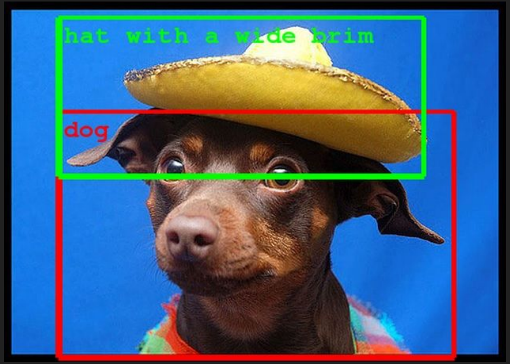
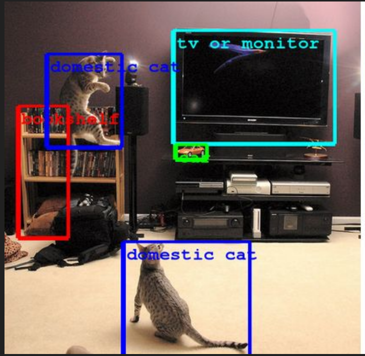
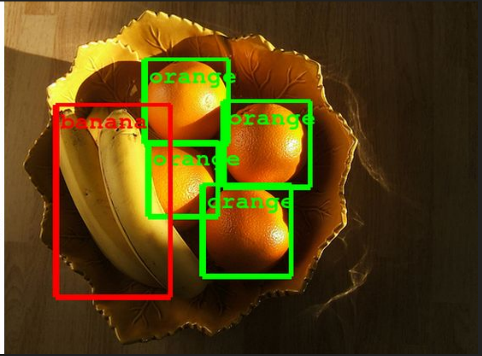
Do deep features retrieve X?

A tool for quick inspection of deep visual similarities

Julieta Martinez
December 2015







ADABAS

AKIBAN
TECHNOLOGIES

seahorse
VISUAL SPARK

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

1 Introduction

Current approaches to object recognition make essential use of machine learning methods. To improve their performance, we can collect larger datasets, learn more powerful models, and use better techniques for preventing overfitting. Until recently, datasets of labeled images were relatively small — on the order of tens of thousands of images (e.g., NORB [16], Caltech-101/256 [8, 9], and CIFAR-10/100 [12]). Simple recognition tasks can be solved quite well with datasets of this size, especially if they are augmented with label-preserving transformations. For example, the current-best error rate on the MNIST digit-recognition task (<0.3%) approaches human performance [4]. But objects in realistic settings exhibit considerable variability, so to learn to recognize them it is necessary to use much larger training sets. And indeed, the shortcomings of small image datasets have been widely recognized (e.g., Pinto et al. [21]), but it has only recently become possible to collect labeled datasets with millions of images. The new larger datasets include LabelMe [23], which consists of hundreds of thousands of fully-segmented images, and ImageNet [6], which consists of over 15 million labeled high-resolution images in over 22,000 categories.

To learn about thousands of objects from millions of images, we need a model with a large learning capacity. However, the immense complexity of the object recognition task means that this problem cannot be specified even by a dataset as large as ImageNet, so our model should also have lots of prior knowledge to compensate for all the data we don't have. Convolutional neural networks (CNNs) constitute one such class of models [16, 11, 13, 18, 15, 22, 26]. Their capacity can be controlled by varying their depth and breadth, and they also make strong and mostly correct assumptions about the nature of images (namely, stationarity of statistics and locality of pixel dependencies). Thus, compared to standard feedforward neural networks with similarly-sized layers, CNNs have much fewer connections and parameters and so they are easier to train, while their theoretically-best performance is likely to be only slightly worse.

CNN Features off-the-shelf: an Astounding Baseline for Recognition

Ali Sharif Razavian Hossein Azizpour Josephine Sullivan Stefan Carlsson
CVAP, KTH (Royal Institute of Technology)
Stockholm, Sweden
{razavian,azizpour,sullivan,stefanc}@csc.kth.se

