Neural Naturalist: Generating Fine-Grained Image Comparisons Appendix

Maxwell Forbes **Christine Kaeser-Chen** **Piyush Sharma** **Serge Belongie**

University of Washington Google Research Cornell University and Cornell Tech

mbforbes@cs.washington.edu

{christinech,piyushsharma}@google.com

sjb344@cornell.edu

https://mbforbes.github.io/neural-naturalist/

A Algorithmic Approach to Dataset Construction

We present here an algorithmic approach to collecting a dataset of image pairs with natural language text describing their differences. The central challenge is to balance empirical desiderata—mainly, sample coverage and model relevance—with practical constraints of data quality and cost. This algorithmic approach underpins the dataset collection we outlined in the paper body.

A.1 Goals

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 textual comparison of them. We can consider each image i as drawn from some domain $\mathcal{D} \in \{\textit{furniture, trees, ...}\}$, or a completely open domain of all concepts. There are several criteria we would like to balance:

1. **Coverage** A dataset should sufficiently cover $\mathcal{D}$ so that generalization across the space is possible.

2. **Relevance** Given the capabilities for models to distinguish i_1 and i_2, t should provide value.

3. **Comparability** Each pair (i_1, i_2) must have sufficient structural similarities that a human annotator can reasonably write t comparing them. Pairs that are too different will yield lengthy and uninteresting descriptions without direct contrasting statements. Pairs that are too similar for human perception may yield *"I can't see any difference."*[1]

4. **Efficiency** Image judgements and textual annotations require human labor. With a fixed budget, we would like to yield a dataset of the largest size possible.

We describe sampling algorithms for addressing these issues given the choice of a domain.

A.2 Pivot-Branch Sampling

Drawing a single image i from domain $\mathcal{D}$, there is a chance $p \in [0, 1]$ that each image is ill-suited for comparisons. For example, i might be out-of-focus or contain multiple instances.

If a pair of images is drawn, and each has probability p of being discarded, then $\frac{1}{(1-p)^2}$ times more pairs must be selected and annotated. For example, if $p = \frac{2}{3}$, then the annotation cost is scaled by 2.25. This severely impacts annotation *efficiency*.

To combat this, we employ a stratified sampling strategy we call *pivot-branch sampling*. Each image on one side of the comparison (say, i_{pivot}) is vetted individually, and k images on the other side (say, i_{branch}) are sampled to produce pairs. With k-times fewer i_{pivot} images, it is feasible to check each instance for usability. This lowers the annotation cost scale to $\frac{1}{1-p}$ (e.g., with $p = \frac{2}{3}$, this is 1.5).

Splitting our selection from $\mathcal{D}$ into two parts allows us to define two distinct sampling strategies. One choice is for $s_{\text{pivot}}(\mathcal{D})$ to select pivot images. The second is for $s_{\text{branch}}(\mathcal{D}, i_{\text{pivot}}, k)$ to sample k images given a single pivot image.

A.3 Designing $s_{\text{pivot}}(\mathcal{D})$

Selecting i_{pivot} are important because each will contribute to k image pairs in a dataset. Here we consider the case where there are class labels $c \in \mathcal{C}$ available for each image in the domain. We propose selecting s_{pivot} to sample uniformly over $\mathcal{C}$. This strategy attempts to provide coverage

[1] This hints at the same sweet spot the fine-grained visual classification (FGVC) community studies, like cars (Krause et al., 2013), aircraft (Maji et al., 2013), dogs (Khosla et al., 2011), and birds (Wah et al., 2011; Van Horn et al., 2018).

over $\mathcal{D}$ using class labels as a coarse measure of diversity. It accounts for category-level dataset bias (e.g., where most images belong to only a few classes). This pushes the need to address *relevance* and *comparability* to the sampling procedure for branched images.

