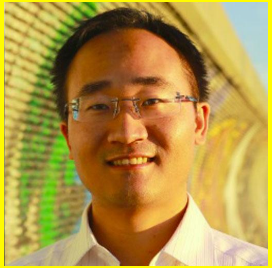
Deep Reinforcement Learning-based Image Captioning with Embedding Reward


Zhou Ren[1] Xiaoyu Wang[1] Ning Zhang[1] Xutao Lv[1] Li-Jia Li[2]

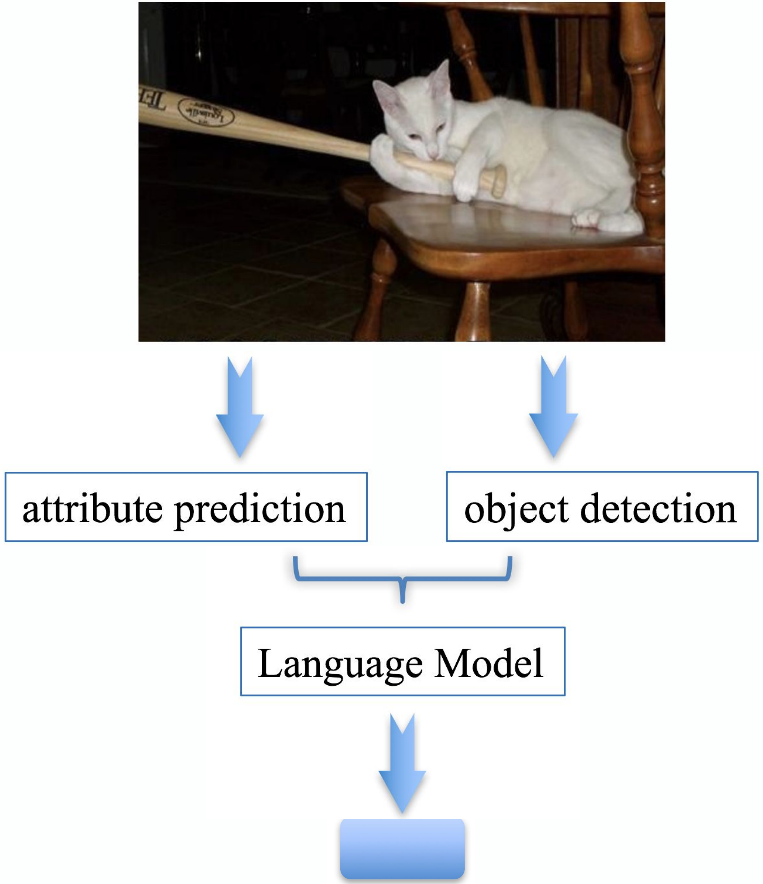
[1] Snap Research [2] Google

Image captioning

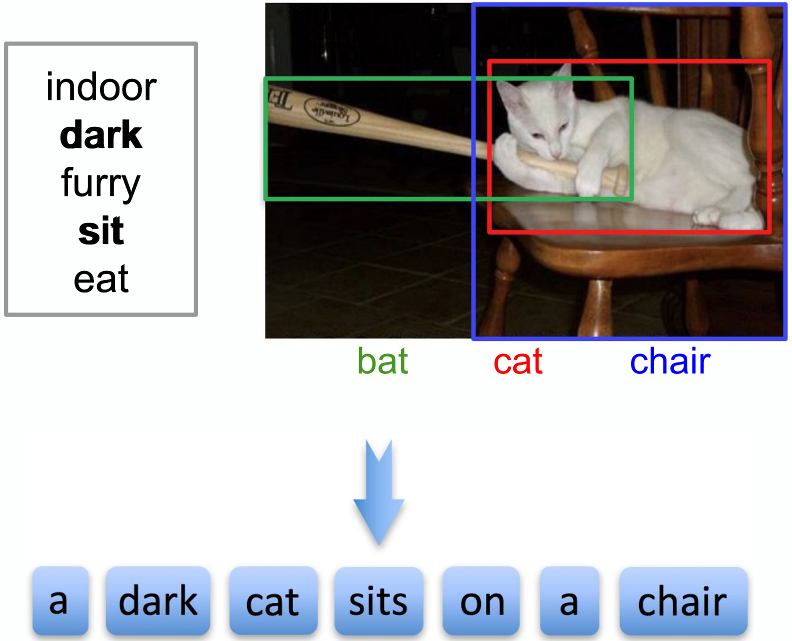
group of people learning how to surf on the beach

[Farhadi *et al*. ECCV 2010]
[Kulkarni *et al.* CVPR 2011]
[Yang *et al*. EMNLP 2011]
[Fang *et al.* CVPR 2015]
[Lebret *et al*. ICLR 2015]
[Mao *et al.* ICLR 2015]
[Vinyals, *et al*. CVPR 2015]
[Karpathy *et al.* CVPR 2015]
[Chen *et al.* CVPR 2015]
[Xu *et al.* ICML 2015]
[Johnson *et al*. CVPR 2016]
[You *et al.* CVPR 2016]
….

Previous work

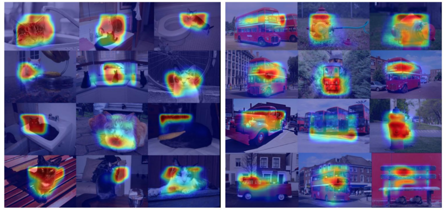


example:

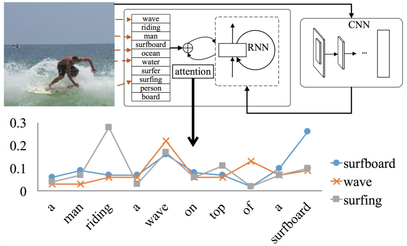


[Farhadi *et al*. 2010; Kulkarni *et al.* 2011; Yang *et al*. 2011]

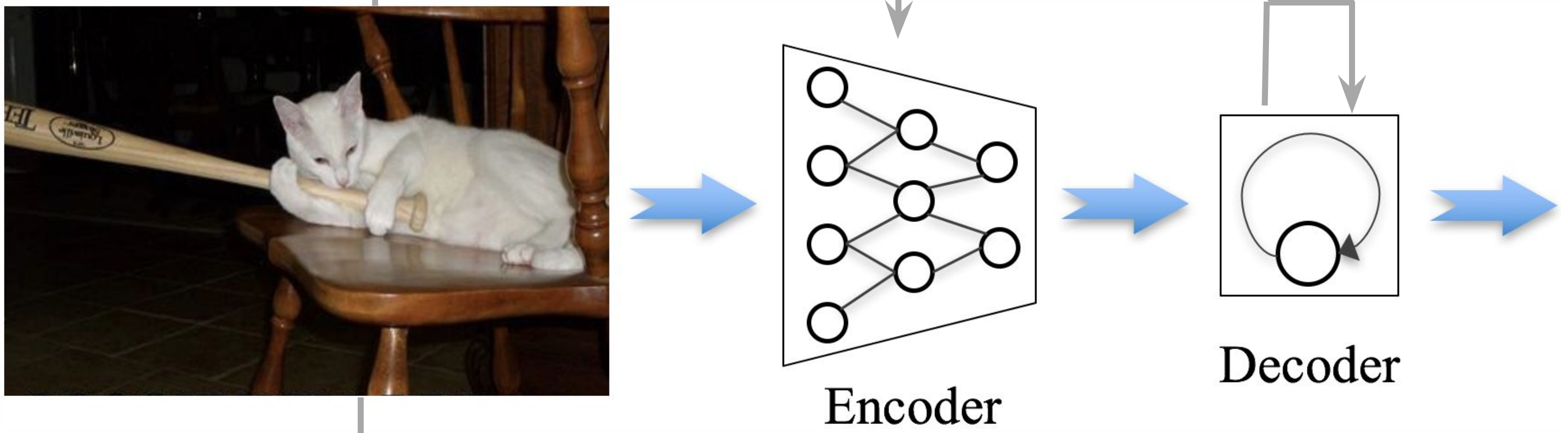
Previous work

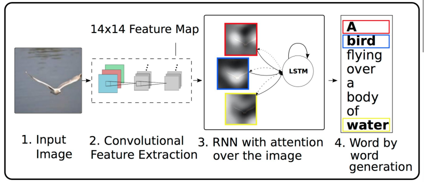


word detection [Fang *et al.* CVPR 2015]



semantic attention [You *et al.* CVPR 2016]





spatial attention [Xu *et al.* ICML 2015]

[Lebret *et al.* 2015; Mao *et al.* 2015; Vinyals, *et al.* 2015; Karpathy *et al.* 2015]

