

# Generating Visual Explanations

By Lisa Anne Hendricks et al.

Report author: Michael Aichmüller

Supervisor: PD Ulrich Köthe

HCI, Heidelberg University, Summer 2018.

**Abstract**—In light of the task set out by the seminar to present the latest research of methods, that try to understand the decision making process of a neural network, the paper called 'Generating Visual Explanations' opted for an automatic sentence generation, that justifies the classification of birds species. By combining deep fine-grained classifiers and LSTM models, the authors were able to generate sentences that justify the decision using features from the image that are both image relevant and also class discerning. An analysis using the latest linguistic metrics shows the strength of the model with respect to being image relevant as well as class relevant. With respect to the class relevance, the main contribution of the paper lies in the introduction of a novel loss function enforcing class discerning quality, i.e the quality of producing sentences that use features more unique to the label at hand.

## I. INTRODUCTION

WHEN the topic of neural networks is discussed, it is easy to point out the effectiveness of these machines. We are now able to create software that can categorize images to high accuracy, detect specific patterns in videos, learn to play games and much more. Especially the task of classification in the field of visual recognition is a great success story, albeit arguably among the easier of the supervised tasks. However, the question of how such a system comes to its conclusions is far from understood, thus they lack the much needed credibility. Without said credibility, we remain hesitant to apply these relatively new systems in sensitive areas - any clinical application comes to mind, military equipment and maybe even more futuristic applications to softer sciences such as judicial sentencing - where wrong labeling, incorrect image segmentation, and lack of understanding of the underlying problem can have drastic, even fatal, consequences. It is therefore obvious, that any such system that can provide explanations, while also performing outstandingly, is preferable to inexplicable systems.

In the following to come, it is important to understand, what aspects the term explanation encapsulates, as there are different forms. The coarse distinction chosen for this setup is the division into the two parts of *introspection* and *justification*. An *introspection* seeks to explain outputs by referring to the specific state the network was in and subsequently how the input traversed the network in terms of its layer activations. For example, an explanation for the classification of an image as 'car' might read: 'The input aggregated to the value  $x$ , the activation of layer 1 equated to  $y$ , and the highest class probability in the output layer was found for the class 'car'. It is clear that such explanations address only people with

technical knowledge. On the other hand, a *justification* tries to correlate the visual evidence with the output, thereby also allowing laymen to understand the explanation. An example of this, again with the 'car' classification, might read: 'The image showed the feature of a bonnet, four wheels, a steering wheel, and windows. It's thus most likely a 'car'. The latter is the approach the authors desired to study.

## II. CLASS AND IMAGE RELEVANCE OF SENTENCES

Each sentence has unique properties that qualify it. It is proposed for sentences to have the two qualities of being *image relevant* and *class relevant* in order to be seen as explaining. Image relevant sentences are descriptions of the image, that

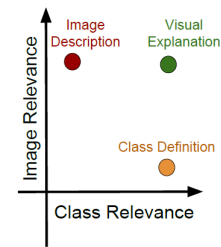


Fig. 1. An explaining sentence should incorporate aspects of both the image and the class.

address the visual evidence at hand by picking specific parts that represent the whole best. Such a sentence can thus only apply to its respective image, if detailed enough. A class relevant sentence on the other hand is a definition of the class itself and by virtue of its nature does not necessarily need to address any part of the image at all. This type of sentence should be applicable to all images of birds belonging to the same class. Logically an explanation entails both dimensions and is graded by its capability of distinguishing with respect to both at the same time.

## III. THE ARCHITECTURE OF THE MODEL

The paper proposed a combination of two pipelines to achieve their goal of generating explaining sentences. Input images are forwarded to a pre-trained deep-fine-grained classifier. Its job is to select detailed features from the images. The architecture used in the paper is that of [2]. One of the issues of fine-grained features is the high dimensionality, usually residing in the hundreds of thousands to the millions.

