Generating Visual Explanations

Lisa Anne Hendricks[1(✉)], Zeynep Akata[2], Marcus Rohrbach[1,3], Jeff Donahue[1], Bernt Schiele[2], and Trevor Darrell[1]

[1] UC Berkeley EECS, Berkeley, CA, USA
{lisa_anne,rohrbach,jdonahue,trevor}@eecs.berkeley.edu
[2] Max Planck Institute for Informatics, Saarbrücken, Germany
{akata,schiele}@mpi-inf.mpg.de
[3] ICSI, Berkeley, CA, USA

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. Through a novel loss function based on sampling and reinforcement learning, our model learns to generate sentences that realize a global sentence property, such as class specificity. Our results on the CUB 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.

Keywords: Visual explanation · Image description · Language and vision

1 Introduction

Explaining why the output of a visual system is compatible with visual evidence is a key component for understanding and interacting with AI systems [4]. Deep classification methods have had tremendous success in visual recognition [8,10,20], but their outputs can be unsatisfactory if the model cannot provide a consistent justification of why it made a certain prediction. In contrast, systems which can justify why a prediction is consistent with visual elements to a user are more likely to be trusted [34]. Explanations of visual systems could also aid in understanding network mistakes and provide feedback to improve classifers.

We consider explanations as determining *why* a decision is consistent with visual evidence, and differentiate between *introspection* explanation systems

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46493-0_1) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part IV, LNCS 9908, pp. 3–19, 2016.
DOI: 10.1007/978-3-319-46493-0_1

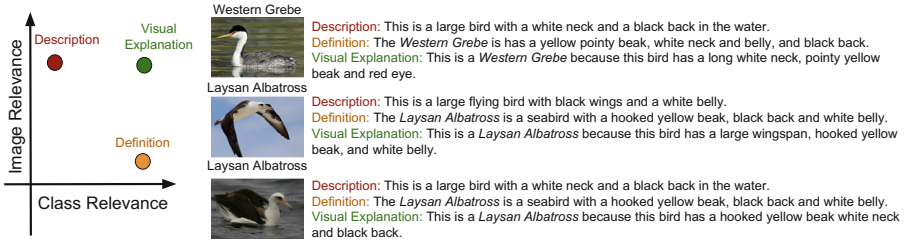

Fig. 1. Our proposed model generates *explanations* that are both image relevant and class relevant. In contrast, *descriptions* are image relevant, but not necessarily class relevant, and *definitions* are class relevant but not necessarily image relevant. (Color figure online)

which explain how a model determines its final output (e.g., "This is a Western Grebe because filter 2 has a high activation...") and *justification* explanation systems which produce sentences detailing how visual evidence is compatible with a system output (e.g., "This is a Western Grebe because it has red eyes..."). We concentrate on justification explanation systems because they may be more useful to non-experts who do not have knowledge of modern computer vision systems [4].

We argue that visual explanations must satisfy two criteria: they must be *class discriminative* and *accurately describe* a specific image instance. As shown in Fig. 1, explanations are distinct from *descriptions*, which provide a sentence based only on visual information, and *definitions*, which provide a sentence based only on class information. Unlike descriptions and definitions, visual explanations detail why a certain category is appropriate for a given image while only mentioning image relevant features. For example, consider a classification system that predicts a certain image belongs to the class "western grebe" (Fig. 1, top). A standard captioning system might provide a description such as "This is a large bird with a white neck and black back in the water." However, as this description does not mention *discriminative* features, it could also be applied to a "laysan albatross" (Fig. 1, bottom). In contrast, we propose to provide *explanations*, such as "This is a western grebe because this bird has a long white neck, pointy yellow beak, and a red eye." The explanation includes the "red eye" property, which is important for distinguishing between "western grebe" and "laysan albatross". As such, our system explains *why* the predicted category is the most appropriate for the image.

We outline our approach in Fig. 2. In contrast to description models, we condition generation on an image and the predicted class label. We also use features extracted from a fine-grained recognition pipeline [10]. Like many contemporary description models [7,18,19,37,40], we use an LSTM [13] to generate word sequences. However, we design a novel loss function which encourages generated sentences to include class discriminative information; i.e., to be class specific. One challenge is that class specificity is a global sentence property: e.g., while

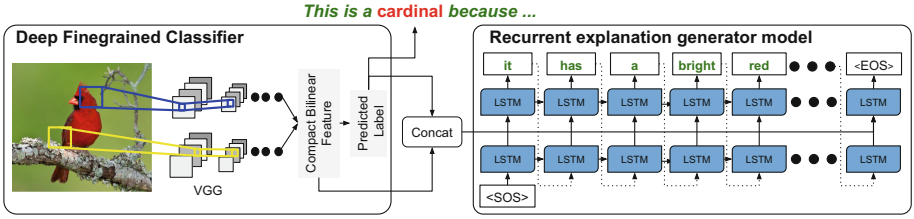

Fig. 2. Our joint classification and explanation model. We extract visual features using a fine-grained classifier before sentence generation and, unlike other sentence generation models, condition sentence generation on the predicted class label. A novel discriminative loss encourages generated sentences to include class specific attributes. (Color figure online)

a sentence "This is an all black bird with a bright red eye" is class specific to a "Bronzed Cowbird", words and phrases in this sentence, such as "black" or "red eye" are less class specific on their own. Our final output is a sampled sentence, so we backpropagate the discriminative loss through the sentence sampling mechanism via a technique from the reinforcement learning literature [39].

