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

Outline

- Introduction
- Related Work
- Our Method
- Experiments and Results

Introduction

❖ Problem Statement

Video Captioning : automatically describing a video in natural language.

Generated sentence:

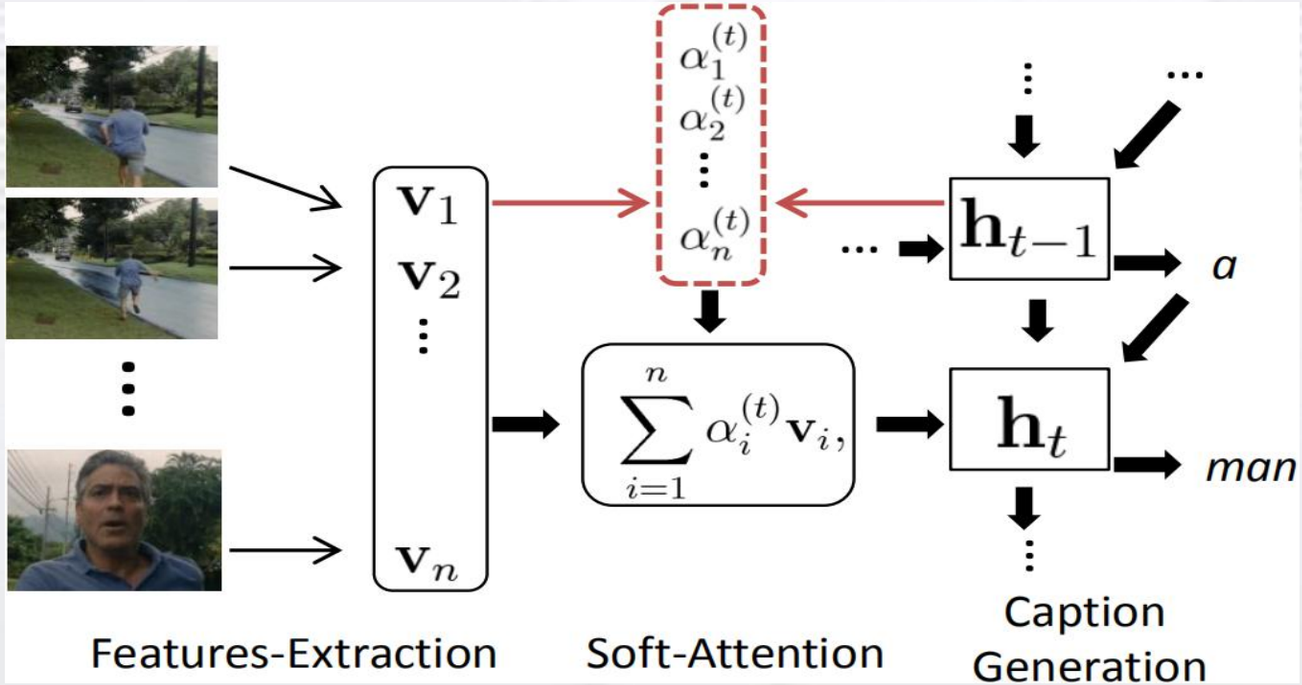
A man is cutting a piece of paper.

