CRIPAC 智能感知与计算研究中心 Center for Research on Intelligent Perception and Computing

中国科学院自动化研究所 Institute of Automation Chinese Academy of Sciences

2018

Multimodal Memory Modelling for Video Captioning

Liang Wang & Yan Huang

Center for Research on Intelligent Perception and Computing (CRIPAC)

National Laboratory of Pattern Recognition (NLPR)

Institute of Automation, Chinese Academy of Sciences (CASIA)

Mar 28, 2018

NVAIL

CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

NVIDIA®

Artificial Intelligence Laboratory

Researches on artificial intelligence and deep learning

Outline

- Introduction
- Model Description
- Experimental Results
- Conclusion

Outline

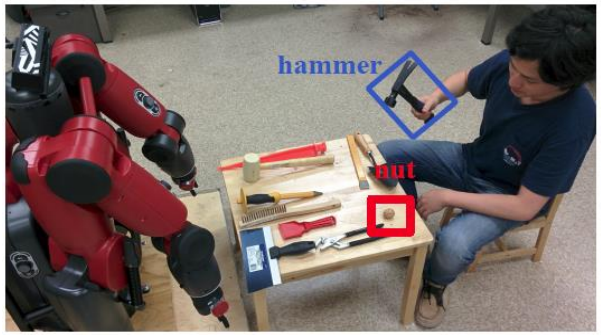
- Introduction
- Model Description
- Experimental Results
- Conclusion

Video Captioning

- Generate natural sentences to describe video content

1. A man and a woman performing a musical.
2. A teenage couple perform in an amateur musical.
3. Dancers are playing a routine.
4. People are dancing in a musical.

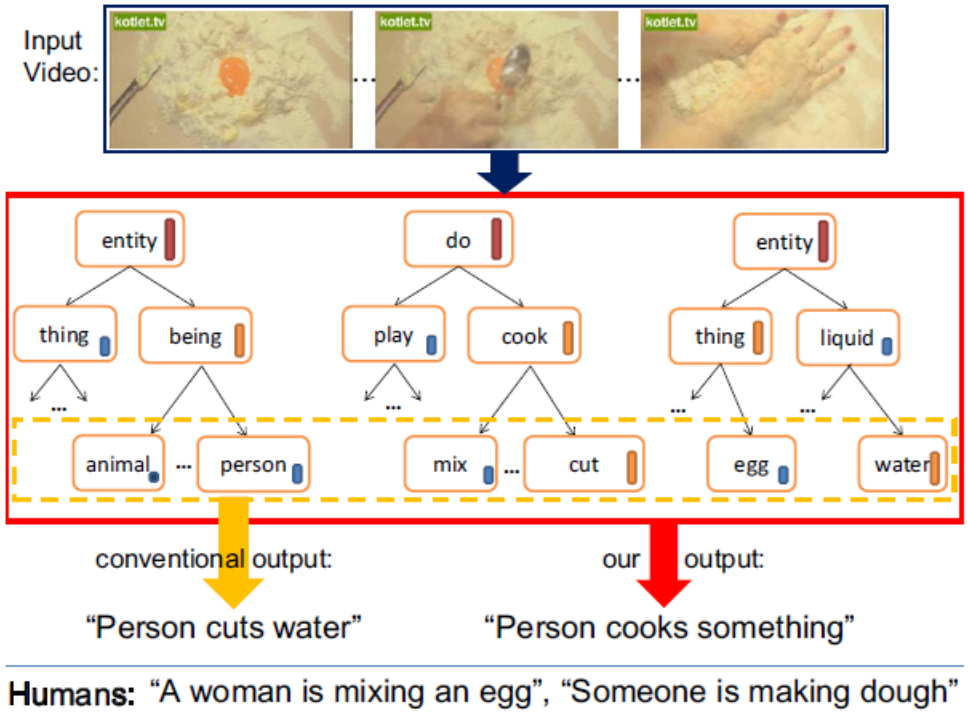
- Potential applications

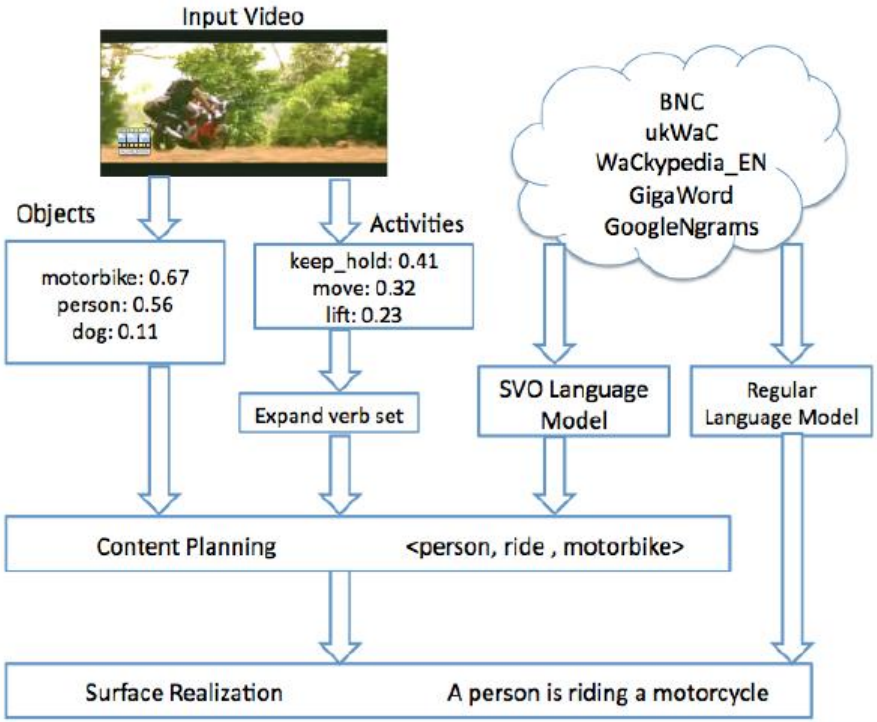


- Challenges
  - ✓ Learning an effective mapping from visual sequence space to language space
  - ✓ The long-term visual-textual dependency modelling

