Measuring Interpretability of Visual Question and Answering Models

Patrick Hayes
UC San Diego
phhayes@ucsd.edu

Abstract

Research in visual question and answer has been criticized as contributing uninterpretable models. In response to this a growing amount of research has claimed to improve the state-of-the-art by increasing interpretability. However these papers define interpretability differently and often provide little to no empirical evidence that their models improve interpretability. *Goldstein et al provide a framework for measuring the interpretability of simple machine learning models. In this paper we extend this framework to measure the interpretability of the variations of the Transparency by Design network. We find that without a network agnostic mechanism for generating characteristic input output examples the framework provide Goldstein et al is a weak measure of a network's interpretability.*

1. Introduction

The accuracy of a machine learning model on a curated test set is currently the primary measure of performance for a machine learning model. The test accuracy gives the user some measure of reassurance that the model will perform well when applied in practice. This assurance is relies on the test set being a large random sample of independent and identically distributed random variables. If the test set satisfies this requirement than statistical learning theory states that the model will perform equally well on random samples that are drawn from the same distribution.

These assumptions rarely hold in practice. The extent to which the assumptions hold relies largely on the domain and the resources available during test time. This leaves the user wondering how much they can trust their model, which becomes particularly problematic for high-risk applications. Interpretability gives the user an additional resource to ensure the model is working as intended.

Additionally accuracy is not the only important metric for evaluating a machine learning model. A machine learning algorithm used to predicted if a person would default on a loan or not was recently accused of discriminating based off of race. The designers of the network were unable to prove that the network was not discriminating off of race so the model was taken off line. Interpretability gives the user a mechanism to explain how the model is working.

With the widespread application of deep learning for computer vision tasks and the difficulty of collecting a test set of iid in computer vision domains, interpretability is particularly important for the field of computer vision.

It is particularly difficult to get a large representative test set for the visual question

and answering task because the domain of possible image question pairs is so large (cite CLEVR paper). Thus many papers have claimed to improve the state of the art by contributing models which improve interpretability (cite all the models here). However it is difficult to compare the various model without first establishing a metric to empirically evaluate the interpretability of the various models. We recognise that interpretability various from user to user and application to application, so instead of defining a metric we provide a protocol for establishing a metric. We also recognize that interpretability is derivative of human behaviour so the protocol for establishing a metric relies on humans ability to interpret the model.

In this paper we:

1. Extend the framework of Goldstein et. al to apply to deep learning visual question and answering models
2. Demonstrate the functionality of the framework by establishing a metric of interpretability and measuring variations of the Transparency by Design network.
3. Discuss the limitations of the Goldstein et al framework.
4. Propose a theoretical, model agnostic mechanism for generating characteristic input output examples.

2. Related Work

The field of visual question and answer brings together advances in cognitive sciences, natural language processing, and computer vision. Research use various different visual question answering datasets to evaluate their models. Each data set has requires different types of knowledge distillation and has its own biases. The CLEVR dataset is a collection of synthetically generated image question pairs. The images are of 3D geometric shapes in various positions, sizes, and colors. The questions ask logical questions such as “How many cubes are there?” or “Is the object on the left larger than the object on the right?”. The CLEVR dataset was designed to avoid biases in question image pairs. For example other visual question answering datasets may have an image of boat floating in the ocean and a plain flying in the sky with the question what is in the sky. A model which is unable to analyze images at all could still answer this question, because boats don’t fly in the sky.

Because visual question answering involves separate tasks comparing models becomes difficult. Two models with similar test accuracies may have very different advantages. One may be much better at parsing the questions while the other is better at extracting features from images. The CLEVR dataset partially refines the to specific tasks but an accuracy on the CLEVR dataset still hides difference in competing models particularly since models which achieve near perfect accuracy on the CLEVR dataset. To differentiate themselves researchers are adding interpretability mechanisms to their models which help them evaluate specific subtasks of the composite visual question answering task. The question then becomes how well do these mechanisms actually evaluate the subtasks.

Transparency by Design networks are one example of a model that has achieved near perfect accuracy on the CLEVR dataset and has differentiated itself from other models by contributing an interpretability mechanism. To verify claims made by the Transparency by Design network we trained our own Transparency by Design network. Using a GeForce GTX 1080 Ti gpu and 32 gigabytes of memory we were able to achieve 95.11%

accuracy after training for five epochs. We used the higher resolution feature maps and regulated the attention maps with an L1 regularization term of 2.5e-07. We also modified the batch size from 128 images to 16 images.