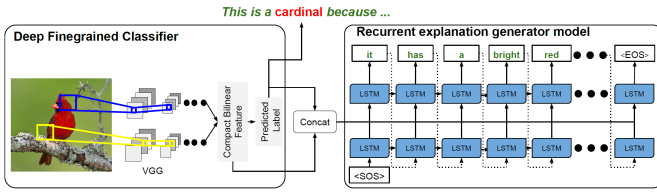


Fig. 2. The model pipeline. The input data is run through a deep fine-grained classifier picking out nuanced details of the image and classifying it. The features and the label are then forwarded to the LSTM stack to produce an explaining sentence.

In order to reduce the dimensionality the features are being approximated by a representation containing negligibly less amount of information at the low cost of merely a few thousand dimensions.

### A. Bilinear Models

Following [4], a bilinear model is defined as the quadruple  $(f_A, f_B, P, C)$ , with  $f_A, f_B$  being feature extractors of the form  $f: L \times I \mapsto \mathbb{R}^{c \times D}$ , taking in a location  $L$  and image  $I$  and returning features of size *feature dimension* times *image dimensions*. In the case of this paper,  $f_A, f_B$  are two convolutional neural networks from VGG (pretrained).  $P$  refers to a pooling operation, and  $C$  to a classification. The bilinear feature is the outer matrix product of the two functions at position  $l$ , namely  $bi(l, I, f_A, f_B) := f_A(l, I)^T f_B(l, I)$ . A pooling operator is then applied onto the bilinear features, which can be any known pooling operation, such as max pooling or aggregating, i.e.  $\phi(I) = \sum_{l \in L} bi(l, I, f_A, f_B)$  for example. This pooling operator was chosen to be among the compact pooling operators of [2], i.e. Random Maclaurin (RM) and Tensor Sketch (TS) algorithm. Both produce low dimensional feature vectors.

Based on these features a classification takes place. The proposed class, as well as the features are afterwards forwarded to a Long Short Term Memory (LSTM) model, whose unrolled structure can be seen in figure 2.

### B. LSTM Models

An LSTM is a variant of the network structure called recurrent neural network (RNN). These structures are meant for sequential data  $(x_t)_{t \in I}$  such as language, as it reuses its output implicitly from step  $t_i$  in step  $t_{i+k}$  for  $k > 0$ . This allows for the modeling of dependencies between the various inputs  $x_i$  and  $x_{i+k}$ . Its logical structure is depicted in figure 3 with its unrolled structure, which shows the inner working more clearly, in 4.

The attractiveness of such a model lies precisely in its capability to recognize the dependency of the data, may it be contextual, syntactical or semantical. It can thus learn the complex and somewhat mathematically hard to grasp nature of language in a natural manner. How this is achieved can be credited to its internal nature of separating the task of forgetting, adding, and outputting into three different parts, called gates. Before diving into the exact mechanism, it is necessary to state the preliminaries for which we follow [5]:

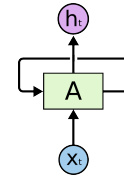


Fig. 3. The logical structure of a recurrent neural network[5]. The neural network A feeds at each timestep  $t$  its output back into itself and adds new input  $x_t$ .

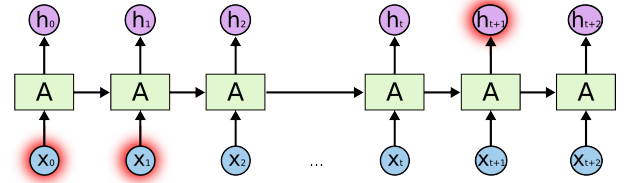


Fig. 4. The unrolled structure of a recurrent neural network[5]. The dependency of the data is incorporated in the parameters of the model and can be detected even across longer steps, hence the name Long Short Term Memory.

