No Metrics Are Perfect: Adversarial REward Learning for Visual Storytelling

Xin (Eric) Wang*, Wenhu Chen*, Yuan-Fang Wang, William Wang

UNIVERSITY OF CALIFORNIA
SANTA BARBARA

Image Captioning

Caption:

Two young kids with backpacks sitting on the porch.

Visual Storytelling

Story:

The brother did not want to talk to his sister. The siblings made up. They started to talk and smile. Their parents showed up. They were happy to see them.

Imagination Emotion Subjectiveness

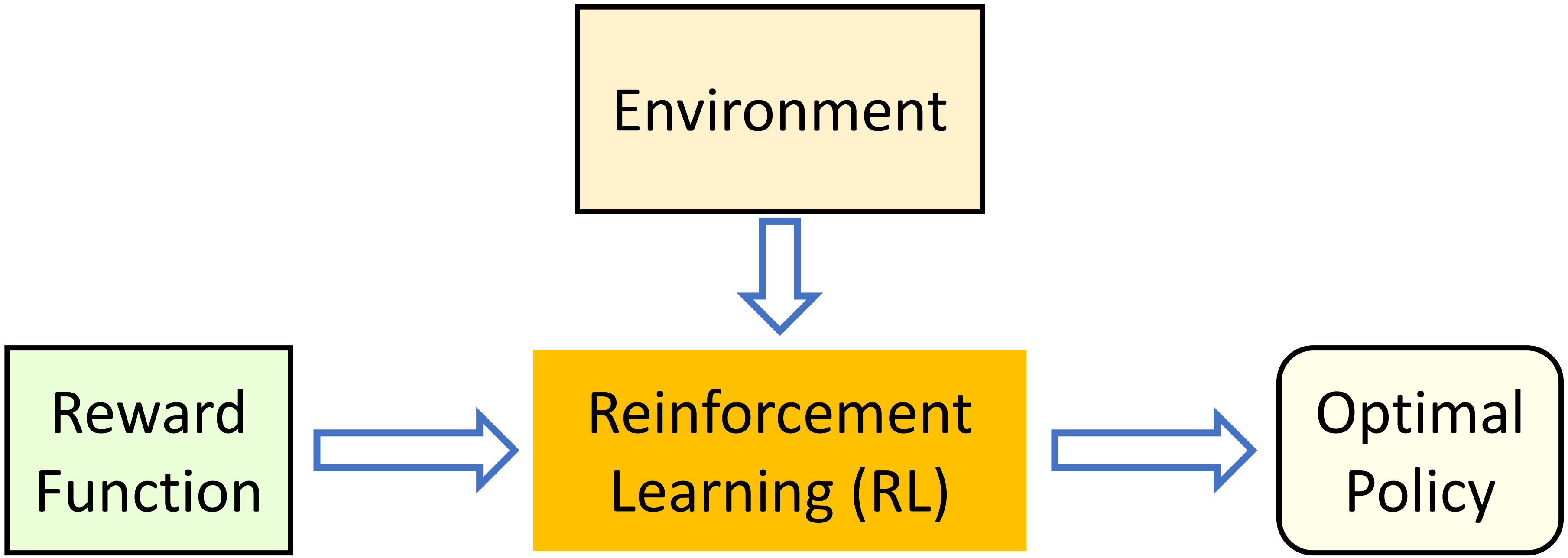
Visual Storytelling

Story #2:
The brother and sister were ready for the first day of school. They were excited to go to their first day and meet new friends. They told their mom how happy they were. They said they were going to make a lot of new friends. Then they got up and got ready to get in the car.

Behavioral cloning methods (*e.g.* MLE) are not good enough for visual storytelling

Reinforcement Learning

- Directly optimize the existing metrics
  - BLEU, METEOR, ROUGE, CIDEr
  - Reduce exposure bias



Rennie 2017, "Self-critical Sequence Training for Image Captioning"

We had a great time to have a lot of the. They were to be a of the. They were to be in the. The and it were to be the. The, and it were to be the.