arXiv:1403.6382v3 [cs.CV] 12 May 2014

Abstract

Recent results indicate that the generic descriptors extracted from the convolutional neural networks are very powerful. This paper adds to the mounting evidence that this is indeed the case. We report on a series of experiments conducted for different recognition tasks using the publicly available code and model of the OverFeat *network which was trained to perform object classification on ILSVRC13. We use features extracted from the* OverFeat *network as a generic image representation to tackle the diverse range of recognition tasks of object image classification, scene recognition, fine grained recognition, attribute detection and image retrieval applied to a diverse set of datasets. We selected these tasks and datasets as they gradually move further away from the original task and data the* OverFeat *network was trained to solve. Astonishingly, we report consistent superior results compared to the highly tuned state-of-the-art systems in all the visual classification tasks on various datasets. For instance retrieval it consistently outperforms low memory footprint methods except for sculptures dataset. The results are achieved using a linear SVM classifier (or $L2$ distance in case of retrieval) applied to a feature representation of size 4096 extracted from a layer in the net. The representations are further modified using simple augmentation techniques e.g. jittering. The results strongly suggest that features obtained from deep learning with convolutional nets should be the primary candidate in most visual recognition tasks.*

1. Introduction

"Deep learning. How well do you think it would work for your computer vision problem?" Most likely this question has been posed in your group's coffee room. And in response someone has quoted recent success stories [29, 15, 10] and someone else professed skepticism. You may have left the coffee room slightly dejected thinking "Pity I have neither the time, GPU programming skills nor large amount of labelled data to train my own network to quickly find out the answer". But when the convolutional neural network OverFeat [38] was recently made publicly available[1] it allowed for some experimentation. In particular we wondered now, not whether one could train a deep network specifically for a given task, but if the features extracted by a deep network - one carefully trained on the diverse ImageNet database to perform the specific task of image classification - could be exploited for a wide variety of vision tasks. We now relate our discussions and general findings because as a computer vision researcher you've probably had the same questions:

Prof: First off has anybody else investigated this issue?

Student: Well it turns out Donahue *et al*. [10], Zeiler and Fergus [48] and Oquab *et al*. [29] have suggested that generic features can be extracted from large CNNs and provided some initial evidence to support this claim. But they have only considered a small number of visual recognition tasks. It would be fun to more thoroughly investigate how

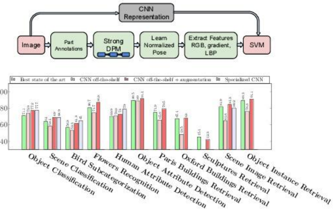
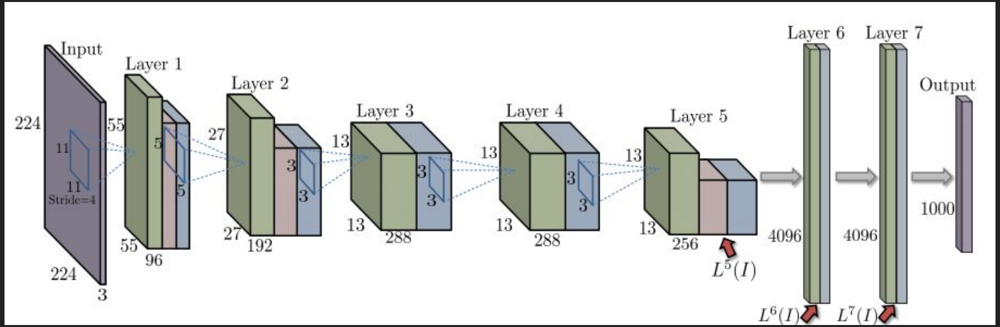


Figure 1: **top**) CNN representation replaces pipelines of s.o.a methods and achieve better results. e.g. DPD [50].
bottom) Augmented CNN representation with linear SVM consistently outperforms s.o.a. on multiple tasks. Specialized CNN refers to other works which specifically designed the CNN for their task

[1] There are other publicly available deep learning implementations such as Alex Krizhevsky's ConvNet and Berkeley's Caffe. Benchmarking these implementations is beyond the scope of this paper.



Daisy flower

$$\overbrace{[0.51, 0, 0.14, -0.34, 0, \ldots., 0.75, -0.29, -0.12]}^{4096}$$

[0.51, 0, 0.14, -0.34, 0, …., 0.75, -0.29, -0.12]

[0.12, -0.4, 0.14, -0.43, 0, …., 0.75, -0.29, -0.19]

[0.51, 0, 0.34, -0.76, 0, …., 0.85, -0.29, -0.52]

