No Metrics Are Perfect: Adversarial REward Learning for Visual Storytelling

Xin Wang*, Wenhu Chen*, Yuan-Fang Wang, William Wang
Image Captioning

Caption:
Two young kids with backpacks sitting on the porch.
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.
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

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)
My brother recently graduated college.

It was a formal cap and gown event.

My mom and dad attended.

Later, my aunt and grandma showed up.

When the event was over he even got congratulated by the mascot.
Reward Model $R_{\theta}$

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**

Approximate data distribution

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:

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

• The objective of Reward Model \( R_\theta \):
  \[
p_e(W) \quad \leftrightarrow \quad p_\theta(W) \quad \leftrightarrow \quad \pi_\beta(W)
\]

• The objective of Policy Model \( \pi_\beta \):
  \[
  \pi_\beta(W) \quad \leftrightarrow \quad p_\theta(W)
  \]
Reward Visualization

![Reward Visualization Diagram](image-url)
# Automatic Evaluation

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

Huang et al. 2016, “Visual Storytelling”
Yu et al. 2017, “Hierarchically-Attentive RNN for Album Summarization and Storytelling”
Human Evaluation

Turing Test

<table>
<thead>
<tr>
<th></th>
<th>XE</th>
<th>BLEU-RL</th>
<th>CIDEr-RL</th>
<th>GAN</th>
<th>AREL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
<td>-17.5</td>
<td>-13.7</td>
<td>-26.1</td>
<td>-6.3</td>
<td></td>
</tr>
</tbody>
</table>

- Win
- Unsure
# Human Evaluation

## Pairwise Comparison

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

**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.
WE-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!

Code: https://github.com/littlekobe/AREL