The LSTM maintains a data stream as the so called memory stream, or cell state denoted by  $(C_t)_t$ . This stream incorporates the whole information learned so far, all the dependencies that have been recognized. It is also this stream that is constantly manipulated with new input  $x_{t+1}$  in order to learn and adapt to new information, and from which the LSTM chooses its output  $h_{t+1}$  at each time-step. Jumping into the finer details of the

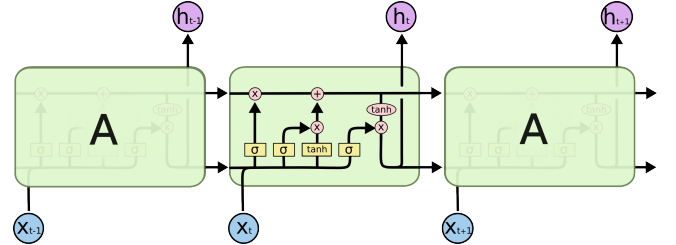


Fig. 5. The internal structure of an LSTM neural network[5]. The manipulation of the memory stream (here the upper line connecting the As) and the output stream (here the lower line, where also the new input  $x_t$  is fed in) is visualized. This manipulation is logically separated into three different segments. In the diagram the layer wise activation is denoted by a yellow box operation, e.g.  $\sigma$  (sigmoid), and element wise operations by a pink box with the operation symbol inside ( $\times$  for multiplication,  $+$  for addition).

system, one now needs to understand the aforementioned gate structure<sup>1</sup>:

- 1) **Forget gate.** In this gate, the decision of which parts of the existing data in the memory  $C_{t-1}$  should be forgotten is made. An example for why this is necessary can be found in the change of the subject in a sentence. The operation performed is that of a weight multiplication  $W_f$  onto the previous output concatenated with the new

<sup>1</sup>It is worthwhile to note, that the explanation in this report is just **one** of the many possible variations proposed and may differ from what the reader may know as an LSTM build. Notable additions to the core structure are so called 'peepholes', which won't be covered by this segment.

input  $x_t$  (e.g. the next word in the sentence) followed by a sigmoid activation. Afterwards we multiply the output onto the cell state element-wise. As the sigmoid renders the values onto the range of  $[0, 1]$ , it directly performs a weighting of the importance of the values, when one associates 0 as unimportant, and 1 as absolutely necessary.

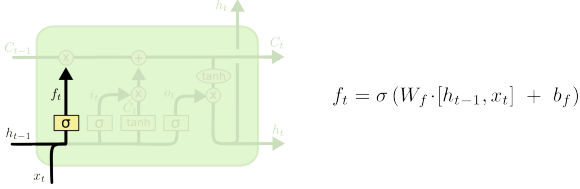


Fig. 6. Forgetting of previous values visualized and the update formula that it begets.

- 2) **Input gate.** After having selected which values are no longer needed, the second step is to include the new information into our existing memory. Another sigmoid activation of previous output and current input selects which values we want to replace, while on the side a tanh, which renders the data into the range of  $[-1, 1]$ , creates candidate values to be added. After element-wise multiplication, an element-wise addition onto the memory stream finalizes the change in values.

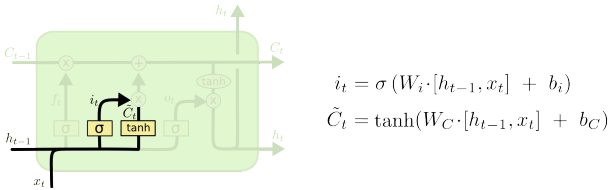


Fig. 7. The input of new values visualized and the update formula that it begets.

- 3) **Output gate.** Lastly, after forgetting and adding of values, the system needs to decide its output at the current step. The cell state undergoes a tanh activation and multiplies the outcome with the sigmoid activation of the previous output and input, resulting in  $h_t$ , which is passed onto the next iteration, as well as returned.

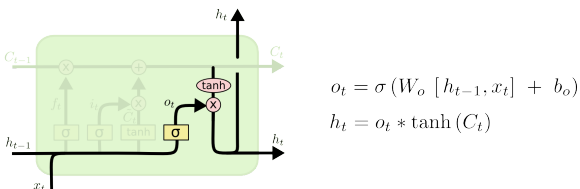


Fig. 8. The output generation visualized and the formula that it begets.

