DEEP-CARVING : Discovering Visual Attributes by Carving Deep Neural Nets

Sukrit Shankar[†], Vikas K. Garg[*], Roberto Cipolla[†]
[†]Machine Intelligence Lab (MIL), Cambridge University [*]Computer Science & Artificial Intelligence Lab (CSAIL), MIT

Most of the approaches for discovering visual attributes in images demand significant supervision, which is cumbersome to obtain. In this paper, we aim to discover visual attributes in a weakly supervised setting that is commonly encountered with contemporary image search engines.

For instance, given a noun (say forest) and its associated attributes (say dense, sunlit, autumn), search engines can now generate many valid images for any attribute-noun pair (dense forests, autumn forests, etc). However, images for an attribute-noun pair do not contain any information about other attributes (like which forests in the autumn are dense too). Thus, a weakly supervised scenario occurs. Let $\mathbf{A} = \{a_1, \ldots, a_M\}$ be the set of M attributes under consideration. We have a weakly supervised training set, $\mathbf{S} = \{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_N}, y_N)\}$ of N images $\mathbf{x_1}, \ldots, \mathbf{x_N} \in \mathbf{X}$ having labels $y_1, \ldots, y_N \in \mathbf{A}$ respectively. Equivalently, segregating the training images based on their label, we obtain M sets $\mathbf{S}_m = \mathbf{X}_m \times a_m$, where $\mathbf{X}_m = \{\mathbf{x} \in \mathbf{X} | (\mathbf{x}, a_m) \in \mathbf{S}\}$ denotes the set of $N_m = |\mathbf{X}_m|$ images each having the (single) positive training label $a_m, m \in \{1, \ldots, M\}$. For a test image $\mathbf{x_t}$, the task is to predict $\mathbf{y_t} \subseteq \mathbf{A}$, i.e. all the attributes present in $\mathbf{x}_t$. The aforementioned weakly supervised problem setting is more challenging for attributes as compared to object and scene detection, because attributes can highly co-occur in the training set, are abstract in nature and cannot be easily separated by well-defined spatial boundaries.

Deep Convolutional Neural Networks (CNNs) [1] have enjoyed remarkable success in vision applications recently. However, in a weakly supervised scenario, widely used CNN training procedures do not learn a robust model for predicting multiple attribute labels simultaneously. Owing to the fact that training of deep CNNs can be modified for learning different kinds of features, and also due to our experiments that show that deep CNNs with conventional training procedures outperform combined low-level features of [2] for attribute prediction even under the weakly supervised problem setting, we endeavor to improve upon the deep CNN architectures.

We thus propose **Deep-Carving**, a novel training procedure with CNNs, that helps the net efficiently carve itself for the task of multiple attribute prediction. During training, the responses of the feature maps are exploited in an ingenious way to provide the net with multiple pseudo-labels (for training images) for subsequent iterations. The process is repeated periodically after a fixed number of iterations, and enables the net carve itself iteratively for efficiently disentangling features.

We take AlexNet [1] as our base architecture. To learn a deep-carved net, we follow the sigmoid cross-entropy loss since it can take into account the probabilities of multiple labels. For a deep-carving iteration c, the following loss is minimized:

$$\mathcal{L}_e^c = -\frac{1}{N}\sum_{r=1}^{N}[\mathbf{p_r^c}\log(\hat{\mathbf{p}}_\mathbf{r}) \;+\; (1-\mathbf{p_r^c})\log(1-\hat{\mathbf{p}}_\mathbf{r})] \qquad (1)$$

where the probability vector $\mathbf{p_r^c}$ is a vector of pseudo-label probabilities, and $\hat{\mathbf{p}}_\mathbf{r}$ is obtained by applying the sigmoid function to the M outputs of the last fully connected layer of AlexNet. For sigmoid cross entropy loss, each image r is expected to be annotated with a vector of label probabilities $\mathbf{p}_r$, having length M. For our weakly supervised case, the vector $\mathbf{p}_r$ is initialized with a very low value of 0.05 for all images, with $\mathbf{p}_r^m = 0.95 \;\; \forall \;\; r \in \mathbf{X}_m$. The paper describes the complete algorithm for generating pseudo-labels during training.

Given a test image, the number of positive labels (say K) is known from the ground-truth. Thus, K denotes the number of correct attributes that need to be predicted for the respective test image. Let $\mathbf{T}$ contain the positive labels for the test image. Given the sorted (in descending order) probabilities for the test image from the prediction model, we pick top K predictions. Let the set $\mathbf{P}$ contain these predicted labels. Both $\mathbf{T}$ and $\mathbf{P}$ have cardinality K. We then calculate the number of true and false positives using $\mathbf{T}$ and $\mathbf{P}$, and use precision as our performance metric.

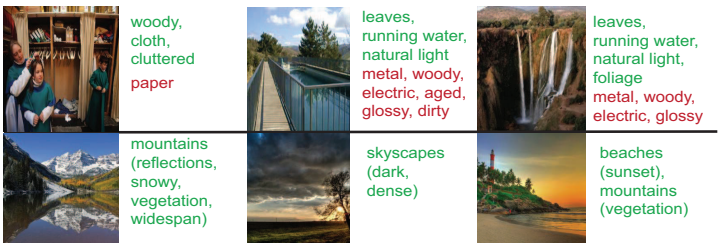

Figure 1: **Attribute Predictions with our Deep-Carved CNNs:** The correctly predicted attributes (true positives) shown in green, and the wrongly predicted ones (false positives) shown in red for various instances in SAD (top row) and CAMIT-NSAD (bottom row) with our deep-carved CNNs. The attributes that are abstract in nature or heavily co-occur with other attributes, are generally predicted with lesser accuracy. *Figure is best viewed in color.*

We contribute a noun-adjective pairing inspired **N**atural **S**cenes **A**ttributes **D**ataset to the research community, **CAMIT - NSAD**, which contains a number of co-occurring attributes within a noun category, and is at least 3 times bigger than the SUN Attributes Dataset [2] in terms of the number of images. We describe, in detail, salient aspects of this dataset in the paper.

For evaluation purposes, we consider three major baselines, viz. Alexnet with softmax loss, Alexnet with Sigmoid Cross Entropy Loss, and fine-tuned AlexNet over the pre-trained model of MIT Places dataset [3]. Our experiments on CAMIT-NSAD and the SUN Attributes Dataset [2], with weak supervision, demonstrate that the Deep-Carved CNNs consistently achieve significant improvement in the precision of attribute prediction over the considered baselines. Fig 1 shows some of the attribute predictions obtained with our deep-carved AlexNet. Fig 2 shows a visualization of convolutional filters learnt with our deep-carved CNNs.

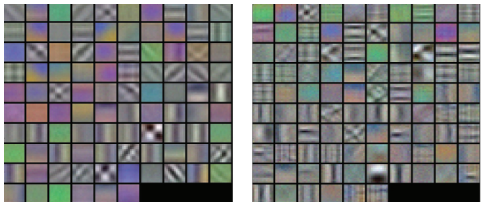

Figure 2: **Visualization of the Filters for the first Convolutional Layer of:** *From Left to Right* - Deep-carved AlexNet for SAD and CAMIT-NSAD. There are 96 filters of sizes 11×11 with 3 channels for each learnt model, and are shown here on a 10×10 grid. *Figure is best viewed in color.*

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, volume 1, page 4, 2012.

[2] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758. IEEE, 2012.

[3] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.