Answerer in Questioner's Mind: Information Theoretic Approach to Goal-Oriented Visual Dialog

Sang-Woo Lee
Clova AI Research
Naver Corp.

Yu-Jung Heo
Seoul National University

Byoung-Tak Zhang
Seoul National University
Surromind Robotics

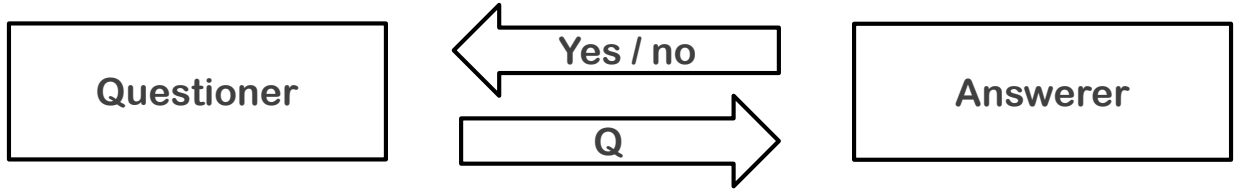
NeurIPS 2018 Spotlight Presentation

Montreal, Canada

Dec 4, 2018

Problem Definition – GuessWhat?!

H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. Guesswhat?! visual object discovery through multi-modal dialogue, CVPR, 2017.



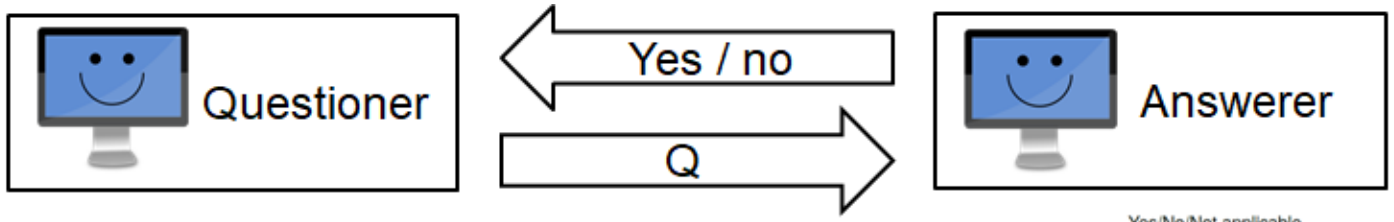
Is it a person?	*No*
Is it an item being worn or held?	*Yes*
Is it a snowboard?	*Yes*
Is it the red one?	*No*
Is it the one being held by the person in blue?	*Yes*

Is it a cow?	*Yes*
Is it the big cow in the middle?	*No*
Is the cow on the left?	*No*
On the right ?	*Yes*
First cow near us?	*Yes*

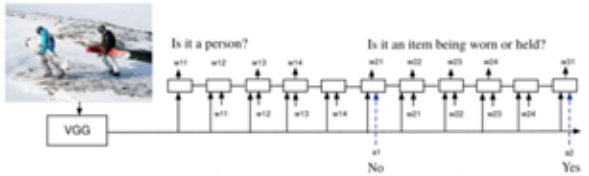
Previous Architectures

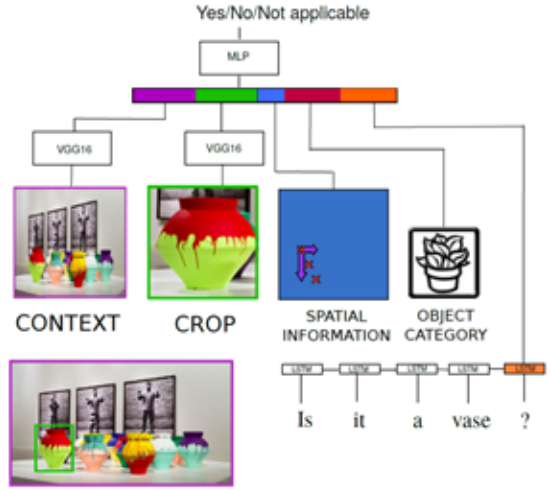
F. Strub, H. de Vries, J. Mary, B. Piot, A. Courville, and O. Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems, IJCAI, 2017.

- The goal of study is to increase the performance of machine-machine game and make emerged dialog from two machines.
- SL and RL are used to train question-generator and guesser.
  - Supervised learning: The questioner and the answerer trains from the training data.
  - Reinforcement learning: The questioner and the answers play a game, and use the dialog log for the training data.