Average METEOR score: 40.2

(SOTA model: 35.0)

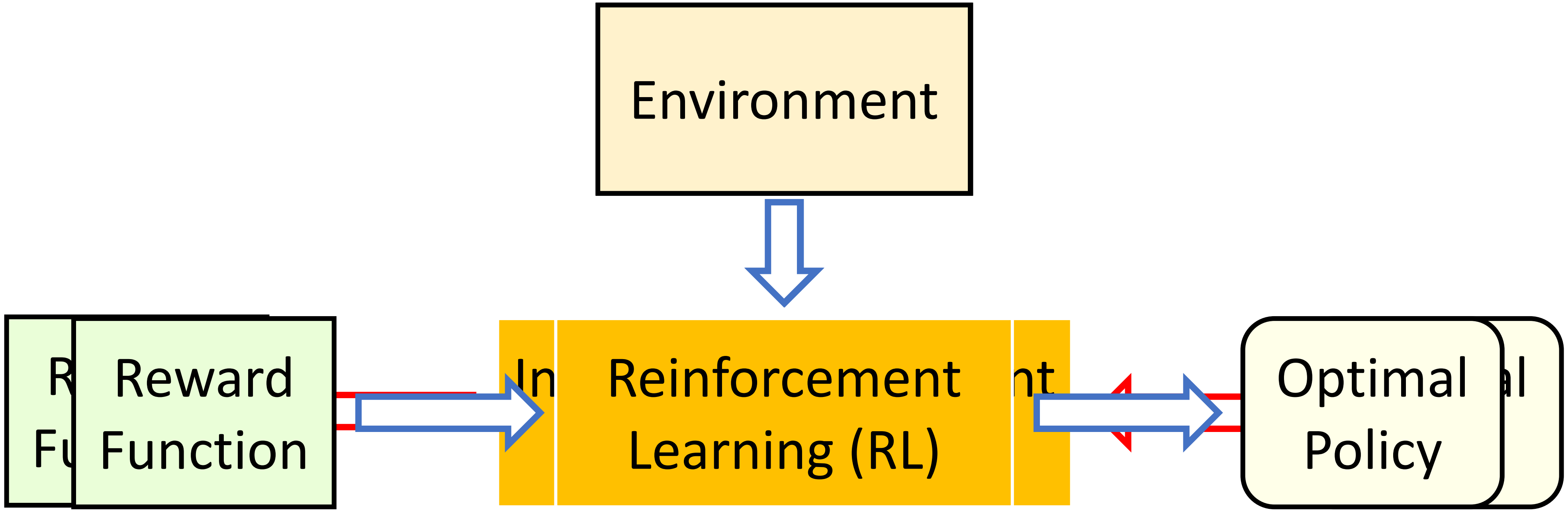
  

I had a great time at the restaurant today. The food was delicious. I had a lot of food. I had a great time.

