Visually Grounded Comparative Language Generation

Anonymous ACL submission

Abstract

Given two distinct stimuli, humans can compare and contrast them using natural language. The comparative language that arises is grounded in structural commonalities of the subjects. We study the task of generating comparative language in a visual setting, where two images provide the context for the description. This setting offers a new approach for aiding humans in fine grained recognition, where a model *explains* the semantics of a visual space by describing the differences between two stimuli. We collect a dataset of paragraphs comparing pairs of bird photographs, proposing a sampling algorithm that leverages both taxonomic and visual metrics of similarity. We present a novel model architecture for generating comparative language given two images as input, and validate its performance both on automatic metrics and via human comprehension.

1 Introduction

Research at the intersection of natural language processing and computer vision has made great strides in recent years. Datasets and models for tasks such as image captioning (Lin et al., 2014; Xu et al., 2015; Anderson et al., 2018; Sharma et al., 2018) and visual question answering (Antol et al., 2015; Andreas et al., 2016; Shah et al., 2019) have driven progress in both areas.

In addition to broad domain image captioning, context sensitive image captioning has emerged as a compelling research direction. These works caption a single image using additional context. One option is to use class labels as context, with the objective of generating captions that distinguish why the image belongs to one class over others (Hendricks et al., 2016; Vedantam et al., 2017). Another choice is to use a second image as context, and generate a caption that distinguishes one image from another (Vedantam et al., 2017; Jhamtani and Berg-Kirkpatrick, 2018).

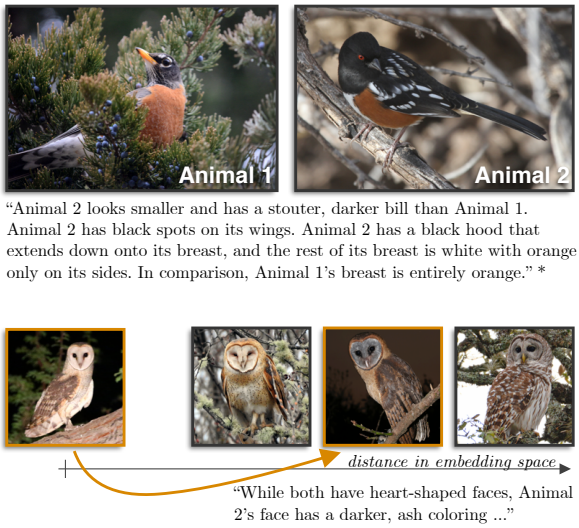

Figure 1: We consider the task of jointly describing two images with a comparative paragraph. TOP: Instance-level descriptions are highly detailed, yet composed of everyday language. BOTTOM: Generated comparisons can serve as a natural language interpretation of deltas in the latent embedding space of neural vision models.

In this work, we consider the task of explaining the difference between two images. Given two images as input, a model must generate a paragraph of comparative language that relates the two images. While closely related to recent research (Jhamtani and Berg-Kirkpatrick, 2018), to our knowledge this is the first work that jointly describes two images with comparative language referencing both images.

Our approach suggests a new technique for explaining visual similarity using natural language. This has implications in fine-grained visual domains where humans require aid to distinguish

* Description adapted from Cornell Lab of Ornithology species comparison at https://allaboutbirds.org.