Motivation

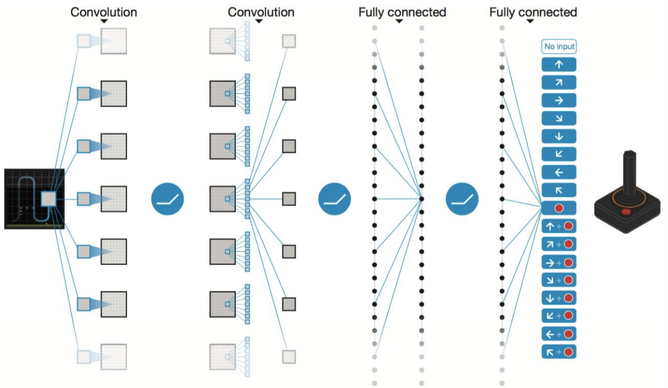
- Limitations of current mainstream framework (encoder-decoder)
    - only **local** information is utilized
    - prone to **accumulate** generation errors during inference
    - **sensitive** to beam sizes during beam search

local

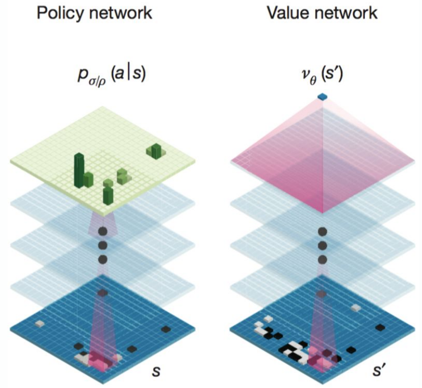
- Our target
    - better at utilizing the **global** information
    - be able to **compensate** errors
    - **less sensitive** to beam sizes

Decision-Making framework with Reinforcement Learning

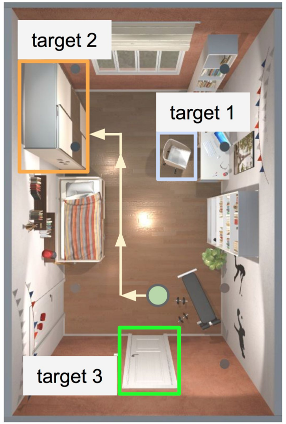
Why using decision-making?



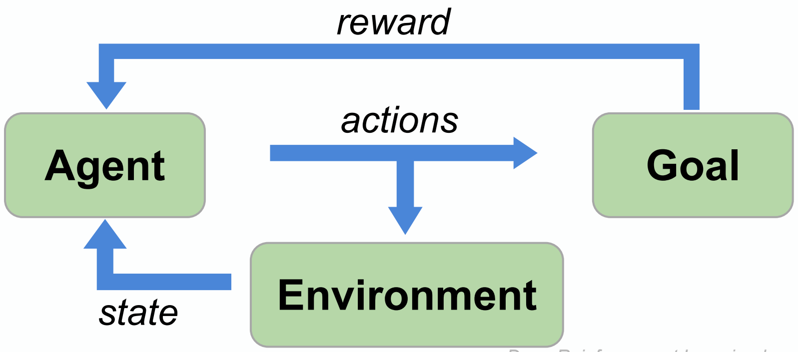
Human-level gaming control
[Mnih *et al.* Nature 2015]



AlphaGo
[Silver *et al.* Nature 2016]

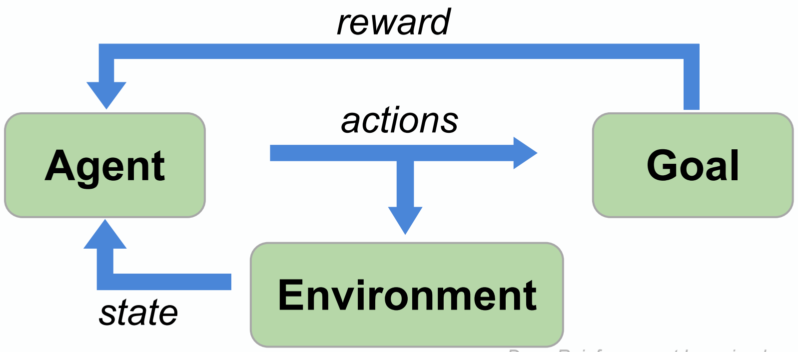


Visual navigation
[Zhu *et al.* ICRA 2017]



Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

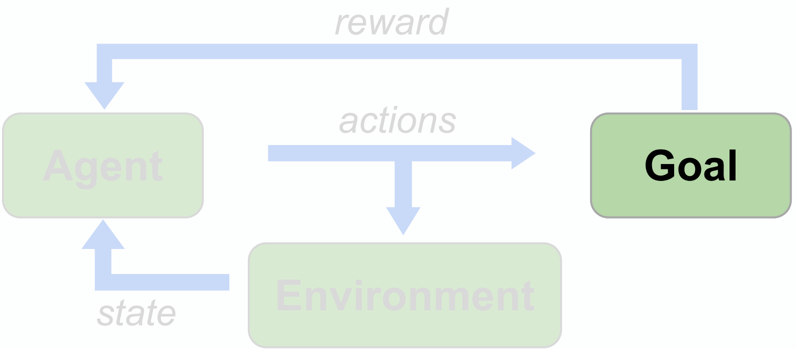
Image captioning reformulation in decision-making



Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

Image captioning reformulation in decision-making

- Goal: to generate a visual description given an image



Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

Image captioning reformulation in decision-making

- Goal: to generate a visual description given an image
- Agent: the image captioning model to learn

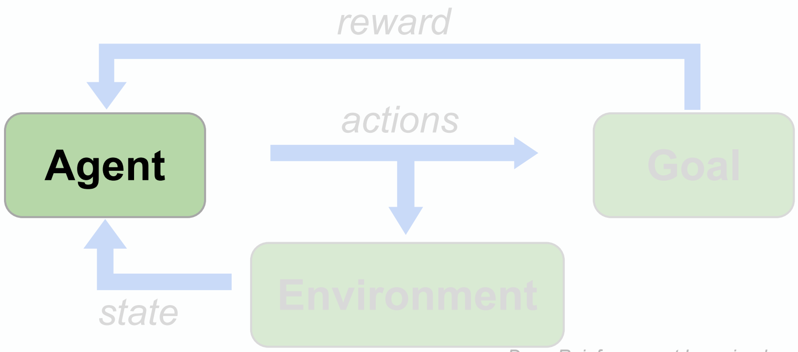
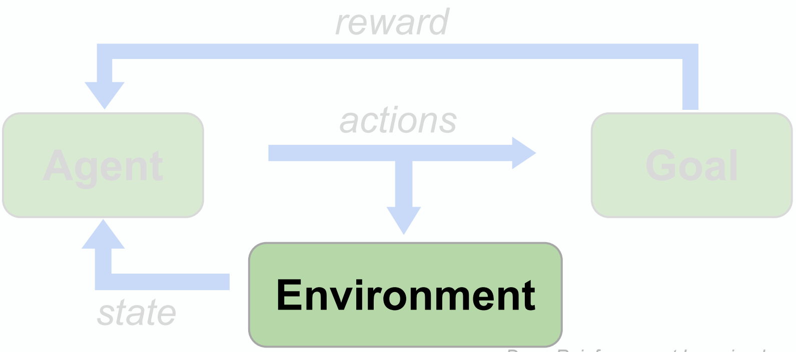


Image captioning reformulation in decision-making

- Goal: to generate a visual description given an image
- Agent: the image captioning model to learn
- Environment: the given image $\mathbf{I}$ + the words predicted so far $\{w_1, ..., w_t\}$



Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

Image captioning reformulation in decision-making

- Goal: to generate a visual description given an image
- Agent: the image captioning model to learn
- Environment: the given image $\mathbf{I}$ + the words predicted so far $\{w_1, ..., w_t\}$
- State: representation of the environment at t, $s_t = \{\mathbf{I}, w_1, ..., w_t\}$

