Encoding for Zero-Example Video Retrieval

Jianfeng Dong, Xirong Li*, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, Xun Wang

Zhejiang Gangshang University

Renmin University of China

Zhejiang University

Alibaba Group

Zero-example Video Retrieval (ZEVR)

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

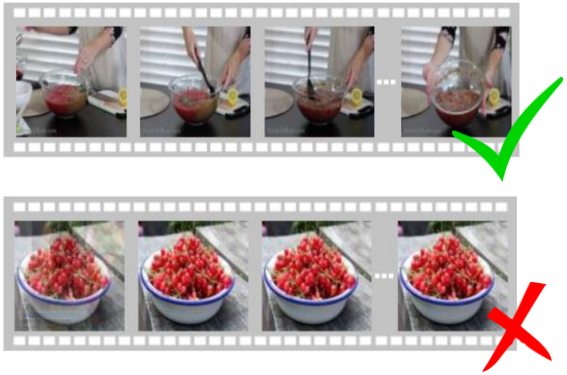
Natural-language query

Someone is making a special fruit punch by adding different types of fruits in a glass bowl

ZEVR

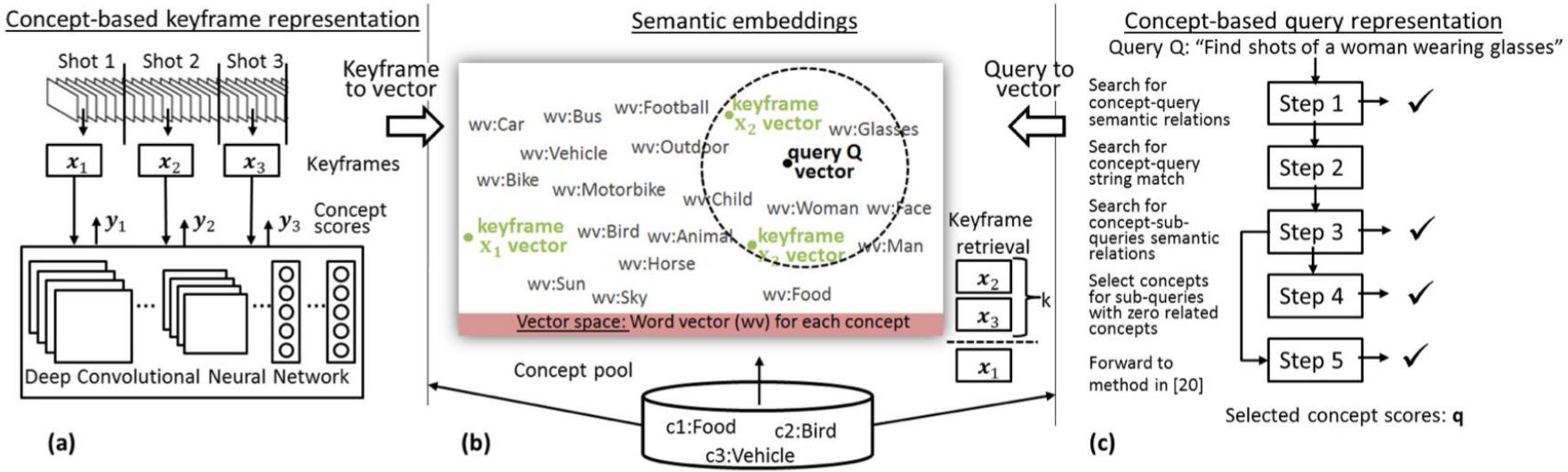
Many unlabeled videos

Retrieved videos



The State-of-the-art (SOTA)

- Representing both video and text by concept vectors
  - Concept detection, selection and representation [Markatopoulou *et al.* ICMR17]



The State-of-the-art (SOTA)

