Neural Naturalist: Generating Fine-Grained Image Comparisons

Anonymous EMNLP-IJCNLP submission

Abstract

We introduce the new *Birds-to-Words* dataset of 41k sentences describing fine-grained differences between photographs of birds. The language collected is highly detailed, while remaining understandable to the everyday observer (e.g., "heart-shaped face," "squat body"). Paragraph-length descriptions naturally adapt to varying levels of taxonomic and visual distance—drawn from a novel stratified sampling approach—with the appropriate level of detail. We propose a new model called *Neural Naturalist* that uses a joint image encoding and comparative module to generate comparative language, and evaluate the results with humans who must use the descriptions to distinguish real images.

Our results indicate promising potential for neural models to explain differences in visual embedding space using natural language, as well as a concrete path for machine learning to aid citizen scientists in their effort to preserve biodiversity.

1 Introduction

Humans are adept at making fine-grained comparisons, but sometimes require aid in distinguishing visually similar classes. Take, for example, a citizen science effort like iNaturalist,[1] 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.

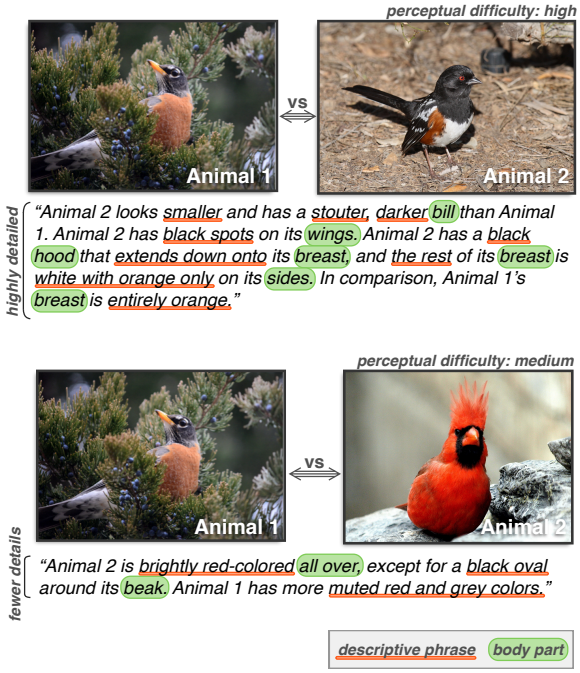


Figure 1: The Birds-to-Words dataset: comparative descriptions adapt naturally to the appropriate level of detail (orange underlines). A difficult distinction (TOP) is given a longer and more fined-grained comparison than an easier one (BOTTOM). Annotators organically use everyday language to refer to parts (green highlights).

Field guides exist for the purpose helping people learn how to distinguish between species. Unfortunately, field guides are costly to create because writing such a guide requires expert knowledge of class-level distinctions.

In this paper, we study the problem of explaining the differences between two images using natural language. We introduce a new dataset called *Birds-to-Words* [2] of paragraph-length descriptions of the differences between pairs of bird photographs. We find several benefits from eliciting comparisons: (a) without a guide, annota-

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

[2] We will release this dataset upon publication.

tors naturally break down the subject of the image (e.g., a bird) into pieces understood by the everyday observer (e.g., head, wings, legs); (b) by sampling comparisons from varying visual and taxonomic distances, the language exhibits naturally adaptive granularity of detail based on the distinctions required (e.g., "red body" vs "tiny stripe above its eye"); (c) in contrast to requiring comparisons between categories (e.g., comparing one species vs. another), non-experts can provide high-quality annotations without needing domain expertise.

We also propose the *Neural Naturalist* model architecture for generating comparisons given two images as input. After embedding images into a latent space with a CNN, the model combines the two image representations with a joint encoding and comparative module before passing them to a Transformer decoder. We find that introducing a comparative module—an additional Transformer encoder—over the combined latent image representations yields better generations.

Our results suggest that these classes of neural models can assist in fine-grained visual domains when humans require aid to distinguish closely related instances. Non-experts—such as amateur naturalists trying to tell apart two species—stand to benefit from comparative explanations. Our work approaches this sweet-spot of visual expertise, where any two in-domain images can be compared, and the language is detailed, adaptive to the types of differences observed, and still understandable by laypeople.

Recent work has made impressive progress on context sensitive image captioning. One direction of work uses 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. Previous work has studied ways to generalize single-image captions into comparative language (Vedantam et al., 2017), as well as comparing two images with high pixel overlap (e.g., surveillance footage) (Jhamtani and Berg-Kirkpatrick, 2018). Our work complements these efforts by studying directly comparative, everyday language on image pairs with no pixel overlap.

