Title

Generating Visual Explanations

1st Author

Lisa Anne Hendricks - Marcus Rohrbach

Submission Date

28 Mar 2016

Publisher

-

Abstract:

Clearly explaining a rationale for a classification decision to an end-user can be as important as the decision itself. Existing approaches for deep visual recognition are generally opaque and do not output any justification text; contemporary vision-language models can describe image content but fail to take into account class-discriminative image aspects which justify visual predictions. We propose a new model that focuses on the discriminating properties of the visible object, jointly predicts a class label, and explains why the predicted label is appropriate for the image. We propose a novel loss function based on sampling and reinforcement learning that learns to generate sentences that realize a global sentence property, such as class specificity. Our results on a fine-grained bird species classification dataset show that our model is able to generate explanations which are not only consistent with an image but also more discriminative than descriptions produced by existing captioning methods

My View:

The paper introduces a model that not only classifies images but also generates explanations for its decisions. It does this by combining image recognition and text generation, creating explanations that justify the model's classifications. The model uses a unique loss function based on sampling and reinforcement learning to produce sentences explaining the reasons behind its predictions. This helps make deep visual models more transparent and understandable. Experiments on a bird species dataset show that the model's explanations are accurate and informative, enhancing the interpretability of deep visual models.

Previous Idea:

- Past AI systems explained using rules or templates, for medical and robotics tasks.
- Some explained decision mechanisms, others justified predictions to non-experts.
- Visual explanation methods focused on image features, not connecting them with language.
- Early image descriptions detected concepts before generating sentences.
- Recent deep models directly generated accurate image descriptions.
- Our model learned explanations without intermediate guidance.
- We introduced a novel loss function for explanation sentence generation.
- Fine-grained classification explanations used natural language descriptions.
- Reinforcement learning applied in vision for tasks like question answering and activity detection, sometimes using sampling.

Previous Idea weakness:

Previous approaches often require expert knowledge, pre-defined structures, and might be complex for non-experts. Early methods lack integration of visual features and language. Attributes are expensive to annotate, lack generalizability, and lack natural language explanations. Traditional LSTM models with cross-entropy loss may not optimize well for desired properties. These weaknesses emphasize the need for more adaptable, user-friendly, and effective explanation methods.

Paper solution:

The paper's solution involves a new model that generates explanations for image classifications. This model learns from data, avoiding the need for expert input. It introduces a unique loss function for sentence generation, improving upon standard methods. The model combines image recognition and text generation, offering coherent justifications for predictions. Using natural language descriptions and optimizing for class specificity, it produces accurate and informative explanations, enhancing interpretability of deep visual models.

Result:

The paper's results confirm that their model generates visual explanations that meet their definition criteria of being image-relevant and class-relevant. Training the model to create class-specific descriptions leads to better sentence quality according to standard metrics for sentence generation.

	Image Relevance		**Class Relevance**		**Best Explanation**
	METEOR	CIDEr	Similarity	Rank (1-200)	Bird Expert Rank (1-5)
Definition	27.9	43.8	42.60	15.82	2.92
Description	27.7	42.0	35.3	24.43	3.11
Explanation-Label	28.1	44.7	40.86	17.69	2.97
Explanation-Dis.	28.8	51.9	43.61	19.80	3.22
Explanation	**29.2**	**56.7**	**52.25**	**13.12**	**2.78**