

# Video Captioning via Hierarchical Reinforcement Learning

---

Xin Wang, Wenhua Chen, Jiawei Wi, Tuan-Fang Wang, William Yang Wang

Publishing year: 2018

Presenter: David Radke

CS885 – University of Waterloo – July, 2020

# Overview



- Problem: automatic video captioning for machines is a challenging problem
  - Past solutions:
    - Image captioning (static scene)
    - Short simple sentences
- Why is this important?
  - Intelligent video surveillance
  - Assistance to visually impaired people

# Related Work



- LSTM for video captioning (seq2seq) [Venugopalan et. al, 2015]
  - Improvements: Attention [Yao et. al, 2015][Yu et. al, 2016], hierarchical RNN [Pan et. al, 2016][Yu et. al, 2016], multi-task learning [Pasunuru et. al, 2017], etc...
  - Most use max-likelihood given previous ground-truth outputs which is not available at test time
- REINFORCE [Ranzato et al, 2015] for video captioning led to highly variant and unstable gradient
  - Could formulate as Actor-Critic, or REINFORCE-baseline
    - Fail to grasp the high-level semantic flow

# High Level Idea

- Generate captions segment-by-segment
- “Divide and conquer” approach by dividing long captions into short segments, allowing different modules to generate short text



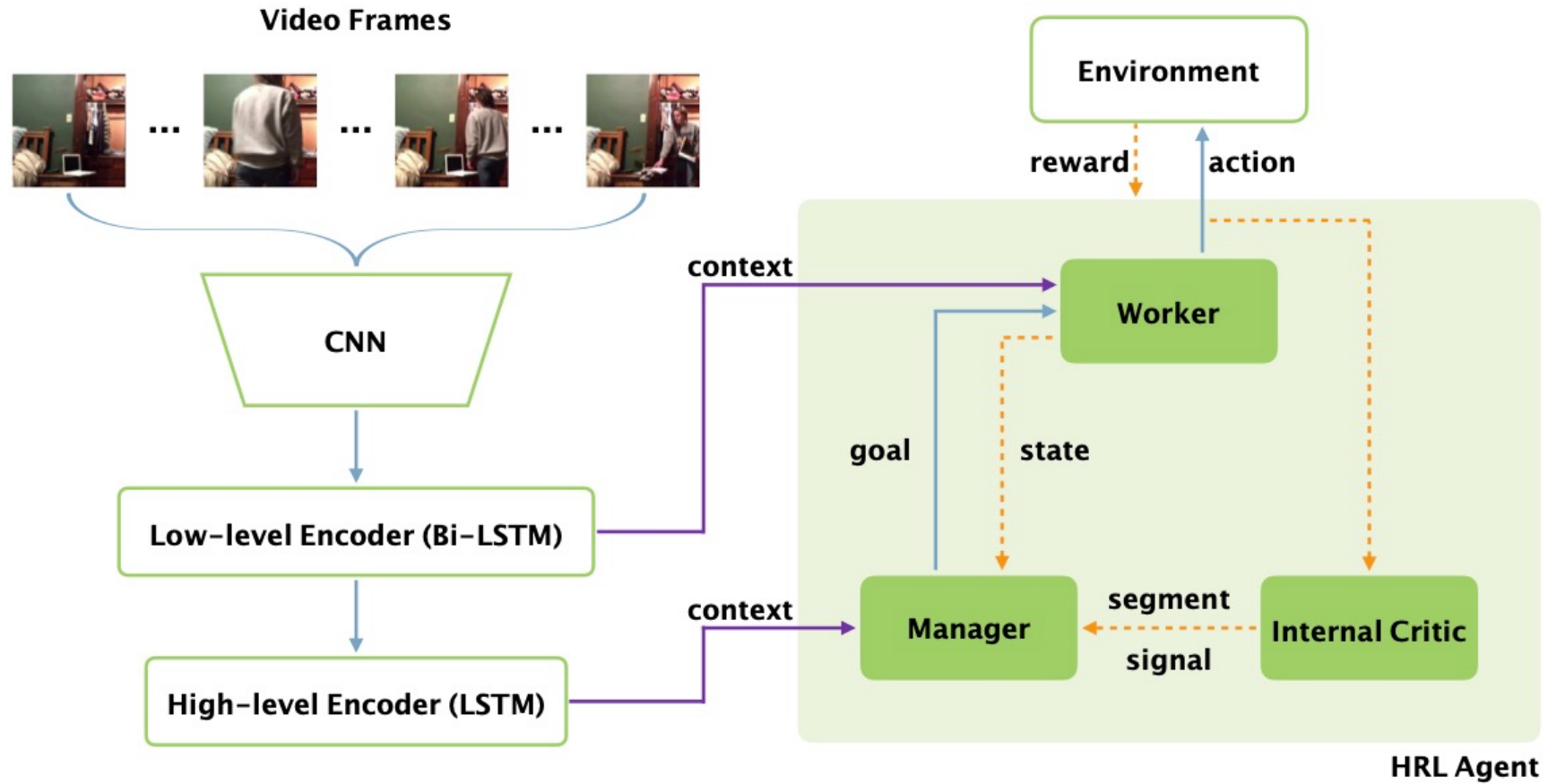
**Caption: A person sits on a bed and puts a laptop into a bag. The person stands up, puts the bag on one shoulder, and walks out of the room.**

