Grounding Visual Explanations

Hendricks, L.A.; Hu, R.; Darrell, T.; Akata, Z.

Link to publication

Grounding Visual Explanations

Lisa Anne Hendricks[1], Ronghang Hu[1], Trevor Darrell[1], and Zeynep Akata[2]

[1] UC Berkeley, [2] University of Amsterdam
{lisa_anne,ronghang,trevor}@eecs.berkeley.edu and z.akata@uva.nl

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

Keywords: Explainability, Counterfactuals, Grounding, Phrase Correction

1 Introduction

Modern neural networks are good at localizing objects [6], predicting object categories [7] and describing scenes with natural language [32]. However, the reasoning behind the decision of neural networks are often hidden from the user. Therefore, in order to interpret and monitor neural networks, providing explanations of network decisions has gained interest [8,22,13].

Ideally, an agent that accurately explains a classifier's decision via natural language, as depicted in Figure 1, is expected to generate explanations such as "This is a *Cardinal* because it is a red bird with a red beak and a black face" where the phrases should be both *class discriminative*, i.e. a red beak is discriminative for cardinals, and *image relevant*, i.e. the image indeed contains a red beak. Moreover, an explainable AI agent should be capable of arguing why the image was not classified as another class such as *Vermilion Flycatcher* by mentioning a class-specific property such as "black wings" that differentiates a cardinal from a vermilion flycatcher. Contrasting two concepts via class-specific attributes provides an additional means of model interpretation.

Often class and image relevance are in opposition. For example, if one attribute frequently occurs within a class, an agent may learn to justify its prediction by mentioning this attribute without even looking at the image. We aim to resolve such conflicts
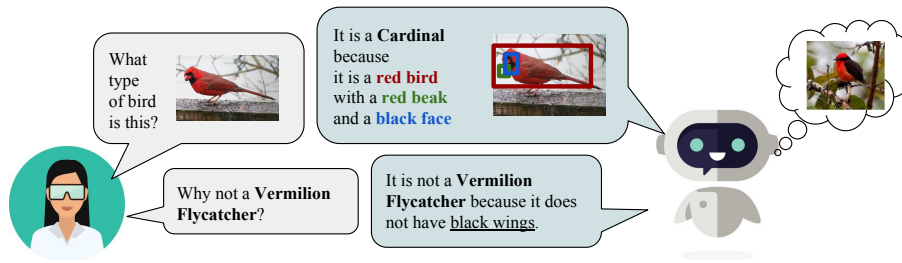
Fig. 1: Our phrase-critic agent considers grounded visual evidence to determine if candidate explanations are image relevant. In this example, as many cardinals are red and have a black patch on their faces, mentioning and grounding those properties constitutes an effective factual explanation, i.e. rationalization. Furthermore, in our framework, informing the user of why an image does not belong to another class via the absence of certain attributes constitutes a counterfactual explanation.

through a phrase-critic which explicitly determines if an explanatory sentence is image relevant by visually grounding discriminative object parts mentioned in an explanation.

One way to design such an agent is to use densely labeled data with ground truth part annotations. However, it can be time consuming to collect densely labeled data for every task. On the other hand, large, diverse datasets such as the Visual Genome [11] with densely labeled out-of-domain data do exist. Detecting visual evidence in a sentence via off-the-shelf grounding models can be unreliable, especially when applied to new domains. Nevertheless, integrating a natural language grounding model [10] trained on auxiliary data and an LSTM-based explanation model [8] via our proposed *phrase-critic* effectively grounds discriminative phrases in generated explanations.

Our phrase-critic integrates a ranking-loss to the explanation model and builds a set of mismatching part-attribute pairs by flipping attributes in the explanation, inspired by a relative attribute paradigm for recognition and retrieval [19]. By positing that a bird with a black beak can not also be a bird with a red beak, our model learns to score the correct attribute higher than automatically generated mutually-exclusive attributes.

We quantitatively and qualitatively show that our *phrase-critic* generates image relevant explanations more accurately than a strong baseline of mean-pooled scores from a natural language grounding model. Furthermore, our framework can easily be extended to other tasks beyond textual explanations. We also show that our phrase-critic framework effectively discerns whether a sentence contains a mistake, points out where the mistake is and fixes it, leading to an impressive performance on FOIL tasks [24].

2   Related Work

In this section, we review recent papers in the context of explanations, mainly focusing on textual and visual explanations and finally we discuss pragmatics oriented language generation papers that are relevant to ours.

Explainability. The importance of explanations for humans has been studied in the field of psychology [14,15], showing that humans use explanations as a guide for learning and understanding by building inferences and seeking propositions or judgments that enrich their prior knowledge. Humans usually seek explanations that fill the requested gap depending on prior knowledge and goal in question. Moreover, explanations are typically contrastive. Much of these ideas are built with careful empirical work, i.e. with human subjects on a specific aspect of explanations [18]. Since explanations are intended for a human understander, we emphasize the importance of human evaluation in evaluating the relevance of textual explanations to the image as well as looking for the criteria for what makes an explanation good, with the goal of training a "critic" that could evaluate explanations automatically.

