Generating Visual Explanations

Lisa et al.

이종진

Seoul National University

ga0408@snu.ac.kr

Nov 15, 2018

Explainable AI; Generating Visual Explanations

- Deep classification methods have had tremendous success in visual reconition.
- Most of them cannot provide a consistent justification of why it made a certain prediction.

Explainable AI; Generating Visual Explanations

- Proposed model predicts a class label(CNN), and explains why the predicted label is appropriate for the image(RNN)
- First method to produce deep visual explanations using language justifications
- Provide an explanation not a description

Visual Explanation

Western Grebe

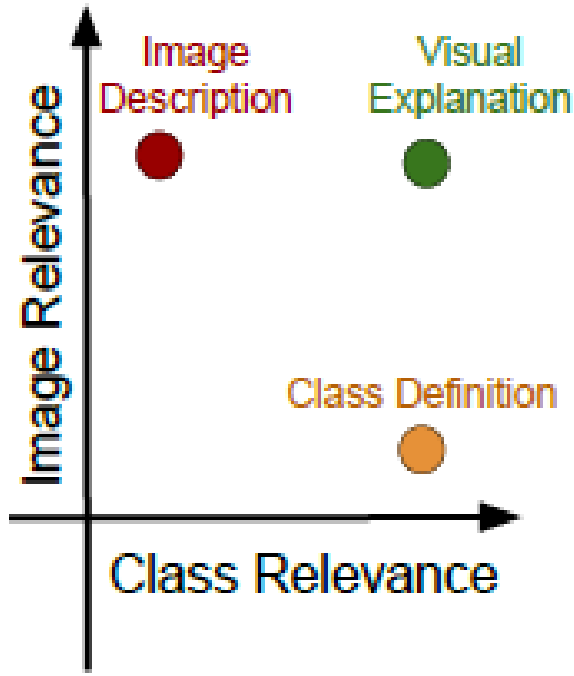
Description: This is a large bird with a white neck and a black back in the water

Class Definition: The Western Grebe is a waterbird with a yellow pointly beak, white neck and belly, and black back.

Explanation: This is a Western Grebe because this bird has a long white neck, pointly yellow beak and red eye.

- Explanation should be class discriminative!!

Visual Explanation



- Visual explanation are both image relevant and class relevant.
- Discriminate class and accurately describe a specific image instance.

  $\rightarrow$ Novel Loss function.

Proposed Model

- Input : Image (+ Descriptive Sentences)
- Output : This is a **CLASS**, because **argument 1** and **argument 2** and...
- Use pretrained CNN(Compact bilinear fine- grained classificaiton model), Sentence classifier(Single Layer LSTM)
- Two contributions are using a predicted label as a input and using novel loss(discrimiative loss) for image relevance and class relevance
  1. Use a predicted label as a input
  2. Propose a novel reinforcement learing based loss for image relevance and class relevance

