Visualizing and Describing Fine-grained Categories as Textures

Tsung-Yu Lin   Mikayla Timm   Chenyun Wu   Subhransu Maji
University of Massachusetts, Amherst
{tsungyulin,mtimm,chenyun,smaji}@cs.umass.edu

We analyze how categories from recent FGVC challenges [4, 5] can be described by their *textural content*. The motivation is that subtle differences between species of birds or butterflies can often be described in terms of the texture associated with them and that several top-performing networks are inspired by texture-based representations. These representations are characterized by orderless pooling of second-order filter activations such as in bilinear CNNs [10] and the winner of the iNaturalist 2018 challenge [8].

Concretely, for each category we (i) visualize the "maximal images" by obtaining inputs $\mathbf{x}$ that maximize the probability of the particular class according to a texture-based deep network $C_\theta(\mathbf{x})$, and (ii) automatically describe the maximal images using a set of texture attributes. We use C_θ as a multi-layer bilinear CNN as described in our prior work on visualizing deep texture representations [9]. The models for texture captioning were trained on our ongoing efforts on collecting a dataset of describable textures building on the DTD dataset[6]. As seen in Figure 1, these visualizations indicate what aspects of the texture is most discriminative for each category while the descriptions provide a language-based explanation of the same.

Visualizing categories as maximal textures. We visualize the categories from Caltech-UCSD birds [14], Oxford flowers [12], FGVC flowers [2], FGVC fungi [3] and FGVC butterflies and moths [1] datasets. Following the approach of [10] we extract the covariance matrix followed by signed square-root and ℓ_2 normalization from *relu*{*2_2,3_3,4_3, 5_3*} layers of VGG-16 network [13] and train a softmax layer to predict class labels. We train the model on the standard training split for birds and Oxford flowers and randomly select 100 images from the 200 categories with the most images for FGVC fungi, flowers, and butterflies.

Let C_i be the predicted probability from layer i. Then the maximal inverse image for a target class $\hat{C}$ is obtained as: $\min_{\mathbf{x}} \sum_{i=1}^{m} L\left(C_i, \hat{C}\right) + \gamma\Gamma(\mathbf{x})$. Here L is the softmax loss and $\Gamma(\mathbf{x})$ is the TV norm that acts as a smoothness prior. This technique was also used to visualize inverse images in [11]. Figure 1 show the maximal images for three categories along with their texture attributes. Additional visualizations selected arbitrarily across datasets are shown in Figure 2 and 3. The maximal images indicate what discriminative texture properties are learned from training images for classification of instances which often appear in clutter, with wide ranges of pose and lighting variations, and under occlusions.

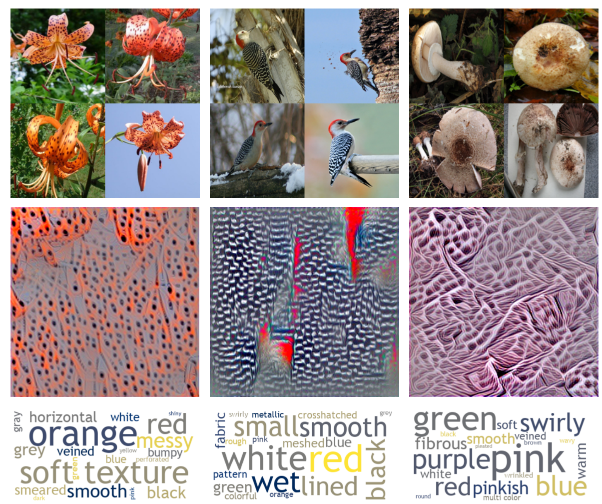

Figure 1. *Tiger Lily* (left), *Red Bellied Woodpecker* (middle) and *Boletus Reticulatus* (right) categories visualized as their training images (top row), maximal texture images (middle row) and texture attributes (bottom row). The size of each phrase in the cloud reflects its likelihood of being associated with the maximal texture.

Describing maximal textures. In addition, we provide the preliminary experiments on describing these textures using attribute phrases that provide a language-based explanation of discriminative texture properties.

We collected a new dataset with natural language descriptions of texture details based on the Describable Textures Dataset (DTD) [6]. For each image from DTD, we ask five human annotators to provide several attribute phrases (e.g., "black and white dots", or "colorful patterns"). We trained linear classifiers based on ResNet-101 [7] activations to predict the probability of each attribute phrase on our collected dataset. For each maximal texture image, the "phrase cloud" shows the top 20 attribute phrases, with the font size proportional to the predicted probability.

