Abstract

This supplement is organized as follows:

- Sec. 1 performs the following ablation study – how well would different models perform if they were provided human dialog as input? Specifically, how good are the different models on the task of human-dialog-based image retrieval?

- Sec. 2 describes details about hyperparameter selection and training.

- Finally, Sec. 3 presents qualitative results on image guessing task and generated language.

It contains 2 figures and 2 tables:

- Fig. 1 presents image-retrieval evaluation.

- Fig. 2 shows image guessing results for our model.

- Tab. 1 highlights example Q-BOT-A-BOT interactions for various models.

- Tab. 2 presents qualitative comparison of answers generated by various models.

1. Dialog-based Image Retrieval

In Section 6 of the main paper, we evaluated how well models learn to cooperate at image guessing by setting up a retrieval task on the test split of VisDial v0.5. Specifically, we allow models to interact for 10 rounds where after each round of dialog exchange, Q-BOT predicts a feature representation $y_t$, and we sort the entire test set in ascending order of distance to this prediction and compute the rank of the true image. Fig. 1a (reproduced from the main paper) shows mean percentile rank of the true image across rounds.

In this section, we conduct ablations to study how well different models perform when provided human dialog as input. Specifically, we study the performance of different models on the task of human-dialog-based image retrieval.

The task of human-dialog-based image retrieval allows us to probe two aspects of our models. First, we study how well RL trained models retain an understanding of human dialog from their SL-pretraining. Second, we can estimate a ‘human upper bound’ on the image guessing task. Specifically, when provided with a human-human dialog, A-BOT has ‘perfect’ perception and Q-BOT always asks consistent and non-repeating questions.

In addition to testing both models from the main paper (SL-pretrained and RL-full-QAf), we also train a supervised model, SL-pretrained-fc7only, specifically for this task of human-dialog-based image retrieval for comparison to those models. Specifically, a model trained to regress to image features given a human dialog. In practice, this amounts to finetuning $f$ in SL-pretrained Q-BOT for image regression only, instead of a question-generation + fc7 regression multitask objective.

Fig. 1b shows results of this experiment as mean percentile rank of true image across rounds of human dialog. We note two observations. First, all models improve at the task of image retrieval when provided human-dialog, suggesting that language generation is a difficult task, and if we are provided ‘good’ dialog – self-consistent, accurate, and free from repetition – all models can improve at image guessing.

Second, we observe SL-pretrained-fc7only performs significantly better than the other models. This is unsurprising as this model is explicitly trained to regress from human dialog and lacks the additional utterance generation loss found in the other models. The strong performance of this model is encouraging because it implies that if we can generate better dialog, we can expect to do better at image retrieval.

*The first two authors (AD, SK) contributed equally.
2. Implementation Details

In this section, we provide additional details about the architecture, hyperparameters, and training procedure for the policy networks described in Section 4.1.

Model Specifics. Both Q-BOT and A-BOT policies are modeled via Hierarchical Recurrent Encoder-Decoder networks [2, 5, 6] and no parameters are shared between them. All LSTMs (fact embedding, history encoder, question/answer decoder, and question encoder) have 2 layers and 512 dimensional hidden states. Word embeddings are 200 dimensional and are shared across question, history and decoder LSTMs. All models are implemented in Torch [1].

Curriculum Training. We use the penultimate layer activations from VGG-16 [7] as image representations and models are pretrained for 15 epochs on VisDial [2]. Afterwards they are transitioned to policy-gradient training by a fixed curriculum. Supervised training is performed for the first $K$ rounds of a dialog and policy-gradient updates are performed for the final $10 - K$. We anneal $K$ down by 1 every epoch (starting from 9). When $K$ is zero, ground-truth captions are swapped for captions generated by an off-the-shelf captioning model [3], so that an infinite number of dialogs can be simulated without any human supervision. Models are trained for 5 epochs on these generated captions. Fig. 1c plots reward vs. epochs during policy learning. We use Adam [4] with a learning rate of $10^{-3}$, and clamp gradients to $[-5, -5]$ to avoid explosion for all settings.

3. Qualitative Results

We provide additional qualitative results in Tab. 1, Tab. 2 and Fig. 2. Tab. 1 compares Q-BOT-A-BOT interactions for SL-pretrained and RL-full-QAf alongside the source images, their captions and human dialogs.

We notice that SL-pretrained agents often repeat questions –

```
what color are his shoes? black;
what color are his shoes? black;
what color ...
```

(see first row).

RL-full-QAf agents do not fall into these loops as often; moreover, when they do repeat themselves, it is later in the dialog and they often sensibly recover from these loops –

```
... can you see any trees? yes, there are lot of trees in background;
can you see any buildings in background? no, I can not;
can you see any trees? ...;
does it look like they are in park?” (row 1).
```

We also find that SL-pretrained agents tend to produce generic, ‘safe’ responses – “I can’t tell; I don’t know; Can’t see”, while RL-full-QAf interactions are noticeably more diverse and image-discriminative – i.e.,

```
what are they wearing? 
they are all wearing snow pants and jackets” (row 1)
```

```
where is man located? 
looks like classroom of some sort” (row 9).
```

Notice that the human dialog does not mention location or classroom, despite these facts being visually discriminative. In Sec. 6 of the main paper, we quantitatively evaluate how well A-BOT mimics human dialog (using the retrieval metrics from [2]). Recall that Frozen-Q-multi outperforms the other approaches on VisDial answering metrics. In Tab. 2, we present qualitative examples of responses generated by SL-pretrained, RL-full-QAf and Frozen-Q-multi A-BOT for ground-truth human-asked questions. We again find that SL-pretrained responses are predictably ‘safe’, while RL-full-QAf responses are more diverse and informative; however, this sometimes leads to RL-full-QAf being inaccurate. Frozen-Q-multi strikes a balance between response specificity and human-ness –

```
how old are men? 
SL-pretrained: I can’t tell, 
RL-full-QAf: they look like teenagers, 
Frozen-Q-multi: middle-aged” (row 2);
```

```
what colors are umbrella? 
SL-pretrained: white, 
RL-full-QAf: rainbow colors, 
Frozen-Q-multi: many different colors” (row 1).
```
Fig. 2 explores how Q-BOT’s image guess changes through rounds of dialog. The left column shows the true image and caption and the middle column shows the text interaction between Q-BOT and A-BOT for select rounds. To show how Q-BOT’s prediction changes through the dialog we show the five nearest neighbors of the prediction in the right column, with the nearest neighbor outlined in red. In the middle column we show where this nearest neighbor image ranks among other images in distance to the true feature, also showing other similarly distant images for context.

References


Table 1: Selected examples of Q-BOT-A-BOT interactions for SL-pretrained and RL-full-QAF. RL-full-QAF interactions are diverse, less prone to repetitive and safe exchanges (“can’t tell”, “don’t know” etc.), and more image-discriminative.
### Table 2: Selected examples of answers generated by SL-pretrained, RL-full-QAf and Frozen-Q-multi A-BOT for ground-truth human-asked questions. Although differences are subtle, SL-pretrained answers are usually ‘safe’ and repetitive, RL-full-QAf answers diverse and sometimes unfaithful, and Frozen-Q-multi answers hit a balance in between.
Figure 2: Qualitative results on predicted fc7-based image retrieval. Left column shows true image and caption, middle column shows dialog exchange, and a list of images sorted by their distance to the ground-truth image, right column shows a list of images sorted by their distance to predicted fc7. The image predicted by Q-BOT is highlighted in red. We can see that the predicted image is often semantically quite similar.