Our approach outlines a new way for models to aid humans in making visual distinctions. The

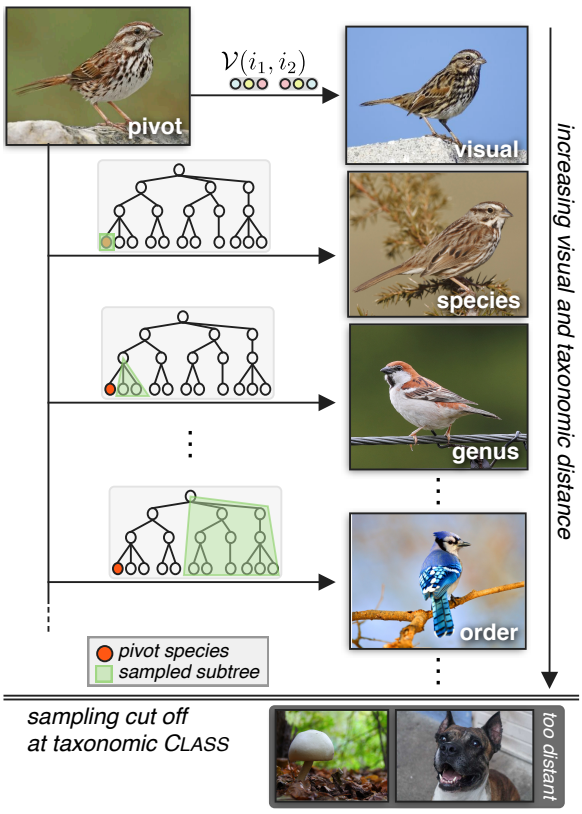


Figure 2: Illustration of pivot-branch stratified sampling algorithm used to construct the Birds-to-Words dataset. The algorithm harnesses visual and taxonomic distances (increasing vertically) to create a challenging task with board coverage.

Neural Naturalist 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.

2 Birds-to-Words Dataset

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 language comparison between the two. Given a domain $\mathcal{D}$, this collection depends critically 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 mush-*

Dataset	Domain	Lang	Images Ctx	Images Cap	Example
CUB Captions *(R, 2016)*	Birds	M	1	1	*"An all black bird with a very long rectrices and relatively dull bill."*
CUB-Justify *(V, 2017)*	Birds	S	7	1	*"The bird has white orbital feathers, a black crown, and yellow tertials."*
Spot-the-Diff *(J&B, 2018)*	Surveilance	E	2	1–2	*"Silver car is gone. Person in a white t shirt appears. 3rd person in the group is gone."*
Birds-to-Words *(this work)*	Birds	E	2	2	*"Animal1 is gray, while animal2 is white. Animal2 has a long, yellow beak, while animal1's beak is shorter and gray. Animal2 appears to be larger than animal1."*

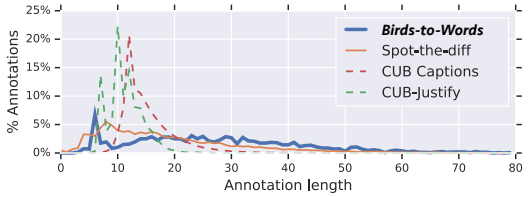
Table 1: Comparison with recent fine-grained language-and-vision datasets. *Lang* values: S = scientific, E = everyday, M = mixed. *Images Ctx* = number of images shown, *Images Cap* = number of images described in caption. Dataset citations: *R* = *Reed et al.*, *V* = *Vedantam et al.*, *J&B* = *Jhamtani and Berg-Kirkpatrick*.

room."). 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.

2.1 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.



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

Figure 3: Annotation lengths for compared datasets (TOP), and statistics for the proposed Birds-to-Words dataset (BOTTOM). The Birds-to-Words dataset has a large mass of long descriptions in comparison to related datasets.

2.2 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. In total, we select 405 species and corresponding photographs to use as i_1 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.