To the best of our knowledge, ours is the first framework to produce deep visual explanations using natural language justifications. We describe below that our novel joint vision and language explanation model combines classification and sentence generation by incorporating a loss function that operates over sampled sentences. We show that this formulation is able to focus generated text to be more discriminative and that our model produces better explanations than a description baseline. Our results also confirm that generated sentence quality improves with respect to traditional sentence generation metrics by including a discriminative class label loss during training. This result holds even when class conditioning is ablated at test time.

2 Related Work

Explanation. Automatic reasoning and explanation has a long and rich history within the artificial intelligence community [4,5,17,22,24,25,33,35]. Explanation systems span a variety of applications including explaining medical diagnosis [33], simulator actions [5,17,24,35], and robot movements [25]. Many of these systems are rule-based [33] or solely reliant on filling in a predetermined template [35]. Methods such as [33] require expert-level explanations and decision processes. As expert explanations or decision processes are not available during training, our model learns purely from visual features and fine-grained visual descriptions to fulfill our two proposed visual explanation criteria. In contrast to systems like [5,22,24,25,33,35] which aim to explain the underlying mechanism behind a decision, Biran et al. [4] concentrate on why a prediction is justifiable to a user. Such systems are advantageous because they do not rely on user familiarity with the design of an intelligent system in order to provide useful information.

Many vision methods focus on discovering visual features which can help "explain" an image classification decision [3,6,16]. Importantly, these models do not link discovered discriminative features to natural language expressions. We believe that the methods discovering discriminative visual features are complementary to our proposed system. In fact, discriminative visual features could be used as additional inputs to our model to produce better explanations.

Visual Description. Early image description methods rely on detecting visual concepts (e.g., subject, verb, and object) before generating a sentence with either a simple language model or sentence template [11,21]. Recent deep models [7,9,18,19,28,37,40] outperform such systems and produce fluent, accurate descriptions. Though most description models condition sentence generation only on image features, [14] condition generation on auxiliary information, such as words used to describe a similar image in the train set. However, [14] does not condition sentence generation on category labels.

LSTM sentence generation models are generally trained with a cross-entropy loss between the probability distribution of predicted and ground truth words [7,18,28,37,40]. Frequently, however, the cross-entropy loss does not directly optimize for properties desirable at test time. [26] proposes a training scheme for generating unambiguous region descriptions which maximizes the probability of a region description while minimizing the probability of other region descriptions. In this work, we propose a novel loss function for sentence generation which allows us to specify a global constraint on generated sentences.

Fine-Grained Classification. Object classification, particularly fine-grained classification, is an attractive setting for explanation systems because describing image content does not suffice as an explanation. Explanation models must focus on aspects that are both class-specific and depicted in the image.

Most fine-grained zero-shot and few-shot image classification systems use attributes [23] as auxiliary information. Attributes discretize a high dimensional feature space into simple and readily interpretable decision statements that can act as an explanation. However, attributes have several disadvantages. They require experts for annotation which is costly and results in attributes which are hard for non-experts to interpret (e.g., "spatulate bill shape"). Attributes are not scalable as the list of attributes needs to be revised to ensure discriminativeness for new classes. Finally, attributes do not provide a natural language explanation like the user expects. We therefore use natural language descriptions [31] which achieved superior performance on zero-shot learning compared to attributes and also shown to be useful for text to image generation [32].

Reinforcement Learning in Computer Vision. Vision models which incorporate algorithms from reinforcement learning, specifically how to backpropagate through a sampling mechanism, have recently been applied to visual question answering [1] and activity detection [41]. Additionally, [40] use a sampling mechanism to attend to specific image regions for caption generation, but use the standard cross-entropy loss during training.

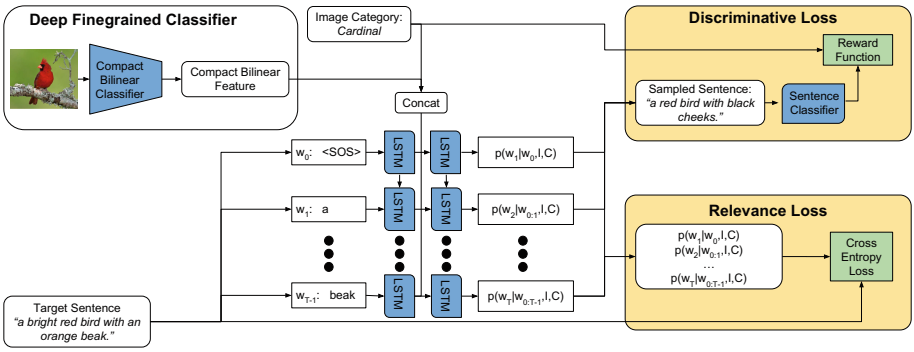

Fig. 3. Training our explanation model. Our explanation model differs from other caption models because it (1) includes the object category as an additional input and (2) incorporates a reinforcement learning based discriminative loss