Related Work

➢ Language template-based approach



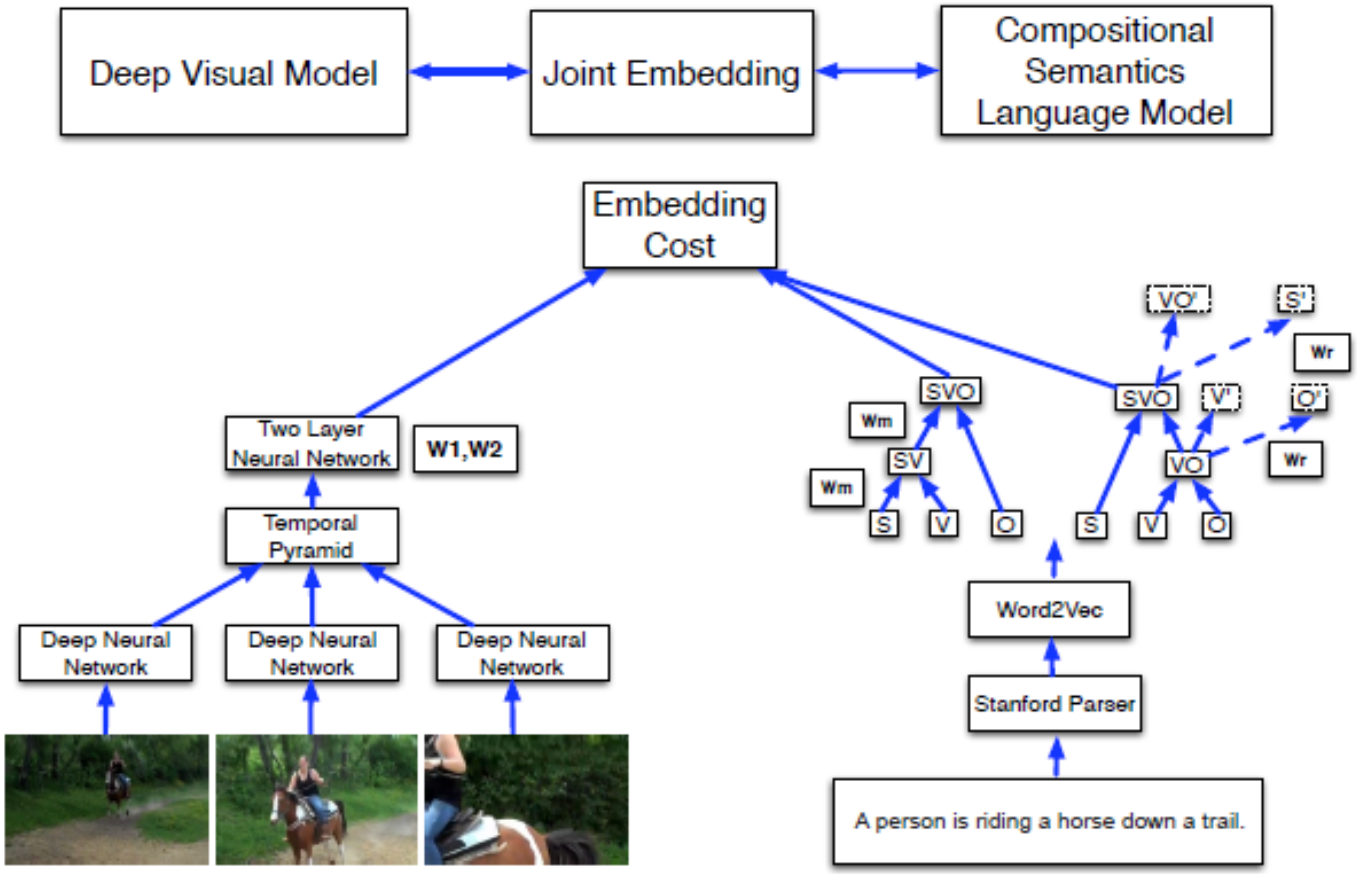


Guadarrama et al. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. ICCV13.

Krishnamoorthy et al. Generating Natural-Language Video Descriptions Using Text-Mined Knowledge. AAAI13.

Related Work

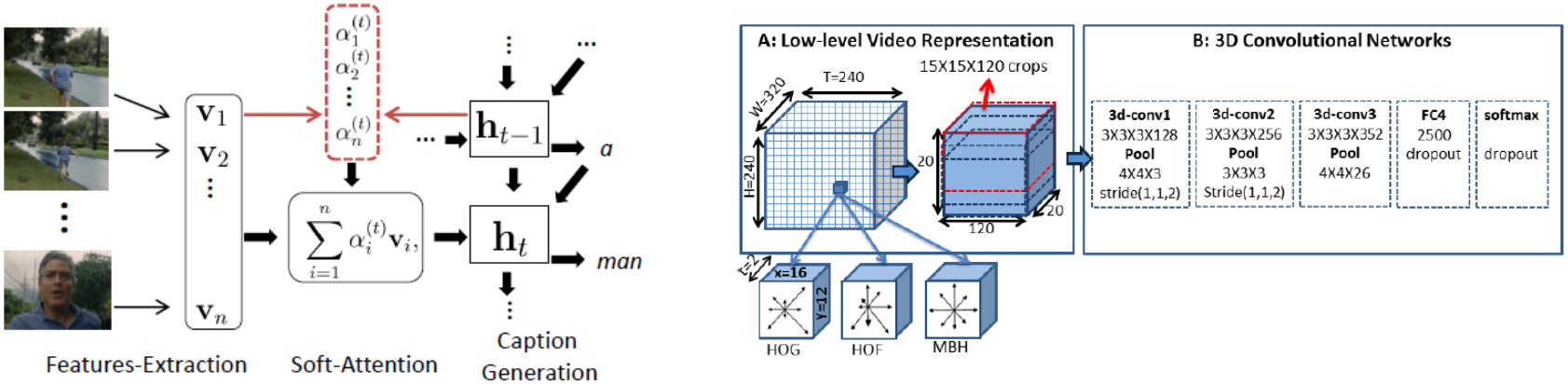
- Search-based approach



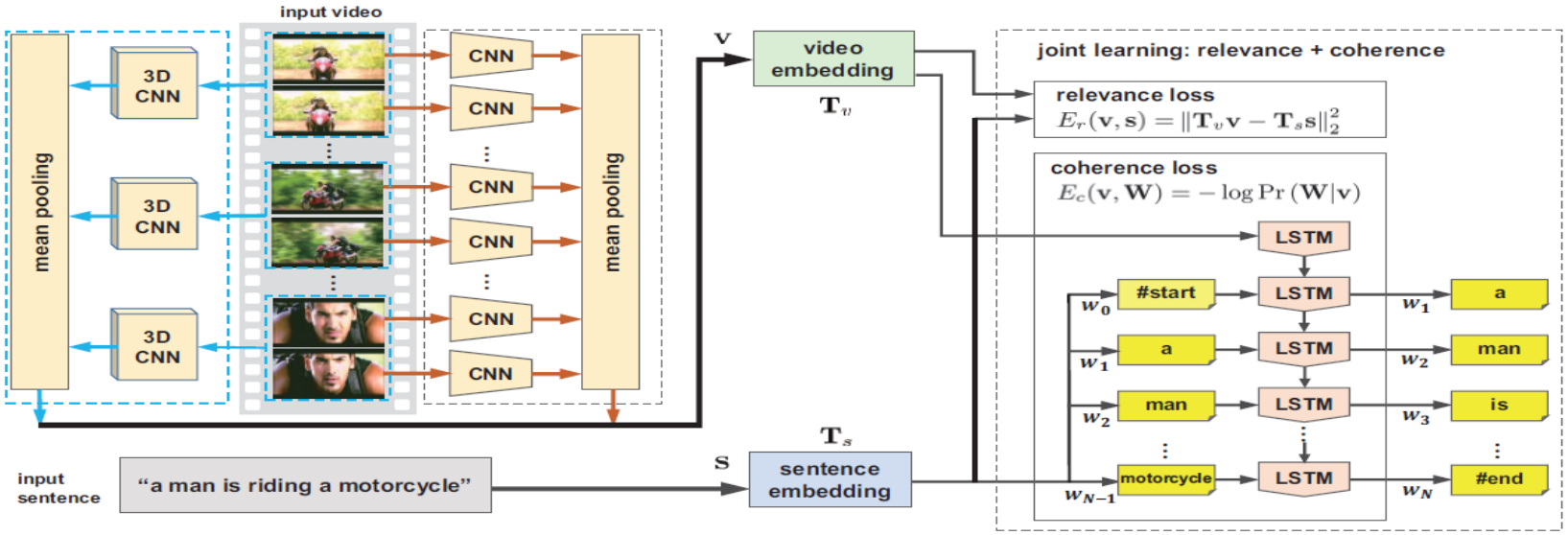
Xu et al. Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework. AAAI15.

