A Statistics of Bad Endings

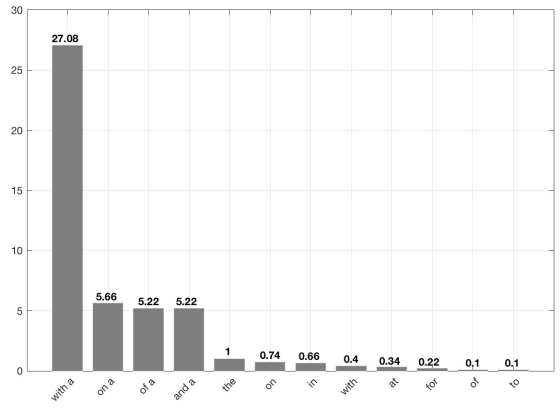

Figure 3: Statistics of bad endings over 5,000 validation instances generated by self-critical method optimized on the CIDEr metric.

B More training details

We drop any words that have appeared less than five times. The vocabulary size is 9,488. We do not rescale or crop the images when extracting CNN features for the attention model. At the beginning of RL training, the learning rate is 5×10^{-5} and we anneal it by a factor of 0.2 when the CIDEr score on validation set has no improvement for over 10 epochs. The CNN weights are fixed during our RL training process. We use a Gumbel sampler and perform our action sampling on GPU devices which is much faster than CPU device. The batch size is set to 50 in all our experiments. In our observation, we discover that a very powerful warm-start model is necessary to maintain the stability and convergence speed of self-critical without our prioritized sampling and n-gram constraints while our methods need not.

We warm-start all models by training them under the cross-entropy objective. We use ADAM (Kingma and Ba, 2015) optimizer with an initial learning rate of 5×10^{-4}. We select the model with best CIDEr scores on the development set to initialize RL training and use a Gumbel sampler (Kusner and Hernndezlobato, 2016) to improve action sampling efficiency. In order to promote captions with higher reward, we sample multiple sequences (we set 10 in our experiments) during the training and update the parameters based on the sample with the highest rewards. We find that this technique empirically helps convergence.

C Qualitative Comparison

Figure 4: Examples generated by our model with attention compared to the self-critical counterpart. After adding N-gram constraints, our results are more accurate and human-readable.

D Human Evaluation Details

We recruit 10 volunteers who are under the correct guidance for finishing the human evaluation process.We implement our human evaluation experiment on a web page (see Figure 5).

Figure 5: Web page for human evaluation

Every time the volunteer must make a choice among these three choices. And "almost same"

means that two captions are both good or both bad for the given images. Each image must be evaluated by at least 5 volunteers.

The order of caption generated from given model is random so the volunteers have no idea where these captions come from which model. Some times the first one comes from Ours-Att-4-gram, and some times from Att-SC.