3 Visual Explanation Model

Our visual explanation model (Fig. 3) aims to produce an explanation which describes visual content present in a specific image instance while containing appropriate information to explain why the image belongs to a specific category. We ensure generated descriptions meet these two requirements for explanation by including both a *relevance loss* (Fig. 3, bottom right) and *discriminative loss* (Fig. 3, top right). We propose a novel discriminative loss which acts on sampled word sequences during training. Our loss enables us to enforce global sentence constraints on sentences. By applying our loss to sampled sentences, we ensure that the final output of our system fulfills our explanation criteria. We consider a sentence to be either a complete sentence or a sentence fragment.

3.1 Relevance Loss

Image relevance can be accomplished by training a visual description model. Our model is based on LRCN [7], which consists of a convolutional network, which extracts high level visual features, and two stacked recurrent networks (specifically LSTMs), which generate descriptions conditioned on visual features. During inference, the first LSTM receives the previously generated word w_{t-1} as input and produces an output l_t. The second LSTM, receives the output of the first LSTM l_t and an image feature f and produces a probability distribution $p(w_t)$ over the next word. The word w_t is generated by sampling from the distribution $p(w_t)$. Generation continues until an "end-of-sentence" token is generated.

We propose two modifications to the LRCN framework to increase the image relevance of generated sequences (Fig. 3, top left). First, category predictions are used as an additional input to the second LSTM in the sentence generation model. Intuitively, category information can help inform the caption generation model which words and attributes are more likely to occur in a description.

For example, category level information can help the model decide if a red eye or red eyebrow is more likely for a given class. We experimented with a few methods to represent class labels, and found that training a language model, e.g., an LSTM, to generate word sequences conditioned on images, then using the average hidden state of the LSTM across all sequences for all classes in the train set as a vectorial representation of a class works best. Second, we use rich category specific features [10] to generate relevant explanations.

Each training instance consists of an image, category label, and a ground truth sentence. During training, the model receives the ground truth word w_t for each time step $t \in T$. We define the relevance loss for a specific image (I) and caption (C) as:

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

where w_t is a ground truth word and N is the batch size. By training the model to predict each word in a ground truth sentence, the model produces sentences which reflect the image content. However, this loss does not explicitly encourage generated sentences to discuss discerning visual properties. In order to generate sentences which are both image relevant and category specific, we include a discriminative loss to focus sentence generation on discriminative visual properties of the object.

3.2 Discriminative Loss

Our discriminative loss is based on a reinforcement learning paradigm for learning with layers which require sampling intermediate activations of a network. In our formulation, we first sample a sentence and then use the sampled sentence to compute a discriminative loss. By sampling the sentence before computing the loss, we ensure that sentences sampled from our model are more likely to be class specific. Our reinforcement based loss enables us to backpropagate through the sentence sampling mechanism.

We minimize the following overall loss function with respect to the explanation network weights W:

$$L_R(I, C) - \lambda \mathbb{E}_{\tilde{w} \sim p(w|I,C)} \left[R_D(\tilde{w}) \right] \tag{2}$$

which is a linear combination of the relevance loss L_R and the expectation of the negative discriminator reward $-R_D(\tilde{w})$ over descriptions $\tilde{w} \sim p(w|I, C)$, where $p(w|I, C)$ is the model's estimated conditional distribution over descriptions w given the image I and category C. Since $\mathbb{E}_{\tilde{w} \sim p(w|I,C)} \left[R_D(\tilde{w}) \right]$ is intractable, we estimate it at training time using Monte Carlo sampling of descriptions from the categorical distribution given by the model's softmax output at each timestep. The sampling operation for the categorical distribution is non-smooth in the distribution's parameters $\{p_i\}$ as it is a discrete distribution. Therefore, $\nabla_W R_D(\tilde{w})$ for a given sample $\tilde{w}$ with respect to the weights W is undefined.

