Can a machine be intelligent? Is the ability to solve a complex arithmetic problem considered intelligence? If so, machines clearly are intelligent, in fact far surpassing the intelligence of most humans. For those who say machines are not, or at least not yet, intelligent, cognition is often a major factor: could the machine understand the posed question, that of its own intelligence? Could it locate itself in a photograph? Creating machines to understand language (first question) and vision (second question) has been the core motivation for the development of Artificial Intelligence, or AI, research. Sixty years ago, Alan Turing proposed a test for language capacity through evaluating how well a machine converses with a human being – The Turing Test. Recently, scientists proposed a Visual Turing Challenge, in which a machine is required to answer basic questions through the use of visual information, such as images. However, the disparity between the ability of humans to evaluate such questions and that of computers is still quite large, causing the questions for the Visual Turing Challenge to be too difficult to provide useful evaluation. However, Computer Vision researchers from the University of Maryland proposed a novel middle layer approach, in which they first allow a computer to generate questions based on images, and then design a method using deep learning to answer those questions.

What is Visual Question Answering?

Humans are generating huge numbers of images every day, through social networks, creation of digital content, etc. With this amount of data at their disposal, recognizing objects from images no longer meets the bar for computer vision. A machine should be able to look into millions of images and digest the information needed to answer important questions, ranging from simple ones like counting objects to far more complicated ones involving inference, such as whether a person sitting alone at a table for two is waiting for someone. Efforts have been made to publish image datasets and the associated questions for researchers to evaluate their algorithms [http://www.visualqa.org](http://www.visualqa.org). However, most recent results on the Visual Question Answering (VQA) benchmarks are solely due to language only information and the contribution of visual recognition is unclear.

Work at UMD

Overcoming the formidable task like the Visual Turing Challenge is nontrivial. The most time consuming (and probably the most expensive) step is to collect vast amount of labeled data (both questions and answers) annotated by humans. Researchers at the University of Maryland - College Park attack this problem by letting the machine generate the
questions on its own, taking only images as input. Their approach simulates a self-talk process in which an intelligent algorithm will look at images and ask human-like questions. The phenomenon of self-talk is well known in the field of Psychology as a special form of intrapersonal communication that facilitates learning and understanding. The model which generates questions from images is called the Visual Question Generation, or VQG - it is built on an image captioning system. A state of the art Visual Question Answering (VQA) such as [2] is then deployed for generating answers given the questions and the original image. These models are detailed below and Figure 1 illustrates the intuition behind these models.

**Visual Question Generation (VQG)**

This module deals with automatically generating questions from images. This extends a data driven image captioning system based on language models from Recursive Neural Networks (RNN) proposed by Karpathy and Li (2014) [1]. In the training step, the RNN will use a small set of images and their questions (a sequence of words) generated by human annotators to learn how to generate a word given that it has seen a previous word and the image pixels. In the generation step, RNN will use the image pixels (from unannotated images) as the starting point and gradually fill up the missing words. Because the model is guessing words with a certain associated probability, many questions can obtained for a given question—similar to what human annotators usually do.

**Visual Question Answering (VQA)**

Given a vast amount of questions generated automatically from VQG, a state of the art VQA [2] deploys a deep learning approach to answer a question given an image. This model uses a 19 layer convolutional neural network (CNN) that is pre-trained and a general purpose skip-gram word embedding. The VQG and VQA are trained simultaneously.

**Experiments**

Experiments were performed on two standard image datasets - the DAQUAR dataset, which has 12,468 human question answer pairs on 1,449 images of indoor scenes and the COCO dataset for VQA - with 369,861 questions and 3,698,610 ground truth answers based on 123,287 MSCOCO images. The evaluation is performed at two levels. On one level, they evaluate the questions generated by the VQG using standard linguistic and image captioning measures. On the other level the researchers use Amazon Mechanical Turk and check the readability, grammar, correctness and accuracy of the description of the image in answers.

**Results**

The team found that the questions generated from the models are similar to the questions asked by humans, based on the linguistic and standard captioning metrics. Even from the Amazon Mechanical Turk evaluation, they find that the questions generated achieved near-human readability. Table 1 summarizes these results. For the self-talk evaluation, turkers found the answers above the “a bit like human being” correctness but below “half-human, half-machine” correctness. In essence, a significant number of these turkers felt that the answers were “roboty” and the very concept of a self-talking robot as humorous. Figures 2 and 3 show what the felt about self-talk. The researchers acknowledge that the project is still a long way before achieving the maturity and accuracy that the project sets out to achieve.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Readability</th>
<th>Correctness</th>
<th>Human-like</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAQUAR</td>
<td>3.35±1.92</td>
<td>2.5±1.03</td>
<td>2.49±1.02</td>
</tr>
<tr>
<td>COCO-VQA</td>
<td>3.39±1.18</td>
<td>2.7±1.29</td>
<td>2.40±1.33</td>
</tr>
</tbody>
</table>

Table 1: Results from Amazon Mechanical Turk evaluation

Conclusions

The Neural Self Talk considers the image understand problem as a sequence of consecutive self-questionings and answer processes. The two key components, Visual Question Generation (VQG) module and Visual Question Answering (VQA) module, enable to construct the question and answer pairs. The evaluation results by users using Amazon Mechanical Turk demonstrate the promising potentials of the research: robots can behave like human being in the near future. (for now, its included as a half human, half robot category) To improve the performance of the intelligent self talk, the research proposed two highly possible ways: it can 1) incorporate with common-sense knowledge and 2) construct a story-line mimicking human communications. The matured result of the improvement will possibly make users feel cure and fondness with robot companions, and it will boost the current robot researches by integrating the intelligent self talk model to them.

References


Contributors

- Raveesh K Mayya (raveesh@rhsmith.umd.edu)
- Rashmi Sankepally (rashmi@mail.umd.edu)
- Richard Neal (richard@cs.umd.edu)
- Sanghyun Hong (shhong@mail.umd.edu)
- Thang Dai Nguyen (daithang@umiacs.umd.edu)