A.4 Designing $s_{\textbf{branch}}(\mathcal{D}, i_{\textbf{pivot}}, k)$

Given each pivot image i_{pivot}, we will choose k images from $\mathcal{D}$ for comparison. We can make use of additional functions and structure available on $\mathcal{D}$:

$\mathcal{V}(i_1, i_2) \rightarrow [0, 1]$
A function that measures the visual similarity between any two images.

$\mathcal{T}(\mathcal{D})$
A taxonomy over $\mathcal{D}$, with image class labels $c \in \mathcal{C}$ as leaves.

We can partition $k = k_v + k_t$ to sample k_v visually-similar images using and k_t taxonomically related images. A simple strategy for visually similar images is to pick

$$\underset{i' \in \mathcal{D}, i' \neq i_{\text{pivot}}}{\operatorname{argmin}} \quad \mathcal{V}(i_{\text{pivot}}, i')$$

k_v times without replacement. This samples the k_v most visually similar images to i_{pivot}, excluding the image itself.

To employ taxonomic information, we propose a walk over mutually exclusive subsets of $\mathcal{T}(\mathcal{D})$. We define a function $a_{\mathcal{T}(\mathcal{D})}(c, \ell)$ that gives the set of other taxonomic leaves that share a common ancestor exactly ℓ taxonomic levels above c, and no levels lower. More formally, if we use $p(c, c', \ell)$ to express that c and c' share a parent ℓ taxonomic levels above c, then we can define:

$$a_{\mathcal{T}(\mathcal{D})}(c, \ell) = \{c' : p(c, c', \ell) \wedge \nexists_{\ell' < \ell}\, p(c, c', \ell')\}$$

The function $a_{\mathcal{T}(\mathcal{D})}(c, \ell)$ partitions the taxonomy $\mathcal{T}(\mathcal{D})$ into disjoint subtrees. For example, $a_{\mathcal{T}(\mathcal{D})}(c, 1)$ are the set of sibling classes to c which share its direct parent; $a_{\mathcal{T}(\mathcal{D})}(c, 2)$ are the set of cousin classes to c which share its grandparent, but *not* its parent.

We can employ $a_{\mathcal{T}(\mathcal{D})}(c, \ell)$ by choosing class c from our pivot image i_{pivot} and varying ℓ. As we increase ℓ, we define mutually exclusive sets of classes with greater taxonomic distance from c.

To sample images using this scheme, we can further split our k_t budget for taxonomically sampled images into $k_t = k_{t_1} + k_{t_2} + \cdots + k_{t_\ell}$ for ℓ different levels. Then, if we write the set of classes $\mathcal{C}_\ell = a_{\mathcal{T}(\mathcal{D})}(c, \ell)$, we can sample k_{t_ℓ} images from $\mathcal{C}$. One scheme is to perform round-robin sampling: rotate through each class $c_\ell \in \mathcal{C}_\ell$ and sample sample one image from each until k_{t_ℓ} are chosen.

A.5 Analyzing $s_{\textbf{branch}}(\mathcal{D}, i_{\textbf{pivot}}, k)$

Given a good visual similarity function $\mathcal{V}$, image pairs will exhibit enough similarity to satisfy requirement that they be semantically close enough to be *comparable*. They may also be so visually similar that comparability is difficult. However, this aspect counter-balances with *relevance*: if $\mathcal{V}(i_1, i_2)$ is small under a visual model, but their differences are describable by humans, their difference description has high value because it distinguishes two points with high similarity in visual embeddings space.

The use of the taxonomy $\mathcal{T}(\mathcal{D})$ complements $\mathcal{V}$ by providing controllable *coverage* over $\mathcal{D}$ while maintaining *relevance* and *comparability*. Tuning the range of ℓ values used in the taxonomic splits $a_{\mathcal{T}(\mathcal{D})}(c, \ell)$ ensures comparability is maintained. Clamping ℓ below a threshold ensures images have sufficient similarity, and controlling the proportion of k_{t_ℓ} for small values of ℓ mitigates the risk of too-similar image pairs.