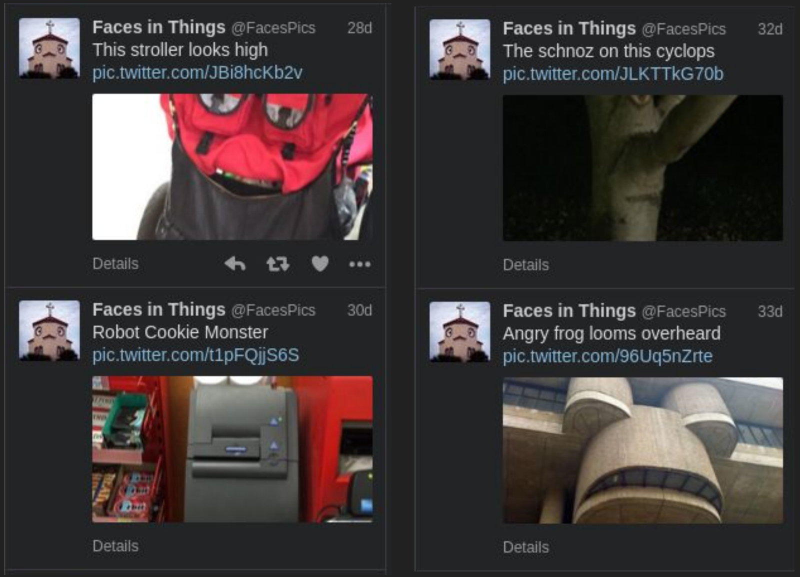
[0.51, 0, 0.14, -0.34, 0, …., 0.75, -0.29, -0.12]

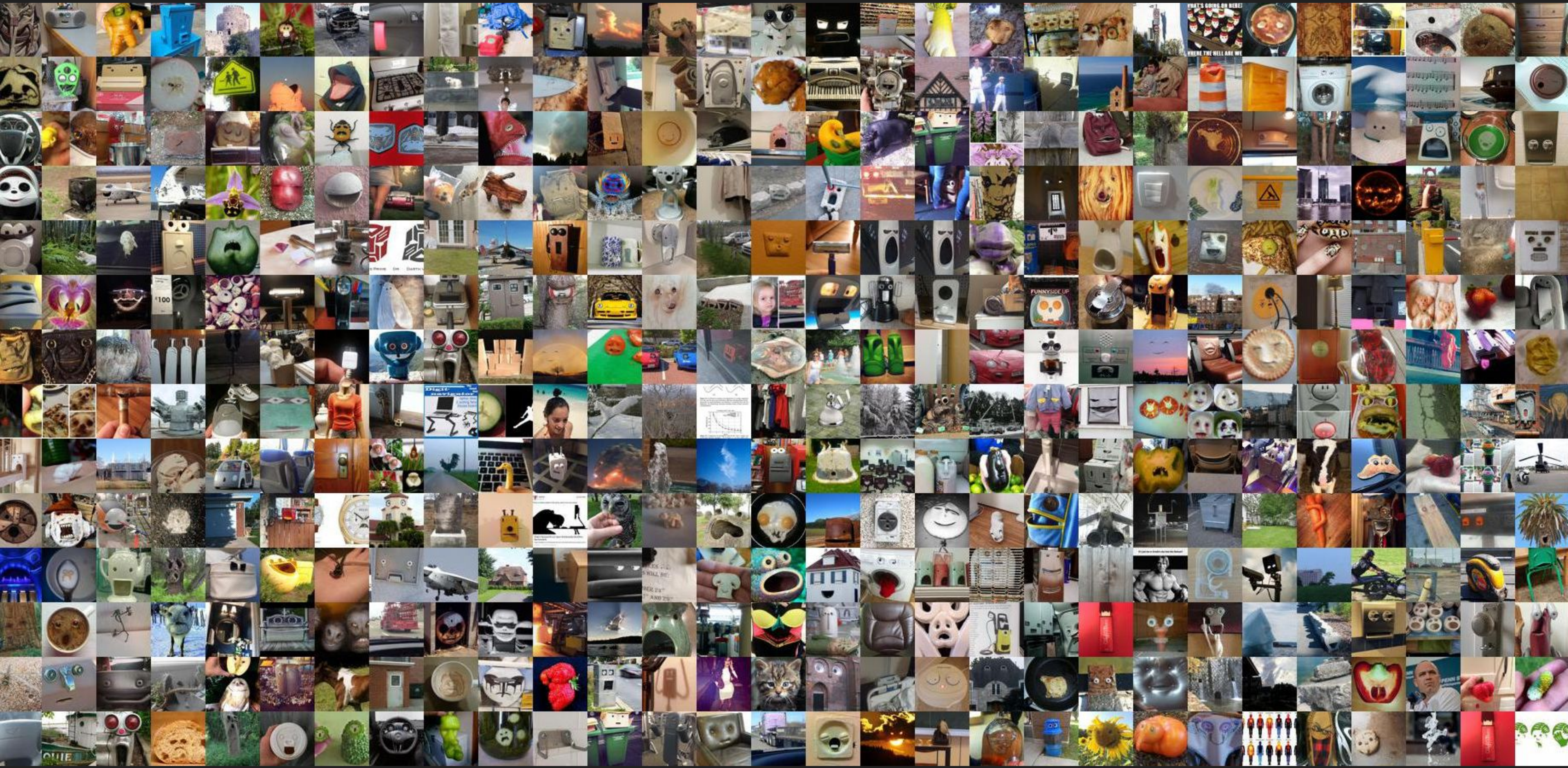
[0.51, 0, 0.23, -0.34, 0, …., -0.25, -0.29, -0.87]

