

The Curious Layperson: Fine-Grained Image Recognition without Expert Labels

Appendix

Subhabrata Choudhury

subha@robots.ox.ac.uk

Iro Laina

iro@robots.ox.ac.uk

Christian Rupprecht

chrisr@robots.ox.ac.uk

Andrea Vedaldi

vedaldi@robots.ox.ac.uk

Visual Geometry Group

University of Oxford

Oxford, UK






Input	Predicted descriptions	Top 5 – retrieval results
	this bird has wings that are blue and has a white belly a small bird with a blue head and blue wings with a grey belly this bird has a blue crown blue primaries and a blue and grey belly this bird has a white belly and breast with a blue crown and wing [...]	<div>Florida Jay ✓</div> <div>Black-throated Blue Warbler ✗</div> <div>Belted Kingfisher ✗</div> <div>Cerulean Warbler ✗</div> <div>Blue Jay ✗</div>
	this is a small bird with a black head and a brown body this bird has wings that are black and has an orange belly this bird is black with red and has a very short beak a small bird with a black head and black nape with yellow and orange covering the rest of its body [...]	<div>Baltimore Oriole ✓</div> <div>Hooded Oriole ✗</div> <div>Orchard Oriole ✗</div> <div>Yellow-headed Blackbird ✗</div> <div>Rufous Hummingbird ✗</div>
	this bird has wings that are black and has a yellow belly a small bird with a light green belly and dark green head this bird has a white belly with a yellow breast and a black cheek patch this bird has wings that are black and white and has a yellow crown [...]	<div>Tennessee Warbler ✓</div> <div>Kentucky Warbler ✗</div> <div>Philadelphia Vireo ✗</div> <div>Myrtle Warbler ✗</div> <div>Orange-crowned Warbler ✗</div>
	this bird is all black and has a long pointy beak this particular bird has a belly that is black with a black bill the bird has a black breast and belly and a black bill a medium sized bird that is all black with a medium sized bill [...]	<div>American Crow ✗</div> <div>Horned Puffin ✗</div> <div>Shiny Cowbird ✓</div> <div>Boat-tailed Grackle ✗</div> <div>Groove-billed Ani ✗</div>
	this bird is brown with white and has a very short beak this bird has wings that are brown and has a short bill this particular bird has a belly that is white and brown a bird with a small pointed bill white eyering and gray plumage [...]	<div>Worm-eating Warbler ✗</div> <div>Winter Wren ✗</div> <div>Mangrove Cuckoo ✗</div> <div>Lincoln Sparrow ✗</div> <div>Field Sparrow ✗</div>

Figure 1: More Qualitative Results — CUB-200. We show several examples of input image, the corresponding predicted descriptions and the top-5 retrieved documents with an example image for each one (for illustration purposes only: the image is not used for matching). Even when the retrieved document is incorrect, often the visual appearance of the bird matches the inputs. This shows that the FGSM module has learned to match descriptions to documents that are visually plausible.

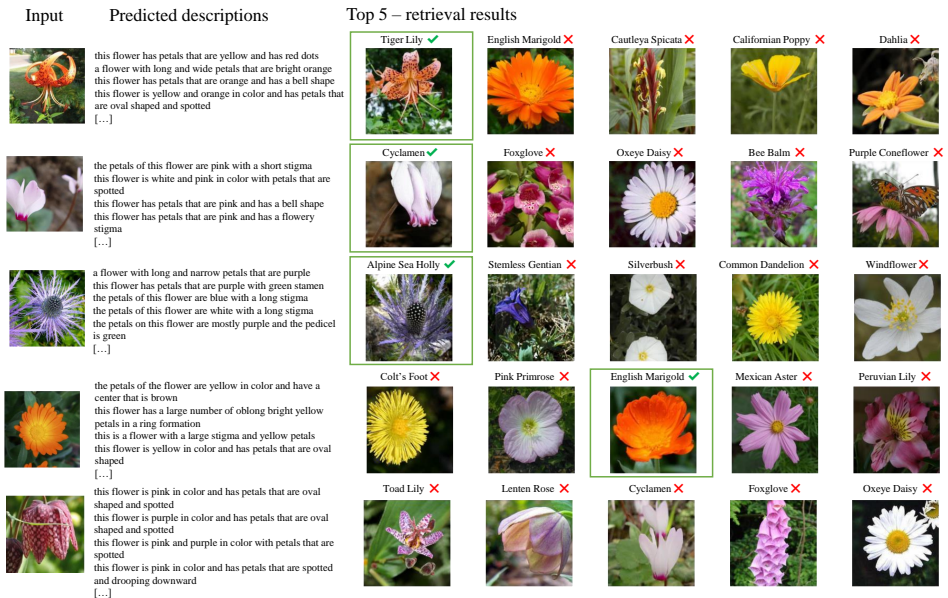


Figure 2: **Qualitative Results — Oxford-102.** We show several examples of input image, the corresponding predicted descriptions and the top-5 retrieved documents with an example image for each one (for illustration purposes only: the image is not used for matching).

