Title

Deep Attentional Representations for Explanations (DARE)

Team Composition

SRI International, Giedrius T. Burachas, Principal Investigator, giedrius.burachas@sri.com
University of Toronto: Richard Zemel, Sanja Fidler, and David Duvenaud
University of California, San Diego: Jurgen Schulze

Problem Statement

End-to-end deep learning-based AI applications for multimodal data analytics have recently produced impressive results - sometimes surpassing human performance - but lack of transparency of their operation raises trust and reliance concerns as users may not be able to understand the reasons and causes for AI application outputs, whether accurate or inaccurate. SRI developed several explainable deep learning models for multimodal data analytics based on a popular Visual Question Answering (VQA) AI task. In order to make the VQA Deep Learning (DL) Neural Network (DNN) systems more transparent to a user, SRI augmented them with multiple modes of explanation.

Approach and Technique

SRI developed an XQA (eXplainable visual Question Answering) system that extends the VQA task with diverse explanation modes, including: causal spatial attention, object attention, and language-based show-and-tell explanations; related question/answer examples; image scene graphs; and counterfactual generative input image editing. The two modes of attention, spatial and object-based, are geared towards focusing on contiguous space features (e.g., “stuff” like grass, pavement, sky, and individual features) and discrete objects, respectively. In addition, SRI’s active attention-based explanations offer the user a capability to interactively modify attention masks for understanding causal effects of attention on answers in VQA tasks. Such active explanations can also be used for improving AI model performance during deployment, a unique feature that present day machine learning (ML) systems are not capable of supporting.

In addition, SRI has explored interpretable hierarchical VQA models that offer greater transparency of model operation at multiple levels of inference. For example, the Progressive Module Network builds a model for the complex VQA task out of simpler submodules trained for such sub-tasks as object detection, attribute and relationship recognition, counting, and captioning. Rich semantic outputs at each submodule level can be used for further enriching explanation modes.

Figure 1 depicts SRI’s XQA model that encodes questions using Long Short Term Memory (LSTM) (top stream), and provides two types of visual attention: 1) spatial-based, using a ResNet (middle stream), and 2) object-based, using Mask Region Convolutional Neural Networks (RCNN) (bottom stream). Spatial and object-based features are concatenated and weighted with corresponding attention weights. The resulting features are used in the final step, an answer classifier that produces a probability distribution over the 3,000 most likely answers. Explanations derived from the model include spatial attention weights (projected back onto image space), object-based attention masks (displayed as object masks), scene graphs derived from Mask RCNN model object detections, and related question/answers. Figure 2 depicts SRI’s XQA interface without explanations on the left, and with explanations on the right; spatial, object-based attention, attention-filtered object bounding boxes, and scene graphs are shown at the bottom.

Approved for Public Release, Distribution Unlimited

Figure 1. XQA model overview

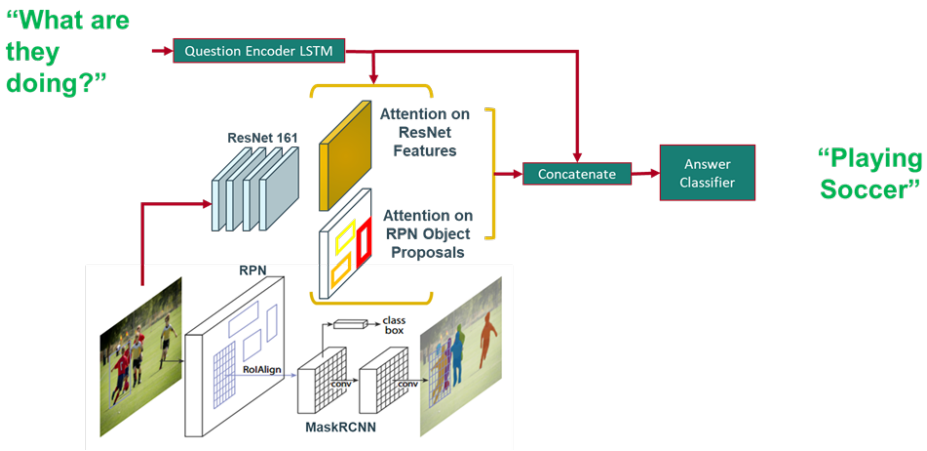
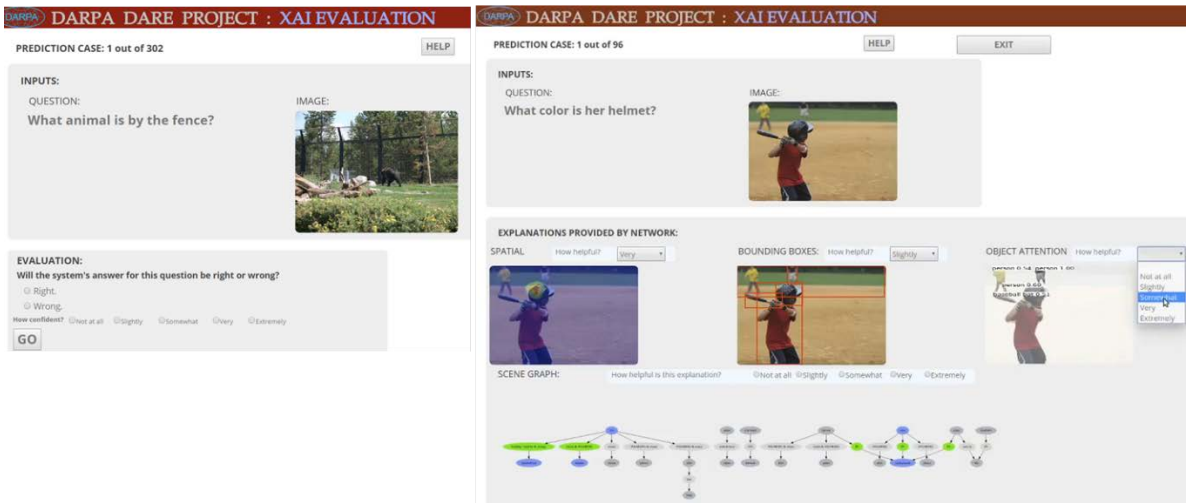


