Title

Deeply Explainable Artificial Intelligence (DExAI): Textual Justification with Grounding

Team Composition

University of California, Berkeley (UCB), Trevor Darrell, Principal Investigator, trevor@eecs.berkeley.edu
UCB, Lisa Anne Hendricks
University of Amsterdam, Zeynep Akata

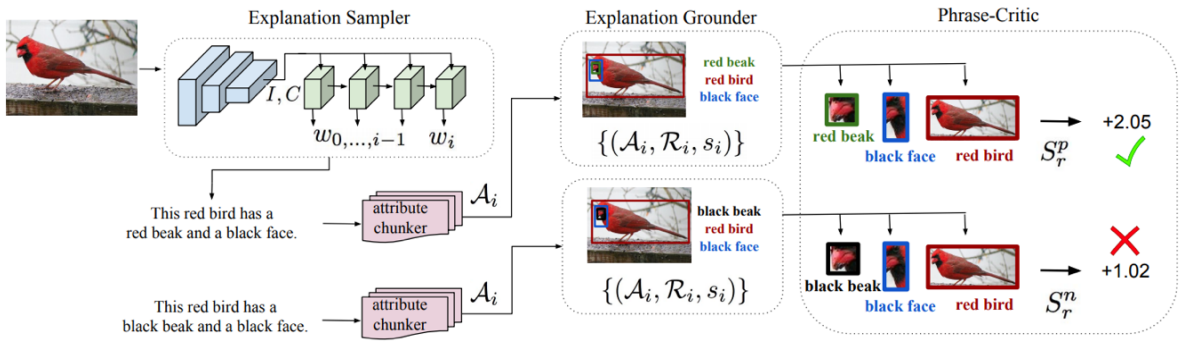
Problem Statement

In UCB's work on textual justification with grounding, they consider justifying fine-grained classification decisions of birds. The focus is on the following criteria: Correctness, Understanding (using a prediction task), Performance, and Satisfaction. To validate the model they consider studies done on Amazon Mechanical Turk (AMT).

Approach and Technique

UCB's approach (Figure 1) combines explanation language generation and natural language grounding models to ensure that explanations are both discriminative for a specific fine grained category and grounded in visual evidence. The input to the system is an image of a bird. First it is classified by a deep classification model. The fine grained image feature and predicted label are concatenated and input into a language generation model. The generation model can output sentences like, "This is a cardinal because it is a red bird with a black cheek patch." They then use a language grounding model to ground constituent visual evidence into the image (e.g., draw a box around the "black cheek patch" in the image). By grounding visual evidence, they add more depth to the explanations (they have a visual component as well as a text component) and they ensure the explanations are more image relevant.

Figure 1. Textual justification with grounding approach



Key Objectives and Current Status

Table 1 below highlights key measures, objectives, and actuals to date.

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
Understanding	Given two images and a class prediction, users better understand the image the predicted label corresponds to when given an explanation.	Generated explanations led to higher understanding than a system with no explanation.

Approved for Public Release, Distribution Unlimited

Performance	Given an explanation and an image, users are able to better determine if they should accept an AI decision when provided with an explanation.	Generated explanations led to higher human-machine performance than a system with no explanation.
Satisfaction	User satisfaction increases when provided with an explanation.	Generated explanations were more satisfactory to humans than no explanation baseline.
Explanation Goodness	Explanations are more grounded in visual evidence when the phrase-critic is used to select possible explanations.	Explanations are more grounded in visual evidence than prior work.

	Numerical Value		**Hypothesis Results**	**User Machine Task**
	With Explanation	**Without Explanation**		
User-Machine Task Performance	72%	67%	Significant, $p < 0.001$	Fine-grained bird classification

	Numerical Value		**Hypothesis Results**	**Mental Model Task**
	With Explanation	**Without Explanation**		
Mental Model Understanding	65%	53%	Significant, $p < 0.001$	Fine-grained bird classification

	Numerical Value		**Hypothesis Results**	**Satisfaction Measure**
	With Explanation	**Without Explanation**		
Explanation Satisfaction	2.57	2.88	Marginal, $0.05 < p < 0.1$	Likert Scale 1 - 5; 1 is most satisfied; 5 is least satisfied

	Numerical Value		**Hypothesis Results**	**Goodness Measure**
	With Explanation	**Without Explanation**		
Explanation Goodness	62%	52%	Significant	Correctness of mentioned attributes in explanations

Test and Evaluation

For additional information see:

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, Zeynep Akata. *Grounding Visual Explanations*. European Conference on Computer Vision, 2018.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, Zeynep Akata. *Generating Counterfactual Explanations with Natural Language*. International Conference on Machine Learning (Workshop on Human Interpretability), 2018.