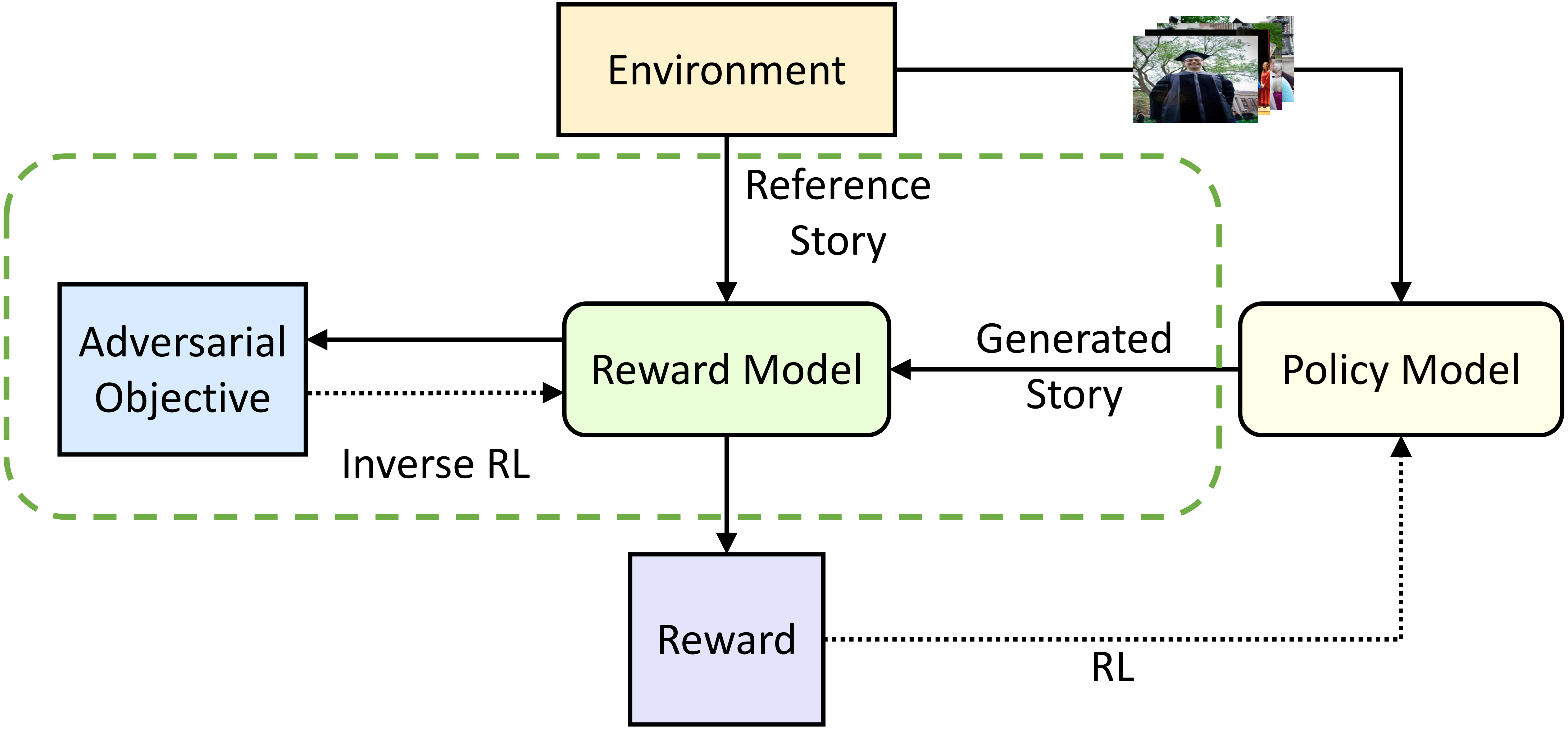
BLEU-4 score: 0

No Metrics Are Perfect!