In the model of the paper there is however not just one LSTM network, but two. For why this is, will become clearer when the training of the model is discussed. However with this system at hand, the model is capable of producing sentences, whose quality depends on the training data at hand. For the

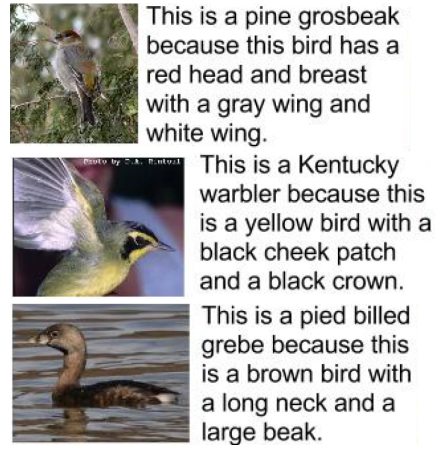


Fig. 9. Example output sentences of the final model.

purpose of the study, the sentences only needed to fulfill a templated form that reads: *This is a CLASS, because argument 1 and argument 2 and...*, which can be seen in some output examples in figure 9.

#### IV. TRAINING THE MODEL

With the basics in mind, it is now time to examine how to train such a model combination to produce sentences as outlined in the examples. As we have discussed in the beginning of this report, we need to ensure both qualities of image and class relevance. One of the possible ways to achieve this task lies in splitting the loss function up into two separate parts as well. The one part, called the *relevance loss* will ensure the image relevance, while the other, called *discriminative loss* provides the class discerning property. The general training schedule sees first to the feature detection and subsequent classification of the image. These two parts of information are then passed onto the LSTM, together with a target sentence. Here is where the fact of two LSTM structures comes into play: The first LSTM, as seen as the left one in figure 10, is given only the target sentence as input. At time  $t$  it is thus provided the word  $w_t$  of the target sentence and passes its output, the hidden output  $h_t$ , on to the second LSTM, and itself to be used in the next timestep. This second LSTM hence receives the output of the first LSTM, but on top of that also the features of the image and its class. Its own hidden output  $h'_t$  in turn is then provided to itself for the timestep  $t + 1$  and also stored as conditional probability of the word at time  $t$  given the previous  $t - 1$  words, the image, and the class, i. e.  $p(w_t | w_{0:t-1}, I, C)$ .

##### A. Relevance Loss

As mentioned, there are sample sentences attached to each image, that describe it. As such, one would want the generated sentences to coalesce with these target sentences. Naturally, this resembles the task of picking the most likely next word  $w_t$ , given the previous chain of words  $w_{0:t-1}$ , the image  $I$ , and the designated class  $C$ , if the probabilities of the words have been correctly adjusted. In other words, one can try to approach the 'true' distribution of the words to come and

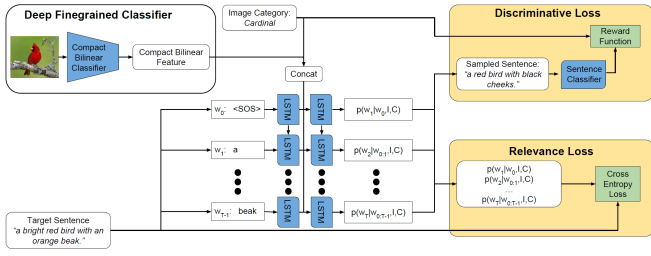


Fig. 10. The training setup for the model. A target sentence is provided with the image, that the LSTM given the features and the class label should aim to incorporate. Afterwards, the posterior probabilities from the LSTMs are check against a *Cross-Entropy-Loss*, while a sampled sentence from the LSTMs will provide the basis for calculating the *Discriminative Loss*.

pick the likeliest accordingly. This scheme is captured by the *Cross-Entropy-Loss*. The paper however merely states that said approach has simply worked best for them to achieve the image relevance, and thus chosen it, without truly explaining their motive behind it. Nevertheless, the form of the loss is given as

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

with  $N$  being the batch size, thus overall an averaging of the log hidden states of the LSTM, which is the approach for a *Cross-Entropy-Loss*, when the ground truth distribution is unknown and needs to be estimated.