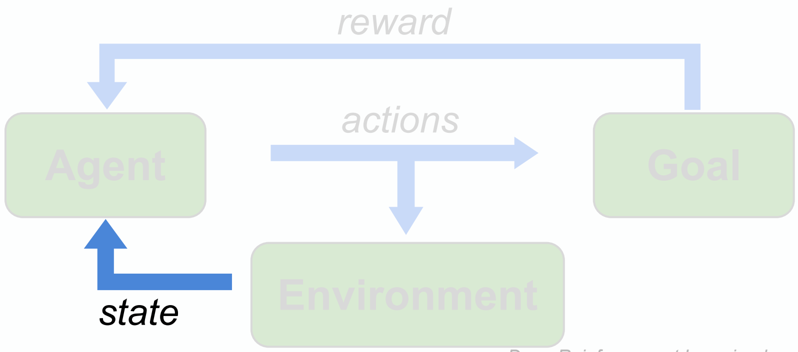
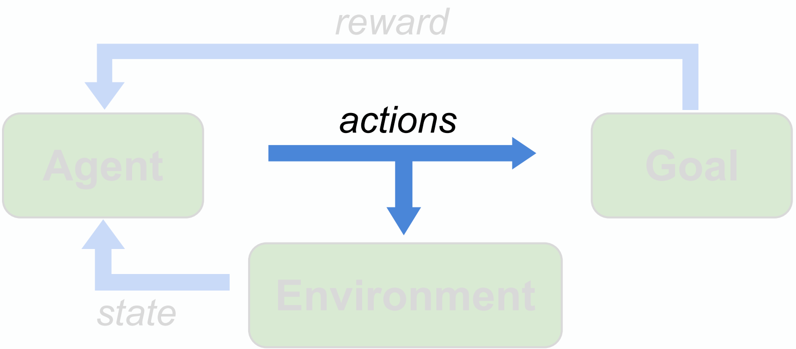


Image captioning reformulation in decision-making

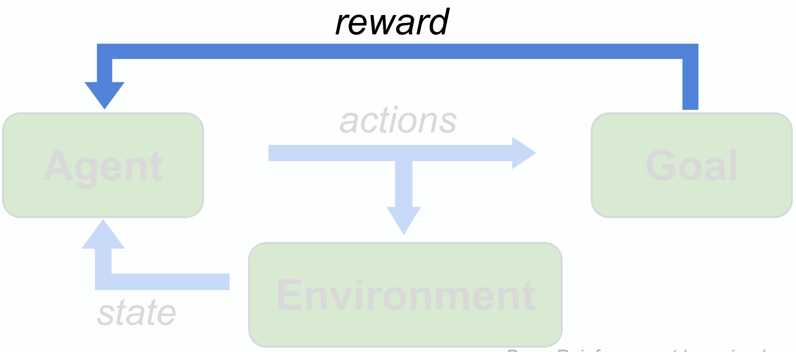
- Goal: to generate a visual description given an image
- Agent: the image captioning model to learn
- Environment: the given image $\mathbf{I}$ + the words predicted so far $\{w_1, ..., w_t\}$
- State: representation of the environment at t, $s_t = \{\mathbf{I}, w_1, ..., w_t\}$
- Action: the word to generate at $t+1$, $a_t = w_{t+1}$



Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

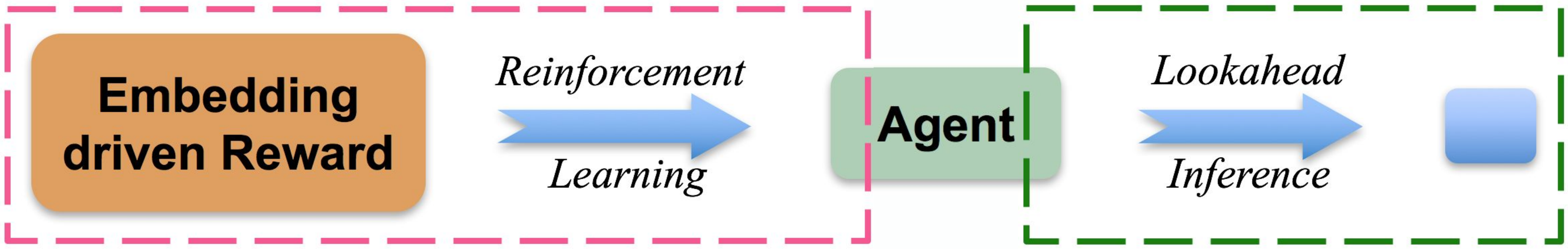
Image captioning reformulation in decision-making

- Goal: to generate a visual description given an image
- Agent: the image captioning model to learn
- Environment: the given image $\mathbf{I}$ + the words predicted so far $\{w_1, ..., w_t\}$
- State: representation of the environment at t, $s_t = \{\mathbf{I}, w_1, ..., w_t\}$
- Action: the word to generate at $t+1$, $a_t = w_{t+1}$
- Reward: the feedback for reinforcement learning



Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

Overview of our approach



- We propose a **decision-making** framework for image captioning
  - ❏ An agent model contains
    - a **policy** network, to capture the **local** information
    - a **value** network, to capture the **global** information
  - ❏ Training using reinforcement learning with **embedding** reward
  - ❏ Testing using **lookahead inference**

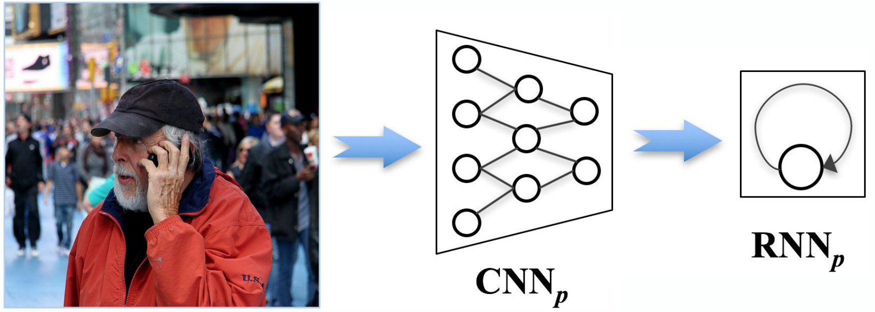
Our approach - agent architecture

- Policy network
- Value network

Our approach - agent architecture

Policy network (local guidance)

Value network



$$p_\pi(a_t|s_t)$$

$$a_t = w_{t+1}$$
$$s_t = \{\mathbf{I}, w_1, ..., w_t\}$$

example:

Current State s_t

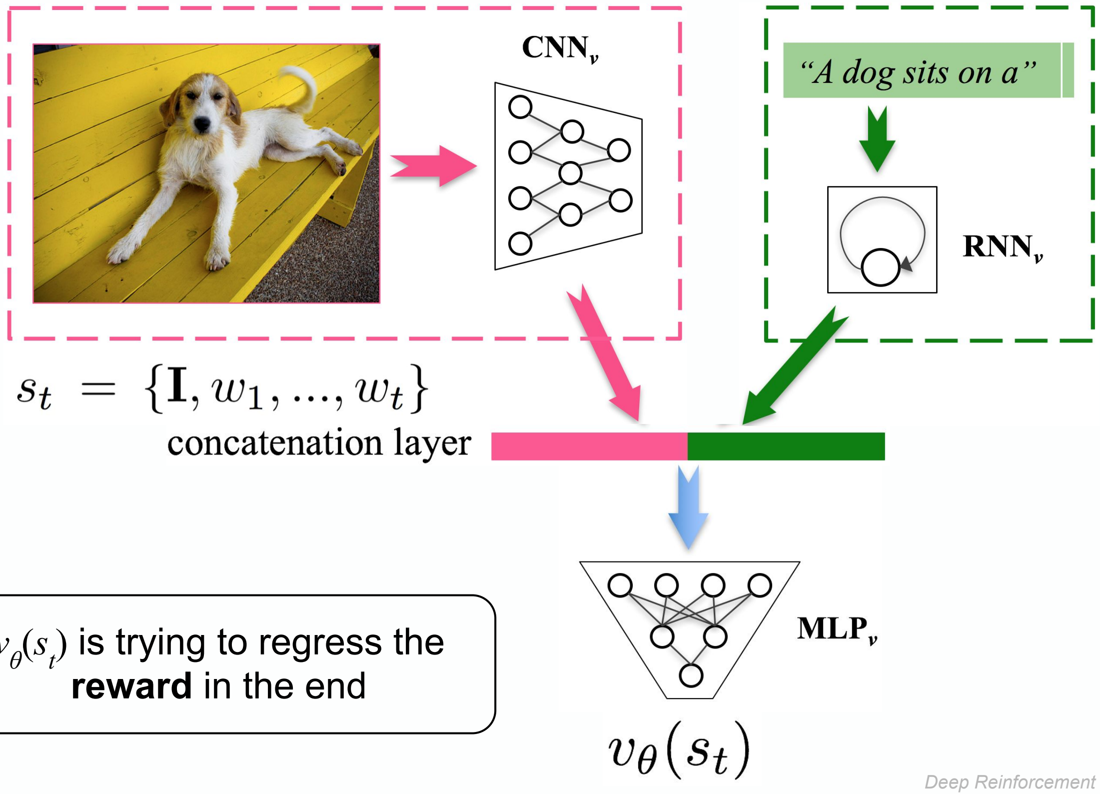
holding	**0.82**
eating	0.56
...	
pretty	0.03