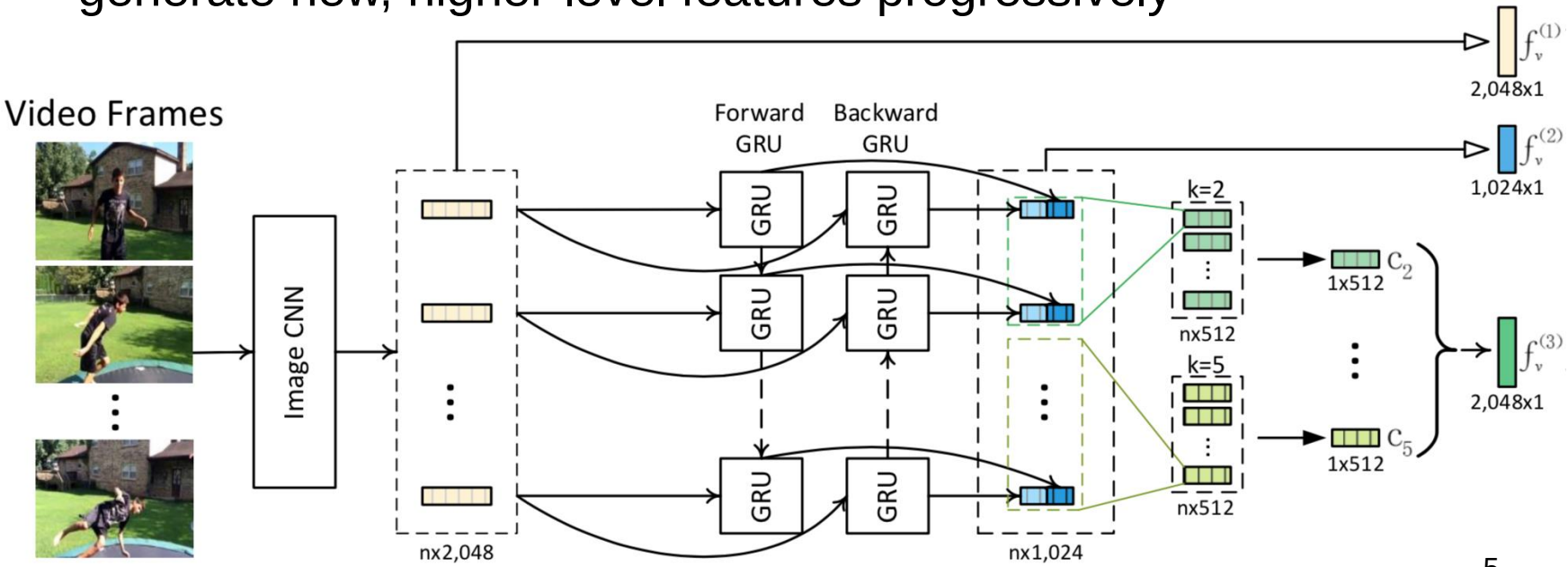
- Few works consider deep learning, but with one encoding strategy

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	Long-Short Term Memory (LSTM)	LSTM
Yu *et al*. ECCV18	CNN	bi-LSTM
Mithun *et al*. ICMR18	Mean pooling	Gated Recurrent Unit (GRU)
Dong *et al.* T-MM18	Mean pooling	[BoW; Word2Vec; GRU]

Mean pooling over frame-level CNN features

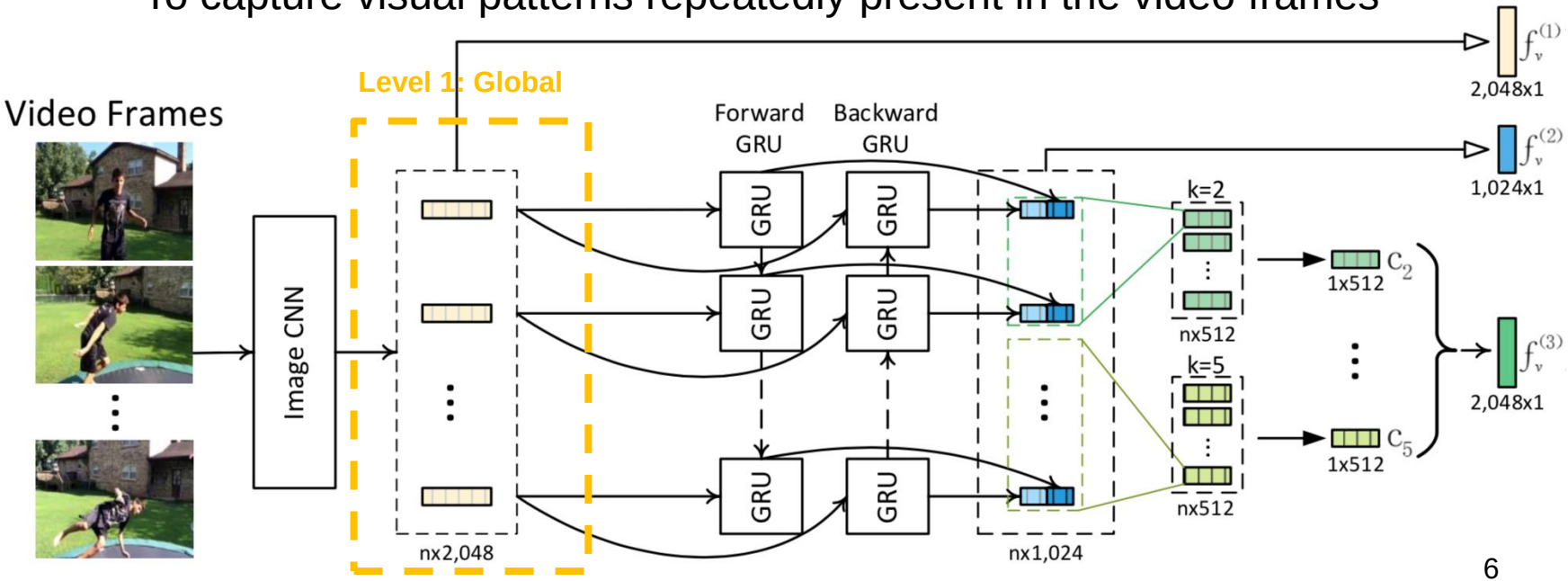
Our Proposal: Video-side Multi-level Encoding