# Framework

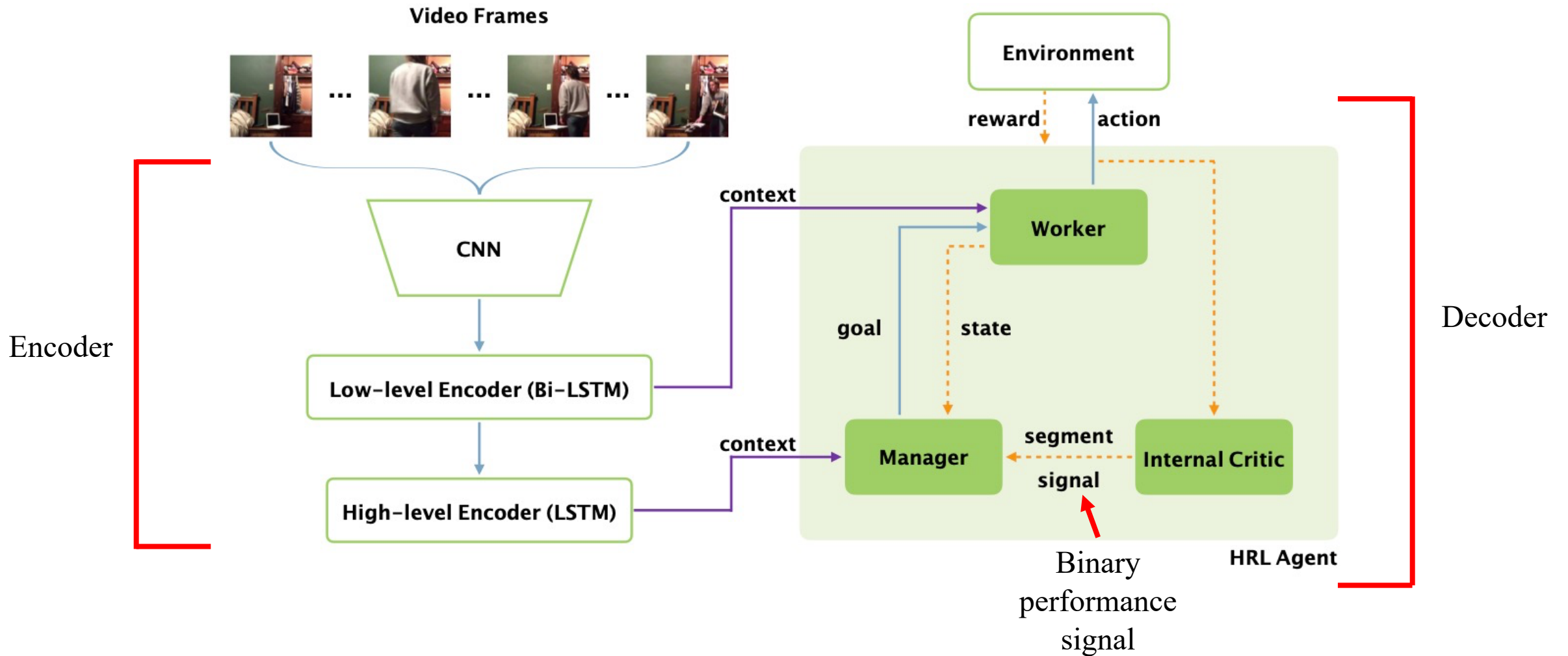


- Environment: textual and video context
- Modules:
  - **Manager**: sets goals at lower temporal resolution
  - **Worker**: selects primitive actions at every step following goals from manager
  - **Internal Critic**: determines if a goal is accomplished by worker
- Actions: worker generating segment of words sequentially
- Details:
  - Manager and worker both have an **attention module** over video frames
  - Exploits the extrinsic rewards in different time spans – first work to consider hierarchical RL in intersection of vision and language