The curious downside of this loss is the fact that any description that describes the image well enough, will ensure a low relevance loss, and thus be seen as a good sentence. However, it is implied that any probability distribution that favours descriptions using class discerning features are equally good to descriptions applying more general features. Therefore, one needs to employ another loss to enforce this quality.

## B. Discriminative Loss

Exactly this second loss function will provide the needed quality. Its computation, however, is not as straightforward as the *Cross-Entropy-Loss*. The idea stems from the branch of *Reinforcement-Learning*, that utilizes a reward function to assess the worth of an action. The specific form of the reward function is thus tailored to each specific problem. In this case, the discriminator reward  $R_D$  of a description  $\tilde{w}$  is chosen to be the posterior probability of the bird class given the description, i.e.

$$R_D(\tilde{w}) := p(C | \tilde{w}) \quad (2)$$

with  $C$  being the class. This probability is given by a pre-trained sentence classifier, i.e. a classifier that produces posterior probabilities over bird species (classes) for a given description sentence  $\tilde{w}$ . Note, that this classifier has an accuracy of only 22% on test sets, yet this doesn't prove cumbersome to the task. Since such sentences need to come from our model in order to produce a meaningful loss, they are sampled from the LSTM. The sampled description is passed onto the sentence classifier returning the reward. This value in itself isn't very meaningful as a loss, since it is simply a value  $\in [0, 1]$  for one

single description, implying it doesn't establish a meaningful connection between the discerning quality of the sentences and the computed value. One measure that does however is the expectation of the class given description, i.e. the expectation of  $p(C | \tilde{w})$  with respect to  $\tilde{w}$ , which is also used as the final, discriminative loss. Finally, the overall loss function is combined by means of adding the two single losses together and multiplying by a hyperparameter  $\lambda$ , resulting in

$$L := L_R - \lambda \mathbb{E}_{\tilde{w} \sim p(w|I,C)} (R_D(\tilde{w})). \quad (3)$$

Note, that the expectation is an intractable one, as there are effectively infinitely many descriptions that need to be considered. As alleviation of this problem, the paper suggests Monte-Carlo sampling from the LSTM model at each timestep to generate descriptions, through which the expectation could be approximated. Having established the loss, one needs to see how to train using it, meaning how to compute the gradient. The paper now argues 'As a discrete distribution, the sampling operation for the categorical distribution is non-smooth in the distribution's parameters  $\{p_i\}$ , so the gradient  $\nabla_W R_D(\tilde{w})$  of the reward  $R_D$  for a given sample  $\tilde{w}$  with respect to the weights  $W$  is undefined.' [3], and goes on to propose a solution to this issue exploiting the relationship of (suppressing the conditional of the probabilities for readability)

$$\nabla_W \mathbb{E}_{\tilde{w} \sim p(w)} [R_D(\tilde{w})] = \nabla_W \int R_D(\tilde{w}) p(\tilde{w}) d\lambda \quad (4)$$

$$= \int \nabla_W R_D(\tilde{w}) p(\tilde{w}) d\lambda \quad (5)$$

$$= \int R_D(\tilde{w}) \nabla_W p(\tilde{w}) d\lambda \quad (6)$$

$$= \int R_D(\tilde{w}) \frac{1}{p(\tilde{w})} (\nabla_W p(\tilde{w})) p(\tilde{w}) d\lambda \quad (7)$$

$$= \int R_D(\tilde{w}) \nabla_W \log p(\tilde{w}) d\lambda \quad (8)$$

$$= \mathbb{E}_{\tilde{w} \sim p(w)} [R_D(\tilde{w}) \nabla_W \log p(\tilde{w})] \quad (9)$$

which has been more thoroughly shown in [9]. There is an unresolved flaw in this argumentation though. First of all, if the gradient of  $R_D$  with respect to the weights  $W$  is undefined, one would need an argument for why the gradient of  $R_D(\tilde{w}) p(\tilde{w})$  is well defined in turn. Hence the legitimacy of this approach appears to be unclear and a more in depth explanation would be welcome, yet the success grants validity. However, with the important gradient relationship stated, the update rule for the model can be stated, which then becomes

$$\nabla_W L_R - \lambda \mathbb{E}_{\tilde{w} \sim p(w)} [R_D(\tilde{w}) \nabla_W \log p(\tilde{w})] \quad (10)$$

In comparison, one will find the paper to state the update rule to be

$$\nabla_W L_R - \lambda R_D(\tilde{w}) \nabla_W \log p(\tilde{w}) \quad (11)$$

which represents stochastic gradient descent with a batch size of 1.