- Given a sequence of frame-level CNN features, we aim to generate new, higher-level features progressively



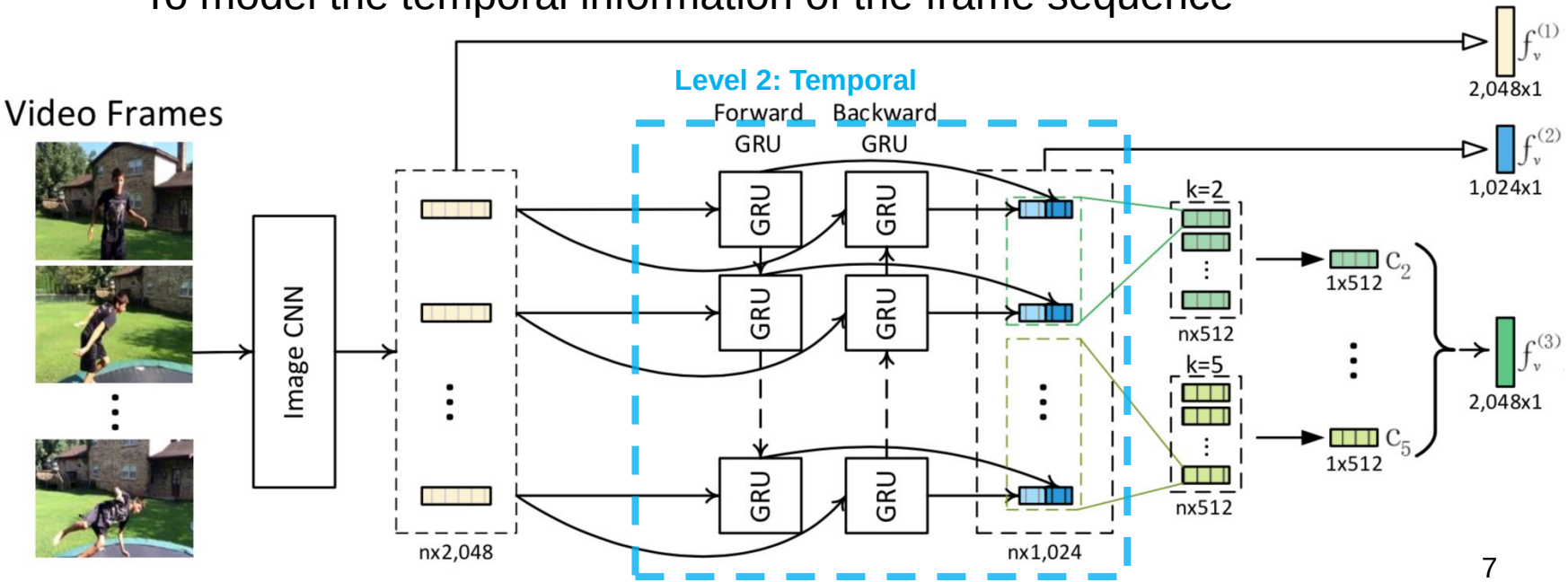
Our Proposal: Video-side Multi-level Encoding

- Level 1: Global encoding by mean pooling
  - To capture visual patterns repeatedly present in the video frames