# Workflow



# Workflow



# Syntax

---

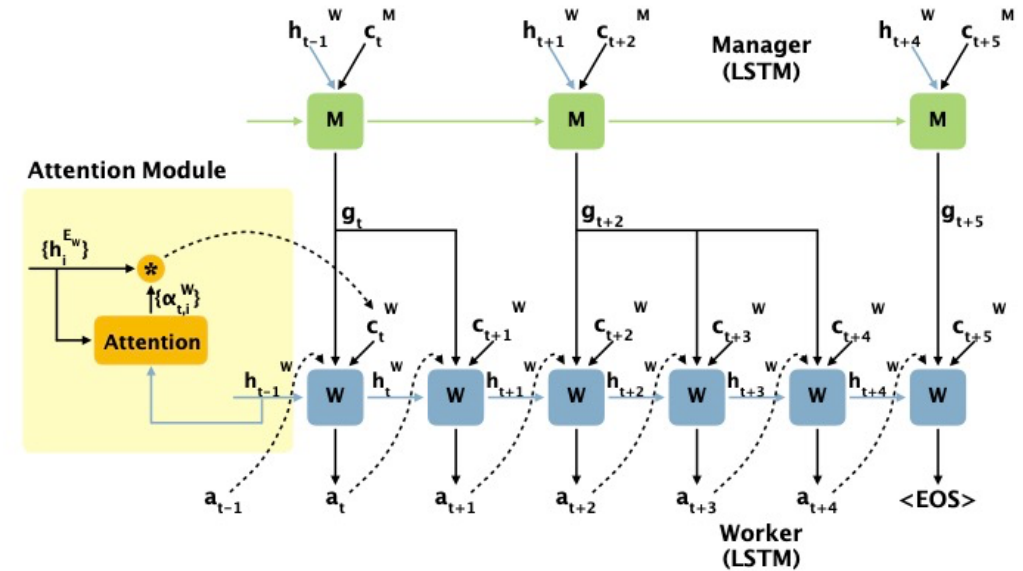
- Video frames:  $v = \{v_i\}$  for times  $i \in \{1, \dots, n\}$
- High and low-level encoder outputs  
for worker:  $h^{E_w} = \{h_i^{E_w}\}$       for manager:  $h^{E_m} = \{h_i^{E_m}\}$
- Decoder output language:  $a_1 a_2 \dots a_T \in V^T$ ; where  $T$  is caption length and  $V$  is the vocabulary set.



# Attention!

- Creates a *context vector* for decoder
  - Bahdanau-style attention (not cited)

$$c_t^W = \sum \alpha_{t,i}^W h_i^{E_w}$$



# Attention!

- Creates a *context vector* for decoder
  - Bahdanau-style attention (not cited)

