On the Relationship between Visual Attributes and Convolutional Networks

Victor Escorcia[1,2], Juan Carlos Niebles[2], Bernard Ghanem[1]
[1]King Abdullah University of Science and Technology (KAUST), Saudi Arabia. [2]Universidad del Norte, Colombia.

The seminal work of Krizhevsky *et al.* [3] that trained a large convolutional network (conv-net) for image-level object recognition on the ImageNet challenge is considered a major stepping stone for subsequent work in conv-net based visual recognition. Such a network is able to automatically learn a hierarchy of nonlinear features that richly describe image content as well as discriminate between object classes. Recent work [4] has shown that features extracted from a conv-net trained on ImageNet are general purpose (or black-box) enough to achieve state-of-the-art results in various other recognition tasks, including scene, fine-grained, and even action recognition. However, unlike hand-crafted features, those learned by a conv-net are usually not visually intuitive and straightforward to interpret. Despite their excellent recognition performance, understanding and interpreting the inner workings of conv-nets remains mostly elusive to the community. It is this lack of deep understanding that is currently motivating researchers to *look under the hood* and comprehend how and why these deep networks work so well in practice. Inspired by recent observations on the analysis of conv-nets [1], this paper takes another step in a similar direction, namely understanding how the inner workings of a conv-net that is trained for a high-level recognition task (object recognition) relate to intuitive and conventional mid-level representations in computer vision.

Despite the insights of recent work, it is still unclear how visual content is represented within the activations of a conv-net. Simply put, a conv-net trained to classify objects in images can be viewed as a deep learning machine that finds the appropriate mapping between input (raw pixel intensities) and output (object labels) layers. It is conceivable to ask here whether this mapping makes use of a mid-level representation for objects, similar in spirit to how the human visual system functions. In fact, the empirically validated and general purposefulness of conv-net activations across different visual recognition tasks [4] suggest that such a shared mid-level representation of the visual world is being automatically learned. More importantly, it is worthwhile to investigate whether this learned mid-level representation is related to (if at all) intuitive mid-level representations (e.g. parts, mid-level patches and visual attributes [2]) innovated by the community before conv-nets were popularized. Since addressing these queries in their entirety is beyond the scope of this paper, we focus on studying the relationship between a conv-net trained to recognize objects in images and object-level visual attributes. This can help us realize, for example, whether a conv-net trained to recognize a ‘dog’ inherently learns what ‘fluffy’ means *without* prior knowledge of the attribute (refer to Figure 1).

Main Findings: In this work, we hypothesize that a sparse number of nodes in a deep conv-net trained for image-level object recognition on ImageNet can reliably predict absolute visual attributes [2]. Through rigorous experimentation, we uncover the following properties of the relationship between such a conv-net and visual attributes.

(1) Visual attributes can be predicted reliably using a sparse number of nodes from the conv-net. This suggests that the conv-net can *indirectly* learn attribute concepts, even though it is trained to recognize objects, as depicted in Figure 1.

(2) Nodes in the conv-net that are used to represent attributes are called Attribute Centric Nodes (ACNs) which are illustrated in Figure 1. The support of ACNs in the network is sparse and unevenly distributed among the different layers (convolutional and fully-connected). On average, these ACNs are concentrated in the top layers of the network; however, their exact locations are attribute dependent. Also, attributes that co-occur in images (e.g. ’furry’ and ’black’/’brown’) share ACNs.

(3) The recognition accuracy of ImageNet objects is significantly reduced when ACNs of all attributes are ablated from the conv-net, significantly more so, than when an equal number of randomly sampled nodes are ablated. This suggests that conv-nets actually make use of learned attribute representations (through ACNs) to recognize objects in images. Interestingly, ablating ACNs corresponding to a specific set of attributes (e.g. ‘furry’, ‘black’, and ‘brown’) has the most effect on object classes that are described by these attributes (e.g. ‘retriever’ and ’gazelle’) and the least effect on classes that are not (e.g. ‘bathtub’ and ‘chain’).

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

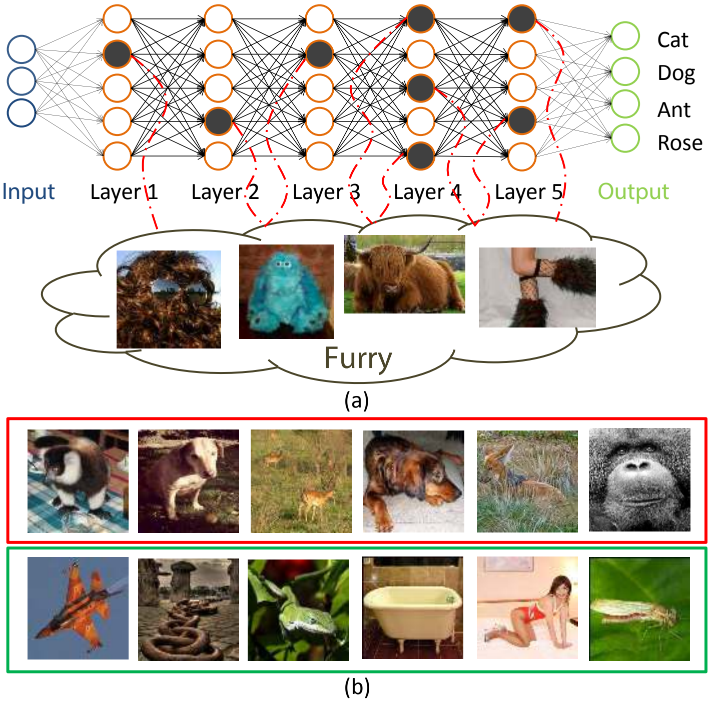

Figure 1: (a) Given the versatility and excellent performance of conv-nets on various visual recognition tasks, it is plausible that mid-level representations such as visual attributes are encoded by the network’s activations. The location and sparsity of this encoding, as well as, the general properties of the relationship between attributes and pre-trained conv-nets are the focus of this work. (b) Interestingly, ablating nodes associated with an specific set of attributes such as *brown, black and furry*, impact the capacity of the network to recognize objects such as ’gazelle’, ’retriever’ (inside the red box), while objects not related with these attributes such as ’maillot’, ’bathtup’ (inside the green box) are less or not affected.

Acknowledgments Research reported in this publication was supported by competitive research funding from King Abdullah University of Science and Technology (KAUST). JCN is supported by a Microsoft Research Faculty Fellowship.

[1] Pulkit Agrawal, Ross Girshick, and Jitendra Malik. Analyzing the performance of multilayer neural networks for object recognition. In *European Conference on Computer Vision (ECCV)*, 2014.

[2] Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[4] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2014.