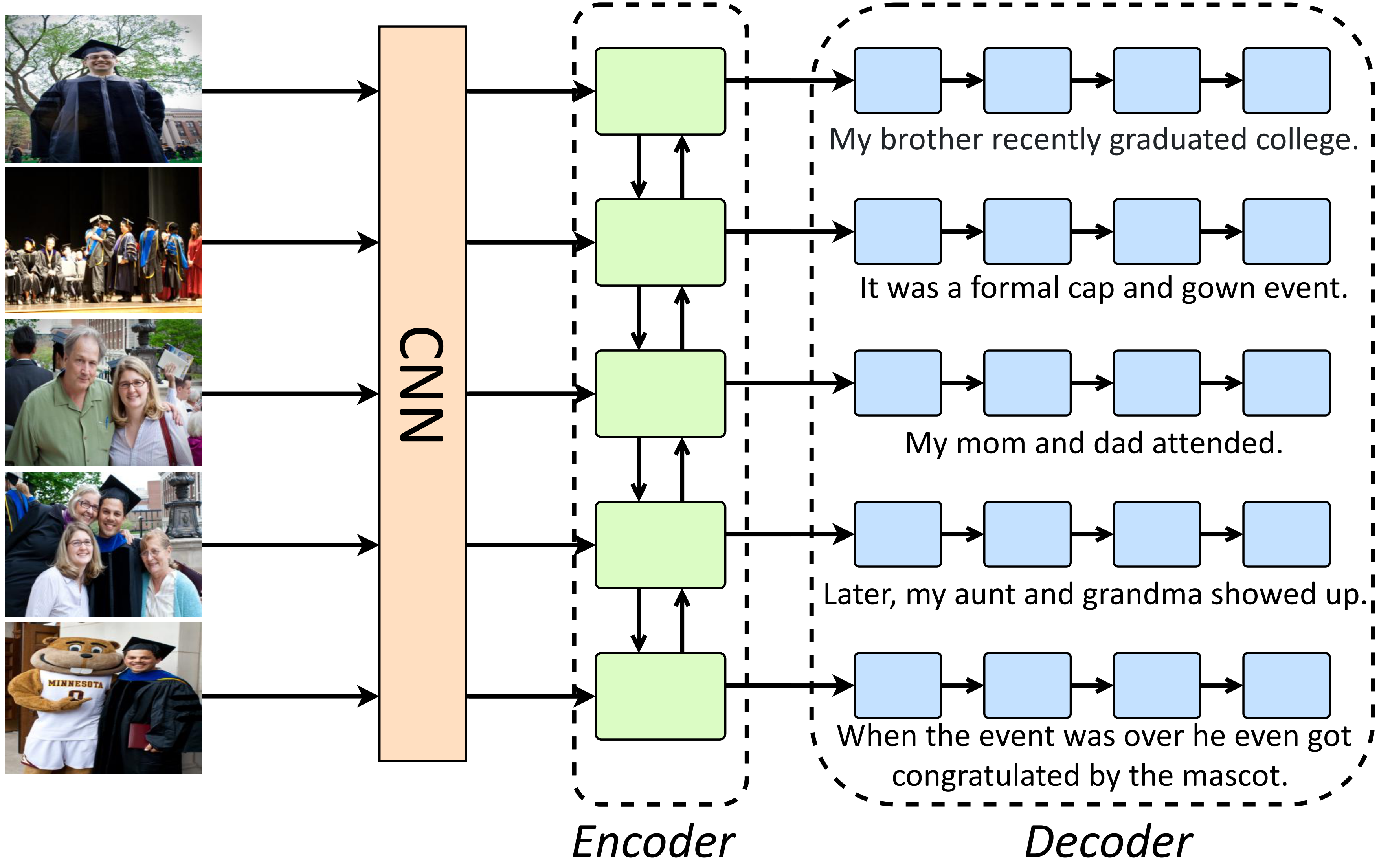
Inverse Reinforcement Learning