Textual and Visual Explanation. In [26], trust is regarded as a primary reason to explore explainable intelligent systems. We argue a system which outputs discriminative features of an object class without being image relevant is likely to lose the trust of users. Consequently, we seek to explicitly enforce image relevance with our model. Like [2], we aim to generate *rationalizations* explaining the evidence for a decision as opposed to introspective explanations which aim to explain the intermediate activations of neural networks.

Early textual explanation models are applied to medical images [25] and developed as a feedback for teaching programs [12,27,3]. These systems are mainly template based. Recently, [8] proposed a deep model to generate natural language justifications of a fine-grained object classifier. However, it does not ground the relevant object parts in the sentence or the image. In [20], although an attention based explanation system is proposed, there are no constraints to ensure the actual presence of the mentioned attributes or entities in the image. Consequentially, albeit generating convincing textual explanations, [8,20] do not include a process for networks to correct themselves if their textual explanation is not well-grounded visually. In contrast, we propose a general process to first check whether explanations are accurately aligned with image input and then improve textually explanations by selecting a better-aligned candidate.

Other work has considered *visual explanations* which visualize which regions of an image are important for a decision[5,23,33,34]. Our model produces bounding boxes around regions which correspond to discriminative features, and is thus visual in nature. However, in contrast to visual explanation work, our goal is to rank generated explanatory phrases based on how well they are grounded in an image.

Pragmatics-Oriented Language Generation. Our work is also related to the recent work of pragmatics-oriented language generation [1] where a describer produces a set of sentences, then a choice ranker chooses which sentence best fulfills a specific goal, e.g. distinguishing one image from another. Similarly, image descriptions are generated to make the target image distinguishable from a similar image in [28], and referential expressions are generated on objects in a discriminative way such that one can correctly localize the mentioned object from the generated expression in [17]. In this work, we generate textual explanation to maximize both class-specificity and image-relevance. Though similar in spirit, part of our novelty lies in how we learn to rank sentences.
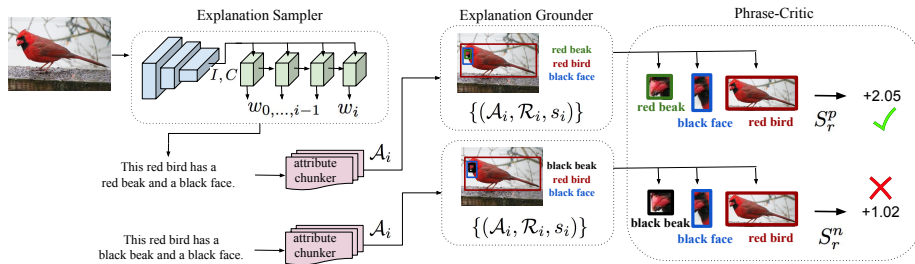
Fig. 2: Our phrase-critic model ensures that generated explanations are both class discriminative and image relevant. We first sample a set of explanations generated by [8], chunk the sentences into noun phrases and visually ground constituent nouns using [10]. Our model assigns a score to each noun phrase-bounding box pair and selects the sentence with the highest cumulative score judging it as the most relevant explanation.

3   Visual Explanation Critic

Our model consists of three main components. First, we train a textual explanation model [8] with a discriminative loss to encourage sentences to mention class specific attributes. Next, we train a phrase grounding model [10] to ground phrases in the generated textual explanations. Finally, our proposed phrase-critic model ranks textual explanations based on how well they are grounded in the image. As shown in Figure 2, our system first generates possible explanations (e.g., "This red bird has a red beak and a black face"), grounds constituent phrases (e.g., "red bird", "red beak","black face") in the image, and then assigns a score to the noun phrase-bounding box pairs based on how well constituent phrases are grounded in the image. The explanation with the highest cumulative score gets selected as the correct explanation.

Phrase-Critic. The phrase-critic model constitutes our core innovation. Given a set $\{(\mathcal{A}_i, \mathcal{R}_i, s_i)\}$, where $\mathcal{A}_i$ is an attribute phrase, $\mathcal{R}_i$ is the corresponding region (more precisely, visual features extracted from the region), and s_i the region score, our phrase-critic model, $f_{critic}(\{(\mathcal{A}_i, \mathcal{R}_i, s_i)\})$, maps them into a single image relevance score S_r. For a given attribute phrase A_i such as "black beak", we ground (localize) it into a corresponding image region R_i and obtain its localization score s_i, using an off-the-shelf localization model from [10]. It is worth noting that the scores directly produced by the grounding model can not be directly combined with other metrics, such as sentence fluency, because these scores are difficult to normalize across different images and different visual parts. For example, a correctly grounded phrase "yellow belly" may have a much smaller score than the correctly grounded phrase "yellow eye" because a bird belly is less well defined than a bird eye. Henceforth, our phrase-critic model plays an essential role in producing normalized, utilizable and comparable scores. More specifically, given an image I, the phrase-critic model processes the list of $\{(\mathcal{A}_i, \mathcal{R}_i, s_i)\}$ by first encoding each $(\mathcal{A}_i, \mathcal{R}_i, s_i)$ into a fixed-dimensional vector x_{enc} with an LSTM and then applying a two-layer neural network to regress the final score S_r which reflects the overall image relevance of an explanation.

