Context Visual Information-based Deliberation Network for Video Captioning.

Min Lu, Xueyong Li and Caihua Liu

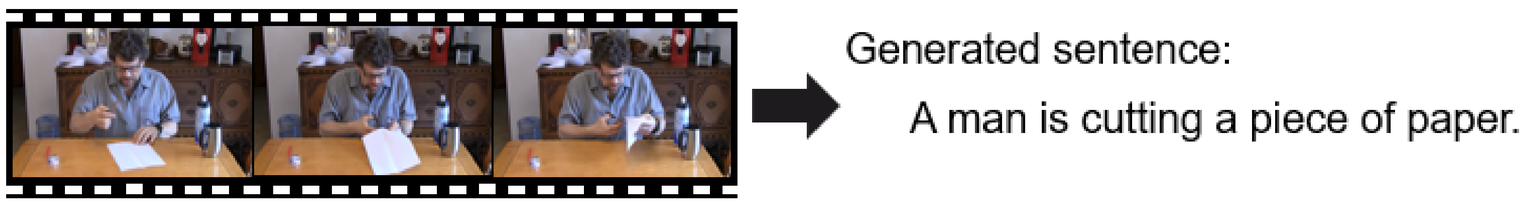
College of Computer Science and Technology, Civil Aviation University of China

mlu, 2018052047, chliu@cauc.edu.cn

1. Introduction

■ **Video Captioning**

· Video captioning aims to automatically generate a natural language sentence to describe a video accurately.