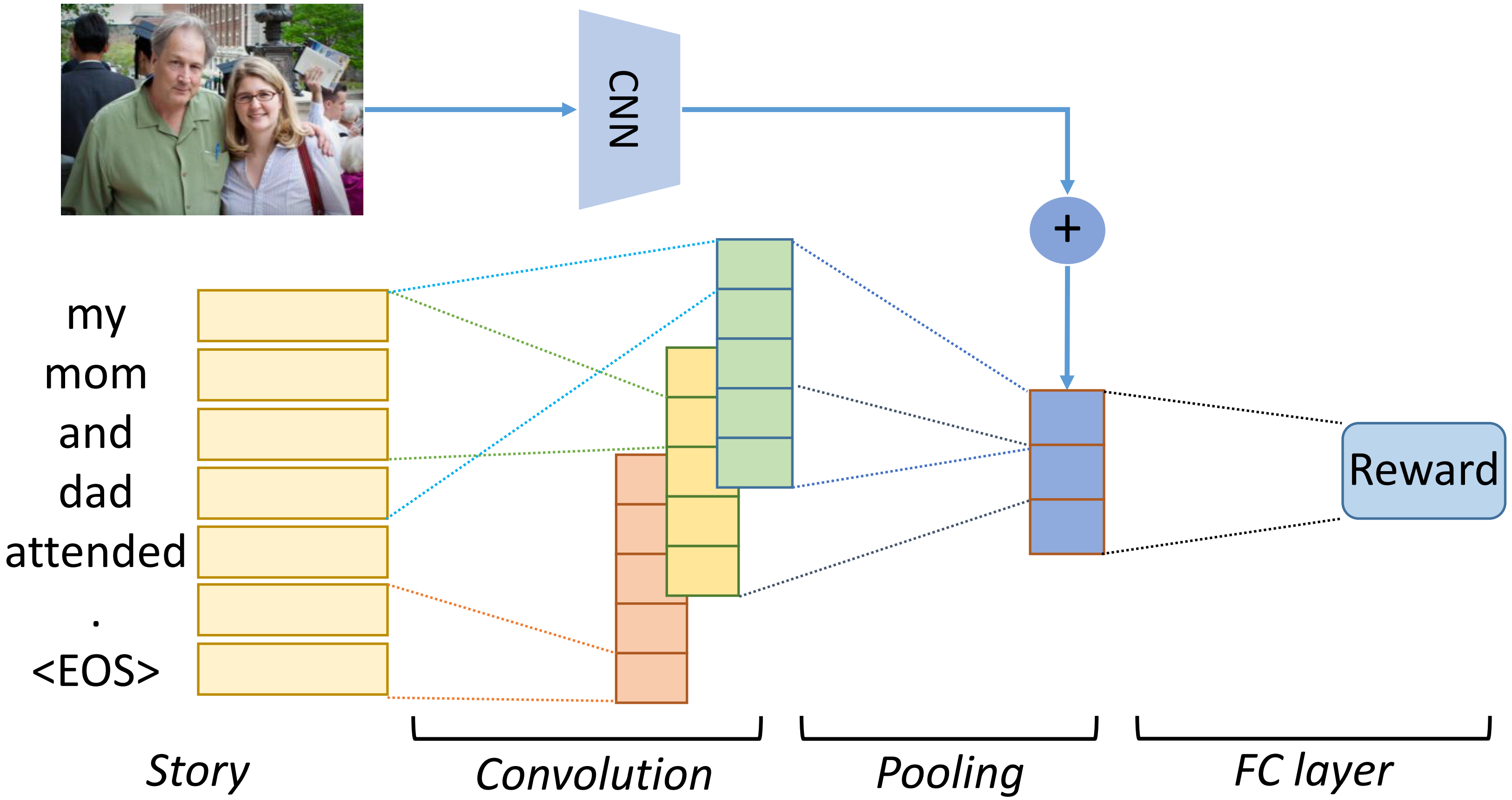
Adversarial REward Learning (AREL)