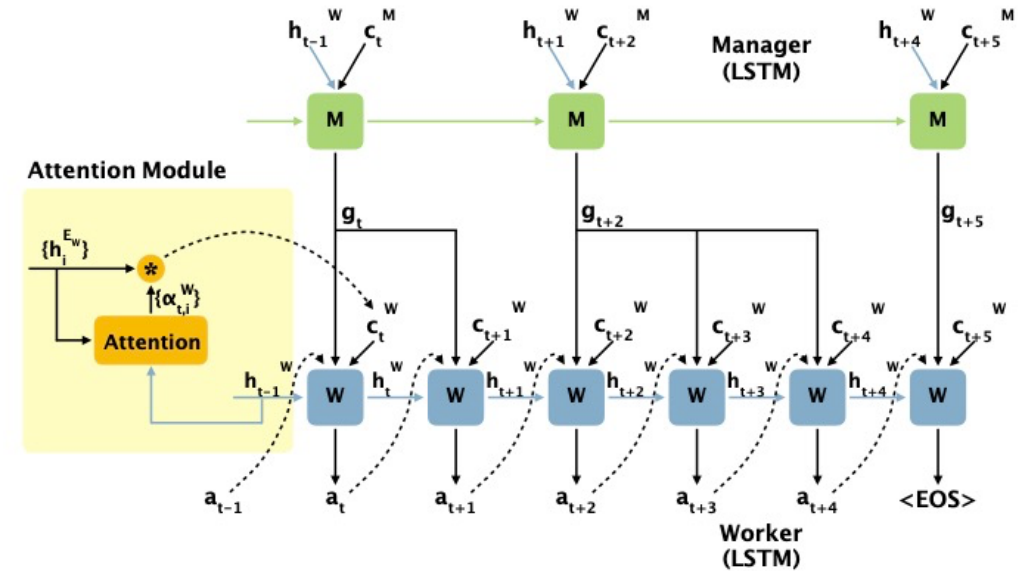
$$c_t^W = \sum \alpha_{t,i}^W h_i^{E_w}$$

- How to find alpha?

$$\alpha_{t,i}^W = \frac{\exp(e_{t,i})}{\sum_{k=1}^n \exp(e_{t,k})}$$

where

$$e_{t,i} = w^T \tanh(W_a h_i^{E_w} + U_a h_{t-1}^W + b_a)$$



# Attention!

- Creates a *context vector* for decoder
  - Bahdanau-style attention (not cited)

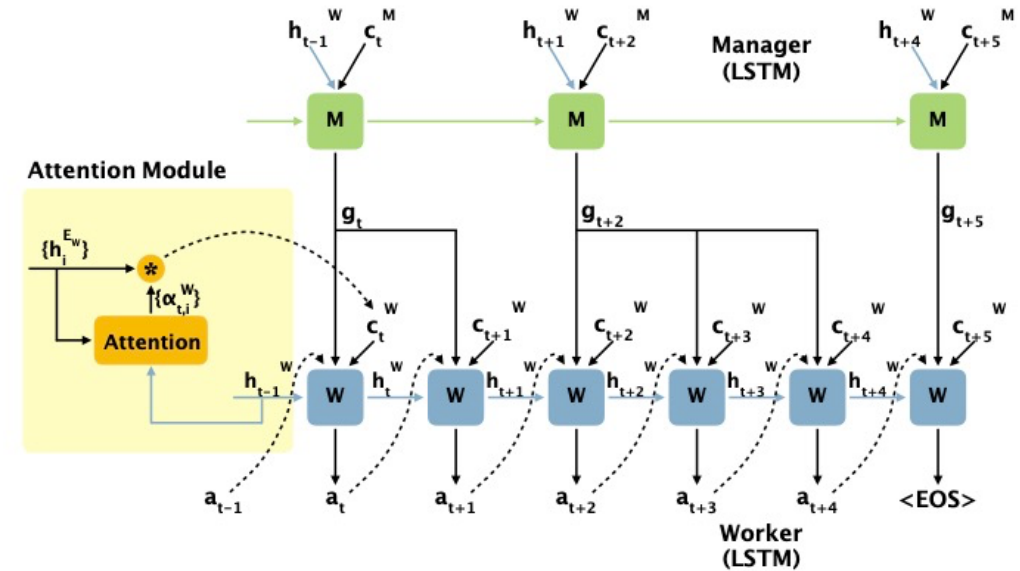
$$c_t^W = \sum \alpha_{t,i}^W h_i^{E_w}$$

- How to find alpha?