Following the REINFORCE [39] algorithm, we make use of the following equivalence property of the expected reward gradient:

$$\nabla_W \mathbb{E}_{\tilde{w}\sim p(w|I,C)}\left[R_D(\tilde{w})\right] = \mathbb{E}_{\tilde{w}\sim p(w|I,C)}\left[R_D(\tilde{w})\nabla_W \log p(\tilde{w})\right] \tag{3}$$

In this reformulation, the gradient $\nabla_W \log p(\tilde{w})$ is well-defined: $\log p(\tilde{w})$ is the log-likelihood of the sampled description $\tilde{w}$, just as L_R is the log-likelihood of the ground truth description. However, the sampled gradient term is weighted by the reward $R_D(\tilde{w})$, pushing the weights to increase the likelihood assigned to the most highly rewarded (and hence most discriminative) descriptions. Therefore, the final gradient we compute to update the weights W, given a description $\tilde{w}$ sampled from the model's softmax distribution, is:

$$\nabla_W L_R - \lambda R_D(\tilde{w})\nabla_W \log p(\tilde{w}). \tag{4}$$

$R_D(\tilde{w})$ should be high when sampled sentences are discriminative. We define our reward simply as $R_D(\tilde{w}) = p(C|\tilde{w})$, or the probability of the ground truth category C given only the generated sentence $\tilde{w}$. By placing the discriminative loss after the sampled sentence, the sentence acts as an information bottleneck. For the model to produce an output with a large reward, the generated sentence must include enough information to classify the original image properly.

For the sentence classifier, we train a single layer LSTM-based classification network to classify ground truth sentences. Our sentence classifier correctly predicts the class of unseen validation set sentences 22 % of the time. This number is possibly low because descriptions in the dataset do not necessarily contain discriminative properties (e.g., "This is a white bird with grey wings." is a valid description but can apply to multiple bird species). Nonetheless, we find that this classifier provides enough information to train our explanation model. Outside text sources (e.g., field guides) could be useful when training a sentence classifier. However, incorporating outside text can be challenging as this requires aligning our image annotation vocabulary to field-guide vocabulary. When training the explanation model, we do not update weights in the sentences classifier.

4 Experimental Setup

Dataset. We employ the Caltech UCSD Birds 200–2011 (CUB) dataset [38] which contains 200 classes of bird species and 11,788 images in total. Recently, [31] collected 5 sentences for each of the images which do not only describe the content of the image, e.g., "This is a bird", but also give a detailed description of the bird, e.g., "red feathers and has a black face patch". Unlike other image-sentence datasets, every image in the CUB dataset belongs to a class, and therefore sentences as well as images are associated with a single label. This property makes this dataset unique for the visual explanation task, where our aim is to generate sentences that are both discriminative and class-specific.

Though sentences collected in [31] were not originally collected for the visual explanation task, we observe that sentences include detailed and fine-grained category specific information. When ranking human annotations by output scores

of our sentence classifier, we find that high-ranking sentences (and thus more discriminative sentences) include rich discriminative details. For example, the sentence "...mostly black all over its body with a small red and yellow portion in its wing" has a score of 0.99 for "Red winged blackbird" and includes details specific to this bird variety, such as "red and yellow portion in its wing". As ground truth annotations are descriptions as opposed to explanations, not all annotations are guaranteed to include discriminative information. For example, though the "bronzed-cowbird" has striking red eyes, not all humans mention this discrimitave feature. To generate satisfactory explanations, our model must learn which features are discriminative from descriptions and incorporate discriminative properties into generated explanations. Example ground truth images and annotations may be found in our supplemental.

Implementation. For image features, we extract 8,192 dimensional features from the penultimate layer of the compact bilinear fine-grained classification model [10] which has been pre-trained on the CUB dataset and achieves an accuracy of 84 %. We use one-hot vectors to represent input words at each time step and learn a 1000 dimensional embedding before inputting each word into an LSTM with 1000 hidden units. We train our models using *Caffe* [15], and determine model hyperparameters using the standard CUB validation set before evaluating on the test set. All reported results are on the standard CUB test set.

Baseline and Ablation Models. We propose two baseline models: a *description* model and a *definition* model. Our description baseline generates sentences conditioned only on images and is equivalent to LRCN [7] except we use image features from a fine-grained classifier [10]. Our definition baseline generates sentences using only an image label as input. Consequently, this model outputs the same sentence for every image of the same class. Our proposed model is both more image and class relevant than either of these baselines and thus superior for the explanation task.

Our explanation model differs from description models in two key ways. First, in addition to an image, generated sentences are conditioned on class predictions. Second, explanations are trained with a discriminative loss which enforces that generated sentences contain class specific information (see Eq. 2). To demonstrate that both class information and the discriminative loss are important, we compare our explanation model to an *explanation-label* model which is not trained with the discriminative loss, and to an *explanation-discriminative* model which is not conditioned on the predicted class.

Metrics. To evaluate our explanation model, we use automatic metrics and two human evaluations. Our automatic metrics rely on the common sentence evaluation metrics (METEOR [2] and CIDEr [36]) and are used to evaluate the quality of our explanatory text. METEOR is computed by matching words in generated and reference sentences, but unlike other common metrics such as BLEU [30], it uses WordNet [29] to also match synonyms. CIDEr measures the similarity of a generated sentence to reference sentence by counting common n-grams which are TF-IDF weighted. Consequently, CIDEr rewards sentences for correctly including n-grams which are uncommon in the dataset.

A generated sentence is *image relevant* if it mentions concepts which are mentioned in ground truth reference sentences for the image. Thus, to measure image relevance we simply report METEOR and CIDEr scores, with more relevant sentences producing higher METEOR and CIDEr scores.