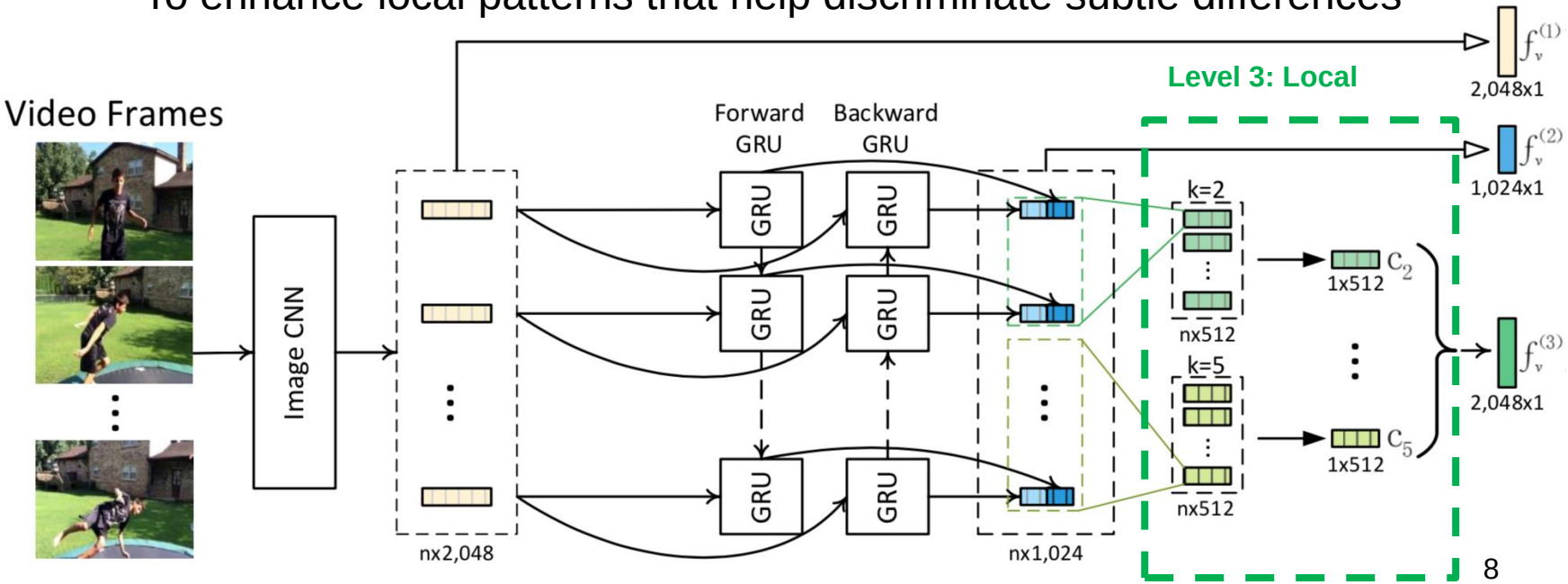
Our Proposal: Video-side Multi-level Encoding