Related Work

➢ Sequence-to-Sequence learning-based approach



Yao et al. Describing Videos by Exploiting Temporal Structure. ICCV15.



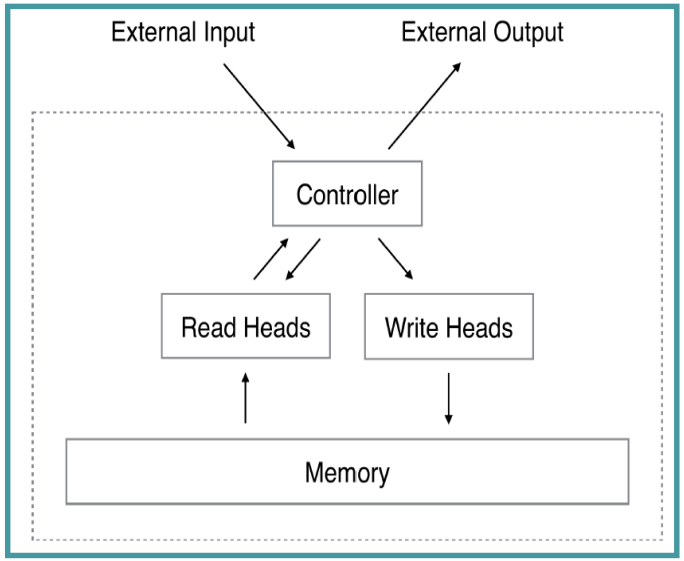
Pan et al. Jointly Modeling Embedding and Translation to Bridge Video and Language. CVPR16.

Outline

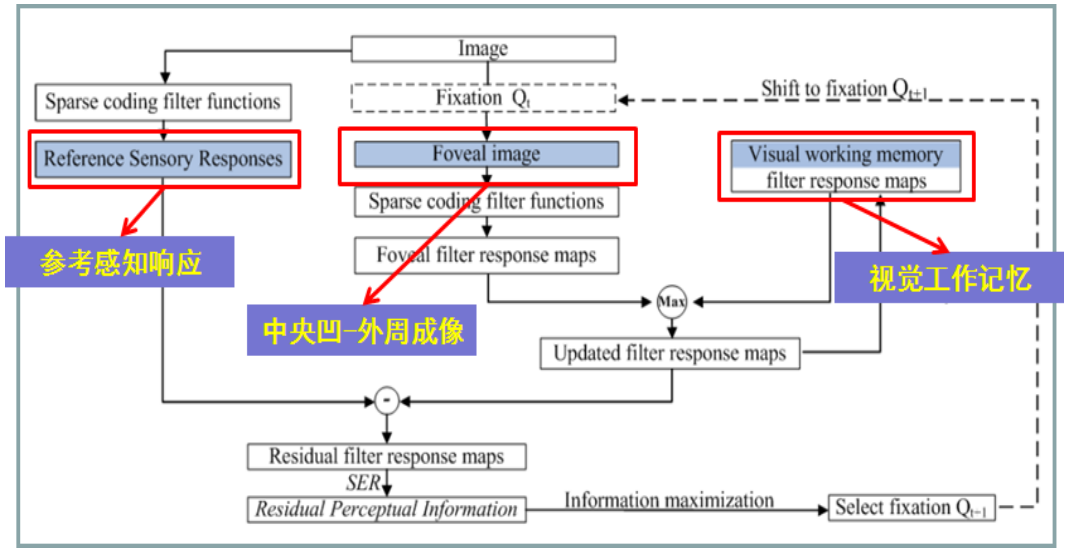
- Introduction
- Model Description
- Experimental Results
- Conclusion

Motivation

- Recent work has pointed out that LSTM doesn't work well when the sequence is long enough.
- Neural memory models have shown great potential to long-term dependency modelling, e.g., QA in NLP.
- Visual working memory is one of the key factors to guide eye movements.



A. Graves, et al. Neural Turing Machines. arXiv:1410.5401



W. Wang, et al. Simulating Human Saccadic Scanpath on Natural Images. CVPR 2011.

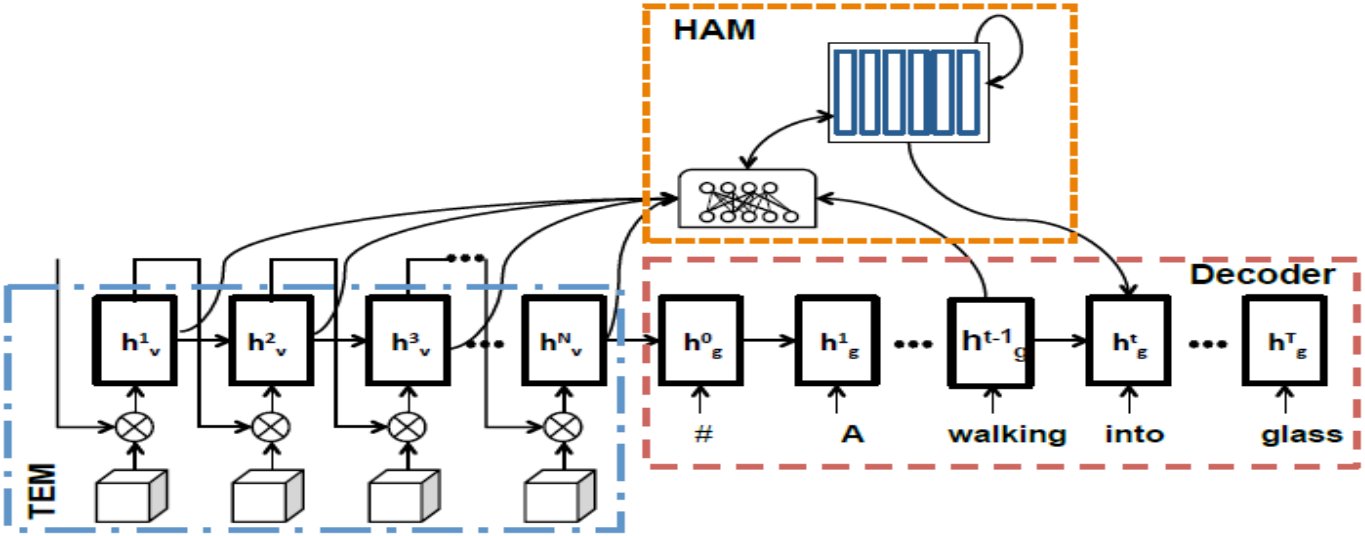
Recent Related Work

Under review as a conference paper at ICLR 2017

MEMORY-AUGMENTED ATTENTION MODELLING FOR VIDEOS

Rasool Fakoor[†], **Abdel-rahman Mohamed,**[††], **Margaret Mitchell**[‡†], **Sing Bing Kang**[††], **Pushmeet Kohli**[††]
[††]Microsoft Research [†]University of Texas at Arlington [‡†]Google
[†]rasool.fakoor@mavs.uta.edu, [††]{asamir, singbing.kang, pkohli}@microsoft.com [‡†]mmitchellai@google.com,

14 Nov 2016



Fakoor et al. Memory-augmented Attention Modelling for Videos. arxiv16.