Outline

- Introduction
- Related Work
- Our Method
- Experiments and Results

Related Work

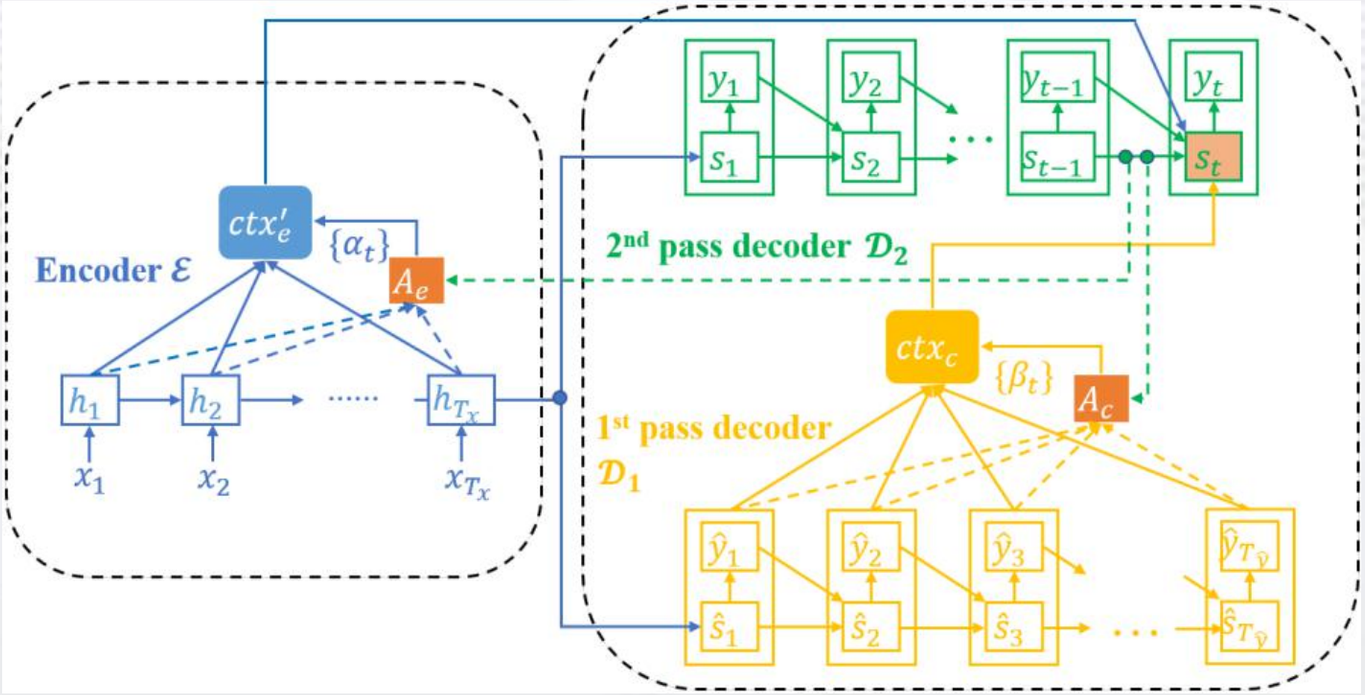
❖ **Encoder-decoder framework with attention mechanisms.**



[1] L. Yao, A. Torabi, K. Cho, N. Ballas et al. " Describing Videos by Exploiting Temporal Structure. " ICCV. 2015.

Related Work

❖ **Deliberation Networks for Neural Machine Tranlation**



[2] Y. Xia, F. Tian, L. Wu, et al. “Deliberation networks: Sequence generation beyond one-pass decoding.” NIPS. 2017.