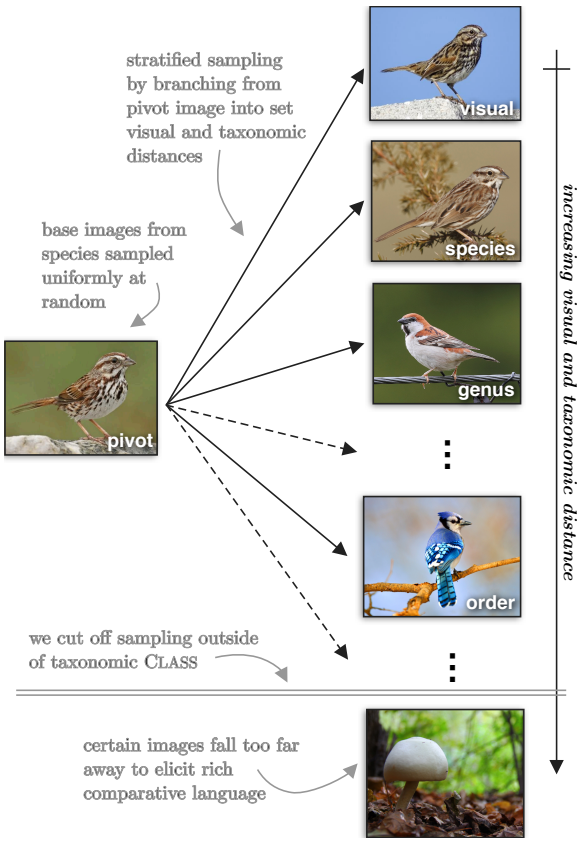

Figure 2: Illustration of spectrum of visual and taxonomic similarity (distance increasing vertically), and of a pivot-branch stratified sampling algorithm layered on top of these metrics.

closely related instances.

Our contributes are as follows. We formulate the task of describing image differences using comparative language and present a new dataset for this challenge. This dataset features fine-grained distinctions of adaptive granularity, uses everyday language, and requires only instance-level judgments.[1] We propose a model that uses multi-headed attention to capture differences in image space and describe them in natural language. Finally, we perform a human evaluation in addition to standard metrics to measure whether generated descriptions contain comprehensible semantics.

2 Explanations of Visual Similarity

The task of generating descriptions between pairs of related images can be partially motivated by a desire for model explainability. Previous work in this area largely focuses on attributing model decisions to input features (Xu et al., 2015; Shrikumar et al., 2016; Lei et al., 2016; Sundararajan et al., 2017). Additional insights may be gained from a model that can explain differences in a shared latent space—for example, visual embeddings.

However, our principal motivation for this task is to assist humans in distinguishing visually similar classes. Take, for example, a citizen science effort like iNaturalist,[2] where everyday people photograph wildlife, and the community reaches a consensus on the taxonomic label for each instance. Many species are visually similar (e.g., Figure 1), making them difficult for a casual observer to label correctly. This puts an undue strain on lieutenants of the citizen science community to curate and justify labels for a large number of instances. While everyone may be capable of making such distinctions visually, non-experts require training to know what to look for.

Field guides exist for the purpose helping people learn how to distinguish between species. Unfortunately, field guides are costly to create. Writing such a guide requires knowledge of class-level distinctions. Experts in a particular domain of biology must craft a guide to individually distinguish each potentially confounding pair of species.

Our approach outlines a new way for models to aid humans in making visual distinctions. The model requires two instances as input; these could be, for example, a query image and an image from a candidate class. By differentiating between these two inputs, a model may help point out subtle distinctions (e.g., one animal has spots on its side), or features that indicate a good match (e.g., only a slight difference in size). These explanations can aid in understanding both differences between species, as well as variance within instances of a single species.

Captioning a single image with a class-discriminative objective is an important related task, which adds significant value to this area (Hendricks et al., 2016; Vedantam et al., 2017). We argue that directly comparing and contrasting two images simultaneously provides complementary means for humans to better understand instance-level distinctions present in complex visual domains.

3 Dataset Construction

Our goal is to collect a dataset of tuples (i_1, i_2, t), where i_1 and i_2 are images, and t is a natural lan-

[1] We will release this dataset upon publication.

[2] https://www.inaturalist.org

Proposed Dataset	
Image pairs	3,347
Paragraphs / pair	4.8
Paragraphs	16,067
Sentences	40,969
Sentences / paragraph	2.6 MEAN
Tokens / paragraph	32.1 MEAN
Clarity rating	$\geq 4/5$
Train / dev / test	80% / 10% / 10%

Table 1: Statistics for the proposed dataset. Image pairs are two photographs of birds drawn from iNaturalist. Each pair is annotated with five natural language paragraphs describing the differences between the animals.

guage comparison between the two. Given a domain $\mathcal{D}$, this collection depends in an important way on the criteria we use to select image pairs.

If we sample image pairs uniformly at random, we will end up with comparisons encompassing a broad range of phenomena. For example, two images that are quite different will yield categorical comparisons (*"One is a bird, one is a mushroom."*). Alternatively, if the two images are very similar, such as two angles of the same creature, comparisons between them will focus on highly detailed nuances, such as variations in pose. These phenomena support rich lines of research, such as object classification (Deng et al., 2009) and pose estimation (Murphy-Chutorian and Trivedi, 2009).