$$\alpha_{t,i}^W = \frac{\exp(e_{t,i})}{\sum_{k=1}^n \exp(e_{t,k})}$$

where

$$e_{t,i} = w^T \tanh(W_a h_i^{E_w} + U_a h_{t-1}^W + b_a)$$



# Attention!

- Creates a *context vector* for decoder
  - Bahdanau-style attention (not cited)

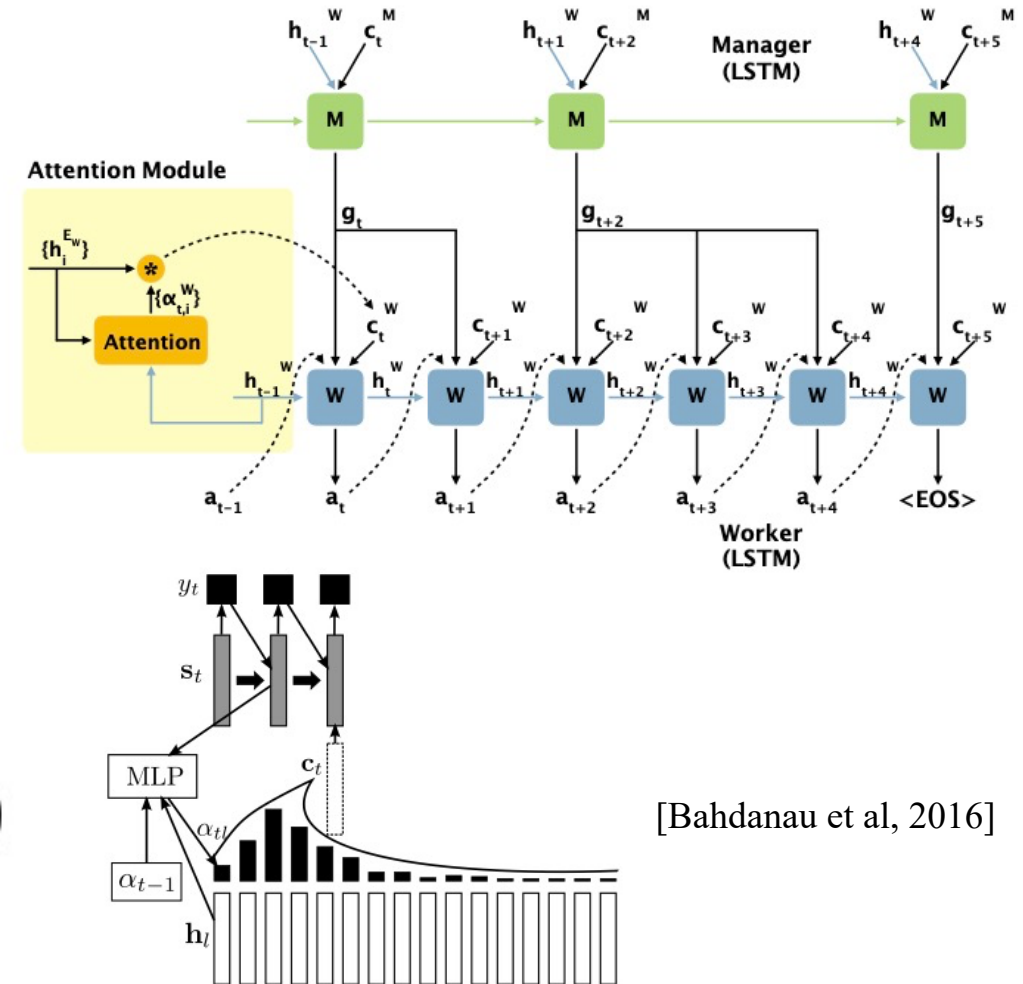
$$c_t^W = \sum \alpha_{t,i}^W h_i^{E_w}$$

- How to find alpha?

$$\alpha_{t,i}^W = \frac{\exp(e_{t,i})}{\sum_{k=1}^n \exp(e_{t,k})}$$

where

$$e_{t,i} = w^T \tanh(W_a h_i^{E_w} + U_a h_{t-1}^W + b_a)$$




[Bahdanau et al, 2016]

# Critic Details

---

- Hidden state:  $h_t^I = RNN(h_{t-1}^I, a_t)$
- Probability of internal critic signal:  $p(z_t) = \text{sigmoid}(W_z h_t^I + b_z)$
- Training goal: maximize likelihood of given ground truth signal  $\{z_t^*\}$

$$\operatorname{argmax} \sum_t \log p(z_t^* | a_1, \dots, a_{t-1})$$


- Note: didn't they criticize past work for doing this same thing?