We construct ten negative explanation sentences for each image as we explain in the next section. Each negative explanation sentence (not image relevant) gets paired with a positive explanation (image-relevant). We then train our explanation critic using the following margin-based ranking loss $\mathcal{L}_{\text{rank}}$ on each pair of positive and negative explanations, to encourage the model to give higher scores to positive explanations than negative explanations:

$$\mathcal{L}_{\text{rank}} = \max(0, \underbrace{f_{critic}(\{A_i^n\}, I; \theta)}_{S_r^n} - \underbrace{f_{critic}(\{A_i^p\}, I; \theta)}_{S_r^p} + 1) \tag{1}$$

where A_i^p are matching noun phrase whereas A_i^n are mismatching noun phrases respectively, therefore S_r^p and S_r^m are the scores of the positive and the negative explanations. In the following, we discuss how we construct our negative image-sentence pairs.

Mining and Augmenting Negative Sentences. The simplest way to sample a negative pair is to consider a mismatching ground truth image and sentence pair. However, we find that mismatching sentences are frequently either too different from ground truth sentences (and thus do not provide a useful training signal) or too similar to ground truth sentences, such that both the positive and negative sentence are image relevant. Hence, inspired by a relative attribute paradigm for recognition and retrieval [19], we create negative sentences by flipping attributes corresponding to color, size and objects in attribute phrases. For example, if a ground truth sentence mentions a "yellow belly" and "red head" we might change the attribute phrase "yellow belly" to "yellow beak" and "red head" to "black head". This means the negative sentence still mentions some attributes present in the image, but is not completely correct. We find that creating hard negatives is important when training our self-verification model.

Ranking Explanations. After generating a set of candidate explanations and computing an explanation score, we choose the best explanation based on the score for each explanation. In practice, we find it is important to rank sentences based on both the relevance score S_r and a fluency score S_f (defined as the $\log P(w_{0:T})$). However, we find that first discarding sentences which have a low fluency score, and then choosing the sentence with the highest relevance works better:

$$S = \mathbb{1}\left(\underbrace{\sum_i \log P(w_i | w_{0,\ldots,i-1})}_{S_f} > T\right) \underbrace{f_{critic}(\{A_i\}, I; \theta)}_{S_r} \tag{2}$$

where S_r is the relevance score and S_f is the log probability of a sentence based on the trained explanation model. $\mathbb{1}(\cdot)$ is the indicator function and T is a fluency threshold. Including S_f is important because otherwise the explanation scorer will rank "This is a bird with a long neck, long neck, and red beak" high (if a long neck and red beak are present) even though mentioning "long neck" twice is clearly ungrammatical. Based on experiments on our validation set, we set T equal to negative five.

Grounding Visual Features. Our framework for grounding visual features involves three steps: generating visual explanations, factorizing the sentence into smaller chunks, and localizing each chunk with a grounding model. Visual explanations are generated

with a recurrent neural network (specifically an LSTM [9]) over the image. Unlike standard visual description models, e.g. [4], the visual explanation generation model [8] is conditioned on the class C predicted by the visual model as well as the image I itself. The explanation model relies on two losses: a relevance loss

$$\mathcal{L}_{\mathrm{rel}} = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{t=0}^{T-1} \log p(w_{t+1}|w_{0:t}, I, C) \tag{3}$$

which corresponds to the standard word level softmax cross entropy loss used to train sentence generation models, and a discriminative loss

$$\mathcal{L}_{\mathrm{discr}} = \mathbb{E}_{\tilde{w} \sim p(w|I,C)} \left[R(\tilde{w}) \right] \tag{4}$$

which assigns a high reward (R) to class discriminative features sentences. REIN-FORCE [30] is used to backpropagate through the sampling mechanism required to generate sentences for the discriminative loss. The visual explanation generation loss ($\mathcal{L}_{\mathrm{VEG}}$) is the linear combination of these two losses:

$$\mathcal{L}_{\mathrm{VEG}} = \mathcal{L}_{\mathrm{rel}} - \lambda \mathcal{L}_{\mathrm{discr}}. \tag{5}$$

In order to verify that explanations are image relevant, for each explanation we extract a list of i attribute phrases ($\mathcal{A}_i$) using a rule-based attribute phrase chunker. Our chunker works as follows: we first use a POS tagger, then extract attribute phrases by finding phrases which syntactically match the structure of attribute phrases. We find that attribute phrases have two basic types of syntactic structure: a noun followed by a verb and an adjective, e.g. "bird is black" or "feathers are speckled", or an adjective (or list of adjectives) followed by a noun, e.g. "red and orange head" or "colorful body". Though this syntactic structure is specific to the bird data, similar methods could be used to extract visual phrases for other applications. Attribute phrases are ordered based on the order in which they occur in the generated visual explanation.