We aim to land somewhere in the middle. We wish to consider sets of distinguishable but intimately related pairs. This sweet spot of visual similarity is akin to the genre of differences studied in fine-grained visual classification (Wah et al., 2011; Krause et al., 2013). We approach this collection with a two-phase data sampling procedure. We first select *pivot* images by sampling from our full domain uniformly at random. We then *branch* from these images into a set of secondary images that emphases fine-grained comparisons, but yields broad coverage over the set of sensible relations. Figure 2 provides an illustration of our sampling procedure.

3.1 Domain

We sample images from iNaturalist, a citizen science effort to collect research-grade[3] observations of plants and animals in the wild. We restrict our domain $\mathcal{D}$ to instances labeled under the taxonomic CLASS[4] *Aves* (i.e., birds). While a broader domain would yield some comparable instances (e.g., *bird* and *dragonfly* share some common body parts), choosing only *Aves* ensures that all instances will be similar enough structurally to be comparable, and avoids the gut reaction comparison pointing out the differences in animal type. This choice yields 1.7M research-grade images and corresponding taxonomic labels from iNaturalist. We then perform pivot-branch sampling on this set to choose pairs for annotation.

3.2 Pivot images

The *Aves* domain in iNaturalist contains instances of 9k distinct species, with heavy observation bias to more common species (such as the mallard duck). We uniformly sample species from the set of 9k to help overcome this bias. For each species, we manually review four images sampled from this species to select the clearest image to use as the pivot image. In total, we select 405 species and corresponding photographs to use as i_1 images.

3.3 Branching images

We use both a visual similarity measure and taxonomy to sample a set of comparison images i_2 branching off from each pivot image i_1. We use a branching factor of $k = 12$ from each pivot image.

To capture visually similar images to i_1, we employ a similarity function $\mathcal{V}(i_1, i_2)$. We use an Inception-v4 (Szegedy et al., 2017) network pretrained on ImageNet (Deng et al., 2009) and then fine-tuned to perform species classification on all research-grade observations in iNaturalist. We take the embedding for each image from the last layer of the network before the final softmax. We perform a k-nearest neighbor search by quantizing each embedding and using L2 distance (Wu et al., 2017; Guo et al., 2016), selecting the $k_v = 2$ closest images in embedding space.

We also use a taxonomy $\mathcal{T}(\mathcal{D})$ to sample images at varying levels of taxonomic distance from i_1. We take as $\mathcal{T}(\mathcal{D})$ the scientific taxonomy used in iNaturalist. For the class c corresponding to image i_1, we split the taxonomic tree into disjoint subtrees rooted $\ell \in \{1..5\}$ taxonomic levels above c. Each higher level excludes the levels beneath it. For example, at $\ell = 1$ we consider all images

[3] Research-grade observations have met or exceeded iNaturalist's guidelines for community consensus of the taxonomic label for a photograph.

[4] To disambiguate *class*, we use CLASS to denote the taxonomic rank in scientific classification, and simply "class" to refer to the machine learning usage of the term as a label in classification.

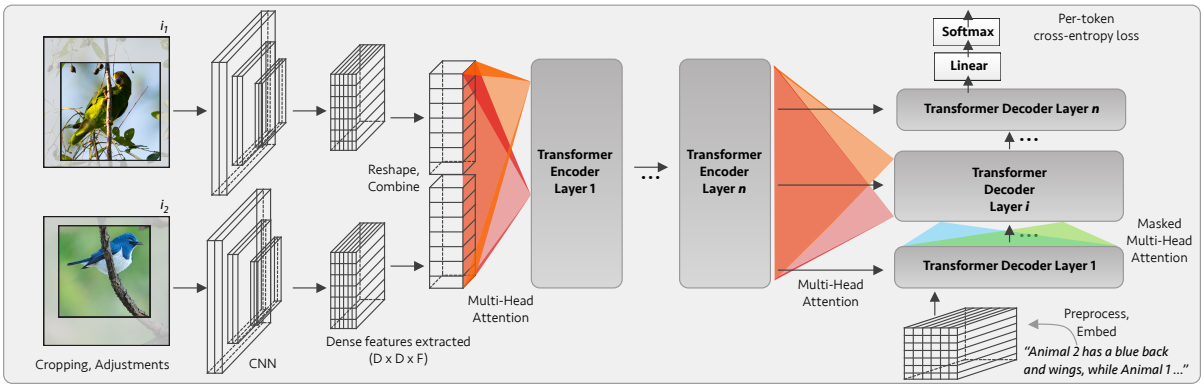

Figure 3: The proposed model architecture. Multi-head attention is applied across both image representations during encoding, as well as across the resulting representation and input tokens during decoding. This provides representational power to generate comparisons between both inputs.