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



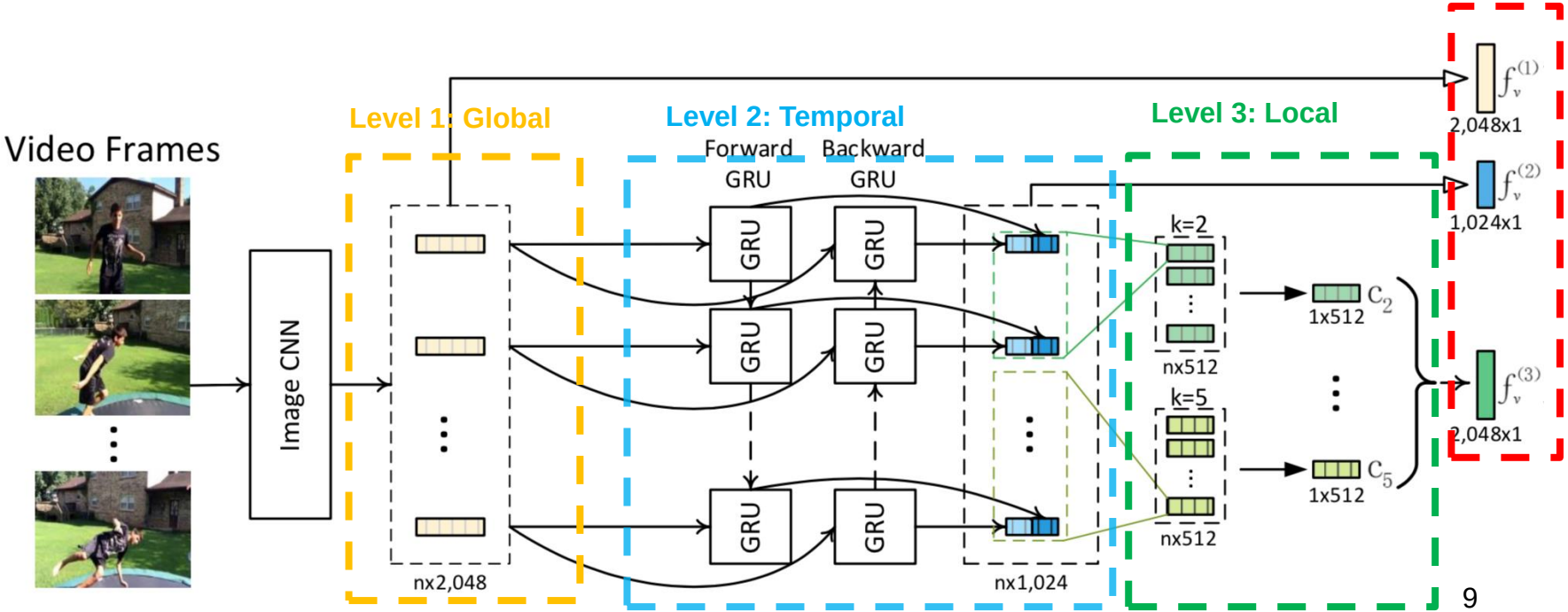
Our Proposal: Video-side Multi-level Encoding

- Level 3: Local-enhanced encoding by biGRU-CNN
  - To enhance local patterns that help discriminate subtle differences



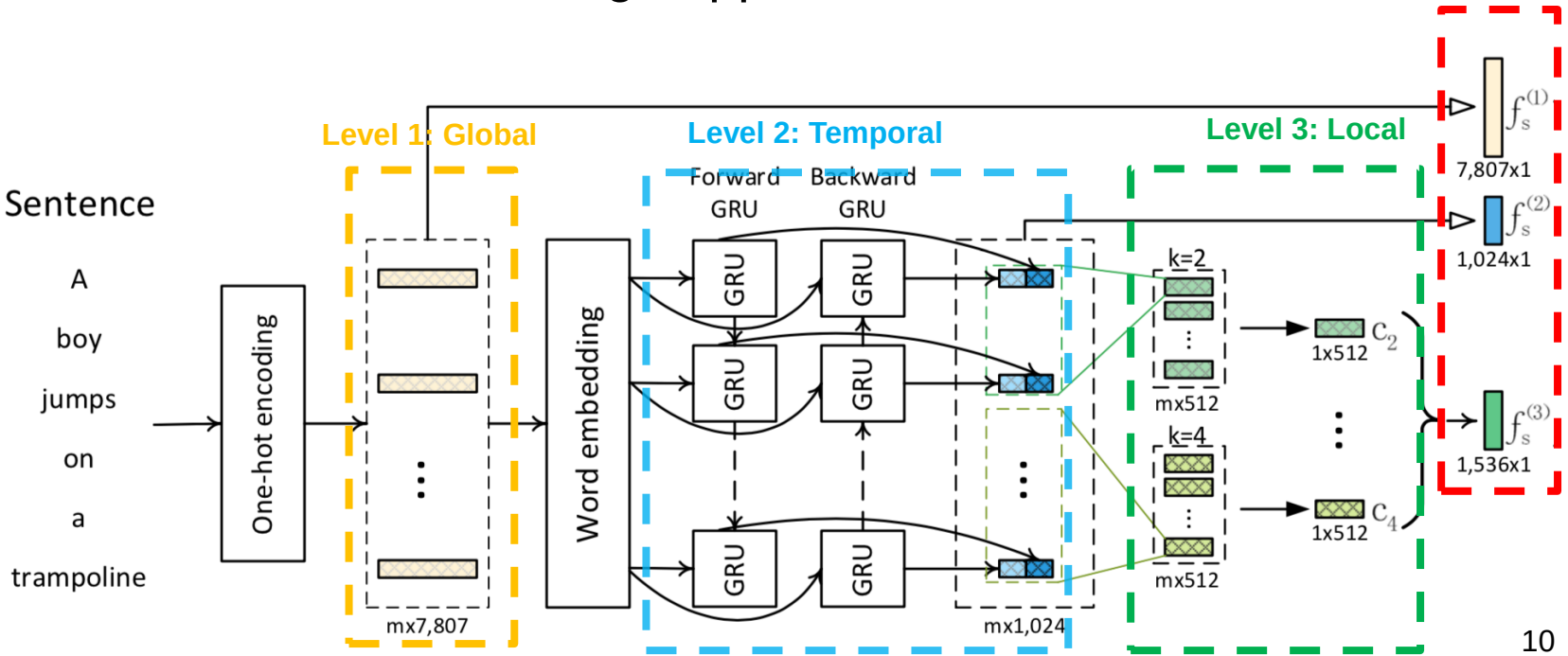
Our Proposal: Video-side Multi-level Encoding