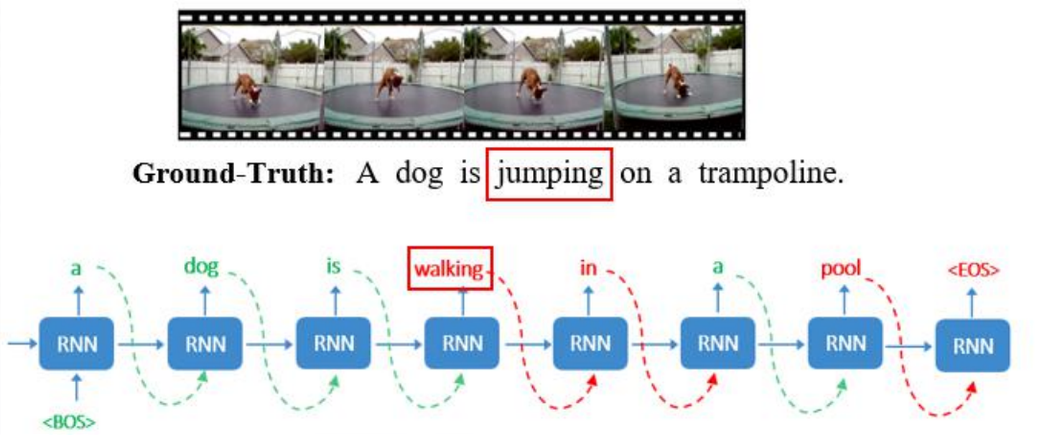
Outline

- Introduction
- Related Work
- Our Method
- Experiments and Results

Our Method

❖ Motivations

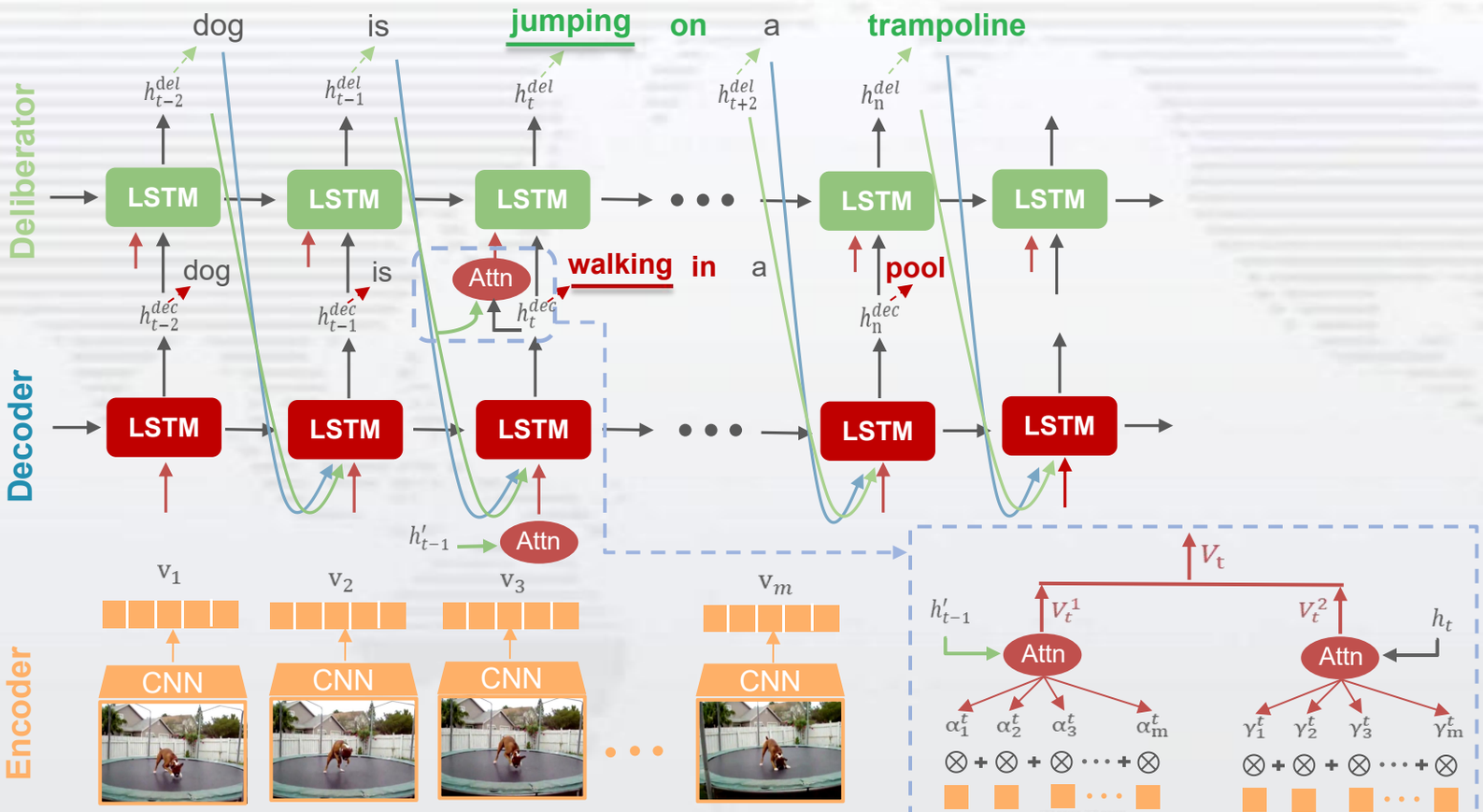
- The hidden states with inaccurate semantic information are not amended before word prediction, which will cause a cascade of errors in predicting words.