## V. EVALUATION PRELIMINARIES

Turning to the quantitative and qualitative evaluation of the proposed model, it is necessary to state the setup, data, and metrics used.

### A. Data

As has constantly been mentioned, the data in question is the so called 'Caltech UCSD Birds 200-2011 (CUB)' dataset [8]. Within this set, one will find 200 classes of North American bird species and 11788 images thereof, with each image, due to a recent extension by [6], containing five detailed description sentences of the form visible in figure 10. Since each image belongs to only one class, and each example sentence being a description of the same and only of the same, this dataset stands out as being unique for the visual explanation task [3]. Note, that the sentences provided do not explain why the bird belongs to its class. Therefore one cannot conclude that the model is trained directly to simply copy explanatory sentences, but rather has to learn this feature itself. As far as the image features are concerned, 8,192 dimensional features are extracted from the penultimate layer of the compact bilinear fine-grained classification model, which has been pre-trained on the CUB dataset and achieves an accuracy of 84%. The words are encoded as one-hot-vectors and a 1000-dimensional embedding is learned before inputting each word into the 1000-dimensional LSTM [3]. Model hyperparameters are chosen by means of the standard CUB validation set before evaluating on the respective test set. All reported results are on the standard CUB test set.

### B. Comparison Models

In order to measure the performance, the model is tested against itself with various parts taken out to highlight their effect, an ablation comparison. Two models are to be seen as baseline, as they should show how each dimension of image and class relevance is independent and necessary to incorporate. The models in question are the

- *Definition model*: Training the model to generate explaining sentences only using the image label as input.
- *Description model*: Training the model by conditioning only on the image features as input.

Furthermore, two key differences to the baseline of the description model, i.e. the addition of the label to condition on and the discriminative loss, are highlighted by comparing the results to two more ablations:

- *Explanation-Label model*: The model trained without the discriminative loss.
- *Explanation-discriminative model*: The model trained without the class label.

### C. Metrics

As sentences are to be evaluated, linguistic metrics are needed. Standard comparison tools in this field are given by the metrics called

- *CIDEr* [7]: Measures similarity by accounting for matching n-grams, a contiguous sequence of n items from text or speech, that are TF-IDF weighted<sup>2</sup>.

<sup>2</sup>TF-IDF, text frequency - inverse document frequency, means that words that are naturally more common (high document frequency), such as 'the', will need a higher text frequency in order to have a big weight.

- *METEOR* [1]: Computed by matching words in two sentences, while also accounting for synonyms.

These measures accurately report the image relevance of sentences as they need to pick out the displayed attributes within the sentence and check it against the templates. If the attributes match, the score needs to be high. Measuring complex properties such as class relevance, and the class discerning quality of sentences proves to be more difficult though. As such, the authors proposed their own composite metrics<sup>3</sup>:

- *Class Similarity*: If a sentence fits the definition of a class well, it would have to score high when matched with the target sentences belonging to its label. Therefore, the CIDEr score of this sentence computed against each target sentence in its class and then added together will provide a measure for the similarity with respect to its own class.
- *Class Rank*: A sentence fitting its class well, doesn't imply that it wouldn't also fit another. Therefore, a class discerning sentence should return low values, when its CIDEr scores against all available sentences of all classes is accumulated. This measure will be called class rank.
- *Human Experts*: A team of bird experts was hired to evaluate the sentences with regards to their explanatory power for the determined bird class.