Once we have extracted attribute phrases $\mathcal{A}_i$, we ground each of them to a visual region $\mathcal{R}_i$ in the original image by using [10] pre-trained on the Visual Genome dataset [11] without any access to task-specific ground truth. For a given attribute phrase $\mathcal{A}_i$, the grounding model localizes the phrase into an image region, returning a bounding box $\mathcal{R}_i$ and a score s_i of how likely the returned bounding box matches the phrase. The grounding model works in a retrieval manner. It first extracts a set of candidate bounding boxes from the image, and embeds the attribute phrase into a vector. Then the embedded phrase vector is compared with the visual features of each candidate bounding box to get a matching score (s_i). Finally the bounding box with the highest matching score is returned as the grounded image region. The attribute phrase, the corresponding region, and the region score form an attribute phrase grounding ($\mathcal{A}_i, \mathcal{R}_i, s_i$). This attribute phrase grounding is used as an input to our phrase-critic.

Whereas visual descriptions are encouraged to discuss attributes which are relevant to a specific class, the grounding model is only trained to determine whether a natural language phrase is in an image. Being discriminative rather than generative, the critic model does not have to learn to generate fluent, grammatically correct sentences, and can thus focus on checking whether the mentioned attribute phrases are image-relevant.

Consequently, the models are complementary, allowing one model to catch the mistakes of the other.

4   Experiments

In this section, we first detail the datasets we use in our experiments. We then present a comparison between different methods to rank sampled sentences. We compare our model to baselines both qualitatively and through a human evaluation. Additionally, we discuss how our model enables *counterfactual* explanations.

CUB. We validate our approach on the CUB dataset [29] which contains 200 classes of fine-grained bird species with approximately 60 images each and a total of 11,788 images of birds. Recently, [21] collected 10 sentences for each image with a detailed description of the bird. This dataset is suitable for our task as every sentence as well as every image is associated with a single label. Note that CUB does not contain ground truth part bounding boxes, however it contains keypoints that roughly fall on each body part. We use them only to evaluate the precision of our detected bounding boxes. To generate sentences, we use random sampling [4] to sample 100 sentences from the baseline model proposed by [8]. We use the set of 100 sentences generated via random sampling as candidate sentences for phrase-critic. As a "Baseline" model, we select sentences based only on the fluency score S_f. In addition to our phrase-critic, we also consider a strong baseline which only considers the score from a grounding model (without the phrase critic), i.e. we call it "Grounding model". In this case, we first ground noun phrases, then rank sentences by the average score of the grounded noun phrases in the image.

FOIL. Our phrase-critic model is flexible and can also be applied to other relevant tasks. To show the generality of our approach, we also consider the dataset proposed in [24] which consists of sentences and corresponding "FOIL" sentences which have exactly one error. [24] proposes three tasks: (1) classifying whether a sentence is image relevant or not, (2) determining which word in a sentence is not image relevant and (3) correcting the sentence error. To use our phrase-critic for (1), we employ a standard binary classification loss. For (2), we follow [24] and determine which words are not image relevant by holding out one word at a time from the sentence. When we remove an irrelevant word, the score from the classifier should increase. Thus, we can determine the least relevant word in a sentence by observing which word (upon removal) leads to the largest score from our classifier. Also following [24], for the third task we replace the foiled word with words from a set of target words and choose a target word based on which one maximizes the score of the classifier. To train our phrase critic, we use the positive and negative samples as defined by [24]. As is done across all experiments, we extract phrases with our noun phrase chunker and use this as input to the phrase-critic.

4.1   Fine-Grained Bird Species Explanation Experiments

In this section, we conduct detailed bird species explanation experiments on the CUB dataset. We first present comparison with baselines qualitatively both for successful
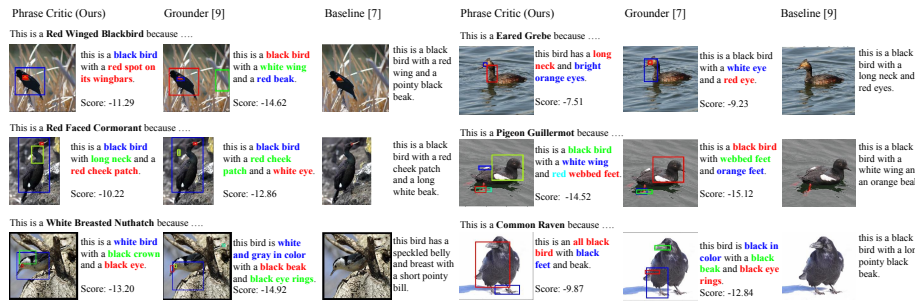
Fig. 3: Our phrase-critic model generates more image-relevant explanations compared to [8] justified by the grounding of the noun phrases. Compared to Grounder [10], our phrase-critic generates more class-specific explanations. The numbers indicate the cumulative score of the explanation computed by our phrase-critic ranker.

and failure cases as well as quantitatively through human judgement. We then present results illustrating the accuracy of the detected bounding boxes. We finally discuss our counterfactual explanation results.

Baseline Comparison. Our aim here is to compare our phrase-critic model with the baseline visual explanation model [8] and the grounding model [10]. In Figure 3, the results on the left are generated by our phrase critic model, the ones in the middle by the grounding model [10] and the ones on the right are by the baseline model [8]. Note that [8] does not contain an attribute phrase grounder, therefore we cannot localize the evidence for the given explanation here. As a general observation, our model improves over both baselines in the following ways. Our critic model (1) grounds attribute phrases both in the image and in the sentence, (2) is in favor of accurate and class-specific noun phrases and (3) provides the cumulative score of each explanatory sentence.