Similarly, we can adjust the relevance of taxonomic sampling by controlling the distribution of $k_{t_1} \ldots k_{t_\ell}$ with respect to the particular structure of the taxonomy $\mathcal{T}(\mathcal{D})$. If the taxonomy is well-balanced, then fixing a constant k_{t_ℓ} will draw proportionally more samples from subtrees close to c. This can be seen by considering that $a_{\mathcal{T}(\mathcal{D})}(c, \ell)$ defines exponentially larger subsets of $\mathcal{T}(\mathcal{D})$ as ℓ increases. Drawing the same number of samples from each subset biases the collection towards relevant pairs (which should be more difficult to distinguish) while maintaining sparse coverage over the entirety of $\mathcal{D}$.

B Details for Constructing Birds-to-Words Dataset

We provide here additional details for constructing the Birds-to-Words dataset. This is meant to link the high level overview in Section 2 with the algorithmic approach presented in the previous section (Appendix A).

B.1 Clarity

To build a dataset emphasizing fine-grained comparisons between two animals, we impose stricter restrictions on the images than iNaturalist research-grade observations (photographs). An iNaturalist observation that is research-grade indicates the community has reached consensus on the animal's species, that the photo was taken in the wild, and several other qualifications.[2] We include four additional criteria that we define together as *clarity*:

1. **Single instance**: A photo must include only a single instance of the target species. Bird photography often includes flocks in trees, in the air, or on land. In addition, some birds appear in male/female pairs. For our dataset, all of those photos must be discarded.

2. **Animal**: A photo must include the animal itself, rather than a record of it (e.g., tracks).

3. **Focus**: A photo must be sufficiently in-focus to describe the animal in detail.

4. **Visibility**: The animal in the photo must not be too obscured by the environment, and must take up enough pixels in the photo to be clearly described.

B.2 Pivot Images

To pick pivot images, we first uniformly sample from the set of 9k species in the taxonomic CLASS *Aves* in iNaturalist. We consider only species with at least four recorded observations to promote the likelihood that at least one image is *clear*. We also perform look-ahead branch sampling to ensure that a species will yield sufficient comparisons taxonomically. For each species, we manually review four images sampled from this species to select the clearest image to use as the pivot image. If none are suitable, we move to the next species. With this manual process, we select 405 species and corresponding photographs to use as pivot i_1 images.

B.3 Branching Images

See Section 2.3 for the description of selecting $k_v = 2$ visually similar branching images using a function $\mathcal{V}(i_1, i_2)$. We highlight here the use of the taxonomy $\mathcal{T}(\mathcal{D})$ to select $k_t = 10$ branching images with varying levels of taxonomic distance.

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 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 set each $k_{t_\ell} = 2$ for a total of $k_t = 10$.

B.4 Annotations

Clarity Annotators first label whether i_1 and i_2 are *clear*. While we manually verified each i_1 is clear, each i_2 must still be vetted.[3] Starting from 405 pivot images i_1, and selecting $k = 12$ branching images i_2 for each, we annotated a total of 4,860 image pairs. After restricting images to have $\geq \frac{4}{5}$ positive clarity judgments, we ended up with the 3,347 image pairs in our dataset, a retention rate of 68.9%.

Quality We vet each annotator individually by manually reviewing five reference annotations from a pilot round, and perform random quality assessments during data collection. We found that manually vetting the writing quality and guideline adherence of each individual annotator vital for ensuring high data quality.

C Model Details

For the image embedding component of our model, we use a ResNet-101 network as our CNN. We use a model pretrained on ImageNet and fix the CNN weights before starting training for our task. We also experimented with an Inception-v4 model, but found ResNet-101 to have better performance.

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 paragraphs is clipped at 64 tokens during training (chosen empirically to cover 94% of paragraphs). The text is preprocessed using standard techniques (tokenization, lowercasing), and we replace mentions referring to each image with special tokens ANIMAL1 and ANIMAL2.