Next Action a_t

Our approach - agent architecture

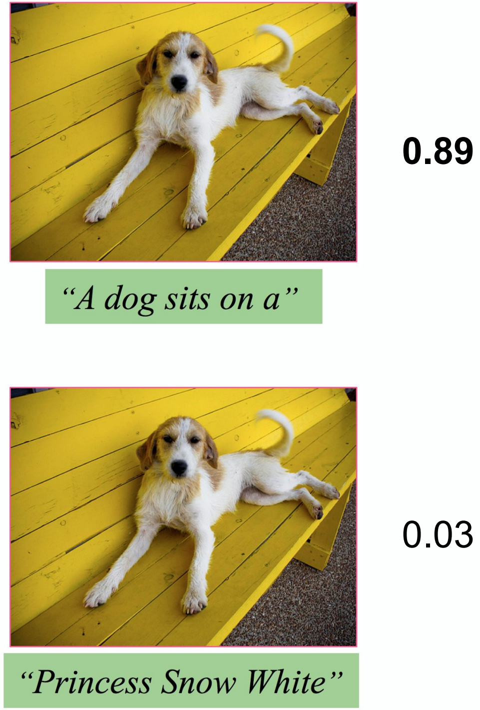
Policy network

Value network (global guidance)



$v_\theta(s_t)$ is trying to regress the **reward** in the end

example:



Our approach - train our agent

- Pretrain policy network p_π with cross entropy loss
- Pretrain value network v_θ with the mean squared loss
- Train p_π and v_θ jointly using deep Reinforcement Learning
  - an Actor-Critic RL model
  - MIXER [Ranzato *et al.* ICLR 2016]

Reinforcement learning - reward definition

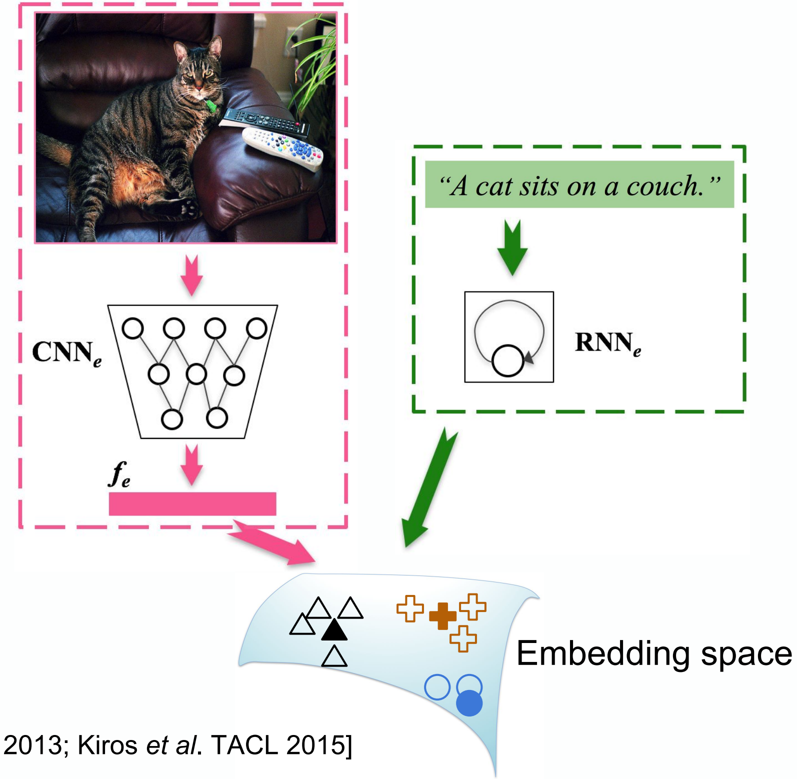
- Literature: metric-driven

  [Ranzato *et al.* ICLR 2016]

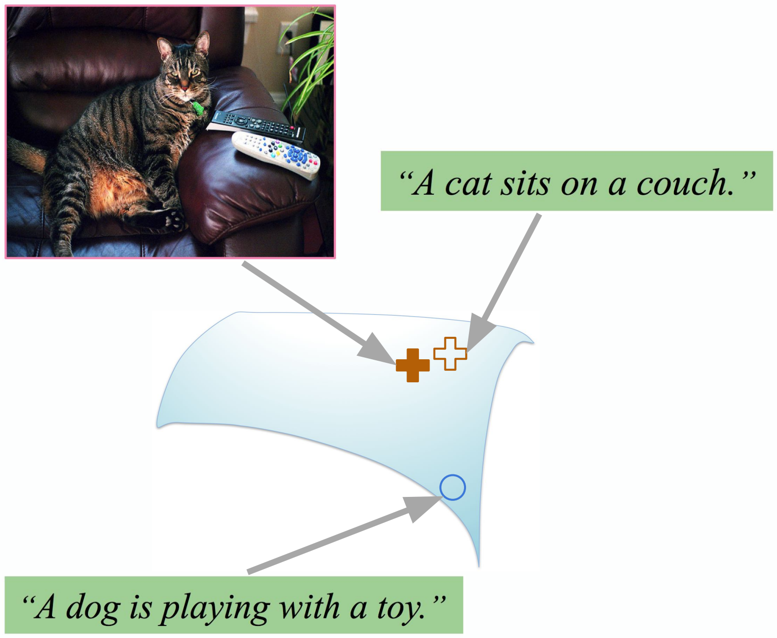
- Limitations:
  - metrics in image captioning are not perfectly defined.
  - it needs to be retrained for each metric in isolation.
  - it doesn't have value network (no global guidance).

Reinforcement learning - reward definition

- Visual-Semantic Embedding



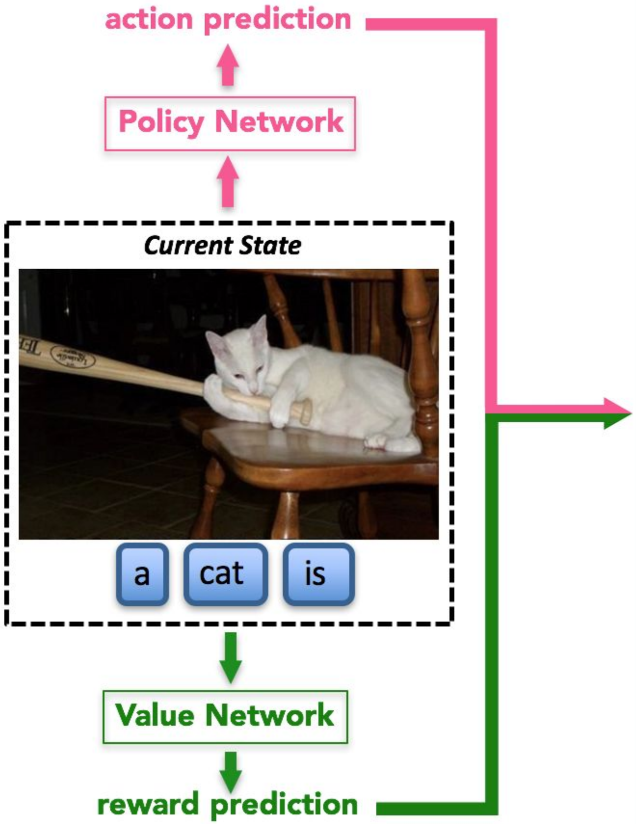
example:



[Frome *et al.* NIPS 2013; Kiros *et al*. TACL 2015]

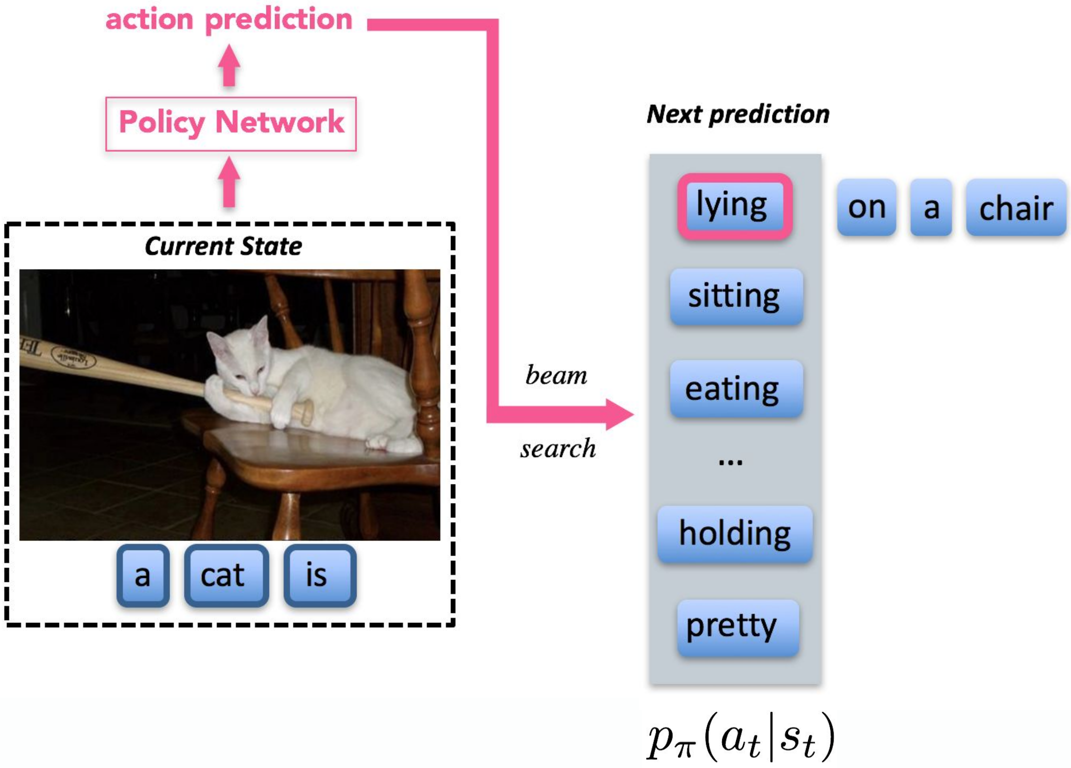
Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