- Multi-level encoding by simple concatenation



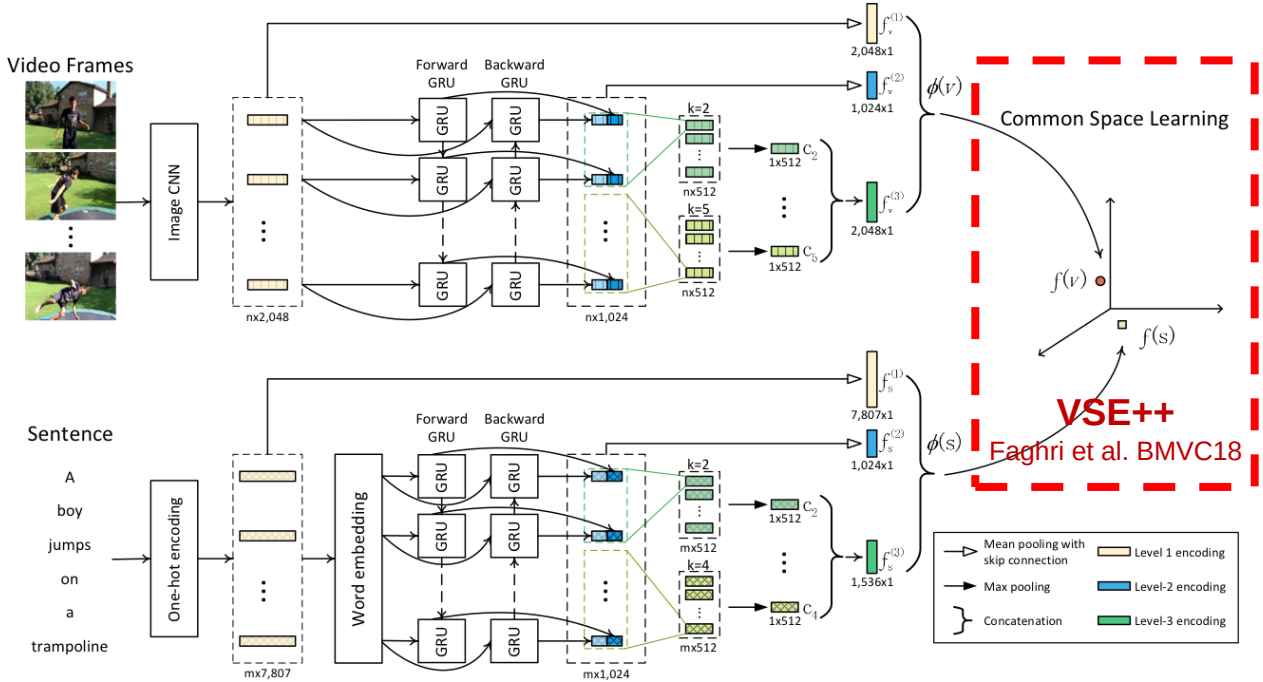
Our Proposal: Text-side Multi-level Encoding

- The same network design applies on the text side



The Dual Encoding Network

- The network encodes a given video / sentence in parallel



+ The same network design for both modalities

+ Three-level encoding for each modality

+ Independent encoding for each modality

+ Any SOTA common space learning can be used

Evaluation

- Questions to answer
  - Is multi-level encoding better than single-level encoding?
  - Is dual encoding better than single-side encoding?