To further emphasize the importance of visual and textual grounding of the noun phrases in evaluating the accuracy of the visual explanation model, let us more closely examine the second row of Figure 3. We note that all models mention a "black bird" and "red cheek patch". As the "Red Faced Cormorant" has these properties, these attributes are accurate. However, the explanation sentence is more trustable when the visual evidence of the noun phrase properly localized, which is not done by the baseline explanation model. To verify our intuition that grounded explanations are more trustable, we ask Amazon Mechanical Turke workers to evaluate whether our explanations with or without bounding boxes are more informative. Our results indicate that bounding boxes are informative (41.9% of the time bounding boxes lead to more informative explanations and 49.3% of the time explanations with and without bounding boxes are equally informative). Therefore, we emphasize that visual and textual grounding is beneficial and important for evaluating the accuracy of the visual explanation model.

Again examining the "Red Faced Cormorant" in the second row of Figure 3, although "red cheek patch" is correctly grounded both by our phrase critic and the baseline phrase grounder, our phrase critic also mentions and grounds an important class-specific attribute of "long neck" while the grounding model mentions a missing "white

This is a **Northern Flicker** because …

… this is a **black and white spotted bird** with **red beak**.

… this bird has a speckled belly and breast with a long pointy bill.

This is a **Pigeon Guillermot** because …

… this is a **black bird with white on the wingbars** and **red feet**.

… this is a black bird with a white wing and an orange beak.

This is a **Pileated Woodpecker** because …

... this is a **black bird** with a **long white neck** and a **red crown**.

… this is a black bird with a white nape and a red crown

Fig. 4: Failure Cases: In some cases our model predicts an incorrect noun phrase and the grounding may reveal the reason. On the other hand, in some cases although the explanation sentences are accurate, the phrases are not grounded well, i.e. the bounding boxes are off. Top: Our phrase-critic, Bottom: Baseline [8].

eye" attribute which it cannot grouned. Thus, the score based ranking of noun phrase and region pairs lead to more accurate and visually grounded visual explanations.

Thanks to the integrated visual grounding capability and phrase ranking mechanism, the critic is able to detect the mistakes of the baseline model and correct them. Some detailed observations from Figure 3 are as follows. "Red Winged Blackbird" having a "red spot on its wingbars" is one of the most discriminative properties of this bird which is mentioned by our critic and also grounded accurately. Similarly, the most important property of "Eared Grebe" is its "red eyes". We see that for "Pigeon Guillermot" our model talks about its "white wing" and "red webbed feet" whereas the grounding model does not mention the "white wing" property and the baseline model does not only ground the phrase but also it does not mention the "red feet". Our model does not only qualitatively generate more accurate explanations, these sentences also get higher cumulative phrase scores as shown beside each image in the figure providing another level of confidence.

Failure Cases. In Figure 4 we present some typical failure cases of our model. In some cases such as the first example, the nouns, i.e. bird and beak, are correctly grounded however the attribute is wrong. Although the bird has a black beak, due to the red color of the fruit it is holding, our model thinks it is a red beak. Another failure case is when the noun phrases are semantically accurate however they are not correctly grounded. For instance, in the second example, both "black bird with white on the wingbars" and "red feet" are correctly identified, however the bounding box of the feet is off. Note that in CUB dataset, the ground truth part bounding box annotations are not available, hence our model figures out the location of a "red feet" by adapting the grounding model trained on Visual Genome, which may not include similar box-phrase combinations. Similarly, in the third example, the orientation of the bounding box of the phrase "long white neck" is inaccurate since the bird is perching on the tree trunk vertically although most of the birds perch on tree branches in a horizontal manner.

Human Evaluation. In this section we discuss the effectiveness of our explanation ranker via human evaluation. We sample 2000 random images from the test set and ask Amazon Mechanical Turkers to annotate whether a noun phrase selected by a model is observed in the image or not. We run this human study three times to eliminate the annotator bias. In this study, we measure the percentage of correct noun phrases, i.e. CNP, and the correct sentences, i.e. CS, that are agreed by at least the two out of

Method	% Correct Noun Phrases	% Correct Sentences
Baseline [8]	76.64	52.10
Grounding model [10]	76.32	49.85
Our Phrase Critic	**77.96**	**61.97**

Table 1: Human evaluations comparing Baseline explanation model [8], Grounding model [10], and our phrase-critic model. CNP: the percentage of correct noun phrases predicted by each model, CS: the percentage of correct sentences where all the phrases are semantically accurate.

three annotators. Our results are presented in Table 1. We first compare our phrase-critic method to the baseline explanation generator [8]. We find that sentences chosen with our phrase critic have improved CNP score (77.96% vs 76.64%) in contrast to our baseline. Next, we find that attributes mentioned by our critic model reflect the images more accurately than the grounding only baseline [10] (77.96% image relevant attributes vs. 76.32%). This result shows that our phrase critic model generates phrases that are more accurate than the baselines as well as being visible in the image.