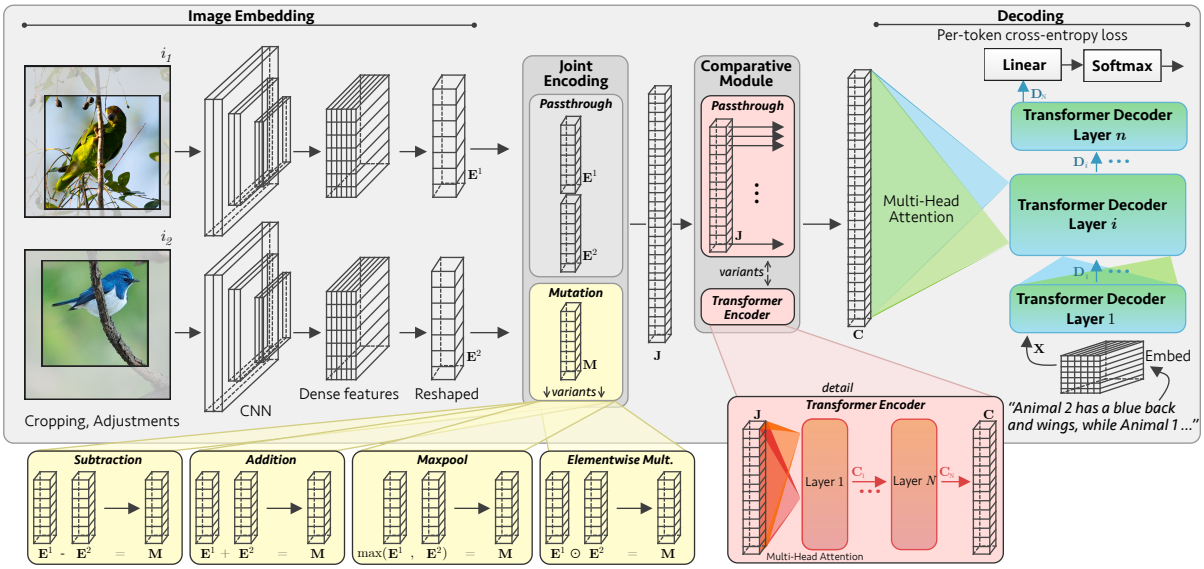


Figure 4: The proposed Neural Naturalist model architecture. The multiplicative joint encoding and Transformer-based comparative module yield the best comparisons between images.

2.3 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 the iNaturalist scientific taxonomy $\mathcal{T}(\mathcal{D})$ to sample images at varying levels of taxonomic distance from i_1. 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).

2.4 Language Collection

For each image pair (i_1, i_2), we elicit five natural language paragraphs describing the differences between them. 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 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 Birds-to-Words dataset are shown in Figure 3, and examples are provided in Figure 5. Further details of our algorithmic approach to image pair selection are given in the supplementary material.

3 Neural Naturalist 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 decoder, typically a recurrent neural network (RNN) (Mikolov et al., 2010) or Transformer (Vaswani et al., 2017). We extend this paradigm with a joint encoding step and comparative module to study how best to encode and transform multiple latent image embeddings. A schematic of the model is outlined in Figure 4, and its key components are described in the upcoming sections.