Our approach - inference with our agent

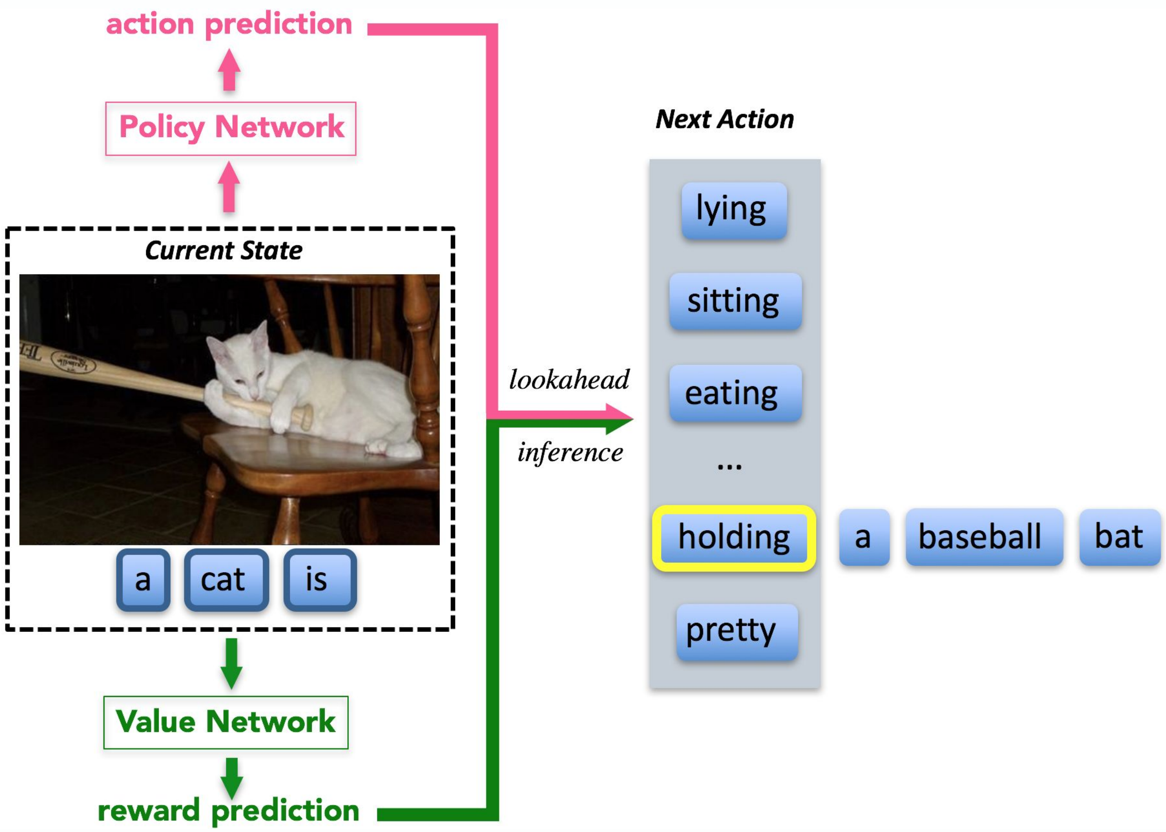


Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

Our approach - inference with our agent

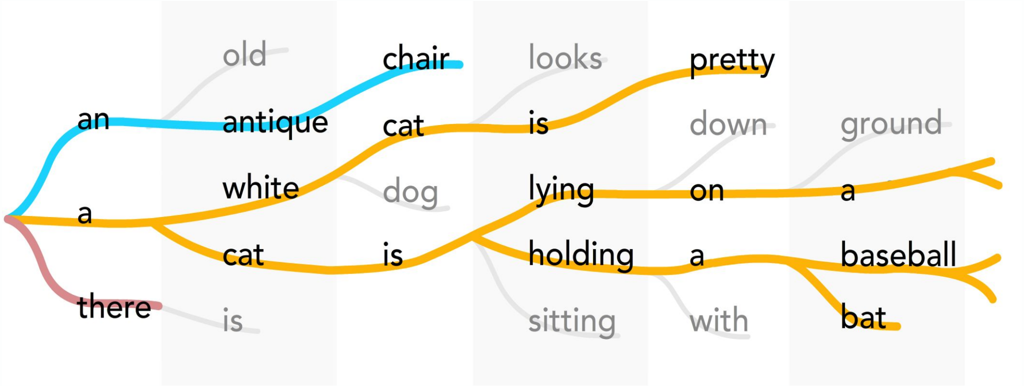


Our approach - inference with our agent



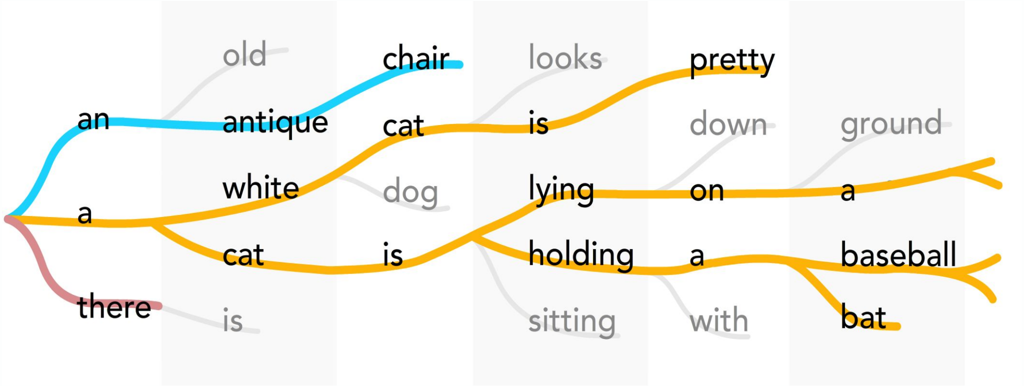
Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

Lookahead inference



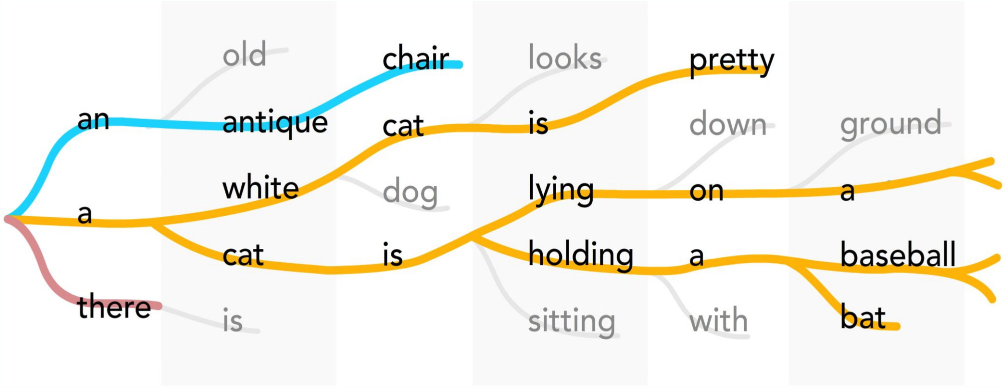
$$W_{\lceil t+1 \rceil} = \underset{\boldsymbol{w}_{b,\lceil t+1 \rceil} \in \mathcal{W}_{t+1}}{\arg topB} \; S(\boldsymbol{w}_{b,\lceil t+1 \rceil})$$

