Title

Deeply Explainable Artificial Intelligence (DExAI): Transparent multi-step reasoning with Stack Neural Module Networks (SNMN)

Team Composition

University of California, Berkeley (UCB), Trevor Darrell, Principal Investigator, trevor@eecs.berkeley.edu
UCB: Ronghang Hu, Anna Rohrbach
Boston University: Kate Saenko

Problem Statement

This effort investigates two distinct, but related explainable AI questions regarding the UCB SNMN model: 1) does the SNMN model improve users' satisfaction of the explanation? And 2) Does the SNMN model allow users to form truthful beliefs about the model's behavior?

Approach and Technique

The UCB SNMN model performs compositional reasoning by automatically inducing a desired sub-task decomposition without relying on strong supervision, being fully differentiable and more interpretable to human evaluators. In the SNMN model, the internal reasoning procedure is transparent to the users (which is unlike black-box neural networks), allowing users to directly see how the final prediction is made based on internal steps. Figure 1 shows an overview of the SNMN model.

Figure 1. SNMN model overview

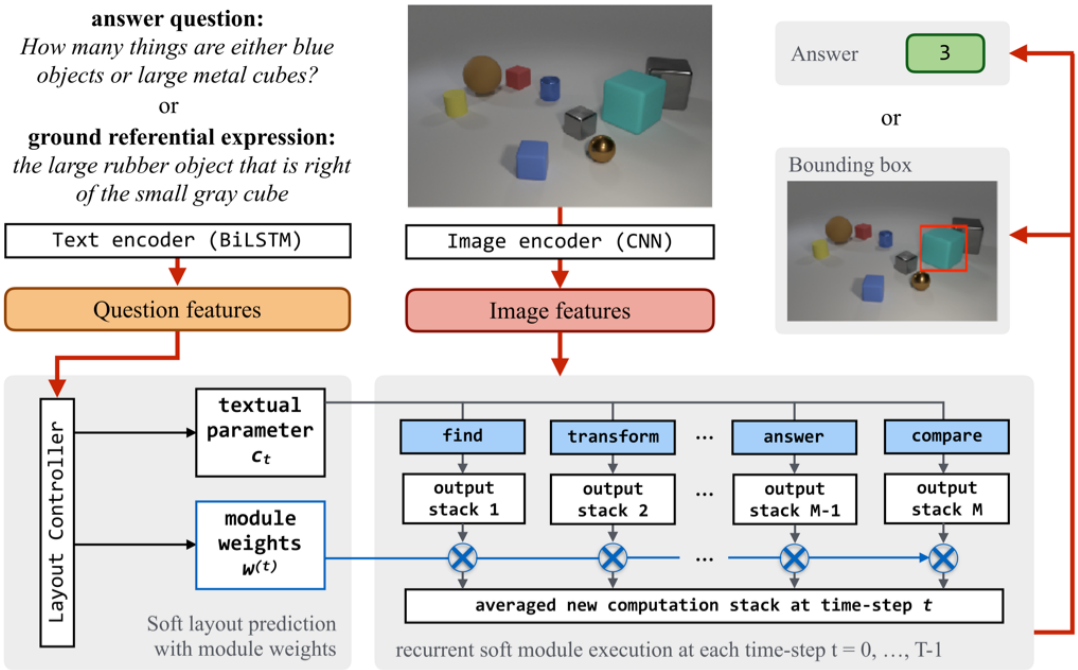
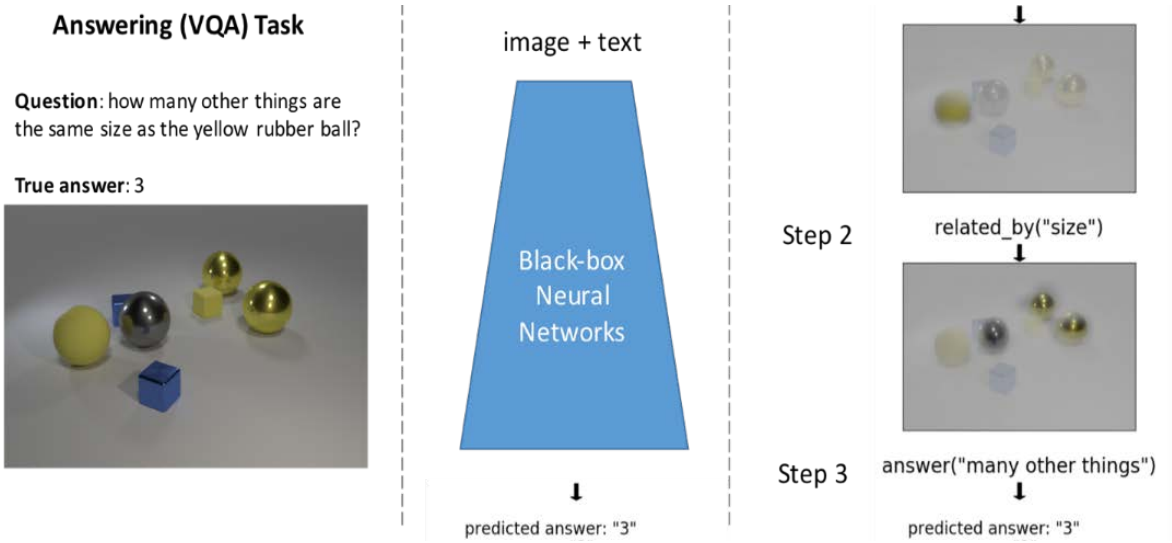


