Visualizing Deep Texture Representations

Tsung-Yu Lin TSUNGYULIN@CS.UMASS.EDU
Subhransu Maji SMAJI@CS.UMASS.EDU
College of Information and Computer Sciences, University of Massachusetts, Amherst

Abstract

A number of recent approaches that combine aspects of deep CNNs and texture models have shown remarkable benefits for various recognition tasks. Variants of these models have also been successfully applied to texture synthesis and style transfer tasks. We aim to understand the categorical properties captured by these deep texture representations by visualizing inverse images across datasets. This is an extended abstract of our CVPR 2016 paper (Lin & Maji, 2016).

Overview

The study of texture has inspired many of the early representations of images. The idea of representing texture using the statistics of image patches have led to the development of "textons", the popular "bag-of-words" models and their variants such as the Fisher vector and VLAD. These fell out of favor when the latest generation of deep CNNs showed significant improvements in recognition performance over a wide range of visual tasks. Recently however, the interest in texture descriptors have been revived by architectures that combine aspects of texture representations with CNNs. For instance, Fisher vectors built on top of CNN activations lead to better accuracy and improved domain adaptation not only for texture recognition, but also for scene categorization, object classification, and fine-grained recognition (Cimpoi et al., 2016).

In this work we visualize how texture representations built on top CNN activations, in particular the bilinear CNN features (Lin et al., 2015), capture invariances at the category level. The model builds an orderless representation by taking the location-wise outer product of CNN activations. The model is closely related to the second-order pooling and Fisher vectors. Moreover, the gradients of the model can be easily computed allowing visualization of categories by approximate inversion. These exact representations have also been shown to be effective for texture synthesis and style transfer tasks (Gatys et al., 2015).

Concretely, given an image $\mathcal{I}$ we compute CNN activations at a given layer r_i to obtain a set of features $F_{r_i} = \{f_j\}$ indexed by their location j. The bilinear feature $B_{r_i}(\mathcal{I})$ is obtained as $B_{r_i}(\mathcal{I}) = \frac{1}{N}\sum_{j=1}^{N} f_j f_j^T$. This is an *orderless representation* of the image and hence is suitable for modeling texture. Let $r_i, i = 1, \ldots, n$, be the index of the i^{th} layer with the bilinear feature representation B_{r_i}. These features can be used for predicting class labels after suitable normalization. Let $l_i : i = 1, \ldots, m$ be the index of the i^{th} layer from which we obtain attribute prediction probabilities C_{l_i}. Given a target class label $\hat{C}$ we can obtain an "inverse image" by solving the following optimization:

$$\min_{\mathbf{x}} \sum_{i=1}^{m} \beta L\left(C_{l_i}, \hat{C}\right) + \gamma\Gamma(\mathbf{x}). \tag{1}$$

L is the *negative log-likelihood* of the label $\hat{C}$; β, γ is a tradeoff parameters; and $\Gamma(\mathbf{x})$ is an image prior, e.g. the TV_β norm with $\beta = 2$. We learn the parameters of softmax layers to predict labels using bilinear features from `relu2_2`, `relu3_3`, `relu4_3`, `relu5_3` layers of the 16-layer VGG-VD network. Fig. 1 shows some example inverse images from the describable texture dataset (DTD), Flickr material dataset (FMD) and MIT indoor dataset. These images were obtained by setting $\beta = 100$, $\gamma = 10^{-6}$, and $\hat{C}$ to various class labels in Eqn. 1.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

References

Cimpoi, Mircea, Maji, Subhransu, Kokkinos, Iasonas, and Vedaldi, Andrea. Deep filter banks for texture recognition, description, and segmentation. *IJCV*, 2016.

Gatys, Leon A, Ecker, Alexander S, and Bethge, Matthias. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

Lin, Tsung-Yu and Maji, Subhransu. Visualizing and understanding deep texture representations. In *CVPR*, 2016.

Lin, Tsung-Yu, RoyChowdhury, Aruni, and Maji, Subhransu. Bilinear CNN Models for Fine-grained Visual Recognition. In *ICCV*, 2015.