# Learning Details

---

- REINFORCE with a baseline for worker:

$$\nabla_{\theta_w} L(\theta_w) \approx -(R(a_t) - b_t^w) \nabla_{\theta_w} \log \pi_{\theta_w}(a_t)$$

- Set worker as static oracle and update manager:

$$\nabla_{\theta_m} L(\theta_m) = -(R(e_{t,c}) - b_t^m) \left[ \sum_{i=t}^{t+c-1} \nabla_{g_t} \log \pi(a_i) \right] \nabla_{\theta_m} \mu_{\theta_m}(s_t)$$

- Gaussian distribution perturbation added to manager policy for exploration

# Reward Details

---

- CIDEr reward

Let  $f(x) = \text{CIDEr}(sent + x) - \text{CIDEr}(sent)$

$$R(a_t) = \sum_{k=0}^{\infty} \gamma^k f(a_{t+k})$$

$$R(e_t) = \sum_{n=0}^{\infty} \gamma^n f(e_{t+n})$$

# Experiments

---

- Datasets:
  - MSR-VTT (10k video clips - Amazon Mechanical Turk (AMT) captions)
  - Charades Captions (~10k indoor activity video clips – also AMT)
- For critic, manually break captions into semantic chunks
- Metrics:
  - BLEU
  - METEOR
  - ROUGE-L
  - CIDEr-D
- Compare with other state-of-the-art algorithms



# Results

- MSR-VTT

Method	BLEU@4	METEOR	ROUGE-L	CIDEr
Mean-Pooling	30.4	23.7	52.0	35.0
Soft-Attention	28.5	25.0	53.3	37.1
S2VT	31.4	25.7	55.9	35.2
v2t_navigator	40.8	28.2	60.9	44.8
Aalto	39.8	26.9	59.8	45.7
VideoLAB	39.1	27.7	60.6	44.1
XE-baseline	<b>41.3</b>	27.6	59.9	44.7
RL-baseline	40.6	28.5	60.7	46.3
HRL (Ours)	<b>41.3</b>	<b>28.7</b>	<b>61.7</b>	<b>48.0</b>

- Charades

Method	B@1	B@2	B@3	B@4	M	R	C
XE-baseline	55.0	36.4	23.6	15.0	18.7	39.0	16.7
RL-baseline	57.6	41.4	28.0	<b>18.8</b>	17.7	39.8	21.6
HRL-16	<b>64.4</b>	<b>44.3</b>	<b>29.4</b>	<b>18.8</b>	<b>19.5</b>	<b>41.4</b>	23.2
HRL-32	64.0	43.4	28.4	17.9	19.2	41.0	21.3
HRL-64	61.7	43.0	28.8	<b>18.8</b>	18.7	31.2	<b>23.6</b>

# Results

- MSR-VTT

Method	BLEU@4	METEOR	ROUGE-L	CIDEr
Mean-Pooling	30.4	23.7	52.0	35.0
Soft-Attention	28.5	25.0	53.3	37.1
S2VT	31.4	25.7	55.9	35.2
v2t_navigator	40.8	28.2	60.9	44.8
Aalto	39.8	26.9	59.8	45.7
VideoLAB	39.1	27.7	60.6	44.1
XE-baseline	<b>41.3</b>	27.6	59.9	44.7
RL-baseline	40.6	28.5	60.7	46.3
HRL (Ours)	<b>41.3</b>	<b>28.7</b>	<b>61.7</b>	<b>48.0</b>