Experiments on MSR-VTT

- MSR-VTT[Xu et al. CVPR16]

	Training	Validation	Test
Videos	6,513	497	2,990
Sentences	On average 20 sentences per video		

- Baselines

Method	Video-side	Text-side	Loss
Mithun et al. ICMR18	Mean pooling	GRU	Improved Marginal Ranking Loss (imrl)
W2VV (Dong et al. T-MM18)	Mean pooling	[BoW; W2V; GRU]	Mean Square Error
$W2VV_{imrl}$	Mean pooling	[BoW; W2V; GRU]	imrl

Experiments on MSR-VTT

- Dual encoding outperforms the SOTA

Table 1. **State-of-the-art on MSR-VTT**. Larger R@{1,5,10}, mAP and smaller Med r indicate better performance. Methods sorted in ascending order in terms of their overall performance. The proposed method performs the best.

Method	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	
W2VV [7]	1.8	7.0	10.9	193	0.052	9.2	25.4	36.0	24	0.050	90.3
Mithun *et al.* [24]	5.8	17.6	25.2	61	-	10.5	26.7	35.9	25	-	121.7
$W2VV_{imrl}$	6.1	18.7	27.5	45	0.131	11.8	28.9	39.1	21	0.058	132.1
Dual encoding	**7.7**	**22.0**	**31.8**	**32**	**0.155**	**13.0**	**30.8**	**43.3**	**15**	**0.065**	**148.6**

Frame-level CNN features: 2,048-dim ResNet-152

Experiments on MSR-VTT

- Ablation study shows that multi-level encoding is the best

Table 2. **Ablation study on MSR-VTT**. The overall performance, as indicated by **Sum of Recalls**, goes up as more encoding layers are added. Dual encoding exploiting all the three levels is the best.

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**

Experiments on MSR-VTT

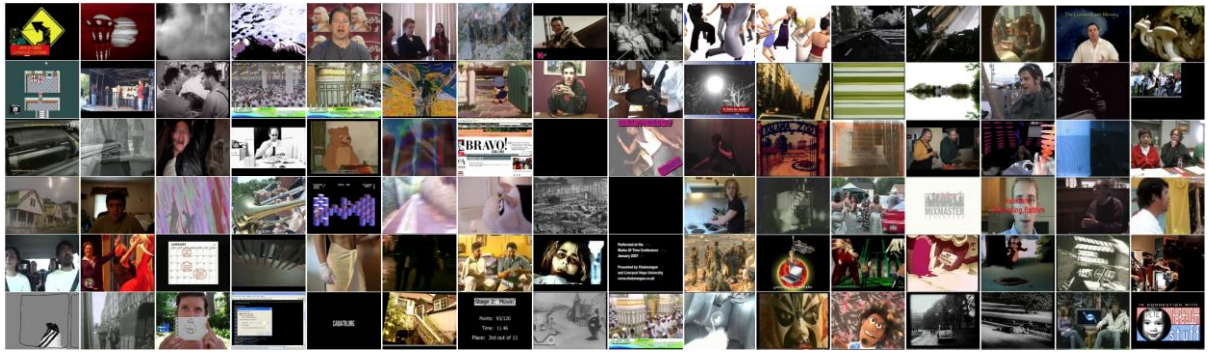
- Dual encoding is better than single-side encoding

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**

Experiments on TRECVID

- TRECVID Ad-hoc Video Search (AVS) Task, since 2016
  - Given the test collection (IACC.3), and a set of 30 ad-hoc queries released by NIST, return for each query a list of at most 1000 shot IDs from the test collection, ranked according to their likelihood of containing the target query.

    IACC.3 : 335,944 shots from 4,593 Internet videos, with very diverse content



  - Performance metric: Inferred AP (infAP)

Experiments on TRECVID

- Dual encoding outperforms the SOTA

Table 3. **State-of-the-art on TRECVID 2016.**

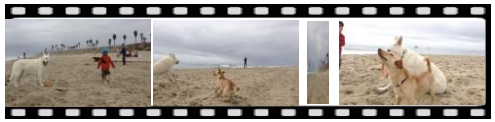
Method	infAP
Top-3 TRECVID finalists:	
Le *et al.* [15]	0.054
Markatopoulou *et al.* [22]	0.051
Liang *et al.* [18]	0.040
Literature methods:	
Habibian *et al.* [10]	0.087
Markatopoulou *et al.* [21]	0.064
$W2VV_{imrl}$	0.132
Dual encoding	**0.159**