We also compute the percentage of correct sentences (CS), i.e. all noun phrases appearing in the sentence correctly match the given image, for each model. We observe the same ordering trend between the three models as the CNP scores, i.e. Grounding model < Baseline < Phrase Critic with the phrase-critic producing correct sentences 59.67% of the time in comparison to 51.49% and 47.74% for the baseline and grounder respectively. We note that the performance gain of our phrase-critic model is larger when considering entire sentences, perhaps because our phrase-critic is specifically optimized to discriminate between sentences. These results indicate that our phrase critic model leads to fewer mistakes overall.

Bounding Box Accuracy. As the CUB dataset does not contain ground-truth bounding boxes, we cannot evaluate the precision of our detected part bounding boxes w.r.t. a ground truth. However, the dataset contains keypoints for 15 body parts, e.g. bill, throat, left eye, nape, etc. and utilizing these keypoint annotations that roughly correspond to "beak", "head", "belly" and "eye" regions, provides us a good proxy for this task. We measure how frequently a keypoint falls into the detected bounding box of the corresponding body part to determine the accuracy of the bounding boxes. In addition, we measure the distance of the corresponding keypoint to the center of the bounding box to determine the precision of the bounding boxes. Note that for the results in the first row, we take the explanation generated by [8] and ground the phrases using the off-the-shelf grounding model [10].

Our results in Table 2 show that while "beak", "head" and "belly" regions are detected with high accuracy (95.88%, 74.06% and 66.65% resp.), "eye" detections are weaker (56.72%). When we look at the distance between the bounding box center and the keypoint, we observe a similar trend. The head region gets detected by our model significantly better than others, i.e. 20.26 vs 46.31 with [10] and 57.56 with [8]. The belly and the beak distances are close to the ones measured by the grounding model

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

Table 2: Evaluating the grounding accuracy for four commonly mentioned bird parts. As we have no have access to ground truth boxes, we measure how frequently the ground truth keypoints fall within a detected bounding box, measuring the % of the keypoints that fall inside the bounding box (left) and the distance between the keypoint and the center of the bounding box (right). The Baseline [8] does not include noun phrase grounding, so we apply [10] to noun phrases extracted from [8].

whereas the eye region gets detected with a lower precision with our model compared to the grounding model.

We closely investigate the accuracy of the predicted noun phrases that fall into the eye region and observe that although the eye regions get detected with a higher precision with the baseline grounding model, the semantic meaning of the attribute gets predicted more accurately with our phrase critic. For instance, our model mentions "red eye" more accurately than the grounding model although the part box is more accurately localized by the grounding model. One example of this can be seen in Figure 3 (top right) where the grounder selects the sentences "... this is a black bird with a white eye and a red eye." Here, the grounding model has selected a sentence which cannot be true (the bird cannot have white eye as well as a red eye). Even though the bounding box around the eye is accurate, the modifying attributes are not both correct.

Counterfactual Explanations. Another way of explaining a visual concept is through generating *counterfactual* explanations that indicate why the classifier does not predict another class label. To construct counterfactual explanations, we posit that if an attribute is discriminative for another class, i.e. a class that is different from the class that the query image belongs to, but not present in the query image, then this attribute is a *counterfactual* evidence. To discuss counterfactual evidence for a classification decision, we first hypothesize which visual evidence might indicate that the bird belongs to another class. We do so by considering explanations produced by our phrase-critic for visually most similar examples from a different, i.e. counterfactual, classes. Our phrase-critic determines which attributes are most class specific for the counterfactual class and most image relevant for the query image while generating factual explanations. While generating counterfactual explanations, our model determines the counterfactual evidence by searching for the attributes of the counterfactual class which lead to the lowest phrase-critic score for the query image. We then construct a sentence by negating counterfactual phrases. For instance, "bird has a long flat bill" is negated to "bird does not have a long flat bill" where the counterfactual phrase is the "long flat bill". Alternatively, we can use the same evidence to rephrase the sentence "If this bird had been a (counterfactual class), it would have had a long flat bill."

This bird is a **Crested Auklet** because this is a black bird with a small orange beak and it is not a **Red Faced Cormorant** because it does not have a long flat bill.

This bird is a **Parakeet Auklet** because this is a black bird with a white belly and small feet and it is not a **Horned Grebe** because it does not have red eyes.

This bird is a **Least Auklet** because this is a black and white spotted bird with a small beak and it is not a **Belted Kingfisher** because it does not have a long pointy bill.

This bird is a **White Pelican** because this is a large white bird with a long orange beak and it is not a **Laysan Albatross** because it does not have a curved bill.

This bird is a **Cardinal** because this is a red bird with a black face and it is not a **Scarlet Tanager** because it does not have a black wings.