[0.51, -0.93, 0.14, -0.34, 0, …., 0.75, -0.29, 0.02]

Do deep features retrieve X?

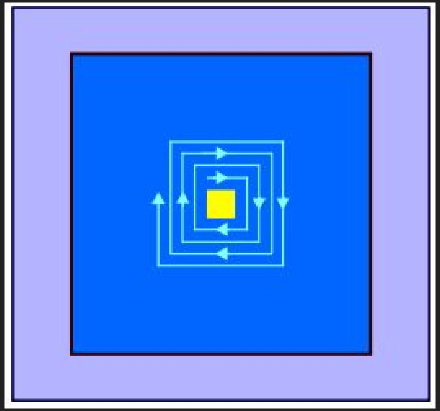
X = faces in things

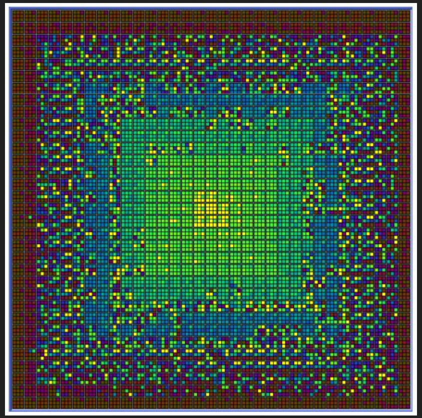




Tasks

- **Browse** a large image dataset
- **Explore** similarity neighbourhoods





Front Row to Fashion Week

By MIKE BOSTOCK, SHAN CARTER, ERIK HINTON, CATHY HORYN and ERIC WILSON | September 12, 2013

Of the more than 300 collections shown during New York Fashion Week, here were the ones that created the most buzz and left the biggest impressions on fashion editors as they headed off to the next round of shows in London, Milan and Paris.

View Full Screen

Calvin Klein

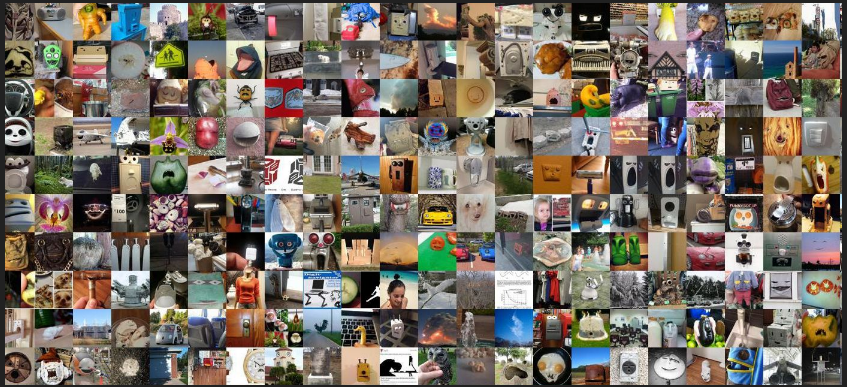
A beautiful, innovative collection in which Francisco Costa layered references to urban tribes, '80s art, handcraft and even, seemingly, radical chicks of the 1920s. It added up to a modern expression of fashion.

Read more: **Calvin Klein in Full Color**

Proenza Schouler

A challenging collection, inspired by the notion of home and interiors, it nonetheless showed the designers in a simpler vein.