Policy Model π_β



Reward Model R_θ



Kim 2014, “Convolutional Neural Networks for Sentence Classification”

Associating *Reward* with *Story*

Energy-based models associate an energy value $E_\theta(x)$ with a sample x, modeling the data as a Boltzmann distribution

$$p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z}$$

Reward Boltzmann Distribution

Story

Reward Function

$$p_\theta(W) = \frac{\exp(R_\theta(W))}{Z_\theta} \qquad \boxed{Z_\theta = \sum_W \exp(R_\theta(W))}$$

Approximate data distribution

Partition function

Optimal reward function $R_\theta^*(W)$ is achieved when $p_\theta(W) = p^*(W)$

LeCun et al. 2006, "A tutorial on energy-based learning"

AREL Objective

Therefore, we define an adversarial objective with KL-divergence

Reward Boltzmann distribution

$$\max_{\beta} \min_{\theta} KL(p_e(W) || p_\theta(W)) - KL(\pi_\beta(W) || p_\theta(W))$$

Empirical distribution

Policy distribution

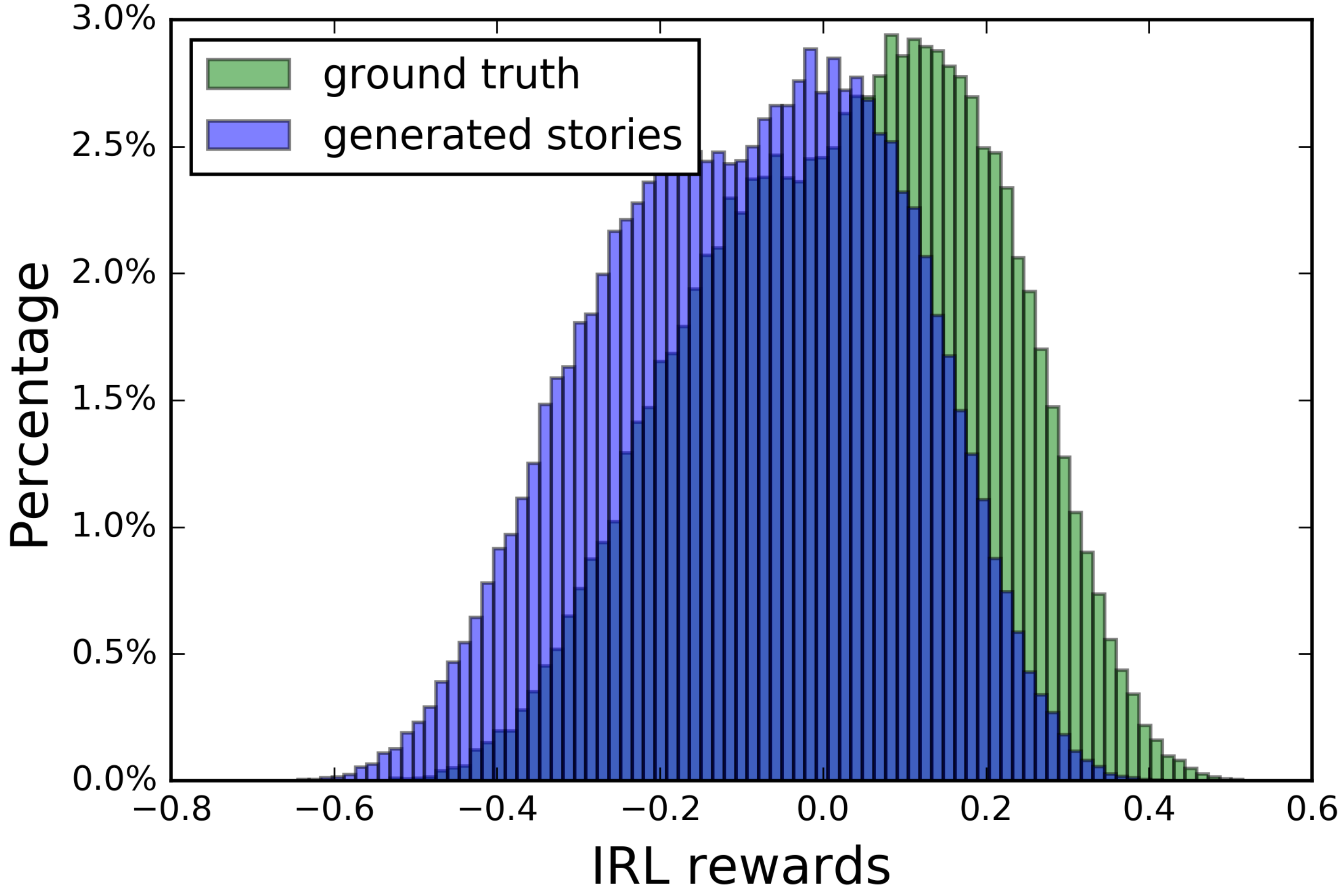
- The objective of Reward Model R_θ:

