Title

Model Explanation by Optimal Selection of Teaching Examples

Team Composition

Rutgers University - Newark, Patrick Shafto, Principal Investigator, patrick.shafto@rutgers.edu
Rutgers University - Newark, Scott Cheng-Hsin Yang

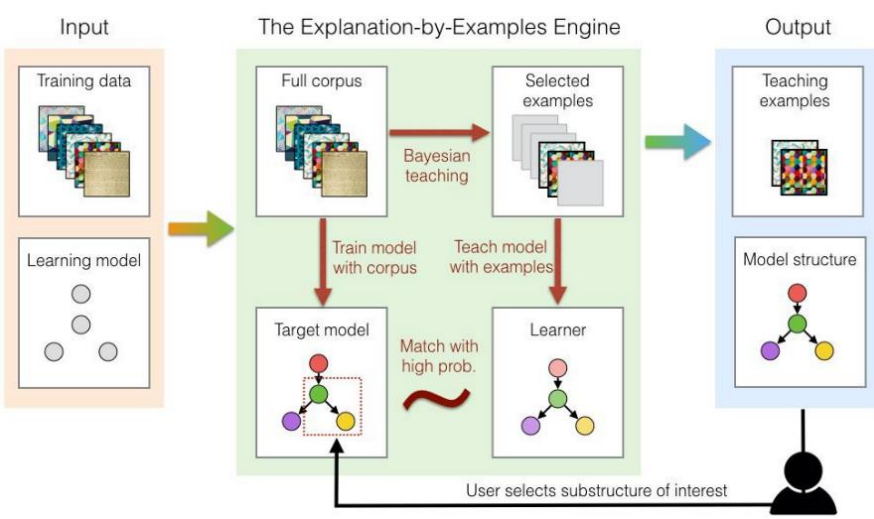
Problem Statement

The Rutgers team works on the challenge problem area of data analytics. The goal is for users of the Explanation-by-Examples system to be able to better understand the inference of black-box machine learning models. This understanding will be reflected in the users' ability to better predict the models' predictions and hence help the users to figure out the strengths and weaknesses of the models. The current focus is on the problem of image classification. Image classification is a core component of computer vision and underlies a wide range of applications, such as facial recognition, disease diagnoses, image search, and surveillance. It has seen great advancement in the past decade due to the development of deep learning and the curation of large, high-quality datasets. Rutgers' Explanation-by-Examples system will be developed on machine learning models ranging from simpler, lower-performing models trained on small datasets and scaling up to state-of-the-art deep learning models trained on state-of-the-art datasets.

Approach and Technique

The engine of the Explanation-by-Examples system (see Figure 1) is based on Bayesian Teaching. In short, Bayesian Teaching is a framework developed in cognitive science for selecting a small subset of the training data that will, with high probability, lead a learner to the inference of a target model (the black-box models trained on the entire training corpus). This framework is model-agnostic and can be applied to any probabilistic model. Deep learning models are included as they can be easily made probabilistic by replacing the model's output layer with a probabilistic model. The explanations, being examples from the training data, are naturally understandable to domain expert users and not restricted to a particular form. The framework also offers a model-based measure of explanation goodness that can be derived from the selection probabilities of examples.

Figure 1. Overview of the Explanation-by-Examples system



Approved for Public Release, Distribution Unlimited

Key Objectives and Current Status

The primary objective is that users of the Explanation-by-Examples system will be able to better predict the prediction of the model of interest. The complementary, higher-level objective is that the system will help users form an understanding of the competencies and incompetencies of the model and thereby foster appropriate trust and usage of the model. The current system can provide explanations for ResNet50, a commonly used high-performing convolutional neural network, by selecting examples from ImageNet, which contains 1,000 categories with roughly 1,000 images per category. The current system is configured in a way that can be easily applied to any pre-trained feedforward and convolution neural networks for classification trained on datasets of similar size to ImageNet.

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
Mental model understanding	Increase user prediction accuracy	System-generated examples increased user prediction accuracy vs. no examples, random examples, or unhelpful examples
Explanation satisfaction	Increase user satisfaction	System-generated examples marginally increased user satisfaction vs. random examples

Note: Explanations and examples are synonymous in the Explanation-by-Examples framework.

	Numerical Value			Hypothesis Results	Mental Model Task
	Helpful Examples	**Random Examples**	**Unhelpful Examples**		
Mental Model Understanding	72.5%	69.3%	66.6%	Significant, p = 0.013 (planned contrast: helpful > random > unhelpful)	Prediction Accuracy (PLDA on CAFE) [1]

Note: Explanation goodness is implicitly reflected in the prediction accuracies.

	Numerical Value					Mental Model Task
	Helpful Examples	**Helpful Examples + Labels**	**Random Examples**	**Random Examples + Labels**	**Labels Only**	
Mental Model Understanding (familiar categories, correct prediction)	93%	91%	76%	75%	93%	Prediction Accuracy (ResNet50 on ImageNet) [2]
Mental Model Understanding (unfamiliar categories, correct prediction)	70%	78%	66%	62%	65%	
Mental Model Understanding (all categories, incorrect prediction)	36%	37%	32%	38%	36%	

Note: Detailed statistical analyses will be reported in [2]; Explanation goodness is implicitly reflected in the prediction accuracies.

Approved for Public Release, Distribution Unlimited

	Numerical Value		Satisfaction Measure
	Helpful Examples	**Random Examples**	
Explanation Satisfaction	2.23/4	2.17/4	Likert Scale with 4 options (ResNet50 on ImageNet) [2]

Note: Detailed statistical analyses will be reported in [2].

	Size	**Length**	**Type**
User Study (PLDA on CAFE) [1]	105 subjects	120 Trials, 3 conditions	AMT (between)
User Study (ResNet50 on ImageNet) [2]	116 subjects	150 Trials, 6 conditions	AMT (between)

Test and Evaluation

Rutgers designed psychophysical experiments to test the explanation effectiveness of the system. The general design is a two-alternative forced choice task where the users are asked to predict the predictions of a target machine learning model on test images. The trials are balanced across the correct and incorrect model predictions. Between-subject conditions include with-explanation vs. without-explanation as well as comparisons between explanations of different model-based goodness. All test evaluations to date have been conducted on Amazon Mechanical Turk (AMT). Evaluation results show the conditions under which explanations generated by the Explanation-by-Examples system improve user prediction accuracy.

[1] Wai Keen Vong, Ravi B. Sojitra, Anderson Reyes, Scott Cheng-Hsin Yang, and Patrick Shafto (2018). *Bayesian teaching of image categories.* In Proceedings of the 40th annual conference of the Cognitive Science Society.

[2] Scott Cheng-Hsin Yang, Wai Keen Vong, Ravi B. Sojitra, and Patrick Shafto (2019). *Explainable artificial intelligence by optimal selection of teaching examples.* In Preparation.

https://imagenet-label-teaching.appspot.com/demo

Approved for Public Release, Distribution Unlimited