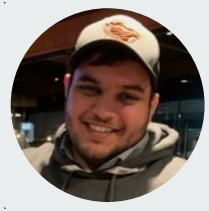
TBD - CLIP

Exploring Visual-Dialog Conversation

Rob

Aayam

Austin

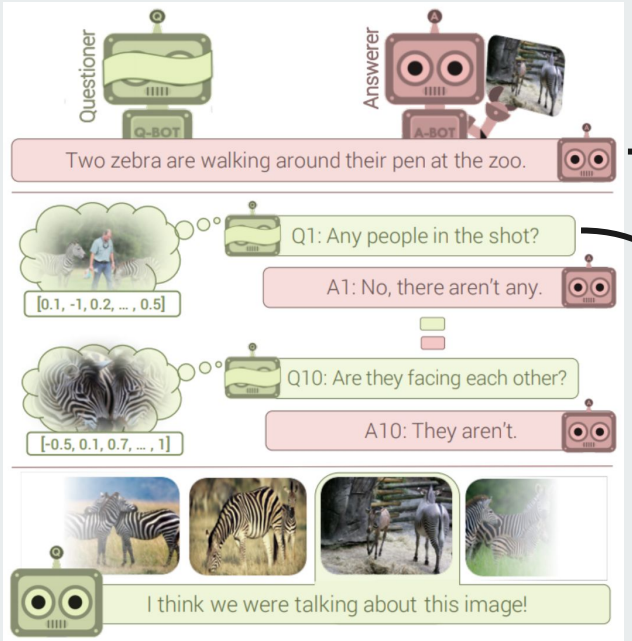


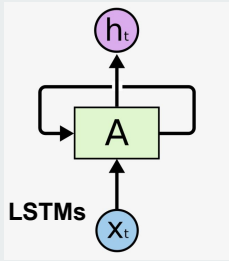
Mazen

Visual Question Answering

A) Supervised Loss
- Image Embedding prediction.
- Word level tokens.
- Diversity Loss

B) RL Fine-tuning
- Distance based rewards.
- Reinforce





Regression Network

Create software that can have meaningful conversations with humans about images

Das, A. et al. "Visual Dialog." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017): 1080-1089.

Murahari, Vishvak S. et al. "Improving Generative Visual Dialog by Answering Diverse Questions." EMNLP/IJCNLP (2019).

War of Embeddings

Caption + QA History

Q-bot

Text Embedding

Q: Are there any animals ?

A: Yes, there are big elephants.