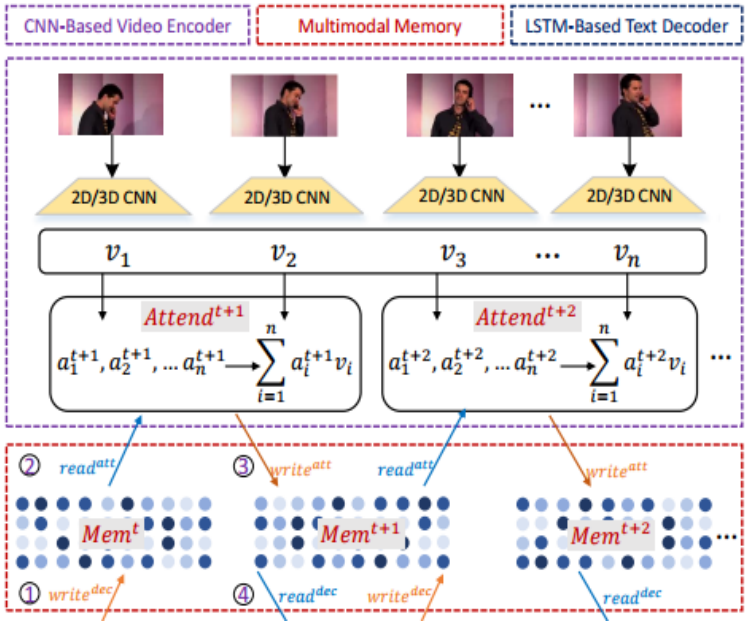
Recent Related Work

Multimodal Memory Modelling for Video Captioning

Junbo Wang[1] Wei Wang[1] Yan Huang[1] Liang Wang[1,2] Tieniu Tan[1,2]
[1]Center for Research on Intelligent Perception and Computing
National Laboratory of Pattern Recognition
[2]Center for Excellence in Brain Science and Intelligence Technology
Institute of Automation, Chinese Academy of Sciences
{junbo.wang, wangwei, yhuang, wangliang, tnt}@nlpr.ia.ac.cn

Abstract

Video captioning which automatically translates video clips into natural language sentences is a very important task in computer vision. By virtue of recent deep learning technologies, e.g., convolutional neural networks (CNNs) and recurrent neural networks (RNNs), video captioning has made great progress. However, learning an effective mapping from visual sequence space to language space is still a challenging problem. In this paper, we propose a Multimodal Memory Model (M^3) to describe videos, which builds a visual and textual shared memory to model the long-term visual-textual dependency and further guide global visual attention on described targets. Specifically, the proposed M^3 attaches an external memory to store and



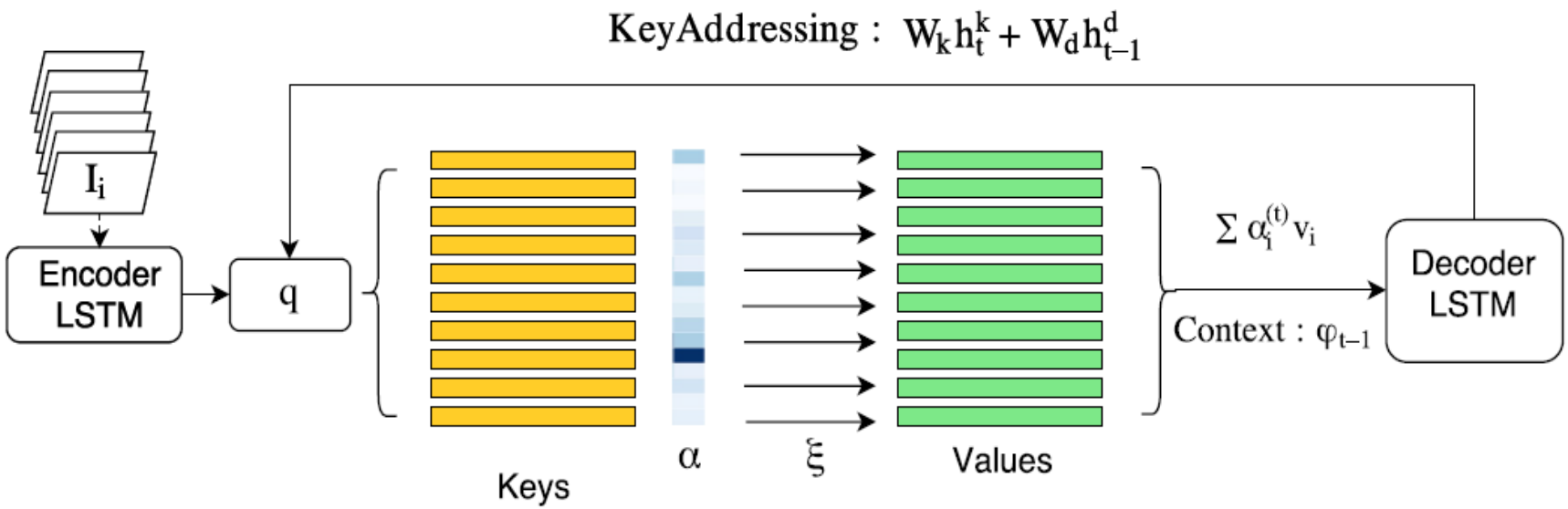
Recent Related Work

Recurrent Memory Addressing for describing videos

Kumar Krishna Agrawal Arnav Kumar Jain Abhinav Agarwalla Pabitra Mitra
Indian Institute of Technology Kharagpur
{kumarkrishna, arnavkj95, abhinavagarawalla, pabitra}@iitkgp.ac.in