Lookahead inference



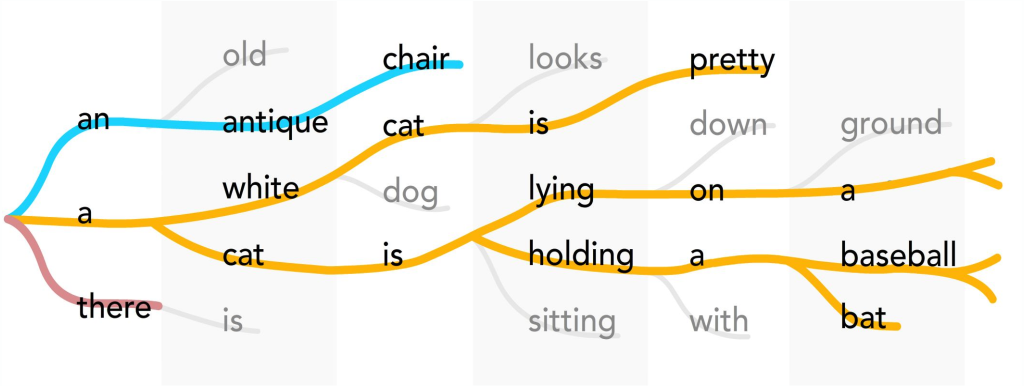
$$W_{\lceil t+1 \rceil} = \underset{\boldsymbol{w}_{b,\lceil t+1 \rceil} \in \mathcal{W}_{t+1}}{\mathrm{arg}topB} \; S(\boldsymbol{w}_{b,\lceil t+1 \rceil})$$

Lookahead inference



$$W_{\lceil t+1 \rceil} = \underset{\boldsymbol{w}_{b,\lceil t+1 \rceil} \in \mathcal{W}_{t+1}}{\arg topB} \; S(\boldsymbol{w}_{b,\lceil t+1 \rceil})$$

Lookahead inference



$$W_{\lceil t+1 \rceil} = \operatorname*{argtop}B_{\boldsymbol{w}_{b,\lceil t+1\rceil} \in \mathcal{W}_{t+1}} \; S(\boldsymbol{w}_{b,\lceil t+1 \rceil})$$

global guidance

$$\begin{aligned} S(\boldsymbol{w}_{b,\lceil t+1 \rceil}) &= S(\{\boldsymbol{w}_{b,\lceil t \rceil}, w_{b,t+1}\}) \\ &= S(\boldsymbol{w}_{b,\lceil t \rceil}) + \lambda \log p_{\pi}(a_t|s_t) + (1-\lambda)\, v_{\theta}(\{s_t, w_{b,t+1}\}) \end{aligned}$$

local guidance

Experimental Results

Results on MS-COCO

Methods	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	Rouge-L	CIDEr
Google NIC [44]	0.666	0.461	0.329	0.246	—	—	—
m-RNN [30]	0.67	0.49	0.35	0.25	—	—	—
BRNN [17]	0.642	0.451	0.304	0.203	—	—	—
LRCN [7]	0.628	0.442	0.304	0.21	—	—	—
MSR/CMU [3]	—	—	—	0.19	0.204	—	—
Spatial ATT [46]	**0.718**	0.504	0.357	0.25	0.23	—	—
gLSTM [15]	0.67	0.491	0.358	0.264	0.227	—	0.813
MIXER [35]	—	—	—	0.29	—	—	—
Semantic ATT [48] *	0.709	0.537	0.402	**0.304**	0.243	—	—
DCC [13] *	0.644	—	—	—	0.21	—	—
Ours	0.713	**0.539**	**0.403**	**0.304**	**0.251**	**0.525**	**0.937**

Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

Results on MS-COCO

Methods	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	Rouge-L	CIDEr
Google NIC [44]	0.666	0.461	0.329	0.246	–	–	–
m-RNN [30]	0.67	0.49	0.35	0.25	–	–	–
BRNN [17]	0.642	0.451	0.304	0.203	–	–	–
LRCN [7]	0.628	0.442	0.304	0.21	–	–	–
MSR/CMU [3]	–	–	–	0.19	0.204	–	–
Spatial ATT [46]	**0.718**	0.504	0.357	0.25	0.23	–	–
gLSTM [15]	0.67	0.491	0.358	0.264	0.227	–	0.813
MIXER [35]	–	–	–	0.29	–	–	–
Semantic ATT [48] *	0.709	0.537	0.402	**0.304**	0.243	–	–
DCC [13] *	0.644	–	–	–	0.21	–	–
Ours	0.713	**0.539**	**0.403**	**0.304**	**0.251**	**0.525**	**0.937**

Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

Results on MS-COCO

	Methods	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	Rouge-L	CIDEr
①	Google NIC [44]	0.666	0.461	0.329	0.246	–	–	–
②	m-RNN [30]	0.67	0.49	0.35	0.25	–	–	–
	BRNN [17]	0.642	0.451	0.304	0.203	–	–	–
	LRCN [7]	0.628	0.442	0.304	0.21	–	–	–
	MSR/CMU [3]	–	–	–	0.19	0.204	–	–
	Spatial ATT [46]	**0.718**	0.504	0.357	0.25	0.23	–	–
	gLSTM [15]	0.67	0.491	0.358	0.264	0.227	–	0.813
③	MIXER [35]	–	–	–	0.29	–	–	–
④	Semantic ATT [48] *	0.709	0.537	0.402	**0.304**	0.243	–	–
	DCC [13] *	0.644	–	–	–	0.21	–	–
	Ours	0.713	**0.539**	**0.403**	**0.304**	**0.251**	**0.525**	**0.937**

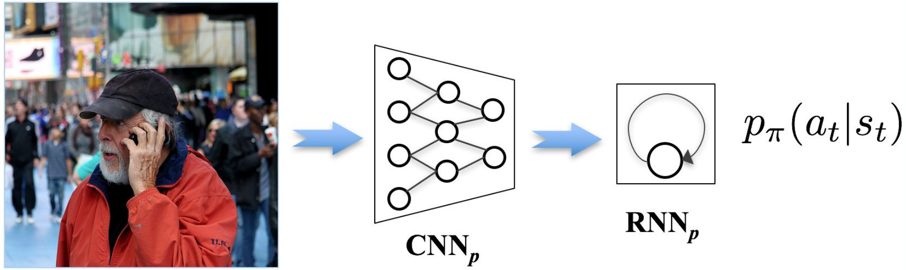
Results on MS-COCO

	Methods
①	Google NIC [44]
②	m-RNN [30]
	BRNN [17]
	LRCN [7]
	MSR/CMU [3]
	Spatial ATT [46]
	gLSTM [15]
③	MIXER [35]
④	Semantic ATT [48] *
	DCC [13] *
	Ours

Results on MS-COCO

	Methods
①	Google NIC [44]
	m-RNN [30]
	BRNN [17]
	LRCN [7]
	MSR/CMU [3]
	Spatial ATT [46]
	gLSTM [15]
	MIXER [35]
	Semantic ATT [48] *
	DCC [13] *
	Ours