of the same species as i_1; at $\ell = 2$, we consider all images of the same genus as i_1, but that have a *different* species. We select $k_t = 10$ taxonomically branched images by sampling two images each from the same SPECIES ($\ell = 1$), GENUS, FAMILY, ORDER, and CLASS ($\ell = 5$) as c. This yields 4,860 raw image pairs (i_1, i_2).

3.4 Language Collection

For each image pair (i_1, i_2), we elicit five natural language paragraphs describing the differences between them. Annotators first label whether i_1 and i_2 are "clear," meaning they contain only one animal, and the animal is photographed with good visibility (not too small or blurry). Note that we have manually verified each i_1 is clear, but each i_2 must be vetted.

An annotator is instructed to write a paragraph (usually 2–5 sentences) comparing and contrasting the animal appearing in each image. We instruct annotators not to explicitly mention the species (e.g., *"Animal 1 is a penguin"*), and to instead focus on visual details (e.g., *"Animal 1 has a black body and a white belly"*). They are additionally instructed to avoid mentioning aspects of the background, scenery, or pose captured in the photograph (e.g., *"Animal 2 is perched on a coconut"*). We vet each annotator individually by manually reviewing five reference annotations from a pilot round, and perform random quality assessments during data collection.

We discard all annotations for an image pair where either image did not have at least $\frac{4}{5}$ positive ratings of image clarity. This yields a total of 3,347 image pairs, annotated with 16,067 paragraphs. Detailed statistics of the dataset are shown in Table 1, and examples are provided in Figure 4. Further details of our algorithmic approach to image pair selection are given in the supplementary material.

4 Model

Recent image captioning approaches (Xu et al., 2015; Sharma et al., 2018) extract image features using a convolutional neural network (CNN) which serve as input to a language generator, typically a recurrent neural network (RNN) (Mikolov et al., 2010) or Transformer (Vaswani et al., 2017). We extend the Transformer-based architecture with a comparative encoder that produces self-attentive mutations of multiple latent image embeddings. A schematic of the model is outlined in Figure 3, and its three key components are described in the upcoming sections.

4.1 Image Feature Extractor

Both input images are first processed using CNNs with shared weights. In this work, we consider ResNet (He et al., 2016) and Inception-v4 (Szegedy et al., 2017) architectures. In both cases, we extract the representation from the deepest layer immediately before the classification layer. This yields a dense 2D grid of local image feature vectors, shaped (d, d, f). For our two models variants, $f = 2048$, ResNet $d = 7$, Inception-v4 $d = 5$. This dense feature extraction is intended to build a spatially-rich representation of each image.

We then flatten each feature grid and combine both into a (d^2, f) shaped matrix:

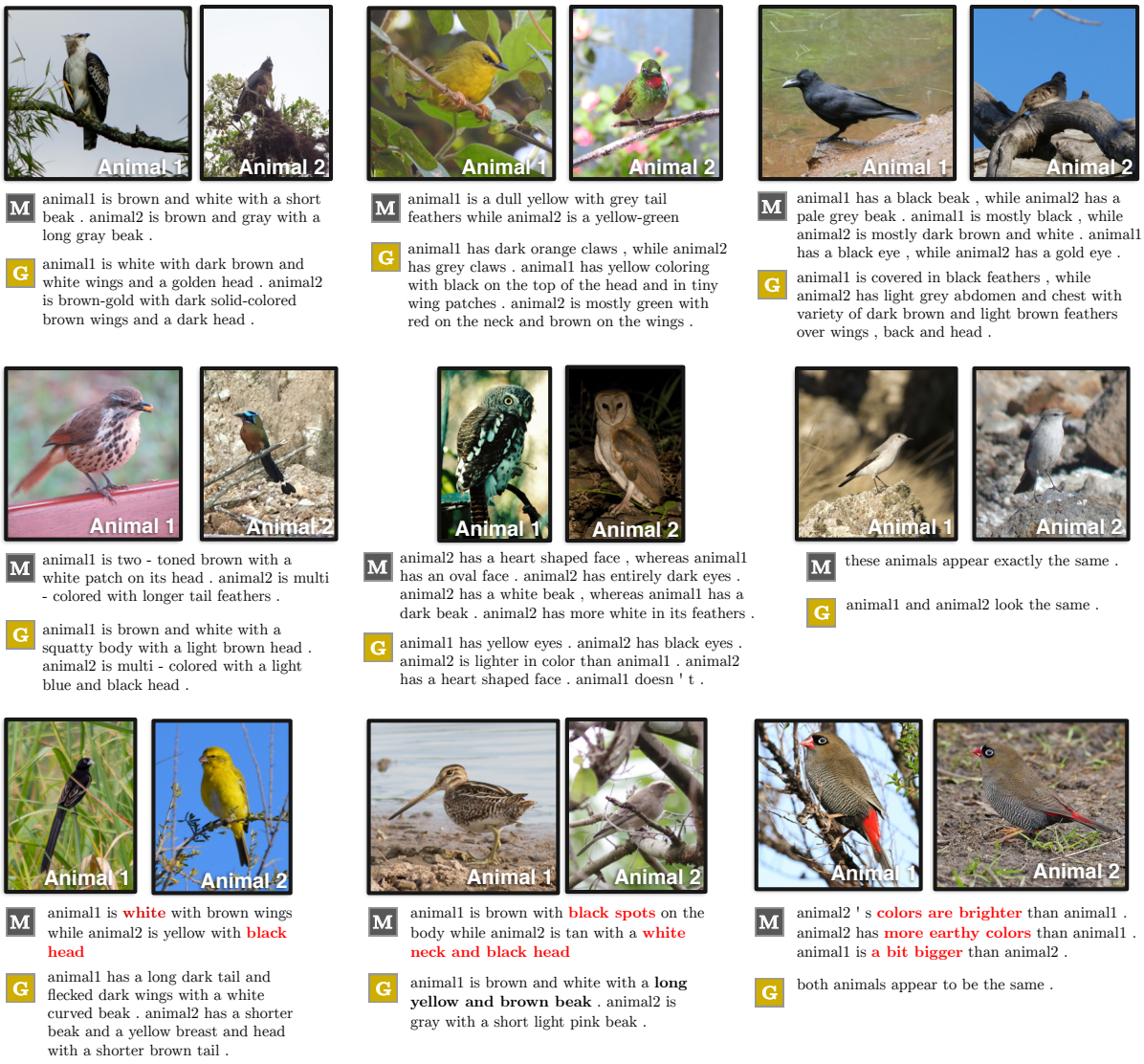

Figure 4: Samples from the dev split of the proposed dataset, along with model output (M) and one of the five ground truth paragraphs (G). The final row shows failure cases, highlighted in red.

$$\mathbf{E}^1 = \langle \mathbf{e}^1_{1,1}, \ldots, \mathbf{e}^1_{d,d} \rangle = \text{CNN}(i_1)$$
$$\mathbf{E}^2 = \langle \mathbf{e}^2_{1,1}, \ldots, \mathbf{e}^2_{d,d} \rangle = \text{CNN}(i_2)$$
$$\mathbf{E} = \langle \mathbf{e}^1_{1,1}, \ldots, \mathbf{e}^1_{d,d}, \mathbf{e}^2_{1,1}, \ldots, \mathbf{e}^2_{d,d} \rangle$$

The matrix $\mathbf{E}$ forms the initial embedding of the images (i_1, i_2), which we then pass to the comparative encoder.

4.2 Comparative Encoder

Given dense feature representations of the input image pair, we want to learn similarities and differences of the feature spaces in order to generate contrastive descriptions. We use an N-layer Transformer encoder to modify these latent representations. Each layer contains multi-headed attention (ATTN_{MH}). Self-attention in Transformer encoder layers guides comparisons across the feature space of both images.

