video link : https://youtu.be/NBxjnRAB3LY

Analysis of Learning Cooperative Visual Dialog Agents

Juanita Ordóñez
Department of Computer Science, Stanford University
ordonez2@stanford.edu

Problem

Visual dialog is a challenging task which requires an agent to answer multi-round question about an image. This work focus on reimplementation goal-driven cooperative multi-agent approach. [1] Previous work has unnaturally treated dialog as a supervised learning problem where the answers are not generated but chosen from a list of possible candidates. [3,4]

Dataset

- Used VisDial [2] dataset
- 20k images and 100 images
- 10 dialog per image

Fig 1) Dataset examples

Game Setup

"Guess What" game environment [3]. Where A-Bot sees the picture and Q-Bot try to guess the image, by a series of question and answer.

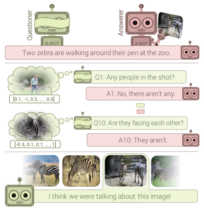

Fig 2) Game setup [1]

Model Architecture

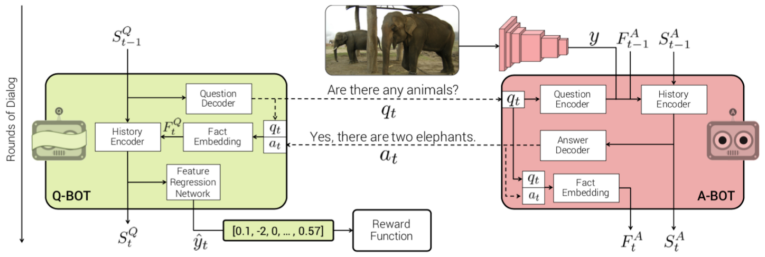

Fig 3) Q-Bot and A-Bot model architecture [1]. Where A-Bot is the only one that sees the picture and Q-Bot tries to figure out what picture A-Bot is looking at by recreating image feature representation.

Training Method

Both Q-Bot and A-Bot are first pre-trained, in a supervised manner using the train split VisDial dialog. This way Q-Bot and A-Bot learns to generate questions and answers respectively. Then fine-tuned using Reinforcement Learning Reward is the how close Q-Bot guess getting to real image per round.

Qualitative Results

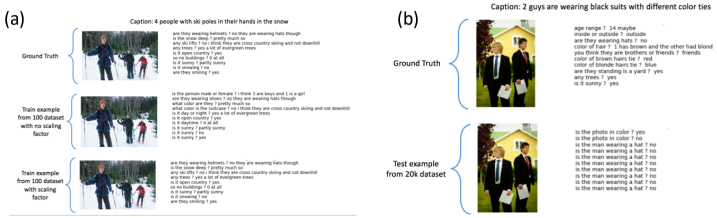

Fig 4) (a) Examples taken from the 100 dataset. Showing the effect of scaling Q-Bot scaling up loss. (b) Result taken from test set in 20k dataset.

Results

Table 1: Overall Performance

Model	Dataset	MRR	Recall@5	Recall@10	Epochs
SL QAbots	Sanity Dataset	44.87	46.00	59.80	1000
SL QAbots	100 Dataset	22.98	29.00	41.00	100
SL QAbots with scaling factor	100 Dataset	19.24	24.00	41.99	100
SL QAbots	20k Dataset	38.11	42.34	50.52	15
SL + RL [1]	VisDial version 0.5	43.8	53.08	60.22	15

Conclusion

In this work, we implemented the goal-driven multi agents for Visual Dialog. Pre-trained and evaluated the model using VisDial 20k train split. We learned that while training A-Bot is straightforward and performed as expected, Q-Bot is harder to train we suspect Q-Bot has more responsibilities.

References

[1] Jose M.F Moura Stefan Lee Abhishek Das, Satwik Kottur and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. *Proceedings of the IEEE International Conference on Computer Vision*, 2017.
[2] Khushi Gupta, Avi Singh, Dshraj Yadav, Jose M. Fmoura, Devi Parikh, Abhishek Das ,Satwik Kottur and Dhruv Batra. Visual dialog. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
[3] Sahar Kazemzadeh. Referitgame: Referring to objects in photographs of natural scenes. *Pro- ceedings of the 2014 conference on emprical methods in natural language processing*, 2014.
[4] Harm De Vries. Guesswhat? visual object discovery through multimodal dialoue. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.