- Charades

Method	B@1	B@2	B@3	B@4	M	R	C
XE-baseline	55.0	36.4	23.6	15.0	18.7	39.0	16.7
RL-baseline	57.6	41.4	28.0	<b>18.8</b>	17.7	39.8	21.6
HRL 16	<b>64.4</b>	<b>44.3</b>	<b>29.4</b>	<b>18.8</b>	<b>19.5</b>	<b>41.4</b>	23.2
HRL 32	64.0	43.4	28.4	17.9	19.2	41.0	21.3
HRL 64	61.7	43.0	28.8	<b>18.8</b>	18.7	31.2	<b>23.6</b>

Dimensionality of  
the latent vectors

# Results

- MSR-VTT

Method	BLEU@4	METEOR	ROUGE-L	CIDEr
Mean-Pooling	30.4	23.7	52.0	35.0
Soft-Attention	28.5	25.0	53.3	37.1
S2VT	31.4	25.7	55.9	35.2
v2t_navigator	40.8	28.2	60.9	44.8
Aalto	39.8	26.9	59.8	45.7
VideoLAB	39.1	27.7	60.6	44.1
XE-baseline	<b>41.3</b>	27.6	59.9	44.7
RL-baseline	40.6	28.5	60.7	46.3
HRL (Ours)	<b>41.3</b>	<b>28.7</b>	<b>61.7</b>	<b>48.0</b>

- Charades

Method	B@1	B@2	B@3	B@4	M	R	C
XE-baseline	55.0	36.4	23.6	15.0	18.7	39.0	16.7
RL-baseline	57.6	41.4	28.0	<b>18.8</b>	17.7	39.8	21.6
HRL-16	<b>64.4</b>	<b>44.3</b>	<b>29.4</b>	<b>18.8</b>	<b>19.5</b>	<b>41.4</b>	23.2
HRL-32	64.0	43.4	28.4	17.9	19.2	41.0	21.3
HRL-64	61.7	43.0	28.8	<b>18.8</b>	18.7	31.2	<b>23.6</b>

# Results

- MSR-VTT

Method	BLEU@4	METEOR	ROUGE-L	CIDEr
Mean-Pooling	30.4	23.7	52.0	35.0
Soft-Attention	28.5	25.0	53.3	37.1
S2VT	31.4	25.7	55.9	35.2
v2t_navigator	40.8	28.2	60.9	44.8
Aalto	39.8	26.9	59.8	45.7
VideoLAB	39.1	27.7	60.6	44.1
XE-baseline	<b>41.3</b>	27.6	59.9	44.7
RL-baseline	40.6	28.5	60.7	46.3
HRL (Ours)	<b>41.3</b>	<b>28.7</b>	<b>61.7</b>	<b>48.0</b>