Transparency by Design networks output a computational graph that demonstrate how the network parsed the natural language question and then for each step of the computational graph they also output an intermediate image that shows what area of the image the model was looking for that computation.

The authors of Transparency by Design (TbD) claim that these attention maps can be used a diagnostic tool. As an informal evaluation of attention maps as a diagnostic tool we asked TbD nets questions about a collection of real world images of 3D geometric shapes. We found that the attention maps allowed us to quickly debug the network. For example the attention map helped us establish that the model mistakes green objects for cyan objects.

This informal evaluation may be sufficient if interpretability is minor consideration, but when interpretability is the main consideration of a networks as it is with Transparency by Design a more formal and reproducible evaluation protocol is required.

3. Experimental Design

Inspired by the formal process of evaluating interpretability proposed by Goldstein et al, we propose a process for evaluating the interpretability of a visual question and answering model.

Because interpretability is inherently a reflection of human behaviour, the method of evaluation should be a reflection of human behavior. The observer must first decided on a set of bugs to use in the process. Once the observer has decided on a set of bugs they generate a set of characteristic input-output pairs half of which show the model making correct predictions on inputs that do not involve the bug and half of which show the model making incorrect predictions on the inputs that do involve the bug.

The observer must then find a population of people they wish to evaluate the model with. An observer may choose to target people from a certain population like regulators or machine learning engineers but these populations may difficult or expensive to reach. Amazon mechanical turk is a large population of people that can be used to test the interpretability of different models. Results collected on Amazon mechanical turk have been shown to be easily reproducible and a large audience can be reached quickly and inexpensively.

Once the observer has decided on a set of bugs to test on, collected characteristic input-output examples and established an population to evaluate on, then they can use our framework to measure the interpretability of their model on their chosen population.

The observer first shows the subject a set of instructions. These instructions tell the subject that will be shown a set of characteristic input - output examples half of which will show the model making a mistake and half of which of show the model making correct predictions. The task of the subject is to infer what type of errors the network makes. The final page of instructions is a short description describing

how the network works or how it use its interpretability mechanisms. These descriptions could have a large influence on the final results so length limits or who is authoring the descriptions should be taken into account and should avoid conflicts of interest.

After the subject has been shown the instructions they will be shown the set of characteristic input - output examples. They will then be shown a set of characteristic inputs and will be asked to predict whether the model will answer correctly or make a mistake. This process is repeated for each bug. Then the subject is asked how confident they were in their predictions and how much they trust the model.

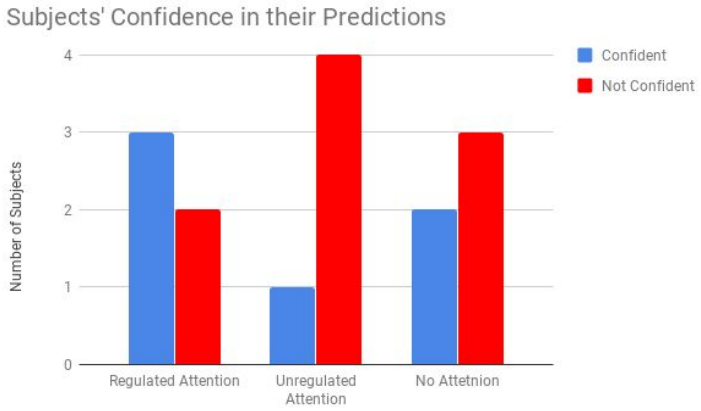
These last questions are designed to measure how much users think they understand how the model works.

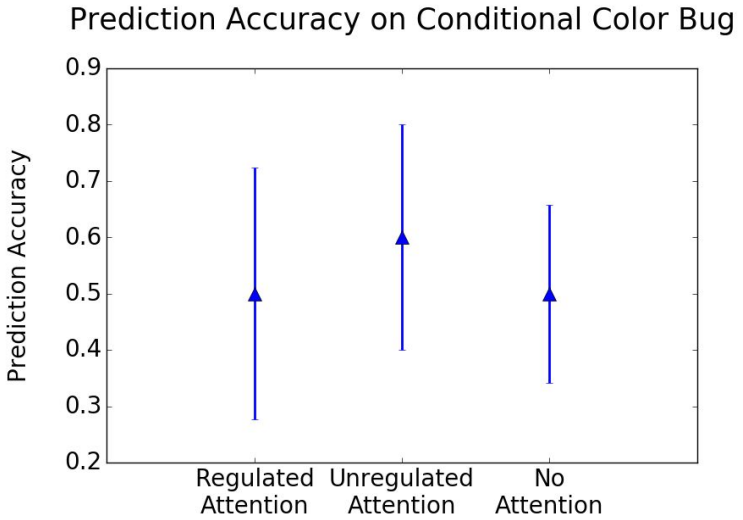
We claim models which users can simulate themselves are more interpretable. We also claim that users trust models that are more interpretable.