This bird is a **Yellow Headed Blackbird** because this is a small black bird with a yellow breast and head and it is not a **Prothonotary Warbler** because it does not have a gray wing.
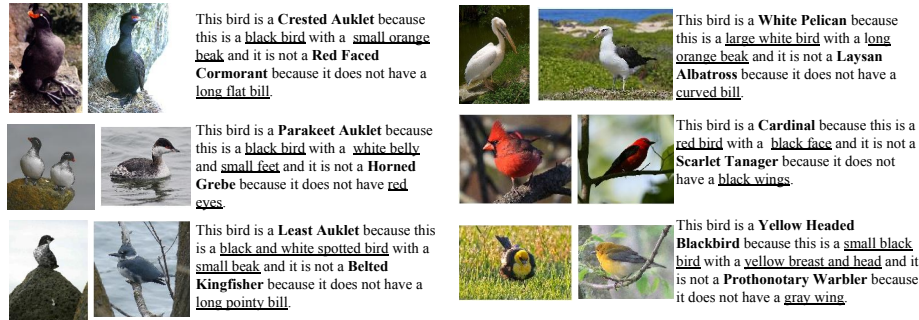
Fig. 5: Our phrase-critic is able to generate factual and counterfactual explanations. Factual explanations mention the characteristic properties of the correct class (left image) and counterfactual explanations mention the properties that are not visible in the image, i.e. non-groundable properties, for the negative class (right image).

To illustrate, we present our results in Figure 5. Note that the figures show two images for each result where the first image is the query image. The second image is the counterfactual image, i.e. the most similar image to the query image from the counterfactual class, that we show only for reference purposes. The counterfactual explanation is generated for this image just for determining the most class-specific noun phrase. Once a list of counterfactual noun phrases is determined, those noun phrases are grounded in the query image and the noun phrase that gets the lowest score is determined as the counterfactual evidence. To illustrate, let us consider an image of a *Crested Auklet* and a nearest neighbor image from another class, e.g., *Red Faced Cormorant*. The attributes "black bird" and "long flat bill" are possible counterfactual attributes for the original crested auklet image. We use our phrase-critic to select the attribute which produces the *lowest* score for the Crested Auklet image.

Figure 5 shows our final counterfactual explanation for why the *Crested Auklet* image is not a *Red Faced Cormorant* (it does not have a long flat bill). On the other hand, when the query image is a *Parakeet Auklet*, the factual explanation talks about "red eyes" which are present for *Horned Grebe* but not for *Parakeet Auklet*. Similarly, a *Least Auklet* is correctly determined to be a "black and white spotted bird" with a "small beak" while a *Belted Kingfisher* is a has a "long pointy bill" which is the counterfactual attribute for *Least Auklet*. On the other hand, a *Cardinal* is classified as a cardinal because of the "red bird" and "black face" attributes while not as a *Scarlet Tanager* because of the lack of "black wings". These results show that our counterfactual explanations do not always generate the same phrases for the counterfactual classes. Our counterfactual explanations talk about properties of the counterfactual class that are not relevant to the particular query image, whose evidence is clearly visible in both the counterfactual and the query images.

In conclusion, counterfactual explanations go one step further in language-based explanation generation. Contrasting a class with another closely related class helps the user build a more coherent cognitive representation of a particular object of interest.

	Classification	Word Detection	Word Correction
IC - Wang [31]	42.21	27.59	22.16
HieCoAtt [16]	64.14	38.79	4.21
Grounding model [10]	56.68	39.80	8.80
Phrase Critic (Ours)	**87.00**	**73.72**	**49.60**

Table 3: Quantitative FOIL results: Our phrase critic significantly outperforms the state-of-the-art [31,16] reported in [24] and the Grounding model [10] on all three FOIL tasks.

Fig. 6: Qualitative FOIL results: We present the image with foil sentence (top) and correct sentence (bottom) as determined by our phrase-critic model. The numbers indicate our phrase-critic score of the given sentence. By design our model grounds all the phrases in the sentence, including the foil phrases.

4.2   FOIL Experiments

Table 3 shows the performance of our phrase critic on the FOIL tasks compared to the best performing models evaluated in [24]. IC-Wang [31] is an image captioning model whereas HieCoAtt [16] is an attention based VQA model. As described above, we follow the protocol of [24] when evaluating our model on the FOIL tasks. To apply the grounding model to the classification task, we determine a threshold score on the train set (i.e., any sentence with an average grounding score above a certain threshold is classified as image relevant).

Our results show that the phrase critic is able to effectively adapt a grounding model in order to determine whether or not sentences are image relevant. We see that our grounding model baseline performs competitively when compared to prior work, indicating that grounding noun phrases is a promising step to determine if sentences are image relevant. However, our phrase critic model outperforms all baselines by a wide margin, outperforming the next best model by over 20 points on the classification task, over 30 points on the word identification task, and close to 30 points on the word correction task. The large gap between the grounding model baseline and the phrase-critic

highlights the importance of our phrase-critic in learning how to properly adapt outputs from a grounding model to our final task.

Figure 6 shows example negative and positive sentences from the FOIL dataset, the grounding determined by our phrase-critic, and the score output by our phrase-critic model. Our general observation is that our phrase critic gives a significantly lower score to FOIL sentences which are not image relevant. In addition, it accurately grounds mentioned objects and accurately scores sentences based on if they are image relevant.

Some detailed observations are as follows. For the first example the score of the FOIL sentence is 0.25 as the sentence contains "a boat" phrase that is inaccurate whereas the sentence with the correct phrase, i.e. "a train", gets the score 0.71 which clearly indicates that this is the correct sentence. Our model is able to ground more than two phrases accurately as well. For the last image in the first row, the phrases "an older man", "green sports coat" and "flower" are correctly predicted and grounded whereas "blue backpack" gets grounded close to the shoulder, which is a sensible region to consider even though there is no backpack in the image. This FOIL sentence gets the score 0.22 whereas the correct sentence that gets the score 0.98 grounds "blue tie" correctly while also correctly grounding all other phrases in the sentence.