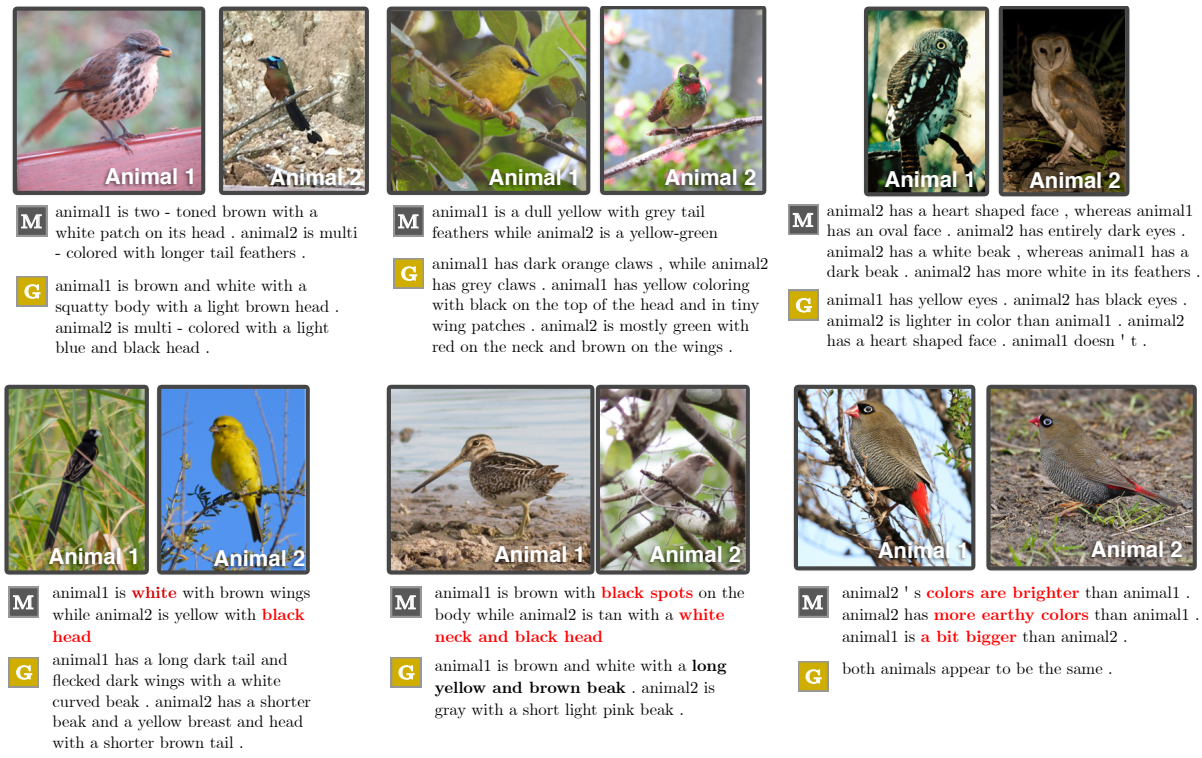


Figure 5: Samples from the dev split of the proposed Birds-to-Words dataset, along with Neural Naturalist model output (M) and one of the five ground truth paragraphs (G). The second row shows failure cases, highlighted in red.

3.1 Image Embedding

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

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

3.2 Joint Encoding

We define a joint encoding $\mathbf{J}$ of the images which contains both embedded images $(\mathbf{E}^1, \mathbf{E}^2)$, a mutated combination $(\mathbf{M})$, or both. We consider as possible mutations $\mathbf{M} \in \{\mathbf{E}^1 + \mathbf{E}^2, \mathbf{E}^1 - \mathbf{E}^2, \max(\mathbf{E}^1, \mathbf{E}^2), \mathbf{E}^1 \odot \mathbf{E}^2\}$. We try these encoding variants to explore whether simple mutations can effectively combine the image representations.

3.3 Comparative Module

Given the joint encoding of the images $(\mathbf{J})$, we would like to represent the differences in feature space $(\mathbf{C})$ in order to generate comparative descriptions. We explore two vari ants at this stage. The first is a direct passthrough of the joint encoding ($\mathbf{C} = \mathbf{J}$). This is analogous to "standard" CNN+LSTM architectures, which embed images and pass them directly to an LSTM for decoding. Because we try different joint encodings, a passthrough here also allows us to study their effects in isolation.

Our second variant is an N-layer Transformer encoder. This provides an additional self-attentive mutations over the latent representations $\mathbf{J}$. Each layer contains a multi-headed attention mechanism (ATTN_{MH}). The intent is that self-attention in Transformer encoder layers will guide comparisons across the joint image encoding.

Denoting LN as *Layer Norm* and FF as *Feed Forward*, with $\mathbf{C}_i$ as the output of the ith layer of the Transformer encoder, $\mathbf{C}_0 = \mathbf{J}$, and $\mathbf{C} = \mathbf{C}_N$:

$$\mathbf{C}^H_i = \text{LN}(\mathbf{C}_{i-1} + \text{ATTN}_{\text{MH}}(\mathbf{C}_{i-1}))$$
$$\mathbf{C}_i = \text{LN}(\mathbf{C}^H_i + \text{FF}(\mathbf{C}^H_i))$$

3.4 Decoder

We use an N-layer Transformer decoder architecture to produce distributions over output tokens. The Transformer decoder is similar to an encoder, but it contains an intermediary multi-headed attention which has access to the encoder's output $\mathbf{C}$ at every time step.

$$\mathbf{D}_i^{H_1} = \text{LN}(\mathbf{X} + \text{ATTN}_{\text{MASK,MH}}(\mathbf{X}))$$
$$\mathbf{D}_i^{H_2} = \text{LN}(\mathbf{D}_i^{H_1} + \text{ATTN}_{\text{MH}}(\mathbf{D}_i^{H_1}, \mathbf{C}))$$
$$\mathbf{D}_i = \text{LN}(\mathbf{D}_i^{H_2} + \text{FF}(\mathbf{D}_i^{H_2}))$$

Here we denote the text observed during training as $\mathbf{X}$, which is modulated with a position-based encoding and masked in the first multi-headed attention.

4 Experiments

We train the Neural Naturalist model to produce descriptions of the differences between images in the Birds-to-Words dataset. We partition the dataset into train (80%), val (10%), and test (10%) sections by splitting based on the pivot images i_1. This ensures i_1 species are unique across the different splits.

We provide model hyperparameters and optimization details in the supplementary material.