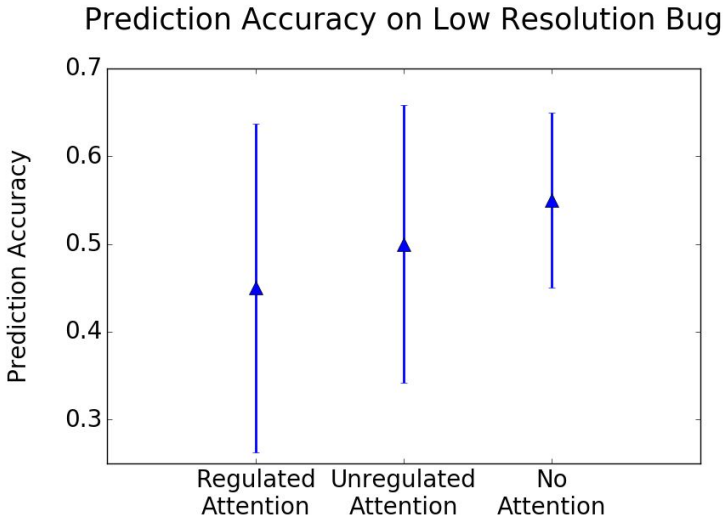
4. Results

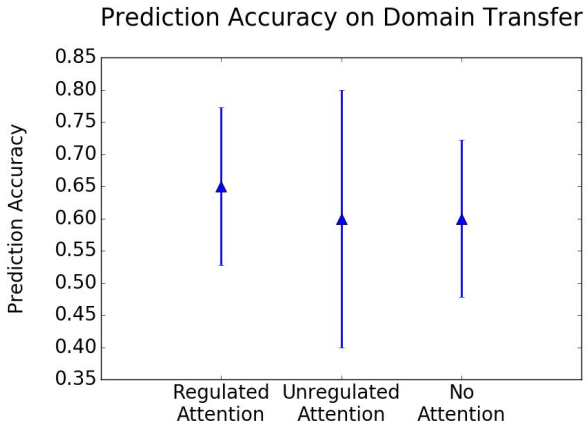
We demonstrate the functionality of the metric by measuring the interpretability of three different versions of Transparency by Design networks. One version shows regulated attention maps, another version shows unregulated attention maps, and the final version shows no attention maps at all.

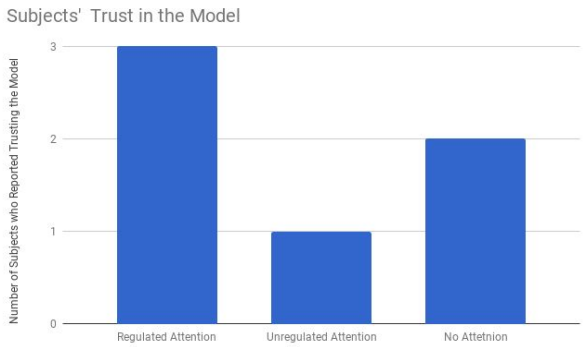
We set up the experiment to be run on Amazon mechanical turk to demonstrate how the experiment could be easily be run on a much larger scale but actually ran the experiment on a convenience sample of 15 UCSD students.











the characteristic examples or to make predictions. The begus we chose could have also been too broad or subtle. The small sample size of our experiment also made it difficult to extract a meaningful signal.

4. Future Work

To collect meaningful results the experiment should be run on a much larger scale with hundreds of participants. The bugs should also be generated by training networks on a training set on which a portion of the training data has been mislabeled.

95% Confidence Interval for Population Mean Accuracy			
	Low Resolution Bug	Conditional Colors Bug	Domain Transfer
Regulated Attention	0.20 - 0.70	0.11 - 0.81	0.48 - 0.82
Unregulated Attention	0.28 - 0.72	0.32 - 0.88	0.32 - 0.88
No Attention	0.42 - 0.68	0.28 - 0.72	0.43 - 0.77

Our t-test showed that regardless of the model users were unable to predict when a model would make a mistake any more than random chance. This could be because we did not give users enough time to observer