■ **Challenges**

· Complex video content understanding

· Heterogeneous gap between vision and language

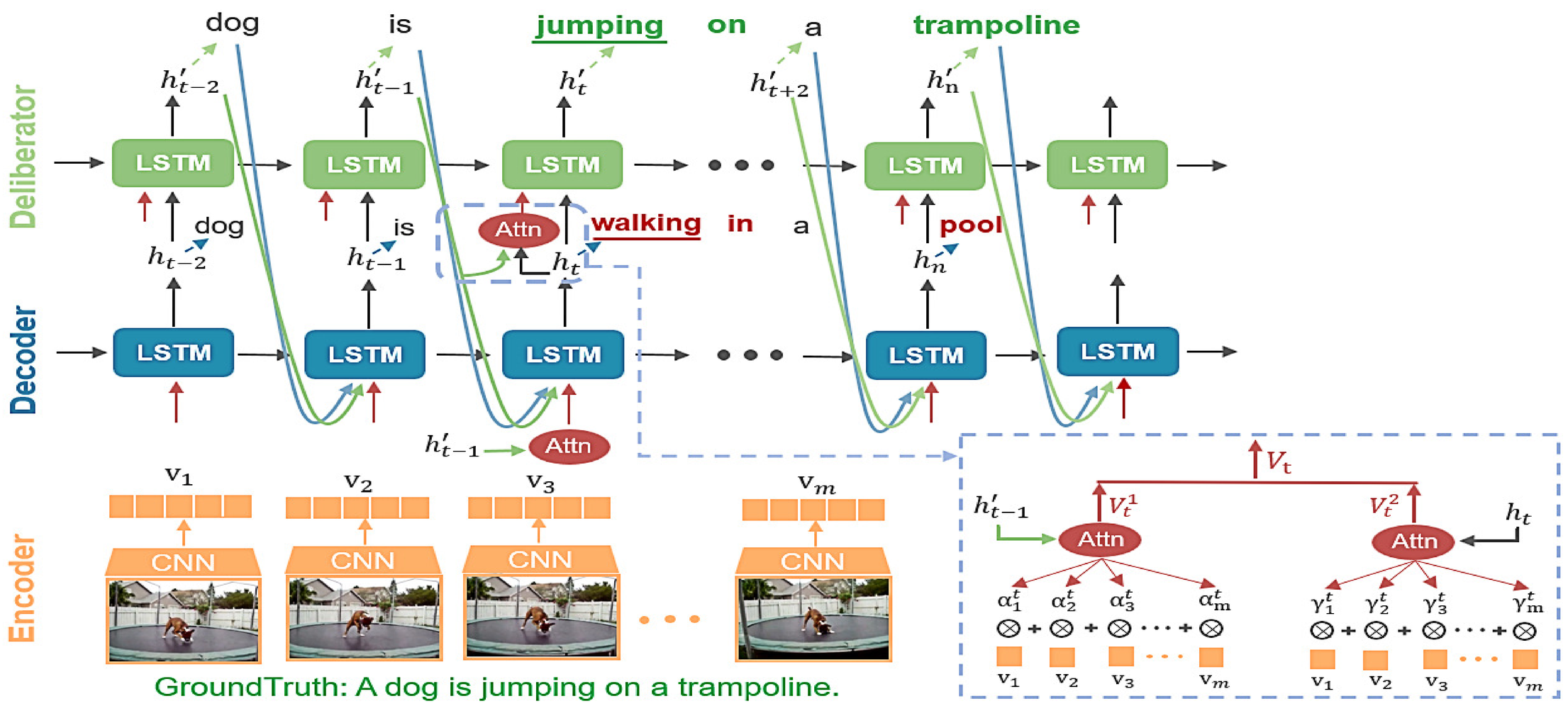
2. Motivation

■ **Motivation**

· The hidden states with inaccurate semantic information are not amended before word prediction, which will cause a cascade of errors in predicting words.

· The attention weights for the current word are normalizing relevance scores between the previous hidden state and visual features. In fact, those weights should be calculated based on the currently hidden state rather than the previous state.

3. Method



■ **The model consists of three components: a CNN-based encoder, an attention-based recurrent decoder and a deliberator.**

· **The encoder** (CNN) takes video frames as input and extracts visual features to capture the high-level semantic information.

$$\mathbf{F} = CNN\left(\mathbf{V}\right) \tag{1}$$

· **The attention-based recurrent decoder** aims to generate a raw hidden state sequence.

$$\mathbf{h}_t = LSTM\left(\mathbf{h}_{t-1},\, \mathbf{inp}_t,\, \mathbf{h}_{t-1}^{'}\right), \quad \mathbf{inp}_t = \left[E\left[y_{t-1}\right],\, \varphi_t^{'}(F)\right], \quad \varphi_t^{'}(F) = \mathbf{f}_{attn}\left(\mathbf{h}_{t-1}^{'}, F\right) \tag{2}$$

· **The deliberator** realized by LSTM focuses on polishing the semantic information contained in the raw hidden state.

$$\mathbf{h}_t^{'} = LSTM\left(\mathbf{h}_{t-1}^{'},\, \mathbf{inp}_t^{'},\, \mathbf{h}_t\right), \quad \mathbf{inp}_t^{'} = \left[E\left[y_{t-1}\right],\, \varphi_t(F)\right], \quad \varphi_t(F) = \left[\mathbf{f}_{attn}\left(\mathbf{h}_{t-1}^{'}, F\right),\, \mathbf{f}_{attn}\left(\mathbf{h}_t, F\right)\right] \tag{3}$$

4. Experiments

■ **Quantitative Analysis**

	MSVD				MSR-VTT			
Methods	*BLEU@4*	*METEOR*	*ROUGEL*	*CIDEr*	*BLEU@4*	*METEOR*	*ROUGEL*	*CIDEr*
Ours	**53.8**	**35.1**	**72.4**	**94.5**	**41.6**	**28.4**	**61.3**	**48.5**
RecNet	52.3	34.1	69.8	80.3	39.1	26.6	59.3	42.7
hLSTMat	53.0	33.6	–	73.8	38.3	26.3	–	–
MARN	48.6	35.1	71.9	92.2	40.4	28.1	60.7	47.1

■ **Qualitative Analysis**

GT: A man is eating spaghetti

Baseline: A man is cooking his kichen

Ours: A man is eating spaghetti

GT: A dog is jumping on a trampoline

Baseline: A dog is walking in a pool

Ours: A dog is jumping on a trampoline