- The attention weights for the current word should be calculated based on the current hidden state rather than the previous state.

Our Method

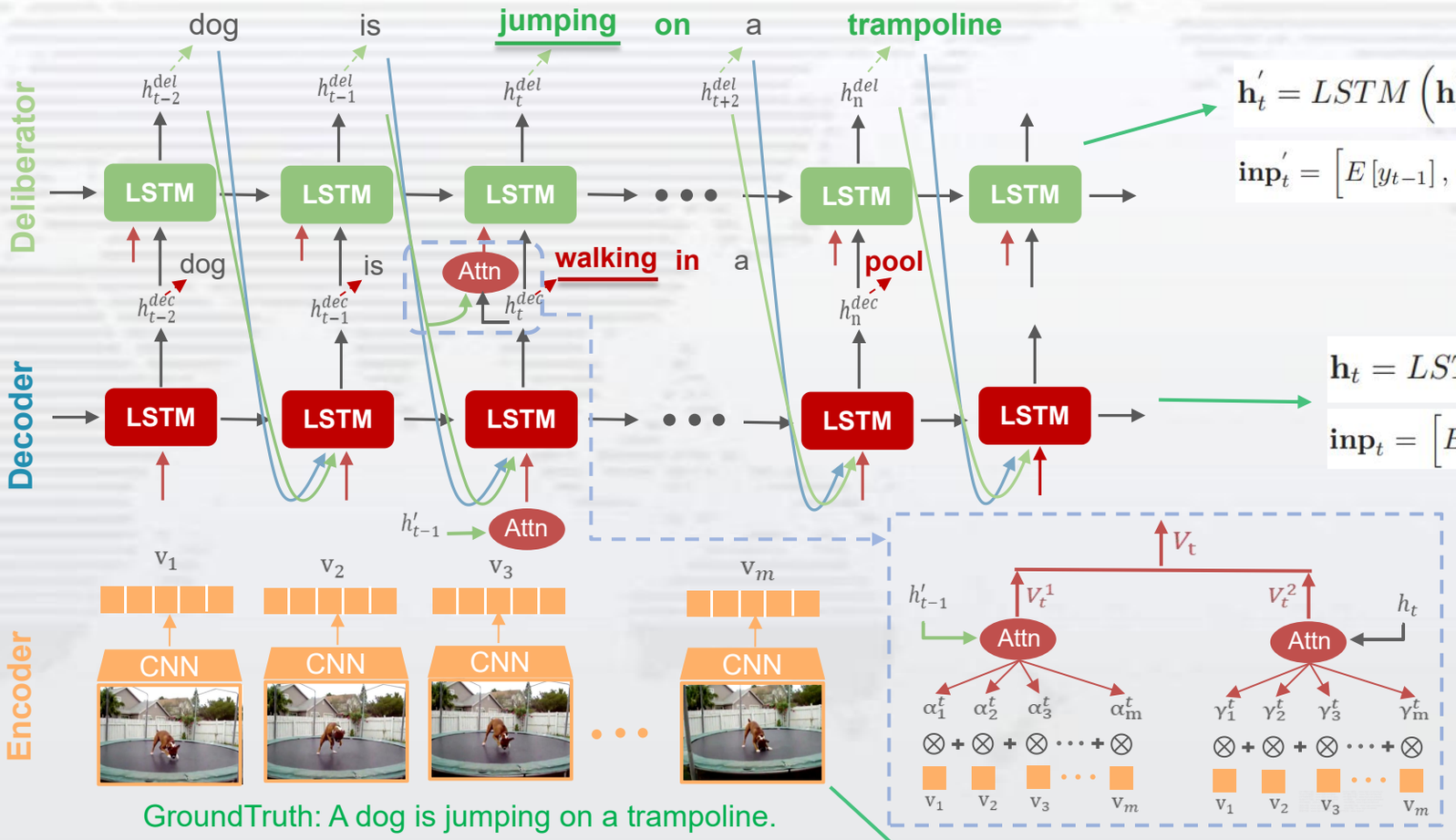
❖ Model Architecture



GroundTruth: A dog is jumping on a trampoline.