## VI. EVALUATION

Table I and II show the results measured by the previously mentioned metrics. In summary, the full model proves to be superior to both baselines in image relevance and class relevance, as well as showing the importance of conditioning on the labels and using the discriminative loss to produce better results overall, demonstrated by its surpassing of both explanation ablations. Columns 1 and 2 show a curious, slight edge the definition model seems to have over the description model in terms of image relevance. Also the Explanation-Label fairs only marginally better than the definition model, whereas the Explanation-Discriminative achieves convincingly higher values in contrast. With respect to class relevance, the definition model trumps the description model as expected, and any addition working with class information improves the model, as seen in the consistently better values from both ablation models. Adding the discriminative loss however doesn't discern between classes as well as when adding the label to the baseline models, as can be seen in column 4 row 4 being worse than row 3. Also surprising is that the raw definition baseline comes second best to the grand model, showing that adding the label and discriminative loss works better in tandem than each alone. The human evaluation by the bird experts again presents a rather stunning verdict as the ablation model with the discriminative loss scores the worst out of all models, even decidingly worse than the definition baseline.

The authors of the paper also present a qualitative analysis of the results using about 18 examples to showcase various

<sup>3</sup>Note, that CIDEr was used for these metrics, because it includes the TF-IDF weighting.

<i>Better is...</i>	Image Relevance		Class Relevance	
	METEOR higher	CIDEr higher	Similarity higher	Rank (1-200) lower
Definition	27.9	43.8	42.60	15.82
Description	27.7	42.0	35.3	24.43
Explanation - Label	28.1	44.7	40.86	17.69
Explanation - Discr.	28.8	51.9	43.61	19.80
<b>Explanation (FULL)</b>	<b>29.2</b>	<b>56.7</b>	<b>52.25</b>	<b>13.12</b>

TABLE I

RESULTS OF THE EVALUATION WITH THE METRICS STATED IN CHAPTER V-C, WITHOUT THE BIRD EXPERT RANK. THE WHOLE MODEL OUTPERFORMS ALL ABLATION MODELS IN THESE CATEGORIES.

<i>Better is...</i>	Best Explanation
	Bird Expert Rank (1-5) lower
Definition	2.92
Description	3.11
Explanation - Label	2.97
Explanation - Discr.	3.22
<b>Explanation (FULL)</b>	<b>2.78</b>

TABLE II

RESULTS OF THE EVALUATION WITH THE METRIC OF BIRD EXPERTS AS STATED IN CHAPTER V-C. AGAIN, THE GRAND MODEL STANDS OUT AGAINST ABLATION MODELS.

differences in the production of the models. The highlights entail a comparison

- 1) between explanation, ablations, and baseline models in figure 14.
- 2) of definitions and explanations in figure 13.
- 3) of descriptions and explanations in figure 11.
- 4) of the effect of conditioning the model on a different label in figure 12.

With respect to 1) it is noticeable how in all six examples the explanation model picked out correct attributes, with the definition of correct and wrong having been deemed by the authors. None of the other models was capable of being this consistent and seemed to pick sometimes wrong characteristics. As for 2) it is unsurprising that the definition model has not changed its output when given a different image with the same label. This is expected, since the definition model is only given the class information. The inclusion of the image relevance by the explanation model is seen in the adaptation of different attributes given a different image. In 3) the strength of class information is visible in the form of more relevant class attributes in the explanation sentences. The descriptions fail to address more important features in comparison. Lastly in 4) the value of information within the class label becomes visible. When conditioned on a different label the model appears to be picking out almost the same ideas of features regardless of the image. This means, that the label itself can have a strong influence over the features that need to be chosen.

## VII. CONCLUSION

The authors have set out to produce explaining sentences that require no technical knowledge of computer science to be understood and succeeded. Related works have produced

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

This is a **Crested Auklet** because...



Description: this bird is black and white in color with a orange beak and black eye rings.

Explanation-Discr.: this is a black bird with a **white eye** and an orange beak.

This is a **Green Jay** because...



Description: this bird has a bright blue crown and a bright yellow throat and breast.

Explanation-Discr.: this is a yellow bird with a **blue head** and a **black throat**.

This is a **White Pelican** because...



Description: this bird is white and black in color with a long curved beak and white eye rings.