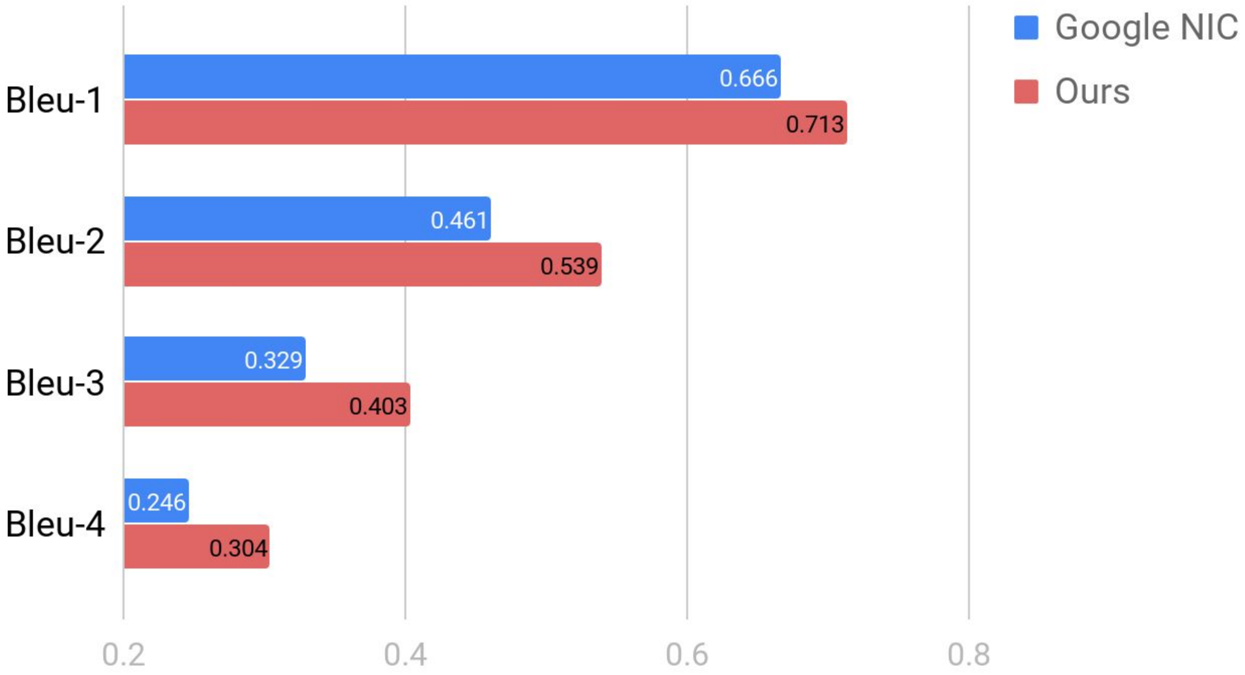
simple net structure



simple policy net

Results on MS-COCO

	Methods
①	Google NIC [44]
	m-RNN [30]
	BRNN [17]
	LRCN [7]
	MSR/CMU [3]
	Spatial ATT [46]
	gLSTM [15]
	MIXER [35]
	Semantic ATT [48] *
	DCC [13] *
	Ours

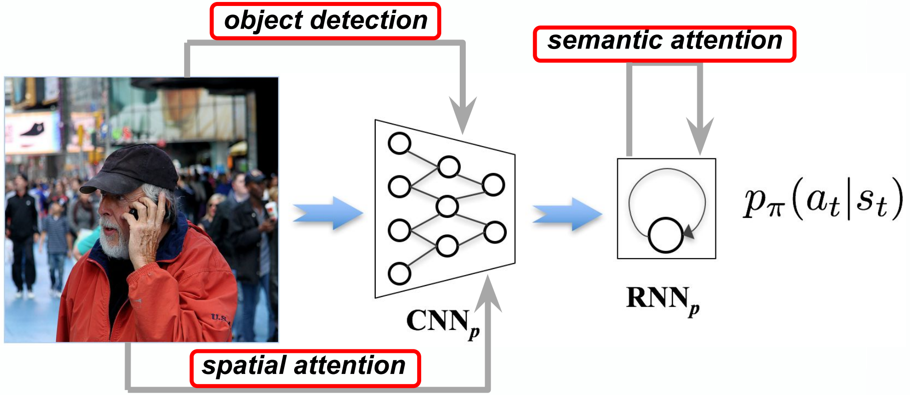


Results on MS-COCO

②

Methods
Google NIC [44]
m-RNN [30]
BRNN [17]
LRCN [7]
MSR/CMU [3]
Spatial ATT [46]
gLSTM [15]
MIXER [35]
Semantic ATT [48] *
DCC [13] *
Ours

advanced net structure

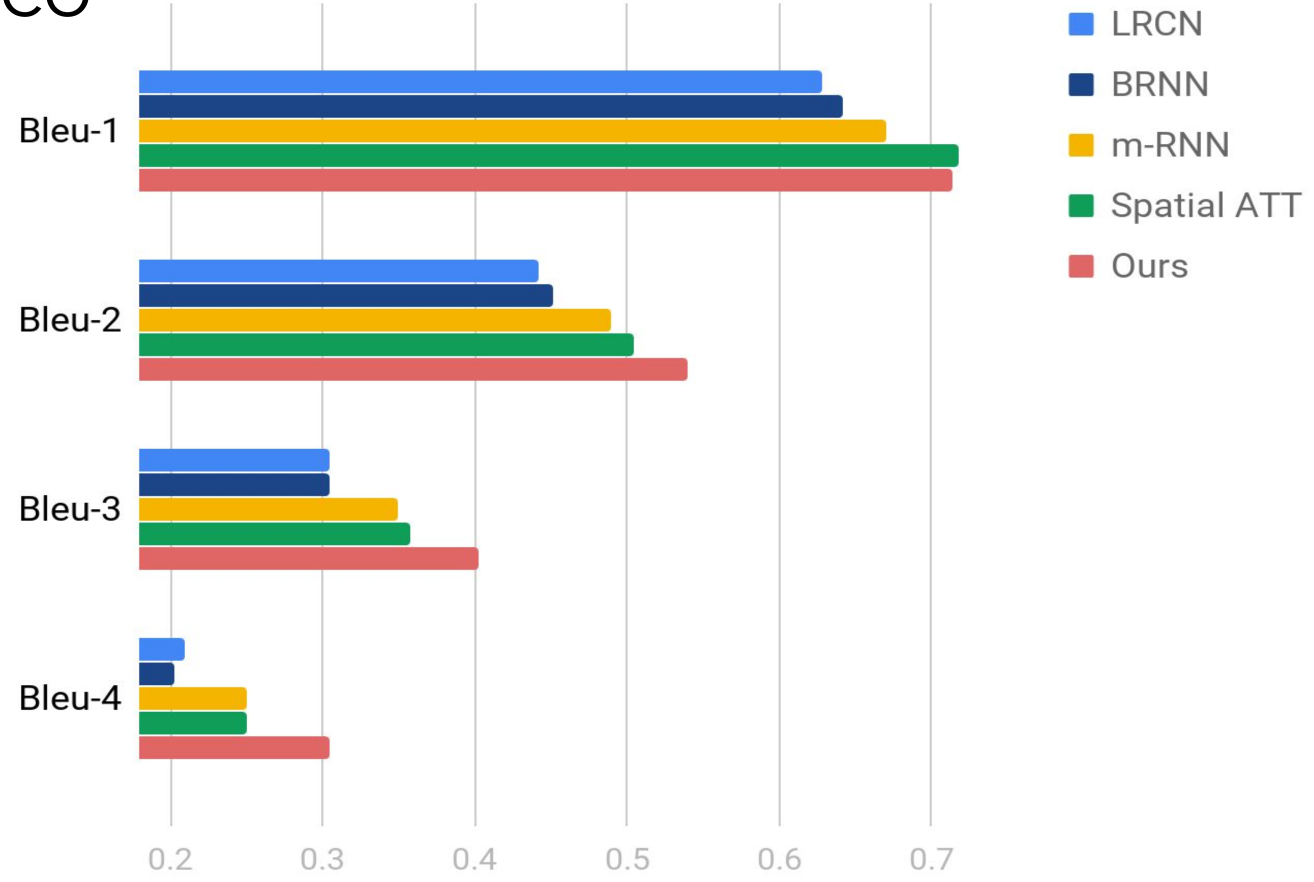


simple policy net

Results on MS-COCO

②

Methods
Google NIC [44]
m-RNN [30]
BRNN [17]
LRCN [7]
MSR/CMU [3]
Spatial ATT [46]
gLSTM [15]
MIXER [35]
Semantic ATT [48] *
DCC [13] *
Ours



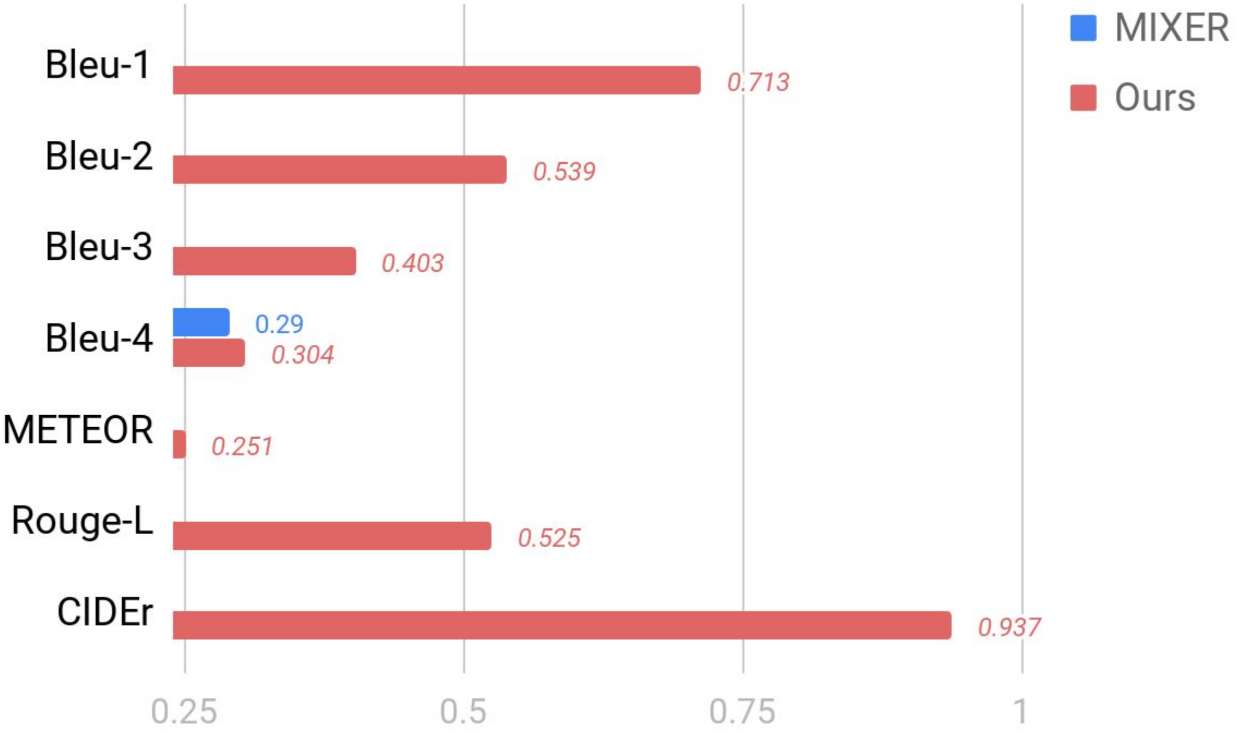
Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