When the FOIL object is one of the many objects in the sentence and occupies a small region in the image, our phrase-critic is also successful. For instance in the third image in the first row, "an suitcase" is grounded in an arbitrary location on the side of an image which leads to an extremely low sentence score, 0.06. In the image relevant sentence, "an umbrella" is grounded correctly leading to a high sentence score, 0.99. In conclusion, our phrase critic accurately grounds the phrases when they are present and assigns scores to the matching phrases and bounding boxes that helps us further understand why a model has taken such a decision.

5   Conclusions

We propose a phrase-critic model which measures the image-relevance of a generated explanation sentence. In this framework, we first generate alternative natural language explanations of a fine-grained visual classification decision, then factorize the explanation sentences into a set of visual attributes and visually ground them. Our phrase critic model (1) assigns normalized scores to noun phrases measuring how well they are grounded in the image, (2) ranks the sentences based on the cumulative score, (3) selects the best explanation that is both image and class relevant.

Our experiments on the CUB dataset shows that this grounding approach forces our explanations to refer to the elements that are present in the image and are class-specific. We evaluate our model on the fine-grained bird explanation task both qualitatively and quantitatively. An intuitive alternative for factual explanations, i.e. explaining the correct class, is counterfactual explanations, i.e. explaining why a certain image is not classified as another class. We show that our model is able to generate accurate counterfactual explanations. Our experiments on the FOIL tasks illustrates that our model is able to determine an inaccurate phrase in the sentence, point to the evidence in the image and also correct it significantly better than the competing models.

References

1. Andreas, J., Klein, D.: Reasoning about pragmatics with neural listeners and speakers. arXiv preprint arXiv:1604.00562 (2016)
2. Biran, O., McKeown, K.: Justification narratives for individual classifications. In: Proceedings of the AutoML workshop at ICML. vol. 2014 (2014)
3. Core, M.G., Lane, H.C., Van Lent, M., Gomboc, D., Solomon, S., Rosenberg, M.: Building explainable artificial intelligence systems. In: NCAI (2006)
4. Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. TPAMI (2016)
5. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. arXiv preprint arXiv:1704.03296 (2017)
6. Girshick, R.: Fast R-CNN. In: ICCV (2015)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: ECCV (2016)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
10. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
11. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowd-sourced dense image annotations. arXiv preprint arXiv:1602.07332 (2016)
12. Lane, H.C., Core, M.G., Van Lent, M., Solomon, S., Gomboc, D.: Explainable artificial intelligence for training and tutoring. Tech. rep., DTIC Document (2005)
13. Letham, B., Rudin, C., McCormick, T.H., Madigan, D., et al.: Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics **9**(3), 1350–1371 (2015)
14. Lombrozo, T.: Explanation and abductive inference. The Oxford handbook of thinking and reasoning (2012)
15. Lombrozo, T.: The structure and function of explanations. Trends in Cognitive Science **10**(10) (2006)
16. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances In Neural Information Processing Systems. pp. 289–297 (2016)
17. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016)
18. Pacer, M., Williams, J., Chen, X., Lombrozo, T., Griffiths, T.: Evaluating computational models of explanation using human judgments. Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (2013)
19. Parikh, D., Grauman, K.: Relative attributes. In: ICCV (2011)
20. Park, D.H., Hendricks, L.A., Akata, Z., Schiele, B., Darrell, T., Rohrbach, M.: Multimodal explanations: Justifying decisions and pointing to the evidence. In: CVPR (2018)
21. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: CVPR (2016)

22. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?": Explaining the predictions of any classifier. arXiv preprint arXiv:1602.04938 (2016)
23. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. See https://arxiv.org/abs/1610.02391 v3 **7**(8) (2016)
24. Shekhar, R., Pezzelle, S., Klimovich, Y., Herbelot, A., Nabi, M., Sangineto, E., Bernardi, R.: Foil it! find one mismatch between image and language caption. arXiv preprint arXiv:1705.01359 (2017)
25. Shortliffe, E.H., Buchanan, B.G.: A model of inexact reasoning in medicine. Mathematical biosciences **23**(3), 351–379 (1975)
26. Teach, R.L., Shortliffe, E.H.: An analysis of physician attitudes regarding computer-based clinical consultation systems. Computers and Biomedical Research **14**(6), 542–558 (1981)
27. Van Lent, M., Fisher, W., Mancuso, M.: An explainable artificial intelligence system for small-unit tactical behavior. In: NCAI (2004)
28. Vedantam, R., Bengio, S., Murphy, K., Parikh, D., Chechik, G.: Context-aware captions from context-agnostic supervision. arXiv preprint arXiv:1701.02870 (2017)
29. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200. Tech. rep., California Institute of Technology (2010)
30. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning **8**(3-4), 229–256 (1992)
31. Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., van den Hengel, A.: Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding **163**, 21–40 (2017)
32. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. pp. 2048–2057 (2015)
33. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014)
34. Zintgraf, L.M., Cohen, T.S., Adel, T., Welling, M.: Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595 (2017)