Denoting LN as *Layer Norm* and FF as *Feed Forward*, each layer of the Transformer encoder computes:

$$\mathbf{H}_1 = \text{LN}(\mathbf{X} + \text{ATTN}_{\text{MH}}(\mathbf{X}))$$
$$\mathbf{H}_2 = \text{LN}(\mathbf{H}_1 + \text{FF}(\mathbf{H}_1))$$

where $\mathbf{X}$ is the input to the layer, and $\mathbf{H}_2$ is the layer's output. In our model, the first layer's input $\mathbf{X}$ is full image representation $\mathbf{E}$. Each subsequent

layer takes $\mathbf{H}_2$ from the previous layer as its input $\mathbf{X}$. We denote the final output from the encoder as $\mathbf{H}$, which is simply the final layer's $\mathbf{H}_2$.

4.3 Decoder

We decode using an N-layer Transformer decoder architecture. Because we decode into text, this module more closely follows the original definition of the Transformer decoder. A Transformer decoder is similar to an encoder, but it contains an intermediary multi-headed attention which has access to the encoder's output at every time step.

$$\mathbf{D}_1 = \text{LN}(\mathbf{X} + \text{ATTN}_{\text{MASK,MH}}(\mathbf{X}))$$
$$\mathbf{D}_2 = \text{LN}(\mathbf{D}_1 + \text{ATTN}_{\text{MH}}(\mathbf{D}_1, \mathbf{E}))$$
$$\mathbf{D}_3 = \text{LN}(\mathbf{D}_2 + \text{FF}(\mathbf{D}_2))$$

Similar to the encoder, each layer of the decoder takes as its input $\mathbf{X}$ the output from the previous layer $\mathbf{D}_3$, and the final output is the output ($\mathbf{D}_3$) of the final layer. Aside from the additional multi-head attention that also uses the encoder's output $\mathbf{E}$, the main difference from the encoder is the treatment of the input $\mathbf{X}$. During training, the initial $\mathbf{X}$ comes from text, so it is first modulated with a position-based encoding. The multi-head attention that observes the $\mathbf{X}$ is masked ($\text{ATTN}_{\text{MASK,MH}}$) so that earlier tokens do not depend on later ones.

5 Experiments

We train our proposed model to produce descriptions of the differences between images in our dataset.

5.1 Task Setup and Model Details

We partition our dataset into train (80%), val (10%), and test (10%) sections by splitting based on the pivot images i_1. The text is preprocessed using standard techniques (tokenization, lowercasing), and we additionally replace mentions referring to each image with special tokens ANIMAL1 and ANIMAL2.

For the image embedding component of our model, we use an Inception-v4 network as our CNN. For both the Transformer encoder and decoder, we use $N = 6$ layers, a hidden size of 512, 8 attention heads, and dot product self-attention. Each paragraph is clipped at 64 tokens during training (chosen empirically to cover 94% of paragraphs). For inference, we experiment with greedy decoding, multinomial sampling, and beam search. Beam search performs best, so we use it with a beam size of 5 for all reported results. We provide model optimization details in the supplementary material.

5.2 Baselines

We consider several baselines and model ablations. Most frequent paragraph (*Freq.*) produces only the most observed description in the training data, which is that the two animals appear to be exactly the same. *Text-only* samples captions from the training data according to their empirical distribution. Our ablations include a variant without the transformer encoder over the image features (*no encoder*), our model early-stopped during training at its highest CIDEr-D score on the development set (*early stopping*), and our model trained until metrics stabilized (*full*).

5.3 Quantitative Results

Automatic Metrics We evaluate our model using three machine-graded text metrics: BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and CIDEr-D (Vedantam et al., 2015). Each generated paragraph is compared to all five reference paragraphs.

For human performance, we use a one-vs-rest scheme to hold one reference paragraph out and compute its metric using the other four. We average this score across twenty-five runs over the entire split in question.

Results using these metrics are given in Table 2. We observe improvement across BLEU-4 and ROUGE-L scores compared to our baselines and model ablations. Curiously, we observe that the CIDEr-D metric is susceptible to common patterns in the data; our model, when stopped at its highest CIDEr-D score, outputs a variant of, *"these animals appear exactly the same"* for 95% of paragraphs, nearly mimicking the behavior of the most frequent paragraph (*Freq.*) baseline. The corpus-level behavior of CIDEr-D gives these outputs a higher score. We observed anecdotally higher quality outputs correlated with ROUGE-L score.

