Teaching Machines To Describe Images With Natural Language Feedback

Huan Ling, Sanja Fidler
{linghuan, fidler}@cs.toronto.edu

Introduction

Overview:
we bring a human in the loop, and enable a human teacher to give feedback to a learning agent in the form of natural language.

Our Contributions:
– A phrase-based captioning model trained with policy gradients.
– A feedback network that provides reward to the learner by conditioning on the human-provided feedback.
– We show that by exploiting descriptive feedback our model learns to perform better than when given independently written human captions.

Human-Provided Language Feedback Collection

Human Feedback: We create a web interface for collecting feedback information on a larger scale via AMT. We collect feedback in the form of natural language as well as information about the type of mistake, mistaken words, and the corrected captions.

Methods

• Phrase-based Image Captioning: Our captioning model, forming the base of our approach, uses a hierarchical Recurrent Neural Network.

• Feedback Network: We design a neural network which will evaluate errors in phrases based on the feedback sentence.

\[
\begin{align*}
\hat{b}_t^{\text{natural}} &= f_{\text{sent}}(b_t^{\text{natural}}, \hat{w}_t) \\
\hat{b}_t^{\text{feedback}} &= f_{\text{sent}}(b_t^{\text{feedback}}, \hat{w}_t) \\
q_i &= f_{\text{fbn}}(w_i, \ldots, \hat{w}_t, q_m) \\
\alpha_i &= f_{\text{fbn}}(b_t, \hat{b}_t, q_i, m)
\end{align*}
\]

Where \( w_t \) and \( \hat{w}_t \) denote the one-hot encoding of words in the sampled caption and feedback sentence, respectively. By \( \hat{w}_t \) we denote words in the \( t \)-th phrase of the sampled caption.

• Policy Gradient Optimization using Natural Language Feedback: We directly optimize for the desired image captioning metrics plus human feedbacks using the Policy Gradient technique. For caption \( w^c = w_1^c \ldots w_n^c \), where \( w_t^c \) denotes the \( t \)-th phrase of the sentence.

\[
r(w^c) = \beta \sum_{i} \lambda_i \cdot \text{BLEU}(w_t^c, \text{ref})
\]

\[
r(w_t^c) = r(w_t) + \lambda_i f_{\text{fbn}}(w_t^c, \hat{w}_t^c)
\]

\[
w_t^c = \text{greedy sampled}_\text{baseline}
\]

\[
\nabla_q L(\theta) = \sum_{t=1} r(w_t^c) - r(w_t) \nabla_q \log p(w_t)
\]

Results

• Caption quality evaluation by the human annotators: Caption quality evaluation by the human annotators. Plot on the left shows evaluation for captions generated with our reference model (MLE). The right plot shows evaluation of the human-corrected captions (after completing at least one round of feedback).

• Quantitative results for feedback network:

<table>
<thead>
<tr>
<th>Feedback network</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE (5 GT)</td>
<td>104.12</td>
</tr>
<tr>
<td>RLB (5 GT)</td>
<td>106.55</td>
</tr>
<tr>
<td>RLF (4GT + FB)</td>
<td>107.67</td>
</tr>
<tr>
<td>RLF (A + FB)</td>
<td>106.12</td>
</tr>
</tbody>
</table>

| Human preferences and clicks: We conducted a human evaluation using AMT and we count how much human interaction is required for either the baseline RL and our approach. In particular, we count every interaction with the keyboard as 1 click.

Examples:

• MLE: (a man) (walking) (in front of a building) (with a cell phone) RLB: (a man) (walking) (in a sidewalk) (with a cell phone) RLF: (a man) (walking) (in a sidewalk)

• MLE: (two girls) (on standing) (in front of a building) RLB: (two girls) (on standing) (in front of a building) RLF: (two girls) (on standing) (in a game field) (in a row)

• MLE: (a clock tower) (with a clock) (on the clock) RLB: (a clock tower) (with a clock) (on top of the clock) RLF: (a clock tower) (with a clock) (on the clock)