Architecture

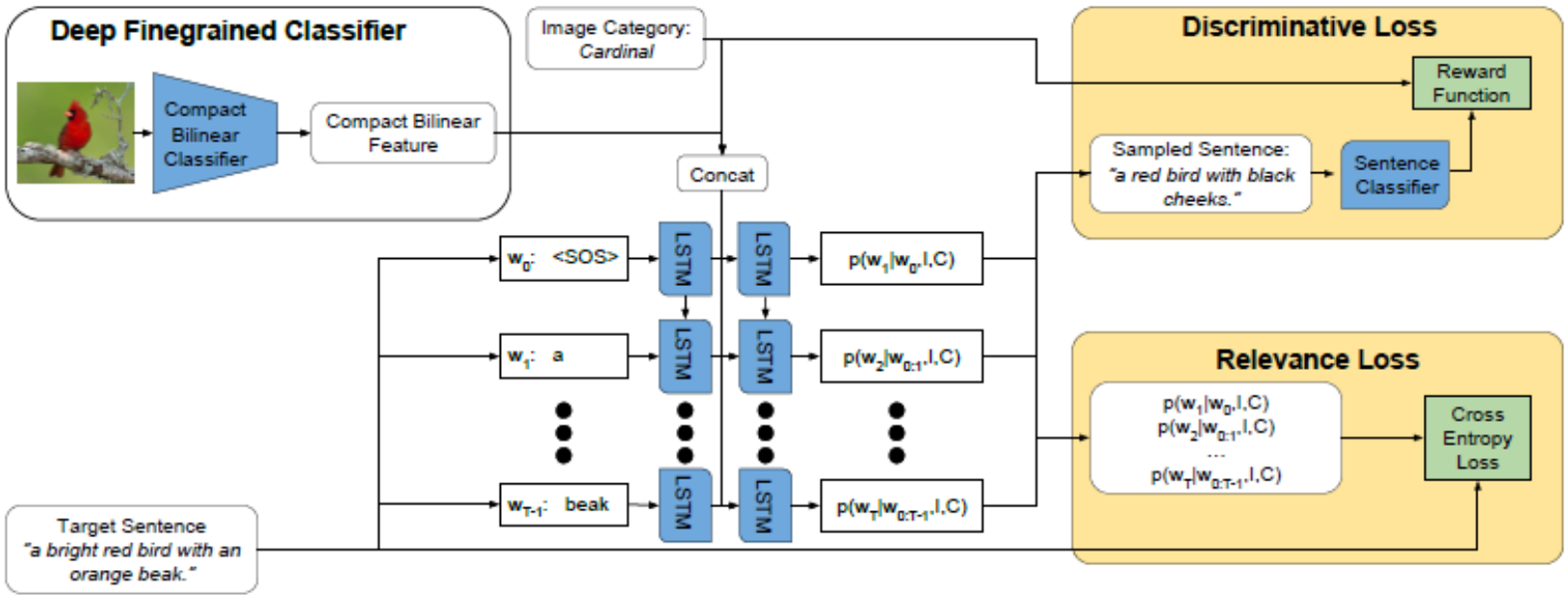
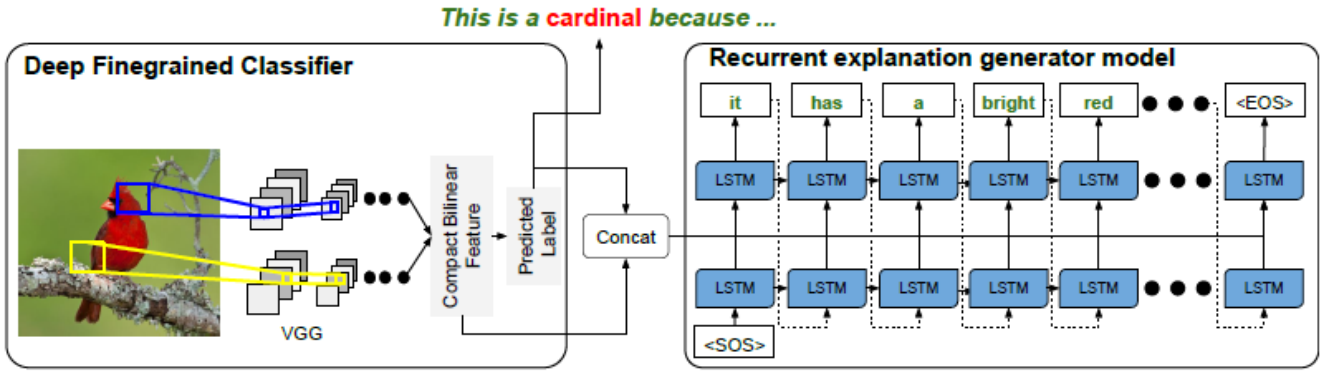


Figure: Architecture

Bilinear Models



- $f : L \times \mathrm{I} \mapsto \mathbb{R}^{c \times D}$, a location L and image I
- f_A, f_B : use pretrained VGG
- Use pooling operation $P(f_A(l, \mathrm{I})^T f_B(l, \mathrm{I}), l \in L)$
- (e.g) $\phi(\mathrm{I}) = \sum_{l \in L} f_A(l, \mathrm{I})^T f_B(l, \mathrm{I})$

Proposed loss

- Proposed loss

$$L_R - \lambda \mathbb{E}_{\tilde{w} \sim p_L(w)}[R_D(\tilde{w})]$$

- Relevance loss(L_R) is related with "Image Relevance"
- Discriminiative loss($\mathbb{E}_{\tilde{w} \sim p_L(w)}[R_D(\tilde{w})]$) is related with "Class Relevance"

Relevance Loss

- Relevance Loss(L_R)

$$L_R = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{t=0}^{T-1} \log p_L(w_{t+1} | w_{o:t}, \mathrm{I}, \mathrm{C})$$

