Visual Oriented Encoder: Integrating Multimodal and Multi-Scale Contexts for Video Captioning

Bang Yang[1] and Yuexian Zou[1,2]*

[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
{yb.ece, zouyx}@pku.edu.cn * (Corresponding author)

Introduction ①

Recent researches have shown that exploiting the multi-modalities of videos significantly promotes captioning performance. However, the downside of most existing methods lies in the neglect of the interaction among multi-modalities and their rich contextual information.

Motivation. People grasp the gist of video content mainly through visual information, supplemented by some other information like motion and audio.

Proposed Solution. To yield better joint representations of video content, a Visual Oriented Encoder (VOE) is proposed to progressively integrate multimodal features in an interactive manner, where global and regional contexts are learnt to assist visual understanding, and a video captioning model VOE-LSTM is developed.

Network architecture ②

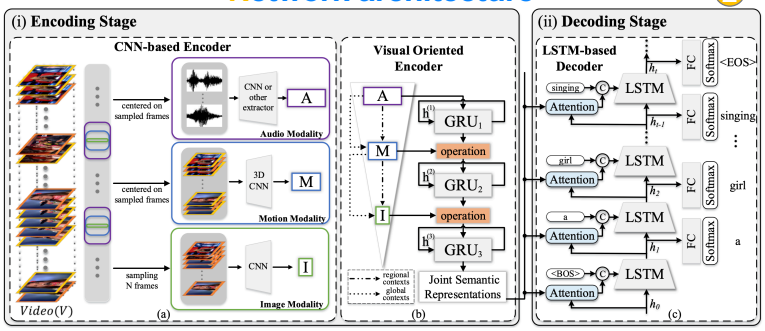

Fig 1. The overall architecture of VOE-LSTM.

VOE-LSTM consists of three parts: (a) CNN-based encoders that extract multimodal features, (b) our proposed VOE that learns joint representations, and (c) a single layer LSTM that generates captions. A, M and I are short for audio, motion and image modalities, respectively.

Proposed Visual Oriented Encoder ③

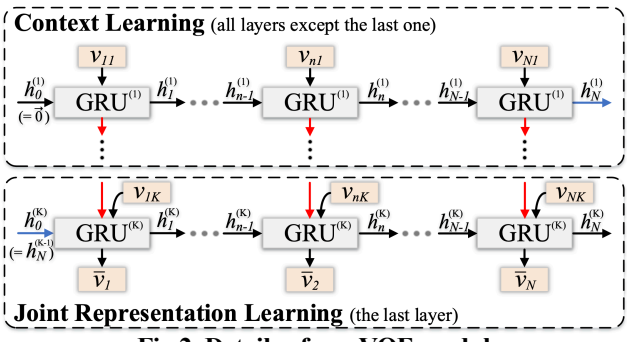

Fig 2. Details of our VOE module.

In our implementations, VOE is essentially a hierarchical GRUs. Given features of K modalities $\mathbf{V}^{(1)}, \mathbf{V}^{(2)}, \ldots, \mathbf{V}^{(K)}$ and $\mathbf{V}^{(k)} = \{v_{nk}\}_{n=1}^{N}$, the goal of VOE is to learn compact joint representations $\overline{\mathbf{V}}$. Specifically, K GRUs are stacked to fuse these features progressively. We treat the hidden state of the i-th GRU at the last time step, *i.e.* $h_N^{(i)}$, as the global context for $\mathbf{V}^{(i+1)}$, while $\{h_n^{(i)}\}_{n=1}^{N}$ as the regional context for $\mathbf{V}^{(i+1)}$.

(1) The first layer
- $h_n^{(1)} = \mathrm{GRU}^{(1)}(v_{n1}, h_{n-1}^{(1)})$
- $h_0^{(1)} = \vec{0}$

(2) The λ-th layer ($\lambda \in [2, K]$)
- $h_n^{(\lambda)} = \mathrm{GRU}^{(\lambda)}([h_n^{(\lambda-1)}, v_{n\lambda}], h_{n-1}^{(\lambda)})$
- $h_0^{(\lambda)} = h_N^{(\lambda-1)}$

(3) Joint representations $\overline{\mathbf{V}} = \{h_1^{(K)}, h_2^{(K)}, \ldots, h_N^{(K)}\}$

Benchmark Comparison ④

Dataset	MSVD				MSR-VTT			
Model	Features	B@4	M	C_D	Features	B@4	M	C_D
HRNE w/ Attention [13]	G+C	46.7	33.9	-	-	-	-	-
hLSTMat [12]	G	48.5	31.9	-	R-152	38.3	26.3	-
MAM-RNN [36]	G	41.3	32.2	53.9	-	-	-	-
DMRM w/o SS [4]	G	50.0	33.2	73.2	-	-	-	-
v2t_navigator [5]	-	-	-	-	C+A	40.8	28.2	44.8
HMVC [17]	V+T	44.3	32.1	68.4	V+T	37.1	26.7	-
TDDF [15]	V+C	45.8	33.3	73.0	V+C	37.3	27.8	43.8
Attentional Fusion[7]	V+C	52.4	32.0	68.8	V+C+A	39.7	25.5	40.0
MA-LSTM [8]	G+C	52.3	33.6	70.4	G+C+A	36.5	26.5	41.0
M^3 [9]	I+C	**52.82**	33.31	-	V+C	38.13	26.58	-
Enhanced TGM [6]	I+C	49.26	33.91	83.02	I+C+A	44.91	29.61	51.80
VOE-LSTM (ours)	I+C	51.44	**34.40**	**84.04**	I+C	41.78	29.22	49.63
					I+C+A	**45.76**	**29.84**	**52.31**

Table 1. Comparison on MSVD and MSR-VTT benchmarks

Ablation Study on MSR-VTT ⑤

Exp	Modality			B@4	M	C_D
	Image	Motion	Audio			
1	✓	-	-	40.67	28.39	48.46
2	-	✓	-	39.03	27.68	43.20
3	-	-	✓	33.37	24.85	29.60
4	✓	✓	-	41.78	29.22	49.63
5	-	✓	✓	43.44	28.94	47.53
6	✓	-	✓	43.56	29.31	49.76
7	✓	✓	✓	**45.76**	**29.84**	**52.31**

Table 2. Performance of different modalities.

- Static appearance plays the main role in video understanding.
- Utilizing multi-modalities is requisite for video captioning.

Features	Context		B@4	M	C_D	Param. (M)
	G	R				
I+C	✓		40.68	28.81	47.92	6.82
		✓	40.42	28.31	47.72	7.61
	✓	✓	**41.78**	**29.22**	**49.63**	7.61
I+C+A	✓		44.56	29.53	51.06	8.60
		✓	43.78	29.07	50.72	10.17
	✓	✓	**45.76**	**29.84**	**52.31**	10.17

Table 3. Effect of global (G) and regional (R) contexts.

- Both kinds of contexts can improve the caption quality.

Qualitative Results ⑥

Fig 3. Four visualized examples of generated captions, where our model captures more precise details from videos.

Main Contributions ⑦

- The proposed Visual Oriented Encoder (VOE) presents an alternative way for multimodal fusion, where inter-modality interaction is highlighted for fully utilization of multi-scale contextual information.
- Extensive experiments with both quantitative and qualitative evaluation demonstrates the effectiveness of our method.