$$p_e(W) \Rightarrow \qquad \Leftarrow p_\theta(W) \Leftarrow \qquad \Rightarrow \pi_\beta(W)$$

- The objective of Policy Model π_β:

$$\pi_\beta(W) \Rightarrow \qquad \Leftarrow p_\theta(W)$$

Reward Visualization

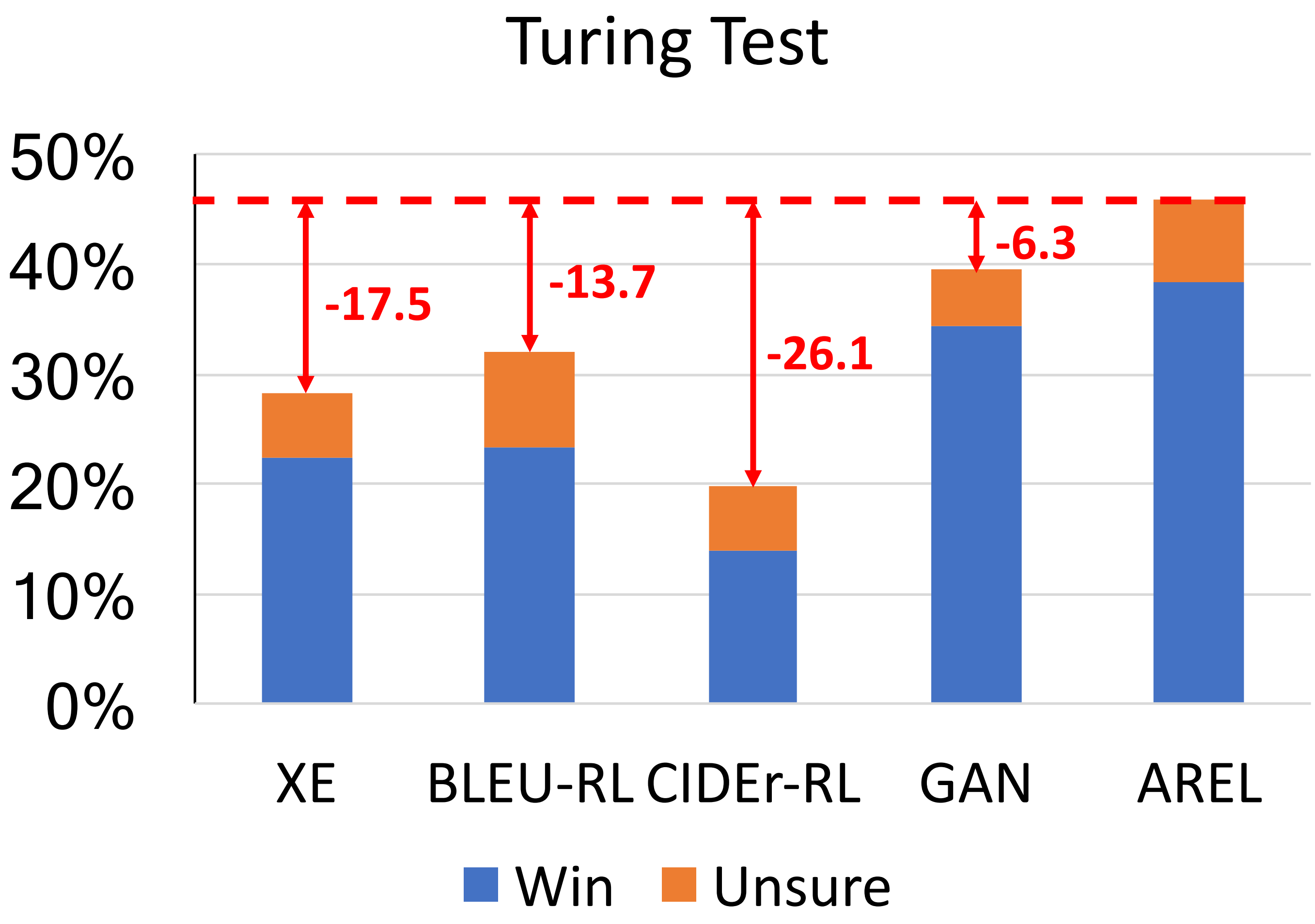


Automatic Evaluation

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
Seq2seq (Huang et al.)	-	-	-	-	31.4	-	-
HierAttRNN (Yu et al.)	-	-	21.0	-	34.1	29.5	7.5
XE	62.3	38.2	22.5	13.7	34.8	**29.7**	8.7
BLEU-RL	62.1	38.0	22.6	13.9	34.6	29.0	8.9
METEOR-RL	68.1	35.0	15.4	6.8	40.2	30.0	1.2
ROUGE-RL	58.1	18.5	1.6	0.0	27.0	33.8	0.0
CIDEr-RL	61.9	37.8	22.5	13.8	34.9	29.7	8.1
GAN	62.8	38.8	23.0	**14.0**	**35.0**	29.5	9.0
AREL (ours)	**63.7**	**39.0**	**23.1**	**14.0**	**35.0**	29.6	**9.5**

Huang et al. 2016, "Visual Storytelling"
Yu et al. 2017, "Hierarchically-Attentive RNN for Album Summarization and Storytelling"

Human Evaluation



Human Evaluation

Pairwise Comparison

	AREL *vs* XE-ss			AREL *vs* BLEU-RL			AREL *vs* CIDEr-RL			AREL *vs* GAN		
Choice (%)	AREL	XE-ss	Tie	AREL	BLEU-RL	Tie	AREL	CIDEr-RL	Tie	AREL	GAN	Tie
Relevance	**61.7**	25.1	13.2	**55.8**	27.9	16.3	**56.1**	28.2	15.7	**52.9**	35.8	11.3
Expressiveness	**66.1**	18.8	15.1	**59.1**	26.4	14.5	**59.1**	26.6	14.3	**48.5**	32.2	19.3
Concreteness	**63.9**	20.3	15.8	**60.1**	26.3	13.6	**59.5**	24.6	15.9	**49.8**	35.8	14.4

Relevance: the story accurately describes what is happening in the photo stream and covers the main objects.
Expressiveness: coherence, grammatically and semantically correct, no repetition, expressive language style.
Concreteness: the story should narrate concretely what is in the images rather than giving very general descriptions.

XE-ss	We took a trip to the mountains.	There were many different kinds of different kinds.	We had a great time.	He was a great time.	It was a beautiful day.
AREL	The family decided to take a trip to the countryside.	There were so many different kinds of things to see.	The family decided to go on a hike.	I had a great time.	At the end of the day, we were able to take a picture of the beautiful scenery.
Human-created Story	We went on a hike yesterday.	There were a lot of strange plants there.	I had a great time.	We drank a lot of water while we were hiking.	The view was spectacular.

Takeaway

- Generating and evaluating stories are both challenging due to the complicated nature of stories
- No existing metrics are perfect for either training or testing
- AREL is a better learning framework for visual storytelling
  - Can be applied to other generation tasks
- Our approach is model-agnostic
  - Advanced models → better performance

Thanks!

Paper: https://arxiv.org/abs/1804.09160
Code: https://github.com/littlekobe/AREL