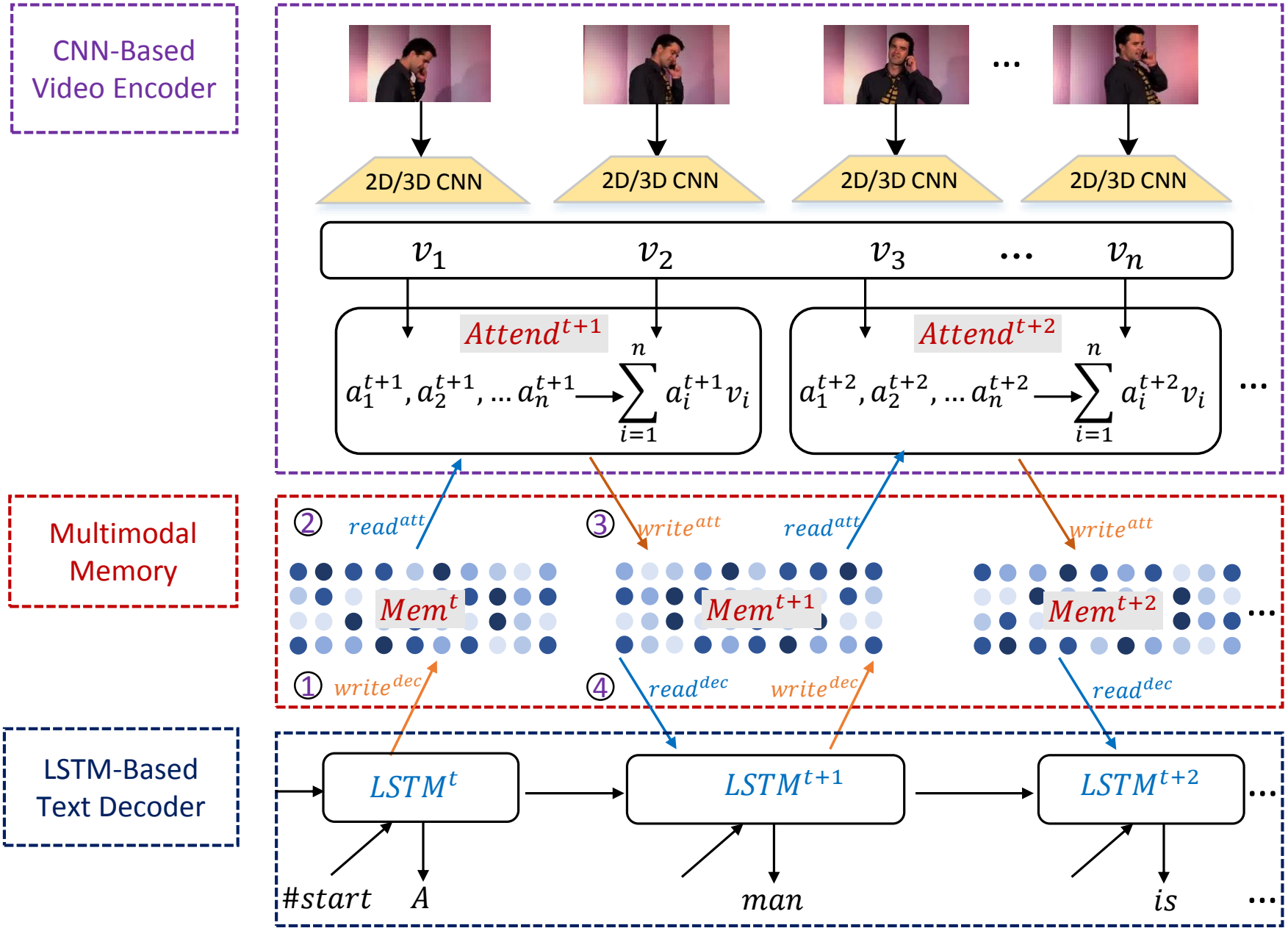
Abstract

Deep Neural Network architectures with external memory components allow the model to perform inference and capture long term dependencies, by storing information ex-

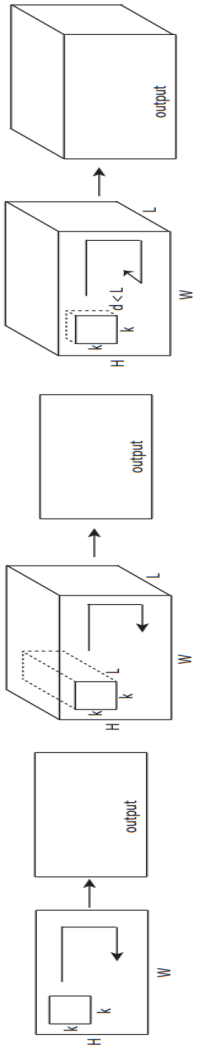


Agrawal et al. Recurrent Memory Addressing for describing videos. arxiv16.