Human Evaluation To verify our observations about model quality, we also perform a human evaluation of the generated paragraphs. We sample 120 instances from the test set, taking twenty

	Dev			Test		
	BLEU-4	ROUGE-L	CIDEr-D	BLEU-4	ROUGE-L	CIDEr-D
Freq.	0.20	0.31	0.42	0.20	0.30	0.43
Text-Only	0.14	0.36	0.05	0.14	0.36	0.07
Ours – Early Stopping	0.21	0.33	0.42	–	–	–
Ours – No Encoder	0.19	0.40	0.17	–	–	–
Ours – Full	0.21	0.43	0.19	0.22	0.43	0.21
Human	0.26 +/- 0.02	0.47 +/- 0.01	0.39 +/- 0.04	0.27 +/- 0.01	0.47 +/- 0.01	0.42 +/- 0.03

Table 2: Experimental results for comparative paragraph generation on the proposed dataset. For human captions, mean and standard deviation are given for a one-vs-rest scheme across twenty-five runs.

each from the six categories for choosing comparative images (visual similarity in embedding space, plus five taxonomic distances). We provide annotators with the two images in a random order, along with the output from the model at hand. Annotators must decide which image contains *Animal 1*, and which contains *Animal 2*, or they may say that there is no way to tell (e.g., for a description like *"both look exactly the same"*).

We collect three annotations per datum, and score a decision only if $\geq$ 2/3 annotators made that choice. A model receives +1 point if annotators decide correctly, 0 if they cannot decide or agree there is no way to tell, and -1 point if they decide incorrectly (confidently label the images backwards). This scheme penalizes a model for confidently writing incorrect descriptions. The total score is then normalized to the range $[-1, 1]$. Note that *Human* uses one of the five gold paragraphs sampled at random.

Results for this experiment are shown in Table 3. In this measure, we see the frequency and text-only baselines now fall flat, as expected. The frequency baseline never receives any points, and the text-only baseline is often penalized for incorrectly guessing. Our model is successful at making distinctions between visually distinct species (GENUS column and ones further right), which is near the challenge level of current fine-grained visual classification tasks. However, it struggles on the two data subsets with highest visual similarity (VISUAL, SPECIES). The significant gap to human performance in these columns indicates ultra fine-grained distinctions are still possible for humans to describe, but a challenge for current models to capture.

5.4 Qualitative Analysis

In Figure 4, we present several examples of the model output for pairs of images in the dev set, along with one of the five reference paragraphs. In the following section, we split an analysis of the model into two parts: largely positive findings, as well as common error cases.

Positive findings We find that the model exhibits *dynamic granularity*, by which we mean that it adjusts the magnitude of the descriptions based on the scale of differences between the two animals. If two animals are quite similar, it generates fine-grained descriptions such as, *"Animal 2 has a slightly more curved beak than Animal 1,"* or *"Animal 1 is more iridescent than Animal 2."* If instead the two animals are very different, it will generate text describing larger-scale differences, like, *"Animal 1 has a much longer neck than Animal 2,"* or *"Animal 1 is mostly white with a black head. Animal 2 is almost completely yellow."*

We also observe that the model is able to produce coherent paragraphs of *varying linguistic structure.* These include a range of comparisons set up across both single and multiple sentences. For example, one it generates straightforward comparisons of the form, *Animal 1 has X, while Animal 2 has Y.* But it also generates contrastive expressions with longer dependencies, such as *Animal 1 is X, Y, and Z. Animal 2 is very similar, except W.* Furthermore, the model will mix and match different comparative structures within a single paragraph.

Finally, in addition to varying linguistic structure, we find the model is able to produce *coherent semantics* through a series of statements. For example, consider the following full output: *"Animal 1 has a very long neck compared to Animal 2. Animal 1 has shorter legs than Animal 2. Animal 1 has a black beak, Animal 2 has a brown beak. Animal 1 has a yellow belly. Animal 2 has darker wings than Animal 1."* The range of concepts in the output covers *neck, legs, beak, belly,*

	VISUAL	SPECIES	GENUS	FAMILY	ORDER	CLASS
Freq.	0.00	0.00	0.00	0.00	0.00	0.00
Text-Only	0.00	-0.10	-0.05	0.00	0.15	-0.15
Ours	0.10	-0.10	0.35	0.40	0.45	0.55
Human	0.55	0.55	0.85	1.00	1.00	1.00

Table 3: Human evaluation results on 120 test set samples, twenty per column. Scale: -1.00 (perfectly wrong) to 1.00 (perfectly correct). Columns denote commonality, and are ordered left-to-right by increasing visual and taxonomic difference. For example, GENUS are pairs of different species with the same genus.

wings without repeating any topic or getting sidetracked.