4.1 Baselines and Variants

The *most frequent* paragraph baseline 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. *Nearest Neighbor* embeds both images and computes the lowest total L_2 distance to a training set pair, sampling a caption from it. We include two standard neural baselines, CNN (+ Attention) + LSTM, which concatenate the images embeddings, optionally perform attention, and decode with an LSTM. The main model variants we consider are a simple joint encoding ($\mathbf{J} = \langle \mathbf{E}^1, \mathbf{E}^2 \rangle$), no comparative module ($\mathbf{C} = \mathbf{J}$), a small (1-layer) decoder, and our full Neural Naturalist model. We also try several other combinations of joint encoding and comparative module, which we report separately.

4.2 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 for the main baselines and variants, and in Table 3 for the extended model variants. We observe improvement across BLEU-4 and ROUGE-L scores compared to baselines. 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 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.

	Dev			Test		
	BLEU-4	ROUGE-L	CIDEr-D	BLEU-4	ROUGE-L	CIDEr-D
Most Frequent	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
Nearest Neighbor	0.18	0.40	0.15	0.14	0.36	0.06
CNN + LSTM (Vinyals et al., 2015)	0.08	0.24	0.02	0.08	0.25	0.02
CNN + Attn + LSTM (Xu et al., 2015)	0.08	0.25	0.02	0.08	0.25	0.01
Neural Naturalist – Simple Joint Encoding	0.23	0.44	0.23	-	-	-
Neural Naturalist – No Comparative Module	0.09	0.27	0.09	-	-	-
Neural Naturalist – Small Decoder	0.22	0.42	0.25	-	-	-
Neural Naturalist – Full	**0.24**	**0.46**	0.28	**0.22**	**0.43**	0.25
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. The Neural Naturalist model benefits from a strong joint encoding and Transformer-based comparative module, achieving the highest BLEU-4 and ROUGE-L scores. We observed that CIDEr-D scores had little correlation with description quality.

Results for this experiment are shown in Table 4. 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.

4.3 Qualitative Analysis

In Figure 5, 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."*

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 5, 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.

We also observe the model sometimes swaps which features are attributed to which animal. This is partially observed in Figure 5 (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.

5 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

Joint Encoding						Comparative Module	Decoder	Decoding Algorithm	Dev		
i_1	i_2	$-$	$+$	max	$\odot$				BLEU-4	ROUGE-L	CIDEr-D
✓	✓					6-Layer Transformer	6-Layer Transformer	Beamsearch	0.23	0.44	0.23
		✓							0.23	0.45	0.27
			✓						**0.24**	0.43	**0.28**
				✓					0.23	0.43	0.24
					✓				**0.24**	**0.46**	**0.28**
✓	✓	✓							0.22	0.44	0.22
✓	✓		✓						0.22	0.42	0.25
✓	✓			✓					0.21	0.42	0.22
✓	✓				✓				0.22	0.43	0.23
✓	✓	✓	✓	✓	✓				0.21	0.43	0.20
					✓	Passthrough	6-Layer Transformer	Beamsearch	0.00	0.02	0.00
					✓	1-L Transformer			**0.24**	0.44	0.27
					✓	3-L Transformer			**0.24**	0.44	0.27
					✓	6-L Transformer			**0.24**	**0.46**	**0.28**
✓	✓	✓				Passthrough	6-Layer Transformer	Beamsearch	0.09	0.27	0.09
✓	✓	✓				1-L Transformer			**0.24**	0.43	0.24
✓	✓	✓				3-L Transformer			0.22	0.42	**0.26**
✓	✓	✓				6-L Transformer			0.22	**0.44**	0.22

Table 3: Variants of the joint encoding and comparative module for the Neural Naturalist model. We find that the elementwise mutation ($\odot$) performs the best of all joint encodings, and that using a Transformer encoder as a comparative module greatly improves model performance.

	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 4: Human evaluation results on 120 test set samples, twenty per column. Scale: -1 (perfectly wrong) to 1 (perfectly correct). Columns are ordered left-to-right by increasing distance. Our model outperforms baselines from intra-Genus distinctions onward, though highly similar comparisons still prove difficult.

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.

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). While dataset captions two images with only a small section of pixels that change (surveillance footage), we consider image pairs with no pixel overlap, which motivates our stratified sampling approach for drawing good comparisons.

6 Conclusion

We present the new Birds-to-Words dataset and Neural Naturalist model for generating comparative explanations of visual distinctions. The dataset—with paragraph-length, adaptively detailed descriptions using everyday language—reflects how humans describe fine-grained visual differences. We hope this line of research will provide assistance to humans in fine-grained classification domains like citizen science.

References

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.

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

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.

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.

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.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

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*.