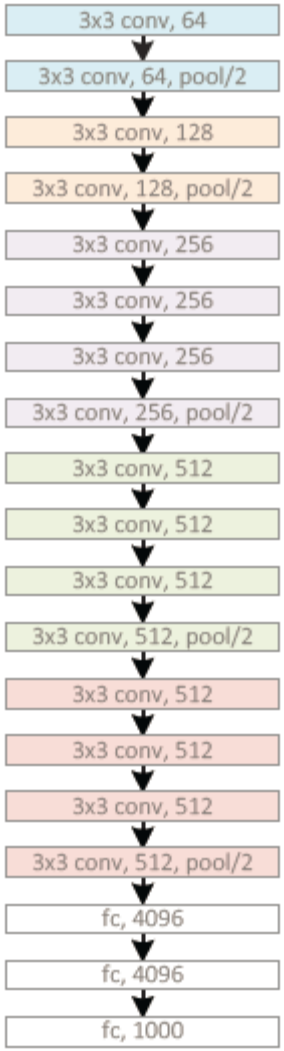
Captioning Framework



CNN-Based Video Encoder

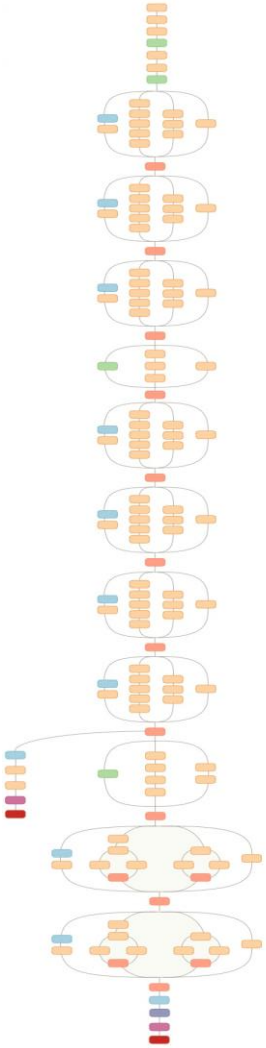


C3D

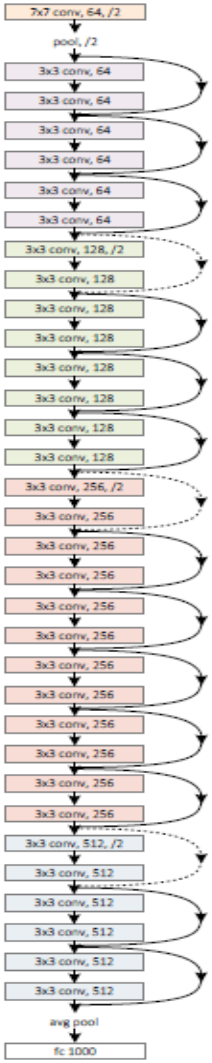


VGG-19

GoogLenet



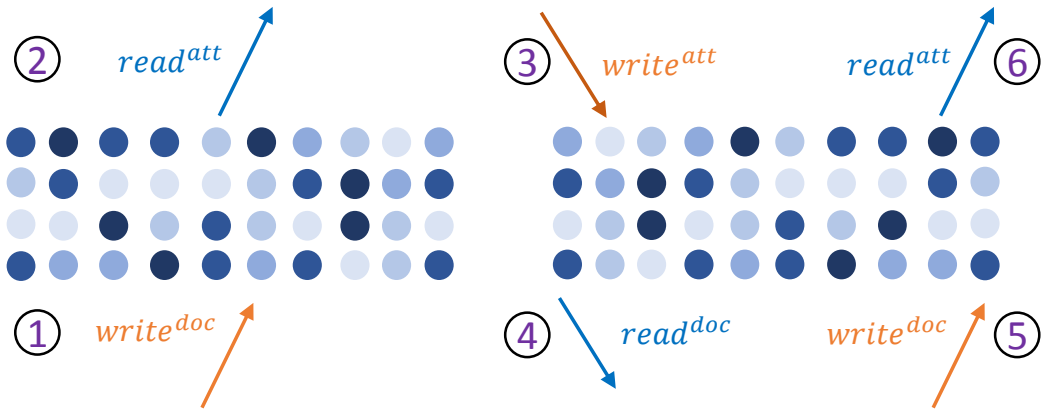
Inception-3



Residual

Multimodal Memory Modelling

- Multimodal Memory
  - $N \times M$ matrix



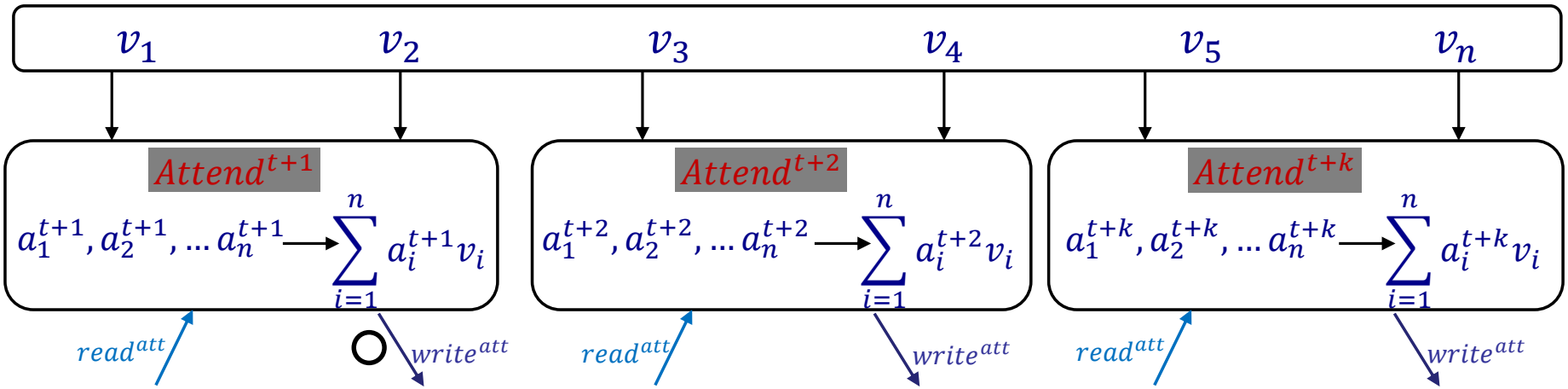
① Writing hidden representations to update memory

$$M_t(i) = M_{t-1}(i)\left[1 - w_t^{sw}(i)\, e_t^{sw}\right] + w_t^{sw}(i)\, a_t^{sw}$$

② Reading updated memory for temporal attention

$$\sum_{i=1}^{N} w_t^{vr}(i) = 1, 0 \le w_t^{vr}(i) \le 1 \qquad r_t^{vr} = \sum_{i=1}^{N} w_t^{vr}(i) M_t(i)$$

Multimodal Memory Modelling



Temporal attention selection for video representation

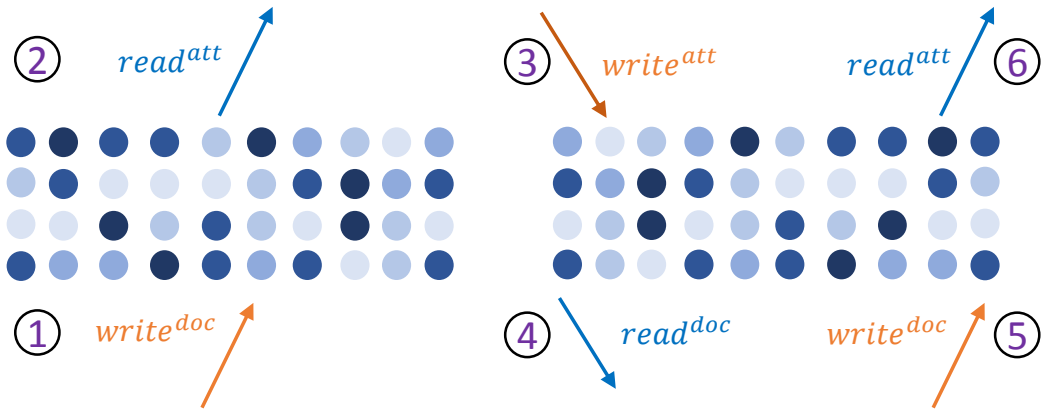
$$e_i^t = w^{\mathrm{T}} \tanh\left(W_r r_t^{vr} + U_\alpha v_i + b_\alpha\right)$$

$$\alpha_i^t = \exp\left\{e_i^t\right\} / \sum_{j=1}^{n} \exp\left\{e_j^t\right\}$$

$$V_t = \sum_{i=1}^{n} \alpha_i^t v_i$$

Multimodal Memory Modelling

- Multimodal Memory
  - $N \times M$ matrix



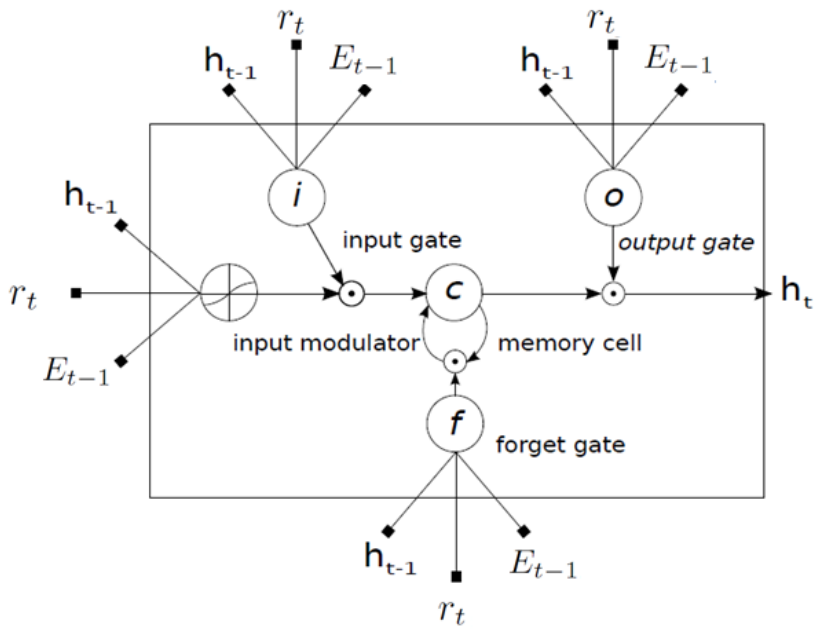
③ Writing selected visual information to update memory

$$M_t(i) = M_t(i)\left[1 - w_t^{vw}(i)\, e_t^{vw}\right] + w_t^{vw}(i)\, a_t^{vw}$$

