Title

Grounding Visual Explanations

1st Author

Lisa Anne Hendricks

Submission Date

25 Jul 2018

Publisher

-

You can download this paper from **here**

Abstract:

Existing visual explanation generating agents learn to fluently justify a class prediction. However, they may mention visual attributes which reflect a strong class prior, although the evidence may not actually be in the image. This is particularly concerning as ultimately such agents fail in building trust with human users. To overcome this limitation, we propose a phrase-critic model to refine generated candidate explanations augmented with flipped phrases which we use as negative examples while training. At inference time, our phrase-critic model takes an image and a candidate explanation as input and outputs a score indicating how well the candidate explanation is grounded in the image. Our explainable AI agent is capable of providing counter arguments for an alternative prediction, i.e. counterfactuals, along with explanations that justify the correct classification decisions. Our model improves the textual explanation quality of fine-grained classification decisions on the CUB dataset by mentioning phrases that are grounded in the image. Moreover, on the FOIL tasks, our agent detects when there is a mistake in the sentence, grounds the incorrect phrase and corrects it significantly better than other models

My View:

The model which is proposed in this paper takes an image and a candidate explanation as input and outputs a score that mean how many the image and explanation are relevant. Furthermore, it detects the mistake in the sentence and corrects it.

Previous Idea:

- Explanations aid human learning; relevance and context matter.
- **Textual/Visual Explanation**: Trust is vital. Earlier models lacked grounding. Proposed aligns and enhances.
- **Textual/Visual Explanation**: Trust is vital. Earlier models lacked grounding. Proposed aligns and enhances.

Previous Idea weakness:

Previous works had weaknesses in grounding explanations accurately to visual evidence, lacking consistent trust-building and image relevance. They often lacked self-correction mechanisms, relied on class priors, and failed to explore counterfactual explanations or detect inaccuracies. Attention-based models focused on language quality, neglecting visual accuracy. Visual models visualized regions instead of ranking grounded phrases, and evaluation lacked comprehensive human assessment, potentially compromising explanation quality.

Paper solution:

The proposed paper addresses these issues through a phrase-critic model. This model evaluates the alignment of generated explanations with image evidence, using both positive and negative examples during training to enhance image relevance. It ranks and selects the best explanations, ensuring they are both image and class relevant. The model's approach enforces accurate grounding of explanations in images, leading to improved textual explanation quality, counterfactual explanations, error correction, and more accurate detection and correction of inaccuracies in sentences.

Result:

This phrase-critic model has more accuracy compared with other models; for example, in Correct Noun Phrase (CNP) or Correct Sentence (CS) is more precise. Or in grounding accuracy for four common mentioned bird parts:

	% Accuracy				Euclidean Distance			
Explanations	Beak	Head	Belly	Eye	Beak	Head	Belly	Eye
Baseline [8]	93.50	58.74	65.58	55.11	24.16	57.56	56.80	76.90
Grounding model [10]	94.30	60.60	65.40	**60.78**	**22.66**	46.31	**52.69**	**57.55**
Phrase Critic	**95.88**	**74.06**	**66.65**	56.72	23.74	**20.26**	**52.75**	69.83

One solution to address the challenge of ensuring explanations are both linguistically fluent and accurately grounded in visual evidence is to incorporate a reinforcement learning approach during training. By using reinforcement learning, the model could receive feedback on the quality of its explanations.