Measuring *class relevance* is considerably more difficult. We could use the LSTM sentence classifier used to train our discriminative loss, but this is an unfair metric because some models were trained to directly increase the accuracy as measured by the LSTM classifier. Instead, we measure class relevance by considering how similar generated sentences for a class are to ground truth sentences for that class. Sentences which describe a certain bird class, e.g., "cardinal", should contain similar words and phrases to ground truth "cardinal" sentences, but not ground truth "black bird" sentences. We compute CIDEr scores for images from each bird class, but instead of using ground truth image descriptions as reference sentences, we pool all reference sentences which correspond to a particular class. We call this metric the *class similarity* metric.

Though class relevant sentences should have high class similarity scores, a model could achieve a better class similarity score by producing better overall sentences (e.g., better grammar) without producing more class relevant descriptions. To further demonstrate that our sentences are class relevant, we compute a *class rank* metric. Intuitively, class similarity scores computed for generated sentences about *cardinals* should be higher when compared to *cardinal* reference sentences than when compared to reference sentences from other classes. Consequently, more class relevant models should yield higher rank for ground truth classes. To compute class rank, we compute the class similarity for each generated sentence with respect to each bird category and rank bird categories by class similarity. We report the mean rank of the ground truth class. We emphasize the CIDEr metric because of the TF-IDF weighting over n-grams. If a bird has a unique feature, such as "red eyes", generated sentences which mention this attribute should be rewarded more than sentences which just mention attributes common across all bird classes. We apply our metrics to instances in which the compact bilinear classifier predicts the correct label as is unclear if the best explanatory text should be more similar to correct or predicted classes. However, the same trends hold if we apply our metrics to all generated sentences.

5 Results

We demonstrate that our model generates superior visual explanations and produces image and class relevant text. Additionally, generating visual explanations results in higher quality sentences based on common sentence generation metrics.

5.1 Quantitative Results

Image Relevance. Table 1, columns 2 & 3, record METEOR and CIDEr scores for our generated sentences. Importantly, our explanation model has higher

Table 1. Comparing our explanation model to our definition and description baseline, as well as the explanation-label and explanation-discriminative (explanation-dis.) ablation models. Our explanations are image relevant, as measured by METEOR and CIDEr scores (higher is better). They are also class relevant, as measured by class similarity metric (higher is better) and class rank metric (lower is better) (see Sect. 4 for details). Finally, our explanations are ranked better by experienced bird watchers.

	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	35.30	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**

METEOR and CIDEr scores than our baselines. The explanation model also outperforms the explanation-label and explanation-discriminative model suggesting that both label conditioning and the discriminative loss are key to producing better sentences. Furthermore, METEOR and CIDEr are substantially higher when including a discriminative loss during training (compare rows 2 and 4 and rows 3 and 5) demonstrating that including this additional loss leads to better generated sentences. Moreover, the definition model produces more image relevant sentences than the description model suggesting that category information is important for fine-grained description. On the other hand, our explanation-label results are better than both the definition and description results showing that the image and label contain complementary information.

Class Relevance. Table 1, columns 4 & 5, report class similarity and class rank metrics (see Sect. 4 for details). Our explanation model produces a higher class similarity score than other models by a substantial margin. The class rank for our explanation model is also lower than for any other model suggesting that sentences generated by our explanation model more closely resemble the correct class than other classes in the dataset. Our ranking metric is quite difficult; sentences must include enough information to differentiate between very similar bird classes without looking at an image, and our results clearly show that our explanation model performs best at this difficult task. The accuracy of our LSTM sentence classifier follow the same general trend, with our explanation model achieves 59.13 % whereas the description model obtains 22.32 % accuracy.

Based on the success of our discriminator, we train a definition model with the discriminative loss and find that our loss does boost performance of the definition model (METEOR: 28.6, CIDEr: 51.7, Similarity: 48.8, Rank: 15.5). Importantly, the explanation still performs best on our evaluation metrics.

User Studies. The ultimate goal of our explanation system is to provide useful information about an unknown object to a user. We therefore also consulted experienced bird watchers to rate our explanations against our baseline and

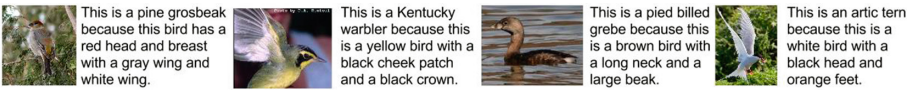

Fig. 4. Visual explanations generated by our system. Our explanation model produces image relevant sentences that also discuss class discriminative attributes.

ablation models. Consulting experienced bird watchers is important because some sentences may provide correct, but non-discriminative properties, which an average person may not be able to properly identify. For example, *This is a Bronzed Cowbird because this bird is nearly all black with a short pointy bill.* is correct, but is a poor explanation as it does not mention unique attributes of a *Bronzed Cowbird* such as *red eye*. Two experienced bird watchers evaluated 91 randomly selected images and answered which sentence provided the best explanation for the bird class (Table 1, column 6). Our explanation model has the best mean rank (lower is better), followed by the definition model. This trend resembles the trend seen when evaluating class relevance.