④ Reading the updated memory for LSTM-based language model

$$\sum_{i=1}^{N} w_t^{sr}(i) = 1, 0 \le w_t^{sr}(i) \le 1 \qquad r_t^{sr} = \sum_{i=1}^{N} w_t^{sr}(i) M_t(i)$$

LSTM-Based Text Decoder



$$i_t = \sigma\left(W_i E_{t-1} + U_i h_{t-1} + M_i r_t + b_i\right)$$

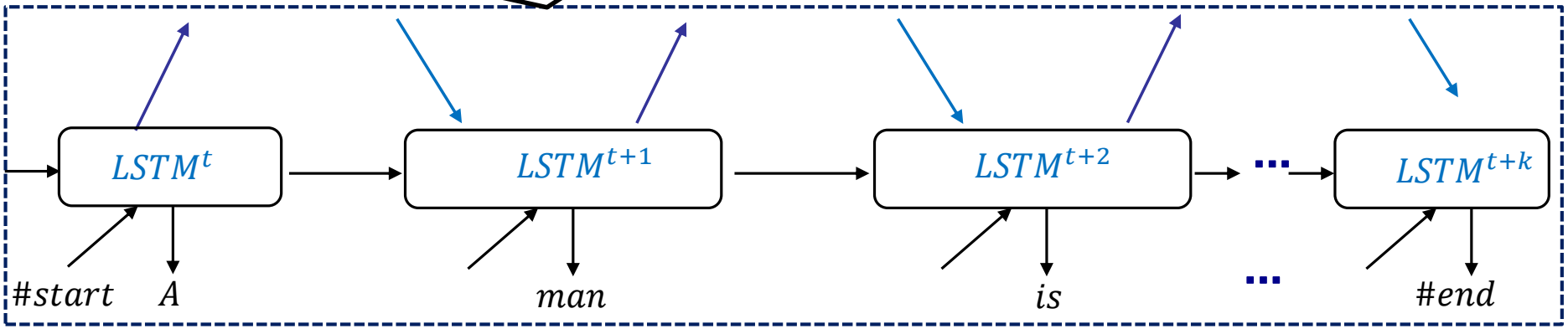
$$f_t = \sigma\left(W_f E_{t-1} + U_f h_{t-1} + M_f r_t + b_f\right)$$

$$o_t = \sigma\left(W_o E_{t-1} + U_o h_{t-1} + M_o r_t + b_o\right)$$

$$\tilde{c}_t = \phi\left(W_c E_{t-1} + U_c h_{t-1} + M_c r_t + b_c\right)$$

$$c_t = i_t \odot \tilde{c}_t + f_t \odot c_{t-1}$$

$$h_t = o_t \odot \phi\left(c_t\right)$$



Memory Addressing & Regularized Loss

- Content-Based Memory Addressing

$$K(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\| + \varepsilon} \qquad z_t(i) = \beta_t K(k_t, M_t(i))$$

$$w_t(i) = \frac{\exp(z_t(i))}{\sum_{j=1}^{N} \exp(z_t(j))}$$

- Regularized Loss

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{t_i} \log \rho\left(y_j^i | y_{1:j-1}^i, x^i, \theta\right) + \lambda \|\theta\|_2^2$$

Outline

- Introduction
- Model Description
- Experimental Results
- Conclusion

Implementation Details

- Variable-length sentences
  - A start tag and an end tag
- Beam search
  - beam size: 5
- LSTM-based decoder
  - visual hidden units:1024,word embedding size: 468
- Memory matrix
  - memory size：(128,512), GlorotUniform
  - read weight and write weight initial with OneHot
- Others
  - Minibatch: 64, optimization algorithm: ADADELTA
  - Dropout with rate of 0.5, gradient norm clipped (-10,10)

Experimental Results

- Microsoft Video Description Dataset

Method	B@1	B@2	B@3	B@4	METEOR
FGM [30]	-	-	-	13.68%	23.90%
LSTM-YT [33]	-	-	-	33.29%	29.07%
SA [41]	-	-	-	40.28%	29.00%
S2VT [32]	-	-	-	-	29.2%
LSTM-E [23]	74.9%	60.9%	50.6%	40.2%	29.5%
p-RNN [42]	77.3%	64.5%	54.6%	44.3%	31.1%
HRNE [22]	79.2%	66.3%	55.1%	43.8%	**33.1%**
BGRCN [2]	-	-	-	49.63%	31.7%
M^3-c3d	77.30%	68.20%	56.30%	45.50%	29.91%
M^3-vgg19	77.70%	67.50%	58.90%	49.60%	30.09%
M^3-google	79.05%	68.74%	60.00%	51.17%	31.47%
M^3-res	80.80%	69.90%	60.40%	49.32%	31.10%
M^3-inv3	**81.56%**	**71.39%**	**62.34%**	**52.02%**	32.18%

Method	B@1	B@2	B@3	B@4	METEOR
SA-G-3C [41]	-	-	-	41.92%	29.60%
S2VT-rgb-flow [32]	-	-	-	-	29.8%
LSTM-E-VC [23]	78.8%	66.0%	55.4%	45.3%	31.0%
p-RNN-VC [42]	81.5%	70.4%	60.4%	49.9%	32.6%
M^3-VC	81.90%	71.26%	62.08%	51.78%	32.49%
M^3-IC	**82.45%**	**72.43%**	**62.78%**	**52.82%**	**33.31%**

CVPR2017

Table 1. Experiment results on the Youtube2Text Dataset.

Model	BLEU@4	METEOR	CIDEr	Feat.	Fine
VGG-Encoder	0.404	0.295	0.515	No	No
GoogLeNet-Encoder	0.427	0.303	0.534	No	No
t-KeyAddressing (G)	0.436	0.308	0.545	No	No
m-KeyAddressing (Memory LSTM) (G)	**0.457**	**0.319**	**0.573**	No	No

ICLR2017

Table 2: Video captioning evaluation on the test set of 670 videos in MSVD.