1 Datasets

CUB-200. The Caltech-UCSD Birds-200-2011 (CUB-200) [18] contains images of 200 different bird species. The train and test set contains 5,994 and 5,794 images respectively. We have collected expert documents—one document corresponding to each of the 200 categories—by crawling AllAboutBirds¹ (AAB), which includes bird identification guides made available by the Cornell Lab of Ornithology. Each document consists of an Overview and ID info sections. We obtain basic description from Overview. From page ID info we use Identification, Size & Shape, Color Pattern, Behavior and Habitat. For Size & Shape key we omit the relative size table. For 17 categories that were not found in AAB, we resorted to Wikipedia articles instead. We replace any mention of the classes in corpus with the word ‘a bird’ so that the model is unable to cheat by using expert labels.

Oxford-102 Flowers. The Oxford-102 Flowers (FLO) dataset [8] contains images of 102 categories of flowers. We use the official train and test set of 1,020 and 6,149 images respectively. Similar to CUB-200, we create an expert document corpus with one document per category by parsing Wikipedia data using the MediaWiki API. We use summary, cultivation, distribution, description, ecology, flowers, habitat sections and ignore the rest. We replace the expert labels in the corpus with the word ‘a flower’.

¹<https://allaboutbirds.com>

2 Implementation Details

Image Captioning. We train Show-Attend-and-Tell (SAT) for 100 epochs using the implementation of [14]. For AoANet [9] we follow the implementation and hyperparameters from the authors’ repository². During inference, we apply beam search with a beam size of 10 to sample multiple captions. For Tables 2 and 3 in the main paper, we have trained the captioning models on the official data splits, reserving 10% of the images from training split for validation. For Table 4, we follow the GZSL split proposed in [14], using the `trainval` set to train the captioning models, with 10% of the images being again kept aside for validation. Therefore, we explicitly avoid using “unseen” categories when training the captioning models.

FGSM. T is a sentence transformer with a RoBERTa-large backbone pretrained on the SNLI [11], Multi-Genre NLI [13] and STS [12] benchmarks. The pretrained model is obtained from the publicly available repository³ of [9]. ϕ is implemented as a two layer MLP with intermediate and output dimensions of 256 and 64 respectively and tanh activation function. For h we use a linear layer with output dimension of 2 (for the binary classification task). During the first stage of training, we use a constant learning rate of $0.5 \cdot 10^{-6}$ for T and 10^{-5} for ϕ and h respectively; weight decay is set to zero for ϕ and h . We follow [9] for rest of the hyper-parameters. During the second stage, we aim to reduce the gap between the data that the model is exposed to for training and the target domain. We add the regularizer R and fix T , pre-computing all embeddings for computational efficiency. We retrain ϕ and h from scratch with the Adam optimizer [8] and an initial learning rate of 10^{-5} . For ϕ we use a three layer MLP with 256, 64, and 32 output dimensions. For h we use a linear layer with an output dimension of 3 to predict positive, negative and neutral sentence pairs, training with a cross-entropy loss and the regularizer with a weight factor $\lambda = 10$. The neutral sentence pairs are either a pair of captions from two different images that have no nouns in common, or a pair containing a image caption and a random sentence from the target corpus that have no nouns in common. We sample with equal probability from these two pools. The reasoning behind common nouns is that sentences containing the same nouns could potentially describe the same parts — *e.g.* head, beak, wings — while adjectives are often used as attributes, *e.g.* red wings, short beak. Pairs of sentences without common nouns contain neither entailing nor contradicting information, *i.e.* they describe different objects/parts, and can be thus safely considered as neutral.

3 Additional Results

Additional Qualitative Results. Fig. 1 shows additional qualitative retrieval results for CUB from the same setting as Fig. 3 in the main paper. As discussed in the main paper, we find that the largest confusion happens between birds of the same family, who often look similar. Fine differences between individual species are sometimes missed by the captioning models. We expect that more detailed descriptions (c.f. user study in the main paper) will help in this regard. In Fig. 2 we show qualitative retrieval results for Oxford-Flowers dataset. We observe the top flowers are in agreement with the predicted descriptions and we expect the retrieval accuracy to improve as the descriptions become more detailed.