We also demonstrate that explanations are more effective than descriptions at helping humans identify different bird species. We ask five Amazon Turk workers to choose between two images given a generated description and explanation. We evaluate 200 images (one for each bird category) and find that our explanations are more helpful to humans. When provided with an explanation, the correct image is chosen (with an image considered to be chosen correctly if 4 out of 5 workers select the correct image) 56 % of the time, whereas when provided with a description, the correct image is chosen less frequently (52 % of the time).

5.2 Qualitative Results

Figure 4 shows sample explanations which first declare a predicted class label ("This is a *Kentucky Warbler* because") followed by the explanatory text produced by the model described in Sect. 3. Qualitatively, our explanation model performs quite well. Note that our model accurately describes fine detail such as *black cheek patch* for *Kentucky Warbler* and *long neck* for *Pied Billed Grebe*.

Comparing Explanations, Baselines, and Ablations. Figure 5 compares sentences generated by our explanation, baseline, and ablation models. Each model produces reasonable sentences, however, we expect our explanation model to produce sentences which discuss class relevant properties. For many images, the explanation model uniquely mentions some relevant properties. In Fig. 5, row 1, the explanation model specifies that the *Bronzed Cowbird* has *red eyes* which is rarer than properties mentioned correctly by the definition and description models (*black*, *pointy bill*). For *White Necked Raven* (Fig. 5 row 3), the explanation model identifies the *white nape*, which is a unique attribute of that bird. Explanations are also more image relevant. For example, in Fig. 5 row 7 the explanation model correctly mentions visible properties of the *Hooded Merganser*, but other models fail in at least one property.

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**.

This is a **Black Billed Cuckoo** *because ...*
Definition: this bird has a **yellow belly** and a **grey head**.
Description: this bird has a **yellow belly** and **breast** with a **gray crown** and **green wing**.
Explanation-Label: this bird has a **yellow belly** and a **grey head** with a **grey throat**.
Explanation-Dis.: this is a **yellow bird** with a **grey head** and a **small beak**.
Explanation: this is a **yellow bird** with a **grey head** and a **pointy beak**.

This is a **White Necked Raven** *because ...*
Definition: this bird is **black in color** with a **black beak** and **black eye rings**.
Description: this bird is **black** with a **white spot** and has a **long pointy beak**.
Explanation-Label: this bird is **black** in color with a **black beak** and **black eye rings**.
Explanation-Dis.: this is a **black** bird with a **white nape** and a **black beak**.
Explanation: this is a **black** bird with a **white nape** and a **large black beak**.

This is a **Northern Flicker** *because ...*
Definition: this bird has a **speckled belly and breast** with a **long pointy bill**.
Description: this bird has a **long pointed bill grey throat** and **spotted black and white mottled crown**.
Explanation-Label: this bird has a **speckled belly and breast** with a **long pointy bill**.
Explanation-Dis.: this is a **grey bird** with **black spots** and a **red spotted crown**.
Explanation: this is a **black and white spotted bird** with a **red nape** and a **long pointed black beak**.

This is a **American Goldfinch** *because ...*
Definition: this bird has a **yellow crown** a **short and sharp bill** and a **black wing** with a **white breast**.
Description: this bird has a **black crown** a **yellow bill** and a **yellow belly**.
Explanation-Label: this bird has a **black crown** a **short orange bill** and a **bright yellow breast and belly**.
Explanation-Dis.: this is a **yellow bird** with a **black wing** and a **black crown**.
Explanation: this is a **yellow bird** with a **black and white wing** and an **orange beak**.

This is a **Yellow Breasted Chat** *because ...*
Definition: this bird has a **yellow belly and breast** with a **white eyebrow** and **gray crown**.
Description: this bird has a **yellow breast and throat** with a **white belly and abdomen**.
Explanation-Label: this bird has a **yellow belly and breast** with a **white eyebrow** and **gray crown.**
Explanation-Dis.: this is a bird with a **yellow belly** and a **grey back and head**.
Explanation: this is a bird with a **yellow breast** and a **grey head and back**.

This is a **Hooded Merganser** *because ...*
Definition: this bird has a **black crown** a **white eye** and a **large black bill**.
Description: this bird has a **brown crown** a **white breast** and a **large wingspan**.
Explanation-Label: this bird has a **black and white head** with a large **long yellow bill** and **brown tarsus and feet**.
Explanation-Dis.: this is a **brown bird** with a **white breast** and a **white head**.
Explanation: this bird has a **black and white head** with a **large black beak**.