Results on MS-COCO

	Methods	
	Google NIC [44]	
	m-RNN [30]	
	BRNN [17]	
	LRCN [7]	
	MSR/CMU [3]	
	Spatial ATT [46]	
	gLSTM [15]	
③	MIXER [35]	**metric**-driven RL
	Semantic ATT [48] *	
	DCC [13] *	
	Ours	**embedding**-driven RL

Results on MS-COCO

Methods
Google NIC [44]
m-RNN [30]
BRNN [17]
LRCN [7]
MSR/CMU [3]
Spatial ATT [46]
gLSTM [15]
③ MIXER [35]
Semantic ATT [48] *
DCC [13] *
Ours

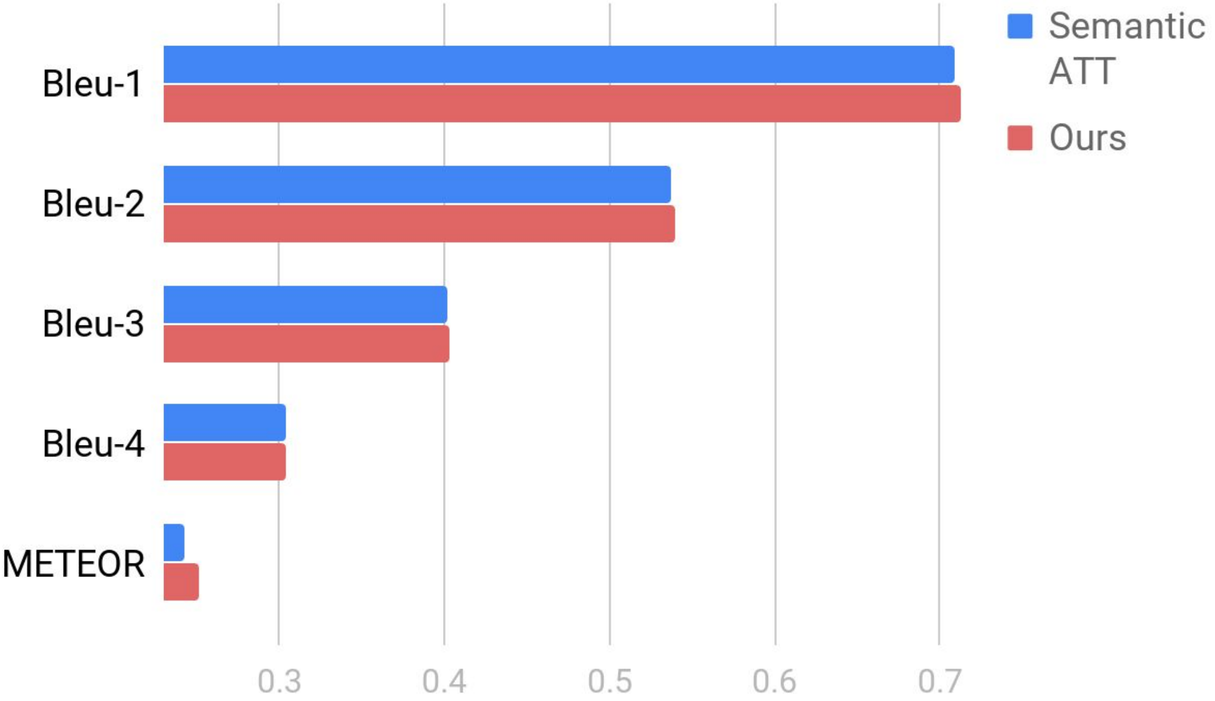


Results on MS-COCO

	Methods	
	Google NIC [44]	
	m-RNN [30]	
	BRNN [17]	
	LRCN [7]	
	MSR/CMU [3]	
	Spatial ATT [46]	
	gLSTM [15]	
	MIXER [35]	
④	Semantic ATT [48] *	**with** external training data
	DCC [13] *	
	Ours	**w/o** external training data

Results on MS-COCO

Methods
Google NIC [44]
m-RNN [30]
BRNN [17]
LRCN [7]
MSR/CMU [3]
Spatial ATT [46]
gLSTM [15]
MIXER [35]
④ Semantic ATT [48] *
DCC [13] *
Ours



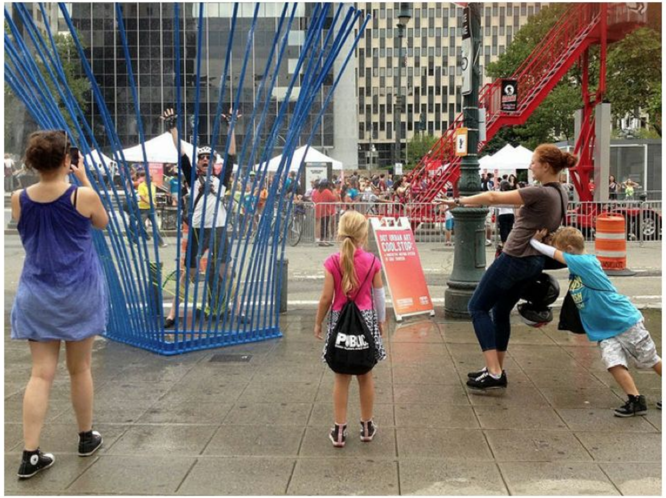
Qualitative results

GT:	the plane is parked at the gate at the airport terminal
SL:	a passenger train that is pulling into a station
Ours:	a white airplane parked at an airport terminal

GT:	a painting of fruit and a candle with a vase
SL:	a table with a vase and flowers on it
Ours:	a painting of a vase sitting on a table

Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

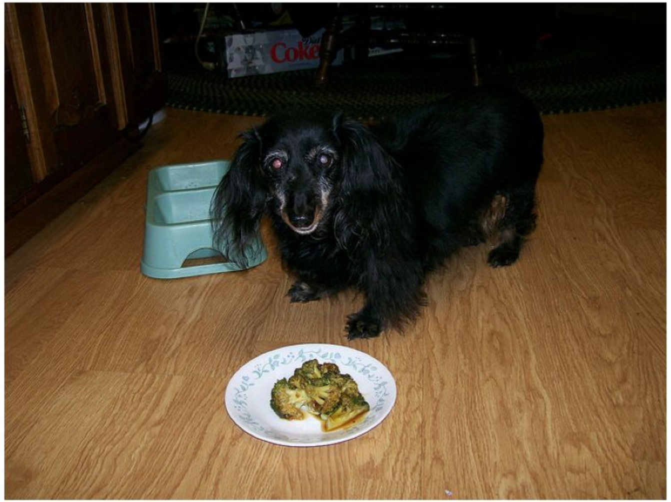
Qualitative results



GT: people are standing outside in a busy city street

SL: a group of young people playing a game of basketball

Ours: a group of people that are standing in the street



GT: a small dog eating a plate of broccoli

SL: a dog that is sitting on a table

Ours: a dog that is eating some food on a table

Deep Reinforcement Learning-based Image Captioning with Embedding Reward. CVPR 2017.

Take Home

- We proposed a novel **decision-making** framework for image captioning.
  - An agent model → a policy network + a value network
  - A training method → Reinforcement Learning with embedding reward
  - An inference method → lookahead inference
- Utilizing both **global** and **local** information is important for sequential generation tasks.
- **Embedding** can capture global information and can serve as a very good global guidance.

Thank you!

Welcome to visit our poster at #9-B

zhou.ren@snap.com