²<https://github.com/husthuanan/AoANet>

³<http://sbert.net/models/>

Image Description Generalization. As an integral part of our approach, we analyze the performance of the captioning module. In particular, we are interested in the degradation (if any) in the capability of the captioning models to describe images of previously unseen categories. To this end, to understand whether the learned image descriptions are dependent on the training categories, we train the captioning model with the zero-shot learning split and compare the validation performance (in terms of common captioning metrics) between seen and unseen classes in Table 1. We report results using common metrics, BLUE1-4 [1], METEOR [2], ROUGE-L [3] and CIDEr-D [4]. Interestingly, we find no significant difference in performance between seen and unseen classes, indicating that the model generalizes well to the appearance of novel categories. This is on par with our intuition and motivation for a layperson-inspired system to describe the appearance of objects without necessarily being able to recognize or name them and even when they have never previously encountered a given object.

Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr-D
CUB-200 (seen)	87.4	72.1	58.0	45.9	31.7	62.6	39.6
CUB-200 (unseen)	86.9	70.9	56.5	44.6	31.1	62.0	38.1
CUB-200 (overall)	87.1	71.4	57.1	45.1	31.4	62.2	38.9
FLO (seen)	88.1	76.3	66.3	58.2	38.6	70.1	54.5
FLO (unseen)	85.6	73.3	63.4	55.1	35.7	67.7	32.6
FLO (overall)	87.0	74.9	65.0	56.8	37.3	69.0	44.5

Table 1: **Captioning performance.** We verify that the captioning model generalizes to unseen classes.

Word Relevance. In Table 2 we show pairs of image descriptions and sentences from the expert corpus, along with the predicted score (after sigmoid). We highlight the importance of individual words which is estimated by masking the word and computing the difference between the new and initial score. The model has learned to pay attention to colors and body parts, which affect its decision the most.

Image Description	Sentence (Expert Corpus)	Score
the bird has wings that are black and has an orange belly.	adult males are flame-orange and black , with a solid-black head and one white bar on their black wings .	0.981
a bird with a white belly and black head and wings .	the underparts are white.	0.729
this is a bird has wings that are brown and has a long bill.	male mallards have a dark , iridescent-green head and bright yellow bill .	0.228

Table 2: **Word relevance visualization.** Words are highlighted in blue (red) to highlight positive (negative) changes in the output when the word is occluded (replaced by [UNK]). The darker the shade, the bigger the change in the output.

Sentence Matching Results. In Tables 3 and 4 we show matching results between a query and a document. We highlight the sentence with the highest matching score within the document, given the query (image description).

query	this bird has a black body yellow head and gray wings.
document	<p>with a golden head, a white patch on black wings, and a call that sounds like a rusty farm gate opening, the bird demands your attention. look for them in wesa bird and prairie wetlands, where they nest in reeds directly over the water. they're just as impressive in winter, when huge flocks seem to roll across farm fields. each bird gleans seeds from the ground, then leapfrogs over its flock mates to the front edge of the ever-advancing troupe. in the midwest and west, look for birds both in freshwater wetlands and in nearby farm fields. though they are striking in appearance, these birds spend a substantial time perched out of view in cattails or reeds, so listen for their harsh check calls and bizarre grinding, buzzing songs in order to pinpoint their location. when searching in farm fields, look for large concentrations of a birds and then scan them carefully. if the bulk of the birds are a bird or some other species, don't despair—focus on finding a white wing patch or yellow head among the other species. birds are fairly large a birds, with a stout body, a large head, and a long, conical bill. males are striking a birds with yellow heads and chests, and black bodies with prominent white patches at the bend of the wing. females and immatures are brown instead of black, with duller yellow heads. immature males show some white at the bend of the wing, while females don't. birds breed in loose colonies, and males mate with several females. during the breeding season, they eat insects and aquatic invertebrates. they form huge flocks in winter, often mixing with other species of a birds, and feed on seeds and grains in cultivated fields. birds breed and roost in freshwater wetlands with dense, emergent vegetation such as cattails. they often forage in fields, typically wintering in large, open agricultural areas.</p>

query	this is a bird with a white belly black back and a red head.
document	<p>the gorgeous bird is so boldly pata birded it's been called a "flying checkerboard," with an entirely crimson head, a snow-white body, and half white, half inky black wings. these birds don't act quite like most other a birds: they're adept at catching insects in the air, and they eat lots of acorns and beech nuts, often hiding away extra food in tree crevices for later. this magnificent species has declined severely in the past half-century because of habitat loss and changes to its food supply. look for birds in scattered, open woodlots in agricultural areas, dead timber in swamps, or pine savannas. walk slowly, listening for tapping or drumming, and keep your eyes alert for telltale flashes of black and white as these high-contrast a birds fly in between perches. the red head can be hard to see in strong glare. raucous, harsh weah!. calls will also give away the presence of a bird. birds are medium-sized a birds with fairly large, rounded heads, short, stiff tails, and powerful, spike-like bills. adults have bright-red heads, white underparts, and black backs with large white patches in the wings, making the lower back appear all white when perched. immatures have gray-brown heads, and the white wing patches show rows of black spots near the trailing edge. in addition to catching insects by the normal a bird method of hammering at wood, birds also catch insects in flight and hunt for them on the ground. they also eat considerable amounts of fruit and seeds. their raspy calls are shriller and scratchier than the red-bellied a bird's. birds live in pine savannahs and other open forests with clear understories. open pine plantations, treerows in agricultural areas, and standing timber in beaver swamps and other wetlands all attract birds. smaller than a northern flicker; about the size of a a bird.</p>