Explanation: this is a large white bird with a **long neck** and a **large orange beak**.

This is a **Geococcyx** because...



Description: this bird has a long black bill a white throat and a brown crown.

Explanation-Discr.: this is a black and white spotted bird with a **long tail feather** and a pointed beak.

This is a **Cape Glossy Starling** because...



Description: this bird is blue and black in color with a stubby beak and black eye rings.

Explanation-Discr.: this is a blue bird with a **red eye** and a blue crown.

Fig. 11. Comparison of sentences generated using description and explanation-discriminative models. The paper argues that while both are capable of accurately describing visual attributes, the explanation-discriminative model captures more 'class-specific' attributes. These features are emphasized in bold.

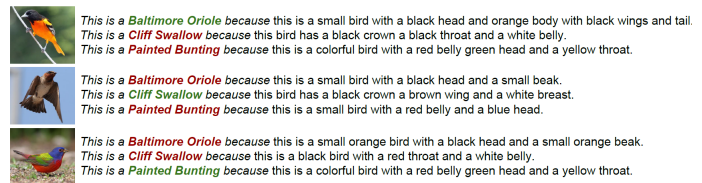



Fig. 12. An example set for the effect of a forced label change on the produced explanation. For some bird classes, such as 'Painted Bunting', the label carries valuable information that impacts the explanation.

explanations that are either rule-based, require filling in a pre-determined template or required expert level knowledge of the design of the system. The main contribution, the introduction of a reinforcement-learning reward function to build a loss function around has also been novel up to the release date of the paper. Through the analysis of the quantitative results an either slightly better, or convincingly better performance of the full model compared to ablations and baselines could be shown, when checked by modern linguistic metrics. The qualitative results indicate some more nuanced details, as seen by the authors. However, it remains uncertain, whether these examples have been cherry-picked to fit the narrative or are simply anecdotal, as is the case for implying a major attribute on merely 18 examples of 200 classes and 11788 images. The performance of the model can't be denied though, and it has


*This is a **Marsh Wren** because...*



Definition: this bird is brown and white in color with a skinny brown beak and brown eye rings.

Explanation: this is a small brown bird with a long tail and a **white eyebrow**.


*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.


*This is a **Shiny Cowbird** because...*



Definition: this bird is black with a long tail and has a very short beak.

Explanation: this is a black bird with a **long tail feather** and a pointy black beak.


*This is a **Marsh Wren** because...*



Definition: this bird is brown and white in color with a skinny brown beak and brown eye rings.

Explanation: this is a small bird with a long bill and brown and black wings.


*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 white bird with a black wing and a black and white striped head.

*This is a **Shiny Cowbird** because...*




Definition: this bird is black with a long tail and has a very short beak.

Explanation: this is a black bird with a small black beak.

Fig. 13. A comparison of generated explanations and definitions. Each explanation on the top mentions an attribute which is not present in the image on the bottom corresponding to it. In contrast to definitions, the explanation model adjusts its output based on visual evidence.

*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 **grey 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 **short orange bill** and a **bright yellow breast** and **belly**.

Explanation-Label: this is a **yellow bird** with a **black wing** and a **black crown**.

Explanation-Dis.: this is a **yellow bird** with a **black and white wing** and an **orange beak**.

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. 14. Example sentences generated by the baseline models, ablation models, and full explanation model. Correct attributes are highlighted in green, mostly correct attributes in yellow, and incorrect attributes in red. The explanation model discusses image relevant and class relevant features in these examples seemingly consistently.

definitely proved to be another step towards more explainable neural network systems. It would surely be helpful if future research were to look further into the introspection of such an approach, covering both grounds of explanations for a system.

REFERENCES

- [1] S. Banerjee and A. Lavie. 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*, pages 65–72, 2005.
- [2] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. *CoRR*, abs/1511.06062, 2015.
- [3] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. *CoRR*, abs/1603.08507, 2016.
- [4] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015.
- [5] C. Olah. Understanding lstm networks, 2018.
- [6] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58, 2016.
- [7] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- [8] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- [9] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.