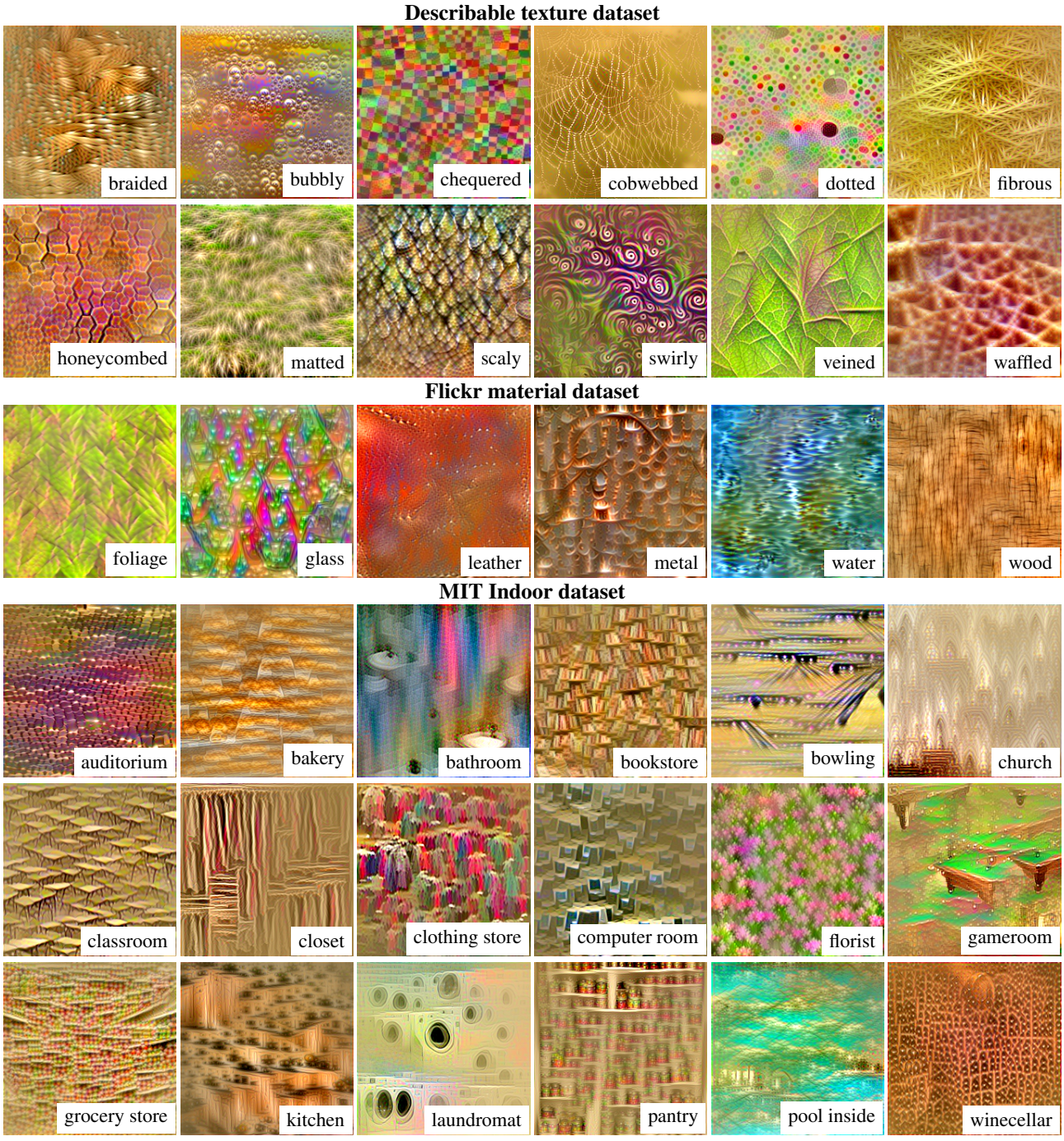

Figure 1. Visualizing categories by inverting the bilinear CNN models trained on DTD, FMD, and MIT Indoor datasets. These images reveal how the model represents texture and scene categories. For instance, the *dotted* category of DTD contains images of various colors and dot sizes and the inverse image is composed of multi-scale multi-colored dots. The inverse images of *water* and *wood* from FMD are highly representative of these categories. Note that these images cannot be obtained by simply averaging instances within a category which is likely to produce a blurry image. The orderless nature of the texture descriptor is essential to produce such sharp images. The inverse scene images from the MIT indoor dataset such as a *bookstore* is visualized as racks of books while a *laundromat* has laundry machines at various scales and locations. More visualizations can be found at http://vis-www.cs.umass.edu/texture