Error analysis We also observe several patterns in the model's shortcomings. The most prominent error case is that the model will sometimes hallucinate differences that do not exist (Figure 4, bottom row). These range from pointing out significant changes that are missing (e.g., *"a black head"* where there is none), to clawing at subtle distinctions where there are none (e.g., *"[its] colors are brighter . . . and it] is a bit bigger"*). We suspect that the model has learned some associations between common features in animals, and will sometimes favor these associations over visual evidence.

The second common error case is missing obvious distinctions. This is observed in the bottom row, middle example of Figure 4, where the extremely prominent beak of Animal 1 is completely ignored by the model in favor of mundane details. While outlying features make for lively descriptions, we hypothesize that the model may sometimes avoid taking them into account given its per-token cross entropy learning objective.

Finally, we also observe that the model will sometimes swap which features are attributed to which animal. This is partially observed in Figure 4 (bottom row, left side), where the *"black head"* actually belongs to Animal 1, not Animal 2. We suspect that mixing up references may be a trade-off for the representational power of attending over both images; there is no explicit bookkeeping mechanism to enforce which phrases refer to which feature comparisons in each image.

6 Related Work

Employing visual comparisons to elicit focused natural language observations was proposed by (Maji, 2012), and later investigated in the context of crowdsourcing by (Zou et al., 2015). We take inspiration from these works.

Previous work has collected natural language datasets captioning photographs of birds: CUB Captions (Reed et al., 2016) and CUB-Justify (Vedantam et al., 2017) are both language annotations on top of the CUB-2011 dataset of bird photographs (Wah et al., 2011). In addition to describing two photos instead of one, the language in our dataset is more complex by comparison, containing a diversity of comparative structures and implied semantics. We also collect our data without an anatomical guide for annotators, yielding everyday language in place of scientific terminology.

Conceptually, our paper offers a complementary approach to works that generate single-image class-discriminative or image-discriminative captions (Hendricks et al., 2016; Vedantam et al., 2017). Rather than discriminative captioning, we focus on comparative language as a means for bridging the gap between varying granularities of visual diversity.

Methodologically, our work is most closely related to the Spot-The-Diff dataset (Jhamtani and Berg-Kirkpatrick, 2018). Their dataset captions two images with only a small section of pixels that change (surveillance footage), and principally caption a single image with respect to another. In contrast, ours describes image pairs which have no pixel overlap, and our descriptions cover both images together. Also contemporary with our work is NLVR2 (Suhr et al., 2018), which introduces a challenging natural language reasoning task using two images as context. Our work instead focuses on generating comparative language rather than reasoning.

7 Conclusion

We present a new approach to generating natural language explanations of visual spaces with comparative language. This line of research may provide assistance to humans in fine-grained classification domains like citizen science. We hope that our proposed dataset and models for this task will motivate new work studying the use of natural language in understanding and describing fine-grained visual distinctions.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database.

Ruiqi Guo, Sanjiv Kumar, Krzysztof Choromanski, and David Simcha. 2016. Quantization based fast inner product search. In *Artificial Intelligence and Statistics*, pages 482–490.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Subhransu Maji. 2012. Discovering a lexicon of parts and attributes. In *European Conference on Computer Vision*, pages 21–30. Springer.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Erik Murphy-Chutorian and Mohan Manubhai Trivedi. 2009. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. *arXiv preprint arXiv:1902.05660*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2556–2565.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

Alane Suhr, Stephanie Zhou, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *CoRR*, abs/1811.00491.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset.

Xiang Wu, Ruiqi Guo, Ananda Theertha Suresh, Sanjiv Kumar, Daniel N Holtmann-Rice, David Simcha, and Felix Yu. 2017. Multiscale quantization for fast similarity search. In *Advances in Neural Information Processing Systems*, pages 5745–5755.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

James Y Zou, Kamalika Chaudhuri, and Adam Tauman Kalai. 2015. Crowdsourcing feature discovery via adaptively chosen comparisons. *arXiv preprint arXiv:1504.00064*.