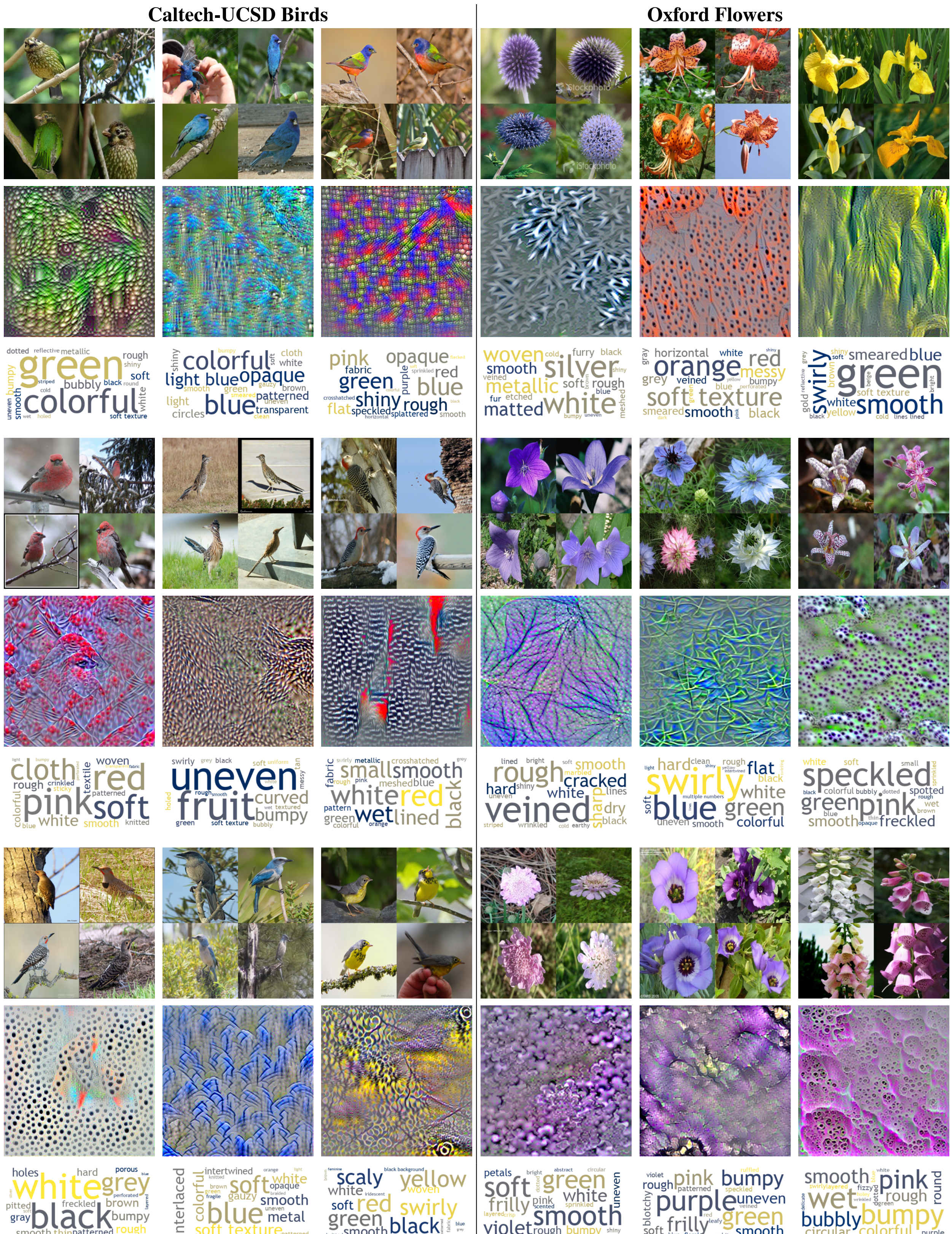

Figure 2. Visualization of fine-grained categories from Caltech-UCSD birds and Oxford flowers. Each example is shown as a column of three images which consists of training examples (top), texture images (middle) and texture attributes as word clouds (bottom). The size of each phrase in the cloud reflects its likelihood of being associated with the maximal texture.

Figure 3. Visualization of fine-grained categories from FGVC butterflies and moths, fungi, and flowers. Each example is shown as a column of three images which consists of training examples (top), texture images (middle) and texture attributes as word clouds (bottom). The size of each phrase in the cloud reflects its likelihood of being associated with the maximal texture.

References

[1] FGVC Butterflies and Moths Dataset, https://sites.google.com/view/fgvc6/competitions/butterflies-moths-2019. 1

[2] FGVC Flowers Dataset, https://sites.google.com/view/fgvc5/competitions/fgvcx/flowers. 1

[3] FGVC Fungi Dataset https://sites.google.com/view/fgvc5/competitions/fgvcx/fungi. 1

[4] The Fifth Fine-Grained Visual Categorization (FGVC) Workshop https://sites.google.com/view/fgvc5. 1

[5] The Sixth Fine-Grained Visual Categorization (FGVC) Workshop https://sites.google.com/view/fgvc6. 1

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[8] Peihua Li, Jiangtao Xie, Qilong Wang, and Zilin Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[9] Tsung-Yu Lin and Subhransu Maji. Visualizing and Understanding Deep Texture Representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2791–2799, 2016. 1

[10] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear Convolutional Neural Networks for Fine-grained Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), volume=40, number=6, pages=1309–1322, year=2018, publisher=IEEE*. 1

[11] Avinash Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision (IJCV)*, 2016. 1

[12] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, Dec 2008. 1

[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[14] Catherine Wah, Steven Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011. 1