Method	METEOR	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr
Venugopalan et al. (2015b)	27.7	–	–	–	–	–
Venugopalan et al. (2015a)	29.2	–	–	–	–	–
Pan et al. (2016b)	29.5	74.9	60.9	50.6	40.2	–
Yu et al. (2016)	31.10	77.30	64.50	54.60	44.30	–
Pan et al. (2016a)	**33.10**	79.20	66.30	55.10	43.80	–
Our Model	31.80	**79.40**	**67.10**	**56.80**	**46.10**	**62.70**

Experimental Results

- Microsoft Research-Video to Text Dataset
  - the largest dataset in terms of sentence and vocabulary, 10,000 video clips and 200,000 sentences
  - each video is labelled with about 20 sentences
  - training set (6513), validation set (497), and test set (2990)

Method	B@1	B@2	B@3	B@4	METEOR
SA-V [41]	67.82%	55.41%	42.90%	34.73%	23.11%
SA-C [41]	68.90%	57.50%	47.00%	37.40%	24.80%
SA-VC [41]	72.20%	58.90%	46.80%	35.90%	24.90%
M^3-V	70.20%	56.60%	44.80%	35.00%	24.60%
M^3-C	**77.20%**	**61.30%**	47.20%	35.10%	25.70%
M^3-VC	73.60%	59.30%	**48.26%**	**38.13%**	**26.58%**

Experimental Results

➢Description Generation

SA： Yao et al. ICCV 2015 M^3： Our model

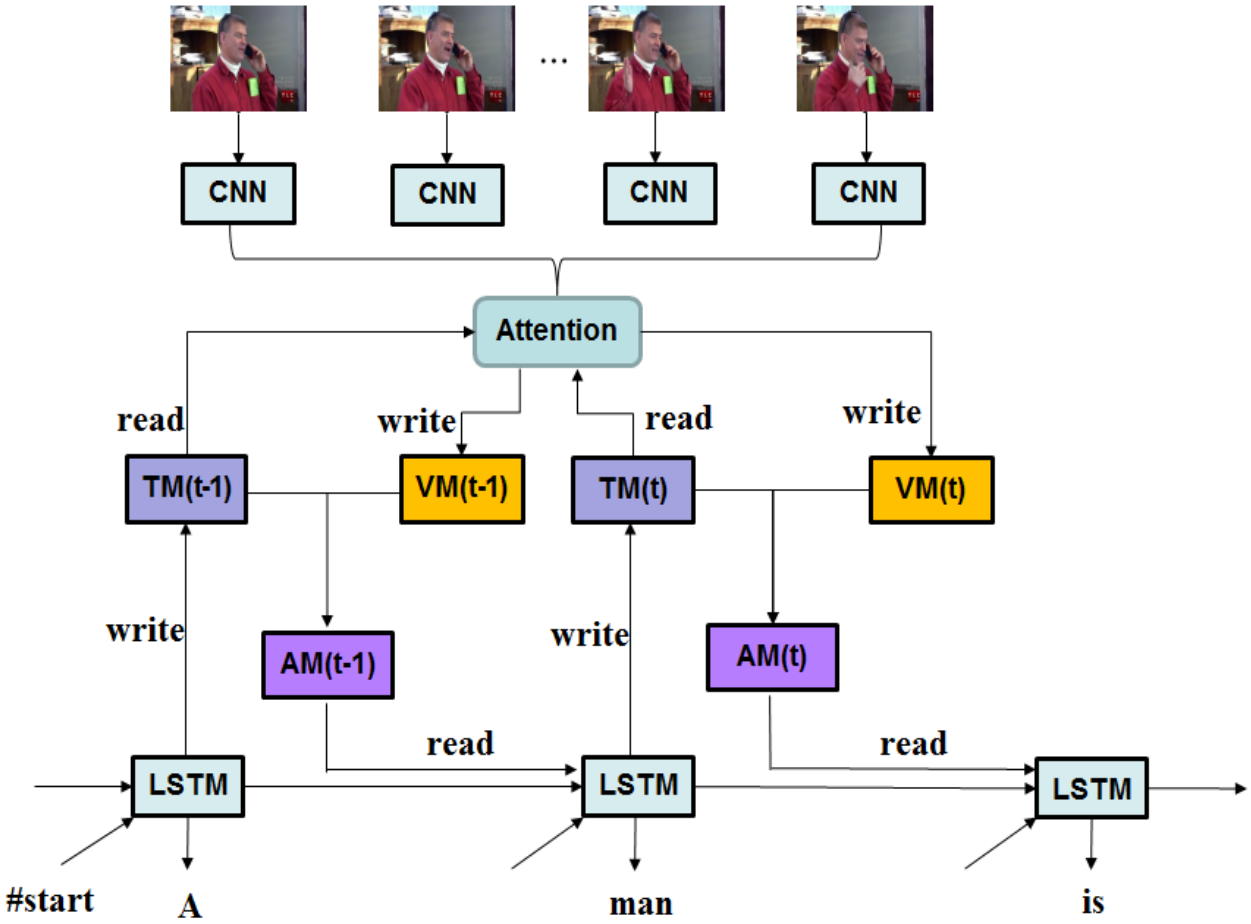
	Generated Sentence: SA: a man is riding a bike M^3:a man is riding a motorcycle	Reference Sentece: 1. a man is riding a motorcycle 2. a man rides a motorcycle on a beach 3. a man riding on his bike in waves
	Generated Sentence: SA: a man is playing a guitar M^3:a man is playing with a dog	Reference Sentece: 1. a man is petting two dogs 2. a man pets some dogs 3. a man is play with pets
	Generated Sentence: SA: men are playing soccer ball M^3: people are playing basketball	Reference Sentece: 1. a basketball game is in play 2. two teams playing basket ball 3. people are playing basketball
	Generated Sentence: SA: a man is riding a horse M^3: a woman is riding a horse	Reference Sentece: 1. a woman is riding her horse 2. a woman riding in horse competition 3. a woman is riding a horse
	Generated Sentence: SA: a woman is peeling a fish M^3: a woman is peeling a shrimp	Reference Sentece: 1. a woman is peeling shrimp 2. the lady peeled the shrimp 3. a woman skins shrimp

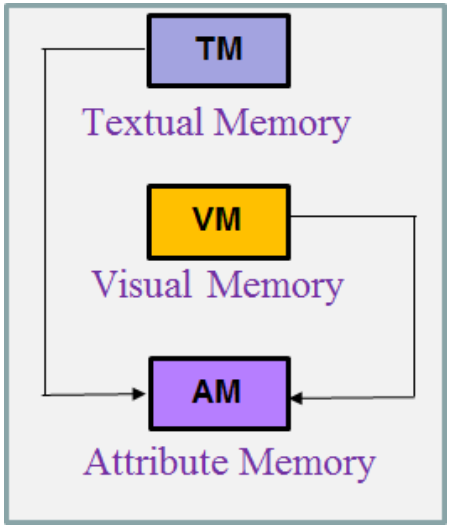
Outline

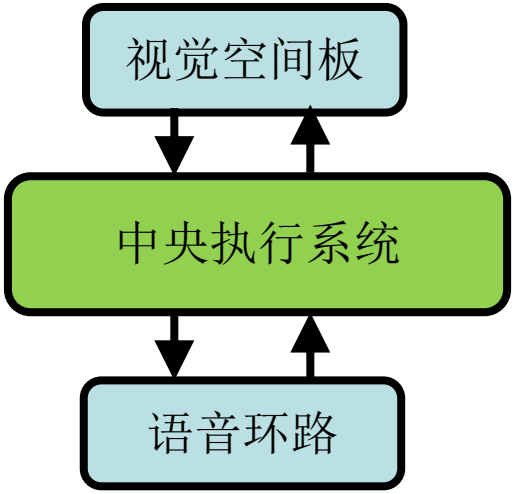
- Introduction
- Model Description
- Experimental Results
- Conclusion

Conclusion & Future Work

- Textual/Visual/Attribute Memory







Working Memory, Baddeley et al.

Acknowledgement

CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

NVIDIA

NVAIL

Artificial Intelligence Laboratory

Sponsor excellent hardware resources

Thank you!

+

(Q/A)

wangliang@nlpr.ia.ac.cn