Image Embedding

Text Embedding

A-bot

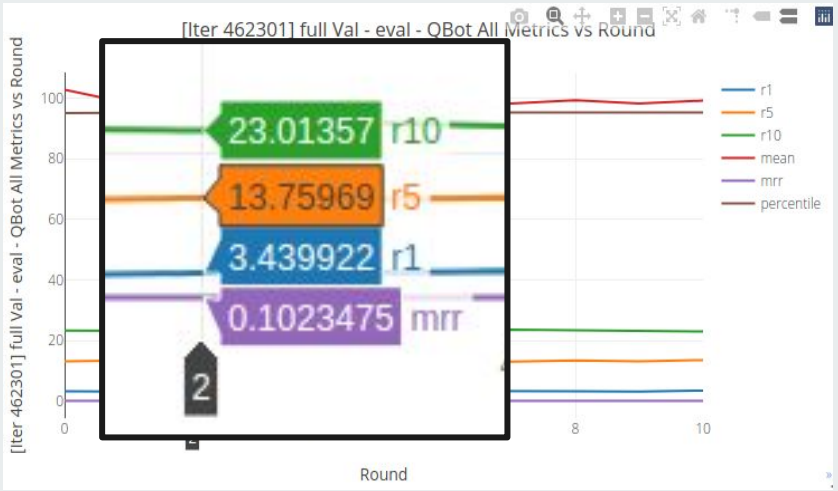
Guess

Image Embedding

NN >

Image Embeddings | VGG -> CLIP

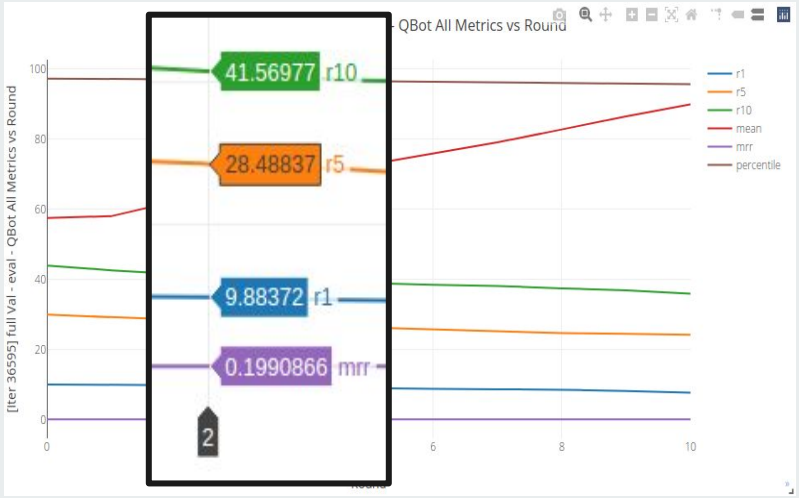
CLIP



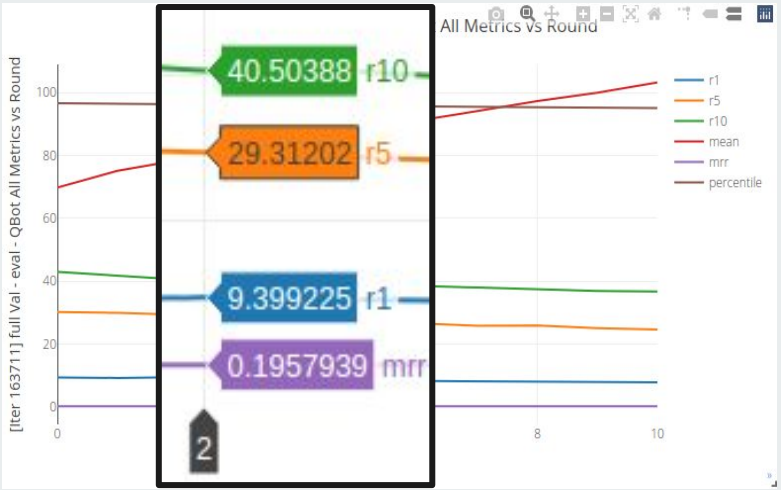
VGG Embeddings
(Murahari et al.)

OpenAI's CLIP uses a Vision Transformer (ViT) to embed images. The old way used embeddings from VGG16.

Text Embeddings | From Scratch -> Pretrained Glove



CLIP + Glove (fixed)



CLIP + (from scratch)

RL Fine-Tuning | Reinforce -> PPO

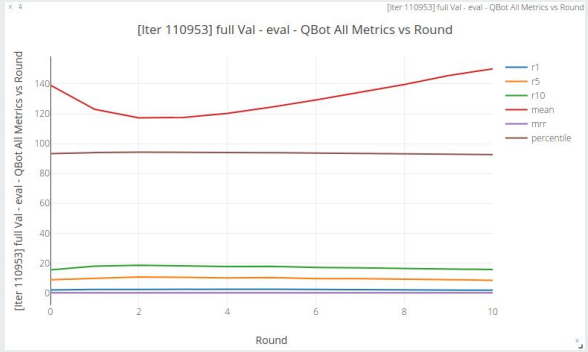
function REINFORCE
Initialise θ arbitrarily
for each episode $\{s_1, a_1, r_2, ..., s_{T-1}, a_{T-1}, r_T\} \sim \pi_\theta$ **do**
for $t = 1$ to $T - 1$ **do**
$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) v_t$
end for
end for
return θ
end function

PPO

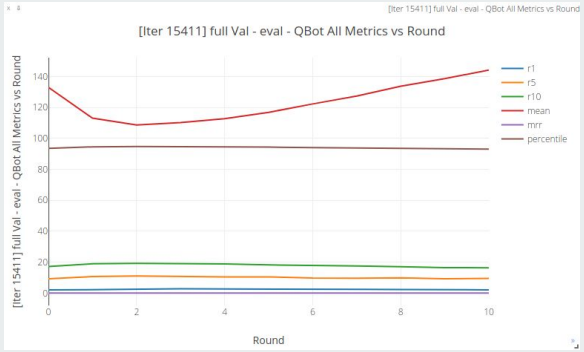
- Make a ratio of word probabilities between the current & old policy.
- Allows for more sensible weight updates.
- Reduces training time.

No Improvements were seen

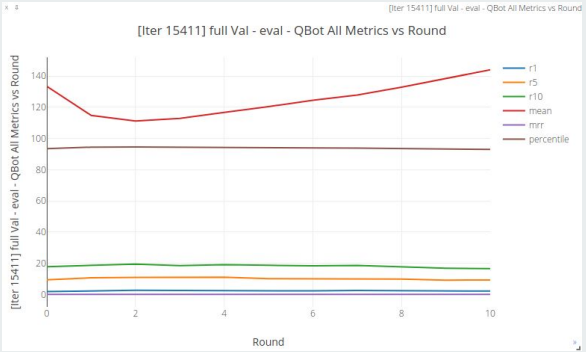
RL Fine-Tuning | Contribution ?



Alpha = 0



Alpha = 1



Alpha = 1e5

Testing for different Coefficient for RL Fine Tuning - Negligible Change in result.

Conclusion

- Visual Dialog Conversation is still in its developmental phase.
- CLIP embeddings do better at relating language with images
  - Big compute + multi-model pre-training is good,
- RL objective does not produce better results.
  - PPO over Reinforce is still an open question.

Exploring Visual-Dialog Conversation = Lots of models, metrics and fun

Demo Time :)

Results and Visualizations

Web Demo

Code

Next up, Questions !