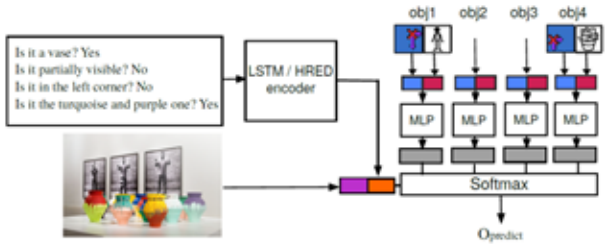
Question-generator



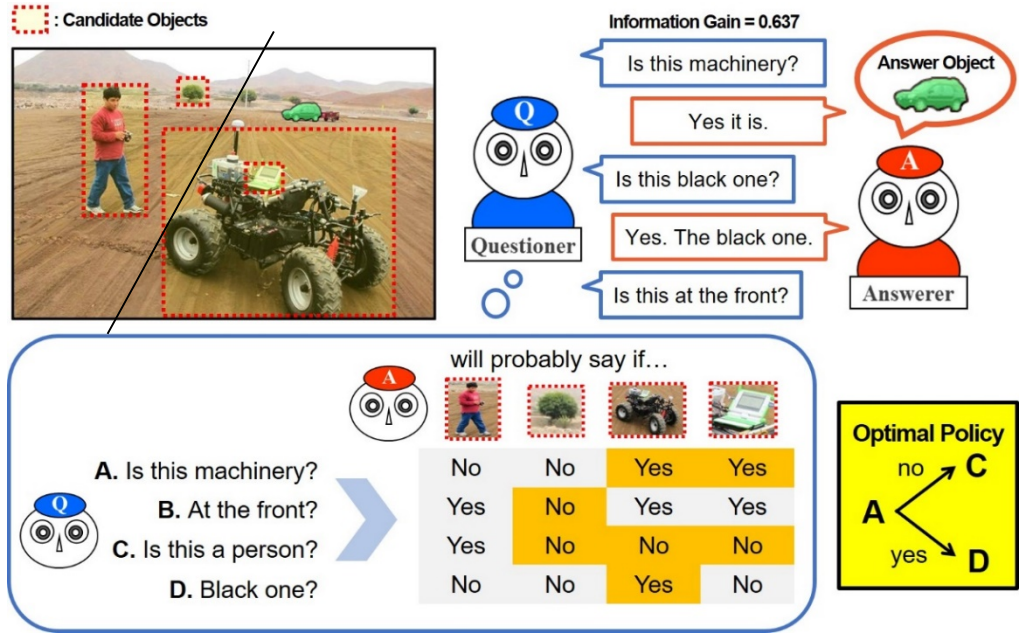


Answer-generator

Guesser



Our Method - AQM (Answerer in Questioner's Mind)

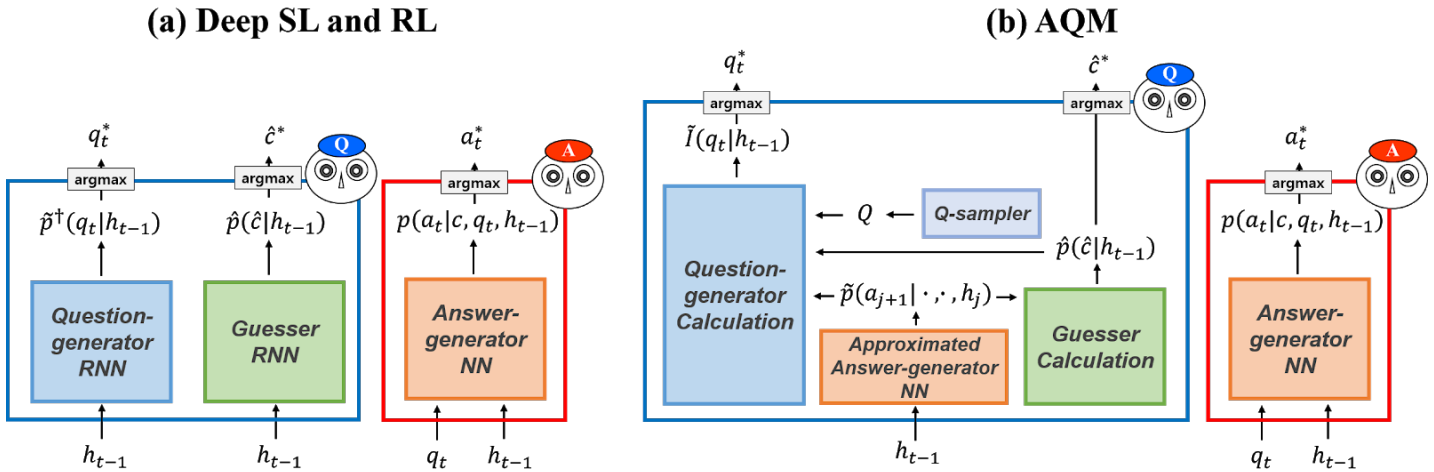


- Our Goal: Making a good questioner.
  - Not an answerer (VQA model).
- Our model asks question as solving 20 questions game.