Our Method

❖ Model Architecture



$$\mathbf{h}_t' = LSTM\left(\mathbf{h}_{t-1}', \mathbf{inp}_t', \mathbf{h}_t\right),$$

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

$$\mathbf{h}_t = LSTM\left(\mathbf{h}_{t-1}, \mathbf{inp}_t, \mathbf{h}_{t-1}'\right),$$

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

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

Outline

- Introduction
- Related Work
- Our Method
- Experiments and Results

Experiments and Results

- **MSVD**
  - It contains 1,200, 100 and 670 videos for training, validation and testing. And each video has 40 english captions annotated by human beings.

- **MSR-VTT**
  - It contains 6,513, 497 and 2,990 videos for training, validation and testing. And each video is associated with 20 descriptions.

Experiments and Results

❖ Quantitative Analysis

➢ MSVD

Models	BLEU@4	METEOR	ROUGE-L	CIDEr
S2VT(I)[19]	39.6	31.2	67.5	66.7
RecNet(I)[23]	52.3	34.1	69.8	80.3
hLSTMat(R-152)[10]	53.0	33.6	-	73.8
TSA-ED(R-152)[39]	51.7	34.0	-	74.9
PickNet(R-152)[13]	46.1	33.1	69.2	76.0
TDConvED(R-152)[40]	53.3	33.8	-	76.4
SCN-LSTM(R-152+C)[12]	50.2	33.4	-	80.5
SA(R-101+RX-101)[7]	52.4	34.3	71.7	89.5
MARN(R-101+RX-101)[11]	48.6	35.1	71.9	92.2
Ours(I)	53.0	33.9	70.8	84.0
Ours(R-152)	53.3	34.1	70.7	84.4
Ours(R-101+RX-101)	**53.8**	**35.1**	**72.4**	**94.5**

➢ MSR-VTT

Models	BLEU@4	METEOR	ROUGE-L	CIDEr
RecNet(I)[23]	39.1	26.6	59.3	42.7
hLSTMat(R-152)[10]	38.3	26.3	-	-
TSA-ED(R-152)[39]	39.5	27.5	-	42.8
VideoLab(R-152+C+A)[41]	39.1	27.7	60.6	44.1
Aalto(G+C)[42]	39.8	26.9	59.8	45.7
v2t_navigator(C+A)[43]	40.8	28.2	60.9	44.8
SA(R-101+RX-101)[7]	39.5	26.4	59.4	45.9
MARN(R-101+RX-101)[11]	40.4	28.1	60.7	47.1
Ours(R-101+RX-101)	**41.6**	**28.4**	**61.3**	**48.5**

Experiments and Results

❖ **Qualitative Analysis**

GroundTruth: a man is eating spaghetti.
Baseline: a man is cooking his kichen.
Ours: a man is eating spaghetti.

GroundTruth: a woman is mixing ingredients in a bowl
Baseline: a man is cooking something
Ours: a woman is **mixing ingredients in a bowl**

GroundTruth: a man is sliding down a railing of the stairs
Baseline: a man is walking down the street
Ours: a man is running down the stairs

GroundTruth: a dog is jumping on a trampoline
Baseline: a dog is walking in a pool
Ours: a dog is jumping on a trampoline

Conclusions

- We propose a novel architecture, Context Visual Information-based Deliberation Network for Video Captioning.
- The proposed method can not only amend the inappropriate hidden state in time but also strengthen the semantic coherence of the adjacent words.
- Experiments on real datasets show that our approach outperforms the state-of-the-art methods.

Q&A