Table 4. **State-of-the-art on TRECVID 2017.**

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**

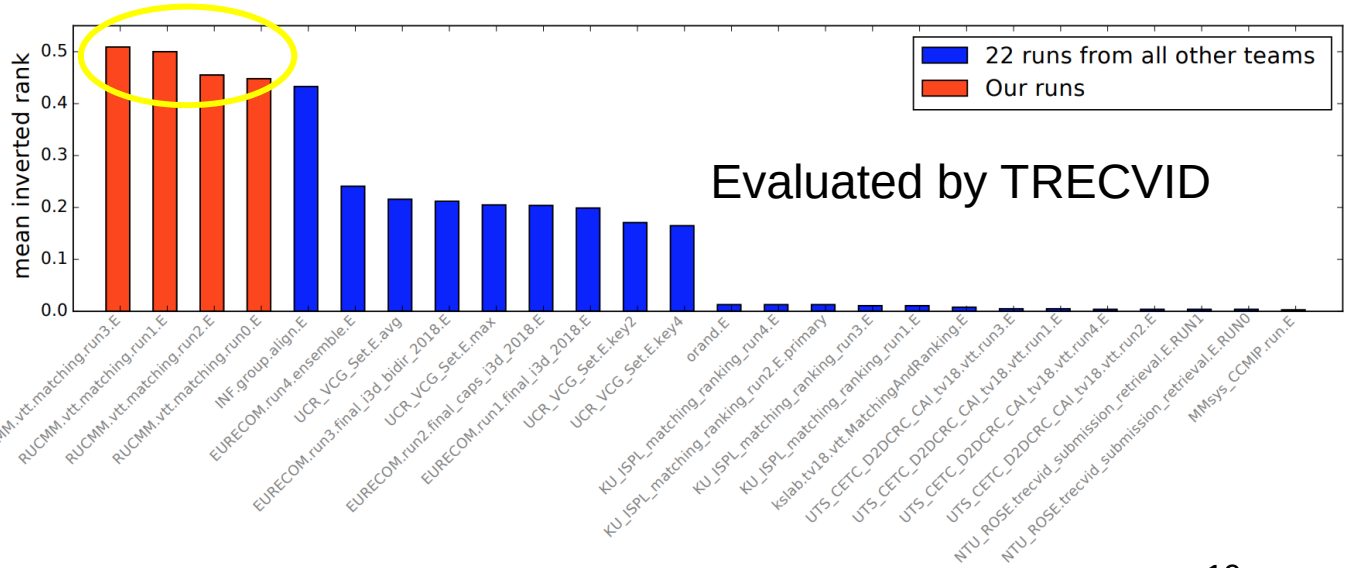
Concept-based methods, mostly

TRECVID 2018 Video-to-Text Matching

Query video  → Retrieved text

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

Dual encoding based solutions perform the best.



Experiments on MSVD and MPII-MD

- MSVD[Chen&Dolan ACL11]
  - similar to MSR-VTT but in a smaller scale
- MPII-MD[Rohrbach et al. CVPR15]
  - A movie description dataset

Table 6. **Performance of zero-example video retrieval, measured by mAP**. Our proposed method is the best.

Model	MSVD	MPII-MD
W2VV	0.100	0.008
$W2VV_{imrl}$	0.230	0.030
VSE++	0.218	0.022
Dual Encoding	**0.232**	**0.037**

Scores appear to be low?

Experiments on MSVD and MPII-MD

- Low scores on MPII-MD are largely due to incomplete ground truth

Query sentence: They wrap their arms around each other (AP=0.25)

Ground truth **Top-5 shots retrieved from the MPII-MD test set by our model**

Query sentence: In a restaurant,Someone sits at a table with the guy (AP=0.031)

Ground truth **Top-5 shots retrieved from the MPII-MD test set by our model**

Experiments on Image-Text Retrieval

- Replace text encoding of VSE++ by our multi-level encoding

Method	Text-to-Image			Image-to-Text		
	R@1	R@5	R@10	R@1	R@5	R@10
On Flickr30k						
VSE++	23.1	49.2	60.7	31.9	58.4	68.0
VSE++, multi-level encoding	**24.7**	**52.3**	**65.1**	**35.1**	**62.2**	**71.3**
On MSCOCO						
VSE++	33.7	68.8	81.0	43.6	74.8	84.6
VSE++, multi-level encoding	**34.8**	**69.6**	**82.6**	**46.7**	**76.2**	**85.8**

Multi-level encoding improves VSE++

Image side: VGGNet features from Faghri et al. BMVC18

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
- New SOTA on multiple benchmarks for zero-example video retrieval

https://github.com/danieljf24/dual_encoding

dongjf24@gmail.com