  - w_t : ground truth word at t, I : image, C : category, N : batch size
  - Average hidden state of the LSTM

- Discriminative Loss

$$\mathbb{E}_{\tilde{w} \sim p_L(w)}[R_D(\tilde{w})]$$

  - Based on a reinforcement learning paradigm.
  - $R_D(\tilde{w}) = p_D(\mathrm{C}|\tilde{w})$
  - $p_D(\mathrm{C}|w)$: pretrained sentence classifier
  - The accuracy of this classifier(pretrained) is not important (22%)
  - $\tilde{w}$: sampled sentences from LSTM($p_L(w)$)

Novel Loss

- Relevance Loss

$$L_R = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{t=0}^{T-1} \log p_L(w_{t+1} | w_{o:t}, \mathrm{I}, \mathrm{C})$$

- Discriminative Loss

$$R_D(\tilde{w}) = p_D(\mathrm{C} | \tilde{w})$$

  - The accuracy of this classifier(pretraine) is not important (22%)

- Proposed Loss

$$L_R - \lambda \mathbb{E}_{\tilde{w} \sim p_L(w)}[R_D(\tilde{w})]$$

Minimizing Loss

- Since expectation over descriptions is intractable, use Monte Carlo sampling from LSTM.
- $\nabla \mathbb{E}_{\tilde{w} \sim p_L(w)}[R_D(\tilde{w})] = \mathbb{E}_{\tilde{w} \sim p_L(w)}[R_D(\tilde{w}) \nabla_{W_L} \log P(\tilde{w})]$
- The final gradient to update the weights W

$$\nabla_{W_L} L_R - \lambda R_D(\tilde{w}) \nabla_{W_L} \log P(\tilde{w})$$

Experiment

- Dataset : Caltech UCSD Birds 200-2011(CUB)
  - Contains 200 classes of North American bird species.
  - 11,788 images
  - 5 sentences for detail description of the bird(These are not collected for the task of visual explanation.)
- 8,192 dimensional features from the classifier
  - Features from the penultimate layer of the compact bilinear fine-grained classification model
  - Pre-trained on the CUB dataset
  - accuracy : 84%
- LSTM
  - 1000-dimensional embedding, 1000 dimensional LSTM

Experiment

- Baseline models : Description model & Definition model
  - Description model : Training the model by conditioning only on the image features as input
  - Definition model : Training the model to generate explaining sentences only using the image label as input
- Abalation models : Explation-label model & Explanation-discriminative model

Measure

- METEOR(Image relevance)
  - METEOR is computed by matching words(synonyms) in generated and reference sentences
- CIDEr(Image relevance)
  - CIDEr measures the similarity of a generated sentence to reference sentence by counting common n-grams which are TF-IDF weighted.
- Similarity(class relevance)
  - Compute CIDEr scores using all reference sentences which correspond to a particular class, instead of using ground truth
- Rank(class relevance)
  - Ranking over similarity of all classes

Experiment : Results

	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**

Figure: Result

Experiment : Results

- Comparison of Explanations, Baselines, and Ablations.

This is a **Bronzed Cowbird** *because ...*

Definition:	this bird is **black** with **blue** on its wings and has a long **pointy beak**.
Description:	this bird is **nearly all black** with a short **pointy bill**.
Explanation-Label:	this bird is **nearly all black** with **bright orange eyes**.
Explanation-Dis.:	this is a **black bird** with a **red eye** and a **white beak**.
Explanation:	this is a **black bird** with a **red eye** and a **pointy black beak**.

  - Green : correct, Yellow : mostly correct, Red : incorrect
  - 'Red eye' is a class relevant attributes.

Experiment : Results

- Comparison of Explanations and Definitions

*This is a **Downy Woodpecker** because...*

Definition: this bird has a white breast black wings and a red spot on its head.

Explanation: this is a black and white bird with a **red spot** on its crown.

  - Definition can produce sentencesd which are not image relevant

Experiment : Results

- Role of Discriminative Loss

*This is a **Black-Capped Vireo** because...*

Description: this bird has a white belly and breast black and white wings with a white wingbar.

Explanation-Dis: this is a bird with a white belly yellow wing and a **black head**.

  - Both models generate visually correct sentences.
  - 'Black head' is one of the most prominent distinguishing properties of this vireo type.