For inference, we experiment with greedy decoding, multinomial sampling, and beam search.

[2] More details on iNaturalist research-grade specification: https://www.inaturalist.org/pages/help#quality

[3] Annotators would occasionally agree that a particular i_1 images was in fact unclear, upon which we removed it and all corresponding pairs from the dataset.

Photograph	Attribution
Fig. 1: Top and bottom left	salticidude *(CC BY-NC 4.0)* https://www.inaturalist.org/observations/20863620
Fig. 1: Top right	Patricia Simpson *(CC BY-NC 4.0)* https://www.inaturalist.org/observations/1032161
Fig. 1: Bottom right	kalamurphyking *(CC BY-NC-ND 4.0)* https://www.inaturalist.org/observations/9376125
Fig. 2: Top left	Ryan Schain https://macaulaylibrary.org/asset/58977041
Fig. 2: Top right	Anonymous eBirder https://www.allaboutbirds.org/guide/Song_Sparrow/media-browser/66116721
Fig. 2: Right, 2nd from top	Garth McElroy/VIREO https://www.audubon.org/field-guide/bird/song-sparrow#photo3
Fig. 2: Right, 3rd from top	Myron Tay http://orientalbirdimages.org/search.php?Bird_ID=2104&Bird_Image_ID=61509&p=73
Fig. 2: Right, 4th from top	Brian Kushner https://www.audubon.org/field-guide/bird/blue-jay
Fig. 2: Bottom, left	А. Щербаков
Fig. 2: Bottom, right	prepa3tgz-11bwv518 *(CC BY-NC 4.0)* https://www.inaturalist.org/observations/23184228
Fig. 4: Top	jmaley *(CC0 1.0)* https://www.inaturalist.org/observations/31619615
Fig. 4: Bottom	lorospericos *(CC BY-NC 4.0)* https://www.inaturalist.org/observations/30605775
Fig. 5: Top left, left	wildlife-naturalists *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/13223248
Fig. 5: Top left, right	Colin Barrows *(CC BY-NC-SA 4.0)* https://www.inaturalist.org/photos/2642277
Fig. 5: Top middle, left	charley *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/13379419
Fig. 5: Top middle, right	guyincognito *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/26314681
Fig. 5: Top right, left	Chris van Swaay *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/18941543
Fig. 5: Top right, right	Jonathan Campbell *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/20120523
Fig. 5: Middle left, left	John Ratzlaff *(CC BY-NC-ND 4.0)* https://www.inaturalist.org/photos/647514
Fig. 5: Middle left, right	Jessica *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/5595152
Fig. 5: Middle middle, left	i_c_riddell *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/1331149
Fig. 5: Middle middle, right	Pronoy Baidya *(CC BY-NC-ND 4.0)* https://www.inaturalist.org/photos/5027691
Fig. 5: Middle right, left	Nicolas Olejnik *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/2006632
Fig. 5: Middle right, right	Carmelo López Abad *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/892048
Fig. 5: Bottom left, left	Luis Querido *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/13052253
Fig. 5: Bottom left, right	copper *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/22043211
Fig. 5: Bottom middle, left	vireolanius *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/13550702
Fig. 5: Bottom middle, right	Mathias D'haen *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/14943695
Fig. 5: Bottom right, left	tas47 *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/10691998
Fig. 5: Bottom right, right	Nik Borrow *(CC BY-NC 4.0)* https://www.inaturalist.org/photos/13776993

Table 1: Attributions for photographs in main paper body.

Beam search performs best, so we use it with a beam size of 5 for all reported results (except the decoding ablations, where we report each).

We train with Adagrad for 700k steps using a learning rate of .01 and batch size of 2048. We decay the learning rate after 20k steps by a factor of 0.9. Gradients are clipped at a magnitude of 5.

D Image Attributions

Table 1 provides attributions for all photographs used in the main paper body.

References

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. 2011. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO.

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.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778.

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