Table 3: **FGSM qualitative results.** We show several examples of query-document pairs and highlight the best matching sentence.

query	this particular bird has a belly that is white with black spots.
document	<p>the active little bird is a familiar sight at backyard feeders and in parks and woodlots, where it joins flocks of a bird and a bird, barely outsizeing them. an often acrobatic forager, this black-and-white a bird is at home on tiny branches or balancing on slender plant galls, sycamore seed balls, and suet feeders. downies and their larger lookalike, the a bird, are one of the first identification challenges that beginning bird watchers master. look for birds in woodlots, residential areas, and city parks. be sure to listen for the characteristic high-pitched pik note and the descending whinny call. in flight, look for a small black and white bird with an undulating flight path. during winter, check mixed-species flocks and don't overlook birds among the a bird and a bird – birds aren't much larger than white-breasted a bird. birds are small versions of the classic a bird body plan. they have a straight, chisel-like bill, blocky head, wide shoulders, and straight-backed posture as they lean away from tree limbs and onto their tail feathers. the bill tends to look smaller for the bird's size than in other a birds. birds give a checkered black-and-white impression. the black upperparts are checked with white on the wings, the head is boldly striped, and the back has a broad white stripe down the center. males have a small red patch on the back of the head. the outer tail feathers are typically white with a few black spots. birds hitch around tree limbs and trunks or drop into tall weeds to feed on galls, moving more acrobatically than larger a birds. their rising-and-falling flight style is distinctive of many a birds. in spring and summer, birds make lots of noise, both with their shrill whinnying call and by drumming on trees. you'll find birds in open woodlands, particularly among deciduous trees, and brushy or weedy edges. they're also at home in orchards, city parks, backyards and vacant lots. about two-thirds the size of a a bird between a bird and a bird.</p>
query	this is a black bird with a white stripe on its face and a red crown.
document	<p>the bird is one of the biggest, most striking forest birds on the continent. it's nearly the size of a a bird, black with bold white stripes down the neck and a flaming-red crest. look (and listen) for birds whacking at dead trees and fallen logs in search of their main prey, carpenter ants, leaving unique rectangular holes in the wood. the nest holes these birds make offer crucial shelter to many species including swifts, owls, a birds, bats, and pine martens. look for birds in stands of mature forest with plenty of dead trees and downed logs—deep excavations into rotten wood are telltale signs of this species. also listen for this bird's deep, loud drumming and shrill, whinnying calls. birds occur at all heights in the forest, and are often seen foraging on logs and near the bases of trees. the bird is a very large a bird with a long neck and a triangular crest that sweeps off the back of the head. the bill is long and chisel-like, about the length of the head. in flight, the wings are broad and the bird can seem a birdlike. birds are mostly black with white stripes on the face and neck and a flaming-red crest. males have a red stripe on the cheek. in flight, the bird reveals extensive white underwings and small white crescents on the upper side, at the bases of the primaries. birds drill distinctive rectangular-shaped holes in rotten wood to get at carpenter ants and other insects. they are loud birds with whinnying calls. they also drum on dead trees in a deep, slow, rolling pata bird, and even the heavy chopping sound of foraging carries well. their flight undulates like other a birds, which helps separate them from a a bird's straight flight path. birds are forest birds that require large, standing dead trees and downed wood. forests can be evergreen, deciduous, or mixed and are often old, particularly in the west. in the east they live in young forests as well and may even be seen in partially wooded suburbs and backyards. nearly the size of an a bird a bird-sized.</p>

Table 4: **FGSM qualitative results.** We show several examples of query-document pairs and highlight the best matching sentence.

References

- [1] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *EMNLP*, 2015.
- [2] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.
- [3] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734. ACL, 2014.
- [4] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.
- [5] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, pages 4634–4643, 2019.
- [6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.
- [7] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [8] Maria-Elena Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *CVPR*, volume 2, pages 1447–1454, 2006.
- [9] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pages 3973–3983, 2019.
- [10] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.
- [11] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *CVPR*, pages 251–260, 2017.
- [12] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- [13] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *NAACL*, 2018.
- [14] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 41(9):2251–2265, 2018.