Title

Faithful Multimodal Explanation for Visual Question Answering

1st Author

Jialin Wu

Submission Date

8 Sep 2018

Publisher

-

Abstract:

AI systems' ability to explain their reasoning is critical to their utility and trustworthiness. Deep neural networks have enabled significant progress on many challenging problems such as visual question answering (VQA). However, most of them are opaque black boxes with limited explanatory capability. This paper presents a novel approach to developing a high-performing VQA system that can elucidate its answers with integrated textual and visual explanations that faithfully reflect important aspects of its underlying reasoning process while capturing the style of comprehensible human explanations. Extensive experimental evaluation demonstrates the advantages of this approach compared to competing methods using both automated metrics and human evaluation.

My View:

This paper introduces an approach which demonstrates the both textual and visual explanations of a VQA system to generate more faithful multimodal explanations. These explanations are also human comprehensible.

Previous Idea:

- Embed images using a CNN and questions using an RNN and then use these embeddings to train an answer classifier to predict answers from a pre-extracted set
- attention mechanisms to recognize important visual features and filter out irrelevant parts
- highlighting relevant image regions. Grad_CAM

Previous Idea weakness:

Attention mechanism use object detection to recognize important visual features which may cause to decrease the accuracy of the model. Visual explanations highlight key image regions behind the decision; however, they do not explain the reasoning process and crucial relationships between the highlighted regions.

Paper solution:

For visual explanation and recognizing the important parts of image, instead of using object detection, this paper purposes segment detection for more accuracy. First, segment the objects in the image and predict the answer using the VQA module, which has an attention mechanism over those objects. Next, the explanation module is trained to generate textual explanations conditioned on the question, answer, and VQA-attended features.

Result:

On average, the model is able to link 1.6 words in an explanation to an image segment, indicating that the textual explanation is actually grounded in objects detected by VQA system.

I think one way to perform the system is to use a visual vocabulary. We prepare this kind of vocabulary to assist the model while training.