Figure 2 below illustrates a comparison between the SNMN model on the far right and the "without explanation" version in the middle (black-box neural networks). The input question and image are shown on the left.

Approved for Public Release, Distribution Unlimited

Figure 2. Illustration of an example visual question answering task (left), a model version without explanation (middle), and the SNMN model version with explanation (right)



Key Objectives and Current Status

Study results show that users can more clearly understand the reasoning procedure in the SNMN model-generated explanations, as shown in Table 1, Explanation Satisfaction. Also, the SNMN model allows the users to form truthful beliefs of the model behavior, and enables them to predict whether the model will get the correct answer or fail, as shown in Table 1, Failure Detection. Table 2 presents numerical results for both Explanation Satisfaction and Failure Detection.

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
Explanation Satisfaction	Generate explanations that are satisfactory to users Satisfactions are subjectively rated on a 4-point Likert scale by users	Users are more satisfied with the explanations from the SNMN model (compared to those from baseline methods), as they can more clearly understand the reasoning procedure from the explanations
Failure Detection	Investigate whether users can form truthful beliefs from the model's explanations, by measuring whether humans can predict the model's answer and detect its failure based on its explanations	The SNMN model enables users to predict whether the model will get the correct answer or fail more accurately than baseline methods

Approved for Public Release, Distribution Unlimited

Table 2. Numerical results for Explanation Satisfaction and Failure Detection

	Numerical Value		Hypothesis Results	Trust Measure
	With Explanation	**Without Explanation**		
Explanation Satisfaction Score	3.33	2.510	The score is significantly higher with explanation	e.g., 1 - 4 (1 is bad, 4 is good)
Failure Detection Accuracy	0.625	0.495	The accuracy should be larger than 0.5 with explanations	Higher accuracy is better (0.5 is random accuracy)

Test and Evaluation

For evaluation, the model's intermediate outputs, such as the image attention and textual attention at each step, are visualized; and the model's final prediction is also shown. The evaluators are then asked to rate the clearness of the explanation.

The failure detection evaluation investigates whether humans can predict the model's answer and detect its failure based on its explanations. The test set is split into half correct and half incorrect model predictions, and the final answer output is not shown, so that human baseline performance should be chance or 50%. The hypothesis is that if humans can predict whether the model succeeds or fails better than chance, they understand something about the model's decision process.

Hu, Ronghang, et al. "Explainable neural computation via stack neural module networks." *Proceedings of the European Conference on Computer Vision (ECCV*). 2018.

For additional information:

https://people.eecs.berkeley.edu/~ronghang/misc/xai_demo_nmn/

Approved for Public Release, Distribution Unlimited