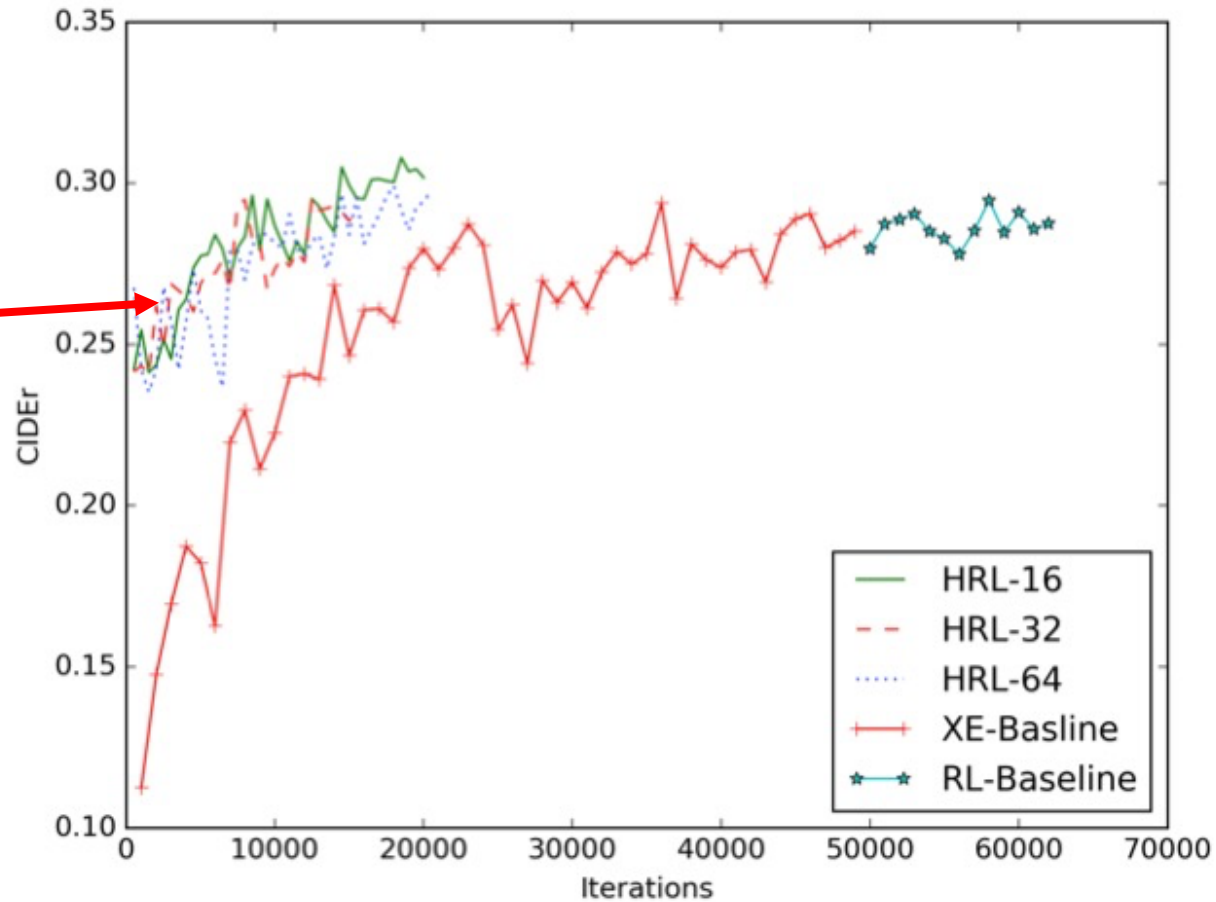
- Charades

Method	B@1	B@2	B@3	B@4	M	R	C
XE-baseline	55.0	36.4	23.6	15.0	18.7	39.0	16.7
RL-baseline	57.6	41.4	28.0	<b>18.8</b>	17.7	39.8	21.6
HRL-16	<b>64.4</b>	<b>44.3</b>	<b>29.4</b>	<b>18.8</b>	<b>19.5</b>	<b>41.4</b>	23.2
HRL-32	64.0	43.4	28.4	17.9	19.2	41.0	21.3
HRL-64	61.7	43.0	28.8	<b>18.8</b>	18.7	31.2	<b>23.6</b>

Charades captions longer, HRL model gains better improvement over baseline for longer videos

# Results – Charades in Detail

No significant difference  
in latent vector size



# Discussion

- First work to consider hierarchical RL in intersection of vision and language
- Good background, a lot of space used for derivations which could have been used to discussed results further
- Would have been nice to include more examples of captions

- i.e.



**GROUND TRUTH:** person walks in room holding phone , sits at table , looks at phone , smiles , put phone down gets up , looks out window and walks out of room .

**XE-BASELINE:** a person is standing in the doorway . the person is standing in the doorway . the person is standing in the doorway . the person is standing in the doorway .

**RL-BASELINE:** a person is sitting on a chair . the person opens the door and walks out .

**HRL:** a person | is sitting in a chair , | and takes a book . | the person | opens the window | and closes the door .

# Future Work

---

- “explore attention space”
  - Leong-style attention
  - Spaciotemporal attention in video frames
    - This paper only uses temporal
- Adversarial game-like training of manager and worker

# References

- L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015
- H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593, 2016
- Y. Yu, H. Ko, J. Choi, and G. Kim. Video captioning and retrieval models with semantic attention. *arXiv preprint arXiv:1610.02947*, 2016
- P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1038, 2016
- R. Pasunuru and M. Bansal. Multi-task video captioning with video and entailment generation. *arXiv preprint arXiv:1704.07489*, 2017
- M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015
- Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2015. End-to-end attention-based large vocabulary speech recognition. *CoRR*, abs/1508.04395