Fig. 5. Example sentences generated by our baseline models, ablation models, and our proposed explanation model. Correct properties are highlighted in green, mostly correct ones are highlighted in yellow, and incorrect ones are highlighted in red. The explanation model correctly mentions image relevant and class relevant properties. (Color figure online)

Comparing Definitions and Explanations. Figure 6 directly compares explanations to definitions for three bird categories. Images on the left include a visual property of the bird species which is not present in the image on the right. Because the definition is the same for all image instances of a bird class, it can produce sentences which are not image relevant. For example, in the second row, the definition model says the bird has a *red spot on its head* which is true for the image on the left but not for the image on the right. In contrast, the explanation model mentions *red spot* only when it is present in the image.

Discriminative Loss. To determine how the discriminative loss impacts sentence generation, we compare the description and explanation- discriminative models in Fig. 7. Neither model receives class information at test time, though the explanation-discriminative model is explicitly trained to produced class

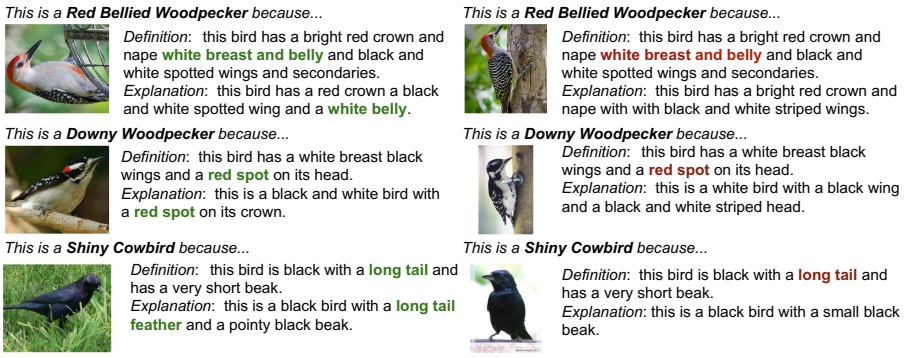

Fig. 6. We compare generated explanations and definitions. All explanations on the left include an attribute which is not present on the image on the right. In contrast to definitions, our explanation model can adjust its output based on visual evidence. (Color figure online)

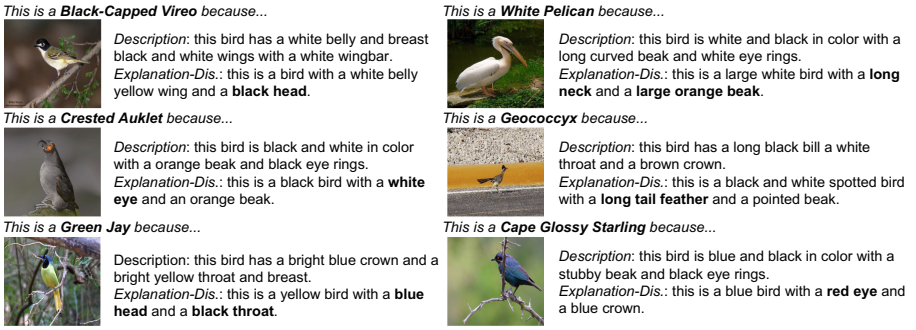

Fig. 7. Comparing sentences generated by description and explanation-discriminative models. Though both are capable of accurately describing visual attributes, the explanation-discriminative model captures more "class-specific" attributes. (Color figure online)

specific sentences. Both models generate visually relevant sentences. However, the model trained with our discriminative loss contains properties specific to a class more often than the ones generated using the description model. For instance, for the class *Black-Capped Vireo*, the explanation-discriminative model mentions *black head* which is one of the most prominent distinguishing properties of this vireo type. For the *White Pelican* image, the explanation-discriminative model mentions highly discriminative features like *long neck* and *orange beak*.

Incorrect Prediction. We qualitatively examine explanations for instances where the incorrect label is predicted (Fig. 8). In these scenarios, explanations are frequently image relevant and mention features common in both the image instance and the predicted class. For example, in the first row of Fig. 8 the model mistakes the "Laysan Albatross" for the "Cactus Wren". The explanation text

Fig. 8. When the model predicts the wrong class, the explanation is image relevant and frequently discusses attributes common between the image and the predicted class.

includes many features also mentioned in the "Cactus Wren" definition (for example color and the spotted feathers) and is relevant to the image.

6 Conclusion

Our work is an important step towards explaining deep visual models, a crucial capability required from intelligent systems. Visual explanation is a rich research direction, especially as the field of computer vision continues to employ and improve deep models which are not easily interpretable. We anticipate that future models will look "deeper" into networks to produce explanations and perhaps begin to explain the internal mechanism of deep models.

We propose a novel reinforcement learning based loss which allows us to influence the kinds of sentences generated with a sentence level loss function. Though we focus on a discriminative loss in this work, we believe the general principle of a loss which operates on a sampled sentence and optimizes for a global sentence property is potentially beneficial in other applications. For example, [12,27] propose introducing new vocabulary words into description systems. Though both models aim to optimize a global sentence property (whether or not a caption mentions a certain concept), neither optimizes for this property directly.