$$I[C, A_t; q_t, a_{1:t-1}, q_{1:t-1}]$$

$$= \sum_{a_t} \sum_{c} p(c \mid a_{1:t-1}, q_{1:t-1}) p(a_t \mid c, q_t, a_{1:t-1}, q_{1:t-1}) \ln \frac{p(a_t \mid c, q_t, a_{1:t-1}, q_{1:t-1})}{p(a_t \mid q_t, a_{1:t-1}, q_{1:t-1})}$$

$$p(c \mid a_{1:t}, q_{1:t}) \propto p(c) \prod_{j}^{t} p(a_j \mid c, q_j, a_{1:j-1}, q_{1:j-1})$$



Experimental Result

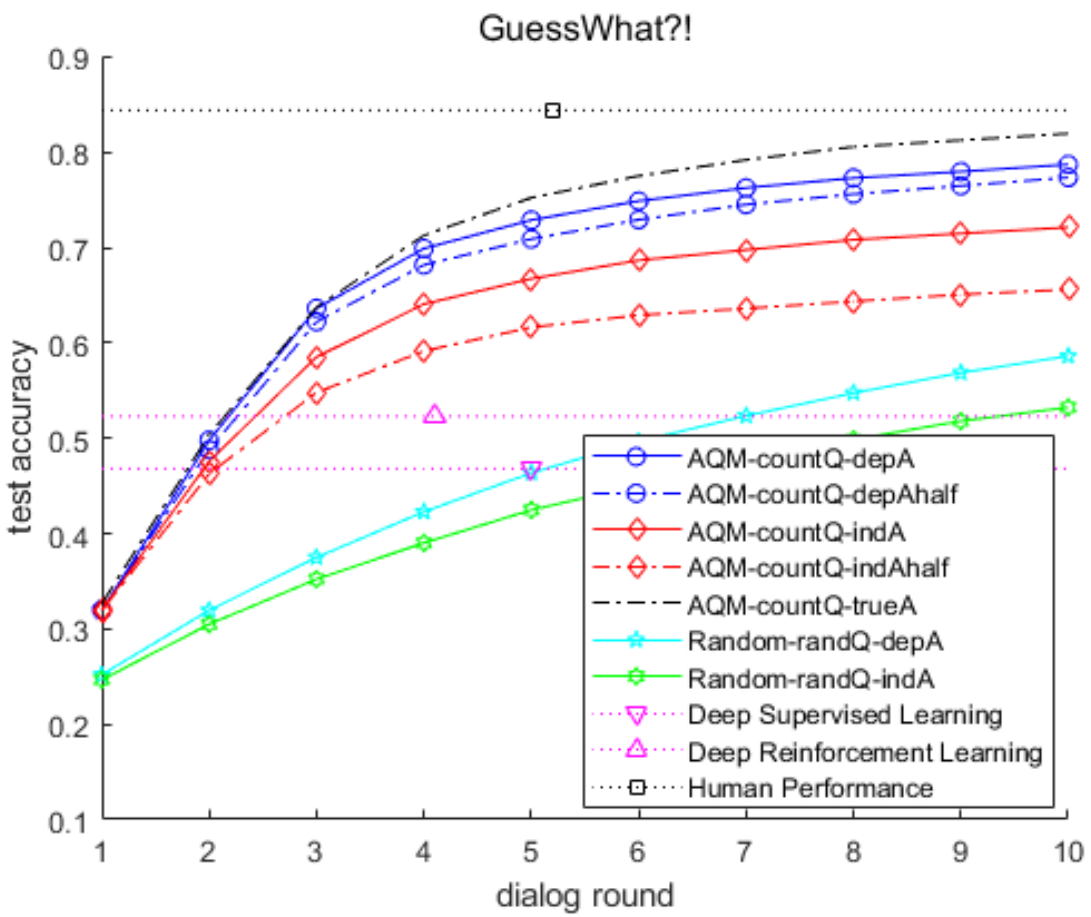


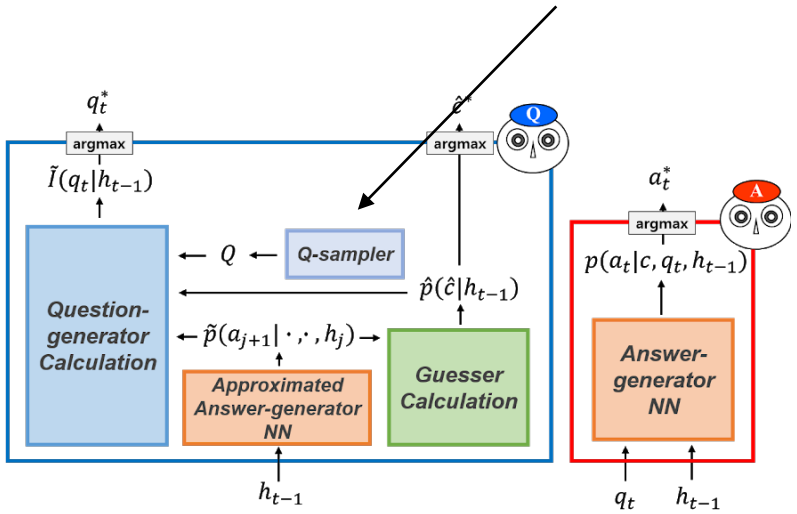
Table 1: Test accuracy from the GuessWhat?!.

Model	Accuracy
Baseline	16.04
Deep SL (5-q) [6]	46.8
Deep RL (4.1-q in Avg) [18]	52.3
Random-randQ-indA (5-q)	42.48 (± 0.84)
Random-randQ-depA (5-q)	46.36 (± 0.91)
AQM-randQ-indA (5-q)	65.66 (± 0.55)
AQM-countQ-indAhalf (5-q)	61.64 (± 0.97)
AQM-countQ-indA (5-q)	66.73 (± 0.76)
AQM-countQ-depAhalf (5-q)	70.90 (± 1.14)
AQM-countQ-depA (5-q)	72.89 (± 0.70)
AQM-countQ-depAhalf (10-q)	77.35 (± 0.85)
AQM-countQ-depA (10-q)	**78.72** (± 0.54)
Human [18]	84.4

Experimental Result

- Retrieve from training data or Generate from SL model

Sample candidate questions
from training dataset
or from SL neural model



Groundtruth	Supervised Learning (SL)	Answerer in Questioner's Mind (AQM, countQ)	Answerer in Questioner's Mind (AQM, gen1Q)
Is it person? Is he in the foreground? Is he wearing blue?	Is it a person? Is it the one in the front? Is it the one in the middle? Is it the one in the middle? Is it the whole person?	Is it a person? Is it a player? Is it in the background? Is it the umpire? Single piece in the bunch?	Is it one of the people? Is it in the background? Is it the person on the right? Is it the person holding the cake? Is it the person or the fire hydrant?