Figure 2. XQA interface overview



Key Objectives and Current Status

The DARE system explanations have proven to be an effective means for improving user performance in a collaborative human-machine task, for inducing meaningful mental models of the XQA system for the user, and for offering satisfactory explanations to the user. Interestingly, the users were able to rate explanation helpfulness in a manner so that explanations perceived as helpful were more effective for a task. Moreover, explanations tended to have greater impact when the system erred. The shortcomings that were revealed during user studies included lack of explanation detail and insufficient information for judging when the algorithm is accurate or trustworthy. Finally, on a small fraction of trials when the system erred but explanations were perceived as highly helpful, the users tended to be more strongly mislead by them.

The key objectives for the next phase of the program are to: improve effectiveness of explanations by improving model interpretability; and provide explanatory/diagnostic information in a more complete manner so that inferences in all XQA model components are explained. For improved model interpretability, Progressive Module Networks that are amenable for providing explanations for multiple

Approved for Public Release, Distribution Unlimited

subtasks contributing to VQA will be employed. Active explanation domain will be expanded by offering interactive tools for additional modules in order to support causal explanation modes. Explanation effectiveness will be further improved through self-diagnostic techniques that track the XQA model's uncertainty in the inference. VQA model performance will be improved through meta-learning techniques that can also readily provide novel explanations by example.

Table 1 summarized the measures, objectives and actuals to date.

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
Explanation effectiveness for collaborative human-machine ML performance	Increase user ability to improve performance measures with explanations	XQA system-generated explanations significantly increased user performance as compared to no explanation baseline
Quality of the XQA system's mental model as measured by prediction task accuracy	Increase functional understanding of the XQA system for both correct and wrong outcomes	Explanations improve user prediction of XQA system correctness only for cases when the answer is wrong
Quality of the mental model as measured by questionnaire	Increase structural understanding of the XQA system	Currently, no structural (*introspective*) explanations are offered
User's trust in the XQA system as measured by questionnaire and spontaneous adoption of explanations	Increase user's ability to diagnose when to trust the XQA system's response	Currently, the evidence for trustworthiness is provided by an observation that users spontaneously opt for explanations; user confidence in their decisions correlate with the XQA system's confidence
User's satisfaction in explanations as measured by questionnaire adjusted for bias	Increase objective user satisfaction with explanations whose perceived satisfactoriness correlates with actual helpfulness	Currently, users tend to prefer intuitive explanation modes like spatial attention even for the trials when they are misleading

	Numerical Value		Hypothesis Results	User Machine Task
	With Explanation	**Without Explanation**		
User-Machine Task Performance	Setting 1 (20 questions): 47.8% Setting 2 (5 questions - attention and rel QA): 69.0%	28.6% 63.2%	Significant, p = 0.019	GuessWhich
	Performance depends on explanation helpfulness rating (most helpful explanations improve performance by 25.0%)		Significant, p < 0.010	

	Numerical Value		Hypothesis Results	Mental Model Task
	With Object Attention Explanation	**Without Explanation**		
Mental Model Understanding	Correct VQA answer: 84.7% Wrong VQA answer: 53.0%	86.5% 35.1%	Not significant Significant, p < 0.05	Prediction Accuracy

	Size	Length	Type
User Study	e.g., 54 subjects	e.g., number of trials, number of conditions	e.g., AMT, etc. e.g., between, within

Approved for Public Release, Distribution Unlimited

GuessWhich AMT Study	69 subjects; 1,469 HITs	>= 20 trials/subject, 4 groups/conditions	AMT, between subjects, mixed
Prediction Task Study	90 subjects; 4,124 trials	6 groups/conditions	Laboratory setting, mixed design

Test and Evaluation

XQA explanation mode evaluations verify that explanations, as intended, improve AI task performance and thus validate the notion that explainable AI is an effective framework for opening up and shedding light on algorithmic AI black boxes. Users can benefit from such explanations in terms of gaining a better understanding of the system (improving mental model) and achieving cognitive satisfaction, which in turn improves their trust in the system. Most importantly, such lucid explanations can improve performance in human-machine tasks.

Arijit Ray, Giedrius Burachas, Yi Yao, and Ajay Divakaran. *Lucid Explanations Help: Using a Human-AI Image-Guessing Game to Evaluate Machine Explanation Helpfulness*, 2019.

Arijit Ray, Yi Yao, Avi Ziskind, Rakesh Kumar, and Giedrius Burachas. *Evaluating Visual-Semantic Explanations using a Collaborative Image Guessing Game*, 2018 (CVPRW 2018).

Kim, Seung Wook, Makarand Tapaswi, and Sanja Fidler. *Visual Reasoning by Progressive Module Networks*, 2018, arXiv:1806.02453.

Kamran Alipour, Jurgen P. Schulze, Yi Yao, Arijit Ray, Avi Ziskind, and Giedrius T. Burachas. *Interactive Explanations for Visual Question Answering.* In preparation.

For additional information:

https://xaidevel.nautilus.optiputer.net/predres, username: guest, password: demo12

https://xaidevel.nautilus.optiputer.net/pred

https://gwapp.nautilus.optiputer.net/ (username: _demo)

Approved for Public Release, Distribution Unlimited