In summary, we have presented a novel image explanation framework which justifies the class prediction of a visual classifier. Our quantitative and qualitative evaluations demonstrate the potential of our proposed model and effectiveness of our novel loss function. Our explanation model goes beyond the capabilities of current captioning systems and effectively incorporates classification information to produce convincing explanations, a potentially key advance for adoption of many sophisticated AI systems.

Acknowledgements. This work was supported by DARPA, AFRL, DoD MURI award N000141110688, NSF awards IIS-1427425 and IIS-1212798, and the Berkeley

Artificial Intelligence Research (BAIR) Lab. Marcus Rohrbach was supported by a fellowship within the FITweltweit-Program of the German Academic Exchange Service (DAAD). Lisa Anne Hendricks is supported by an NDSEG fellowship. We thank our experienced bird watchers, Celeste Riepe and Samantha Masaki, for helping us evaluate our model.

References

1. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Learning to compose neural networks for question answering. In: NAACL (2016)
2. Banerjee, S., Lavie, A.: Meteor: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, vol. 29 (2005)
3. Berg, T., Belhumeur, P.: How do you tell a blackbird from a crow? In: ICCV (2013)
4. Biran, O., McKeown, K.: Justification narratives for individual classifications. In: Proceedings of the AutoML Workshop at ICML 2014 (2014)
5. Core, M.G., Lane, H.C., Van Lent, M., Gomboc, D., Solomon, S., Rosenberg, M.: Building explainable artificial intelligence systems. In: Proceedings of the National Conference on Artificial Intelligence, vol. 21. AAAI Press, Menlo Park (1999). MIT Press, Cambridge (2006)
6. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.: What makes Paris look like Paris? ACM Trans. Graph. **31**(4), 101:1–101:9 (2012). doi:10.1145/2185520.2185597
7. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR (2015)
8. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: a deep convolutional activation feature for generic visual recognition. In: ICML (2013)
9. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: CVPR (2015)
10. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In: CVPR (2016)
11. Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., Saenko, K.: YouTube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: ICCV (2013)
12. Hendricks, L.A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., Darrell, T.: Deep compositional captioning: Describing novel object categories without paired training data. In: CVPR (2016)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
14. Jia, X., Gavves, E., Fernando, B., Tuytelaars, T.: Guiding long-short term memory for image caption generation. In: ICCV (2015)
15. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia. ACM (2014)

16. Jiang, Z., Wang, Y., Davis, L., Andrews, W., Rozgic, V.: Learning discriminative features via label consistent neural network (2016). arXiv preprint arXiv:1602.01168
17. Johnson, W.L.: Agents that learn to explain themselves. In: AAAI (1994)
18. Karpathy, A., Li, F.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
19. Kiros, R., Salakhutdinov, R., Zemel, R.: Multimodal neural language models. In: ICML (2014)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
21. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A., Berg, T.: Baby talk: understanding and generating simple image descriptions. In: CVPR (2011)
22. Lacave, C., Díez, F.J.: A review of explanation methods for Bayesian networks. Knowl. Eng. Rev. **17**(02), 107–127 (2002)
23. Lampert, C., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. In: TPAMI (2013)
24. Lane, H.C., Core, M.G., Van Lent, M., Solomon, S., Gomboc, D.: Explainable artificial intelligence for training and tutoring. Technical report, DTIC Document (2005)
25. Lomas, M., Chevalier, R., Cross II., E.V., Garrett, R.C., Hoare, J., Kopack, M.: Explaining robot actions. In: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction. ACM (2012)
26. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016)
27. Mao, J., Wei, X., Yang, Y., Wang, J., Huang, Z., Yuille, A.L.: Learning like a child: fast novel visual concept learning from sentence descriptions of images. In: ICCV (2015)
28. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.L.: Explain images with multimodal recurrent neural networks. In: NIPS Deep Learning Workshop (2014)
29. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to wordnet: an on-line lexical database*. Int. J. Lexicogr. **3**(4), 235–244 (1990)
30. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL (2002)
31. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: CVPR (2016)
32. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML (2016)
33. Shortliffe, E.H., Buchanan, B.G.: A model of inexact reasoning in medicine. Math. Biosci. **23**(3), 351–379 (1975)
34. Teach, R.L., Shortliffe, E.H.: An analysis of physician attitudes regarding computer-based clinical consultation systems. Use and Impact of Computers in Clinical Medicine. Springer, New York (1981)
35. Van Lent, M., Fisher, W., Mancuso, M.: An explainable artificial intelligence system for small-unit tactical behavior. In: Proceedings of the National Conference on Artificial Intelligence. AAAI Press, Menlo Park (1999). MIT Press, Cambridge (2006)
36. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDEr: consensus-based image description evaluation. In: CVPR (2015)
37. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR (2015)

38. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Technical report CNS-TR-2011-001, California Institute of Technology (2011)
39. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. **8**, 229–256 (1992)
40. Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: ICML (2015)
41. Yeung, S., Russakovsky, O., Jin, N., Andriluka, M., Mori, G., Fei-Fei, L.: Every moment counts: dense detailed labeling of actions in complex videos. In: CVPR (2016)