Is it an animal? Is it a device? Is it silver in color?	Is it the cat? Is it the cat? Is it the chair? Is it the book? Is it the book?	Is it alive? A laptop? Is it the shelf portion on the right? Left side of green color? Is a table?	Is it on the left half of the screen? Is it a book? Is it one of the bottles? Is it one of the cats? Is it on the table?
Groundtruth	**Reinforcement Learning (RL)**	**Answerer in Questioner's Mind (AQM, countQ)**	**Answerer in Questioner's Mind (AQM, gen1Q)**
Is it a person?	Is it a person? Is it girl in white?	Is it a fork? Is it a chair? Is it a food? Is it a person? Is it alive?	Is it on the table? Is the object human? Is it human? Is it one of the people? Is it a human?
Is a person? Is it tree? Is an object? The bat playing? The shoes on food? Is the red seat?	Is it a person? Is it a love? Is an umbrella? Is in the middle? On a person? Is it on on far right?	Is it a person? Is it the chair? Is it a traffic signal light? Is it the phone? Is it alive?	Is it human? Is it the man or part of the ball? Is it the person on the right? Is it a person or something they are holding? Is it being held by a person?

Conclusion & Argument

- Conclusion
  - We propose a practical goal-oriented dialog system motivated by theory of mind.
  - We test our AQM on two goal-oriented visual dialog tasks, showing that our method outperforms comparative methods.
  - We use AQM as a tool to understand existing deep learning methods in goal-oriented dialog studies.
  - We extend AQM to generate questions, in which case AQM can be understood as a way to boost the existing deep learning method.
- Argument
  - The objective function of AQM is indeed similar to RL in our task.
  - Learning both agents with RL in self-play in our task basically means that training the agent to fit the distribution of the other agent, making their distribution for from human's distribution.

See you at Poster session Tue Afternoon 95 & ViGIL workshop Fri for a future work of AQM!