

Appendix

A Error Analysis

Failure Case in Turing Test In Figure 7, we presented a negative example that failed the Turing test (4 out of 5 made the correct decision). Compared with the human-generated story, our AREL story lacked emotion and imagination and thus can be easily distinguished. For example, the real human gave the band a nickname “very loud band” and told a more amusing story. Though we have made encouraging progress on generating human-like stories, further research of creating diversified stories is still needed.

Data Bias From the experiments, we observe that there exist some severe data bias issues in the VIST dataset, such as gender bias and event bias. In the training set, the ratio of male and female’s appearances is 2.06:1, and it is 2.16:1 in the test set. the models aggravate the gender bias to 3.44:1. Besides, because all the images are collected from Flickr, there is also an event bias issue. We count three most frequent events: party, wedding, and graduation, whose ratios are 6.51:2.36:1 on the training set and 4.54:2.42:1 on the test set. However, their ratio on the testing results is 10.69:2.22:1. Clearly, the models tend to magnify the influence of the largest majority. These bias issues remain to be studied for future work.

B Training Details

Our model is implemented on PyTorch and consists of two parts – a policy model and a reward model. The policy model is implemented with a multiple-RNN architecture. Each RNN model is responsible for generating a sub-story for each photo in the stream. But the weights are tied to minimize the memory consumption. The image features are extracted from the pre-trained ResNet-152 model⁸. The visual encoder receives the ResNet-152 features and uses recurrent neural network to understand the temporal dynamics and represents them as hidden state vectors, which is further fed into the decoder to generate stories. The reward model is based on convolutional neural network and uses convolution kernels to extract semantic features for prediction. Here we give the detailed description of our system:

⁸<https://github.com/KaimingHe/deep-residual-networks>

- **Visual Encoder:** the visual encoder is a bi-directional GRU model with hidden dimension of 256 for each direction. we concatenate the bi-directional states and form a 512 dimension vector for the story generator. The input album is composed of five images, and each image is used as separate input to different RNN decoders.
- **Decoder:** The decoder is a single-layer GRU model with hidden dimension of 512. The recurrent decoder model receives the output from visual encoder as the first input, and then at the following time steps, it receives the last predicted token as input or uses the ground truth as input. During scheduled sampling, we use a sampling probability to decide which action to take.
- **Reward Model:** we use a convolutional neural network to extract n-gram features from the story embedding and stretch them into a flattened vector. The embedding size of input story is 128, and the filter dimension of CNN is also 128. Here we use three kernels with window size 2, 3, 4, each with a stride size of 1. We use a pooling size of 2 to shrink the extracted outputs and flatten them as a vector. Finally, we project this vector into a single cell indicating the predicted reward value.

During training, we first pre-train a schedule-sampling model with a batch size of 64 with NVIDIA Titan X GPU. The warm-up process takes roughly 5-10 hours, and then we select the best model to initialize our AREL policy model. Finally, we use alternating training strategy to optimize both the policy model and the reward model with a learning rate of $2e-4$ using Adam optimization algorithm. During test time, we use a beam size of 3 to approximate the whole search space, we force the beam search to proceed more than 5 steps and no more than 110 steps. Once we reach the EOS token, the algorithm stops and we compare the results with human-annotated corpus using 4 different automatic evaluation metrics.

C Amazon Mechanical Turk

We used AMT to perform two surveys, one picks a more human-like story. We asked the worker to answers 8 questions within 30 minutes, and we pay 5 workers to work on the same sheet to

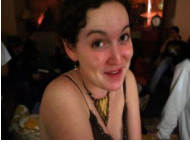

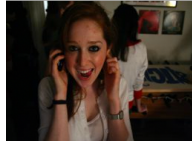


					
XE-ss	I went to the party last week.	The band played a lot of music.	[female] and [female] were having a great time.	[male] and [male] are having a great time at the party.	We had a great time at the party.
AREL	My friends and I went to a party.	The band played a lot of music.	[female] and [male] were having a good time .	[male] and [male] are the best friends in the world.	After a few drinks, everyone was having a great time.
Human-created Story	My first party in the dorm!	There was a very loud band called "very loud band".	my friend [female] had enough. She took my hand and led me to the kitchen where we couldn't hear.	[male] and [male] cornered me and asked me out on a date with them both .	Party! We all danced until passed out .

Figure 7: Failure case in Turing test. 4 out of 5 workers correctly recognized the human-created story and 1 person mistakenly chose AREL story.

eliminate human-to-human bias. Here we demonstrate the Turing survey form in [Figure 8](#). Besides, we also perform a head-to-head comparison with other algorithms, we demonstrate the survey form in [Figure 9](#).

Survey Instructions (Click to expand)

Read the following image streams and compare two stories in the aspect of matching, coherence, and concreteness.

Given a photo stream, select a story which is more likely to be generated by human

Q1 Read the following image stream to answer the questions



A. the park was so crowded in the morning . the venue was filled with antsy people . the graduates word glossy black gowns . this faculty member gave a excited speech . we gathered together to share roses and balloons .

B. today was the day of the graduation ceremony . there were a lot of people there . everyone was very excited . the dean gave a speech to the graduates . everyone was very happy to be there .

Which story is generated by human?

A

B

Unsure

Figure 8: Turing Survey Form

Survey Instructions (Click to expand)

Read the following image streams and compare two stories in the aspect of matching, coherence, and concreteness.



Relevance: the story **accurately describes what is happening** in the image stream and covers the main objects appearing in the images.

Expressiveness: coherence, grammatically and semantically correct, **no repetition, expressive language style**

Concreteness: the story should **narrate concretely what is in the image** rather than giving very general descriptions.

Good example: the students gathered to listen to the presenters give lectures . there was several presenters on hand to speak . they spoke to the crowd with new ideas . the students listened with interest . some of the students took notes as the presenters spoke .

Bad example (repetition): today was the day . i was very happy to see them . she was very happy to be there . they were all very happy to see him . this is a picture of a group .

Bad example (too abstract): this is a picture of a speaker . the speaker was very good . everyone is happy to be there . everyone was very happy . everyone was very happy .

Q1 Read the following image stream to answer the questions



A. the graduation ceremony was held in the auditorium . there were a lot of people there . i was so proud of me . the dean of the school gave a speech to the graduates . everyone was so happy to be married .

B. today was the day of the graduation ceremony . there were a lot of people there . everyone was very excited . the dean gave a speech to the graduates . everyone was very happy to be there .

Which story better describe the images? A B Tie

Which story is more coherent? A B Tie

Which story is more concrete? A B Tie

Figure 9: Pairwise Comparison Form