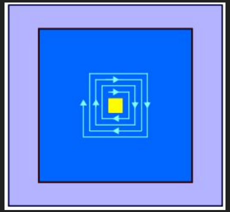
Read more: **Pleats and Prints**



+


Calvin Klein
A beautiful, innovative collection in which Francisco Costa layered references to urban tribes, '80s art, handcraft and even, seemingly, radical chicks of the 1920s. It added up to a modern expression of fashion.
Read more: Calvin Klein in Full Color

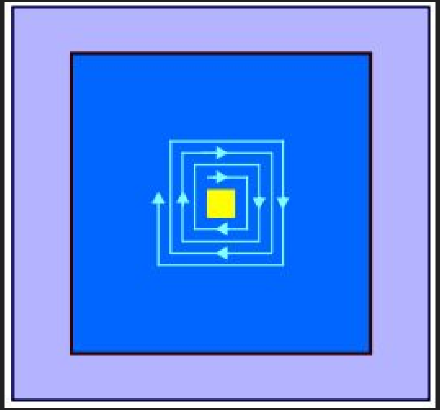
+

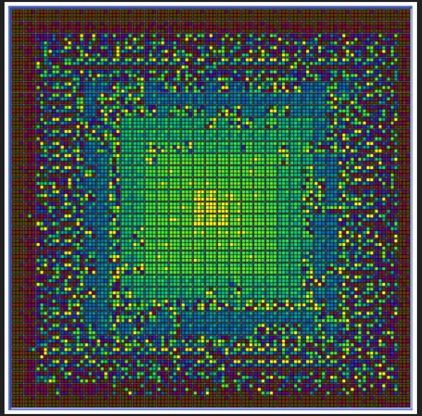


=

Tasks

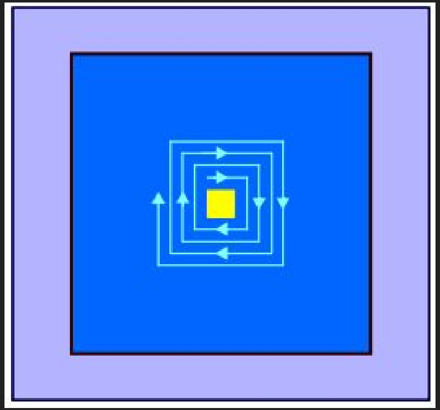
- **Browse** a large image dataset
- **Explore** similarity neighbourhoods

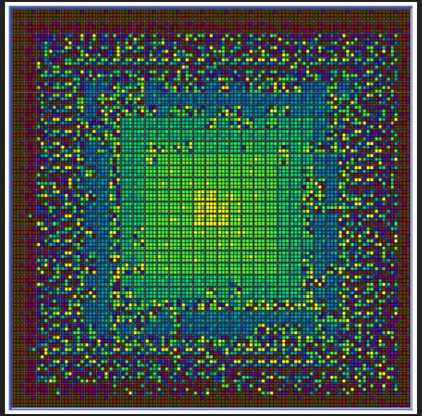




Tasks

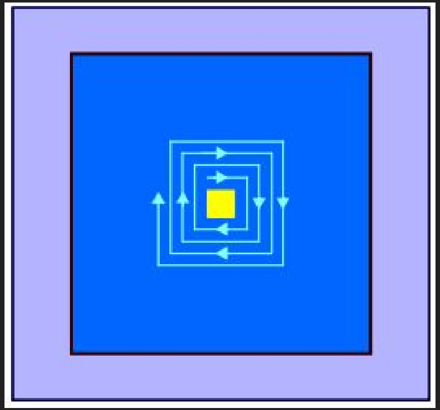
- **Browse** a large image dataset
- **Explore** similarity neighbourhoods
- **Explore** similarity distributions

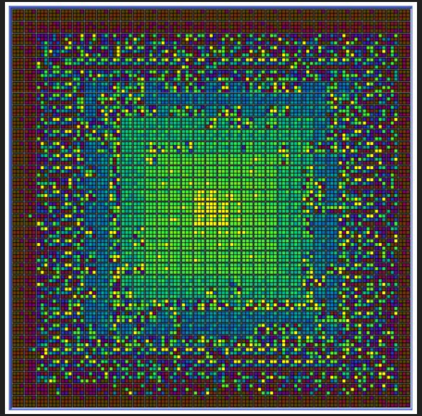


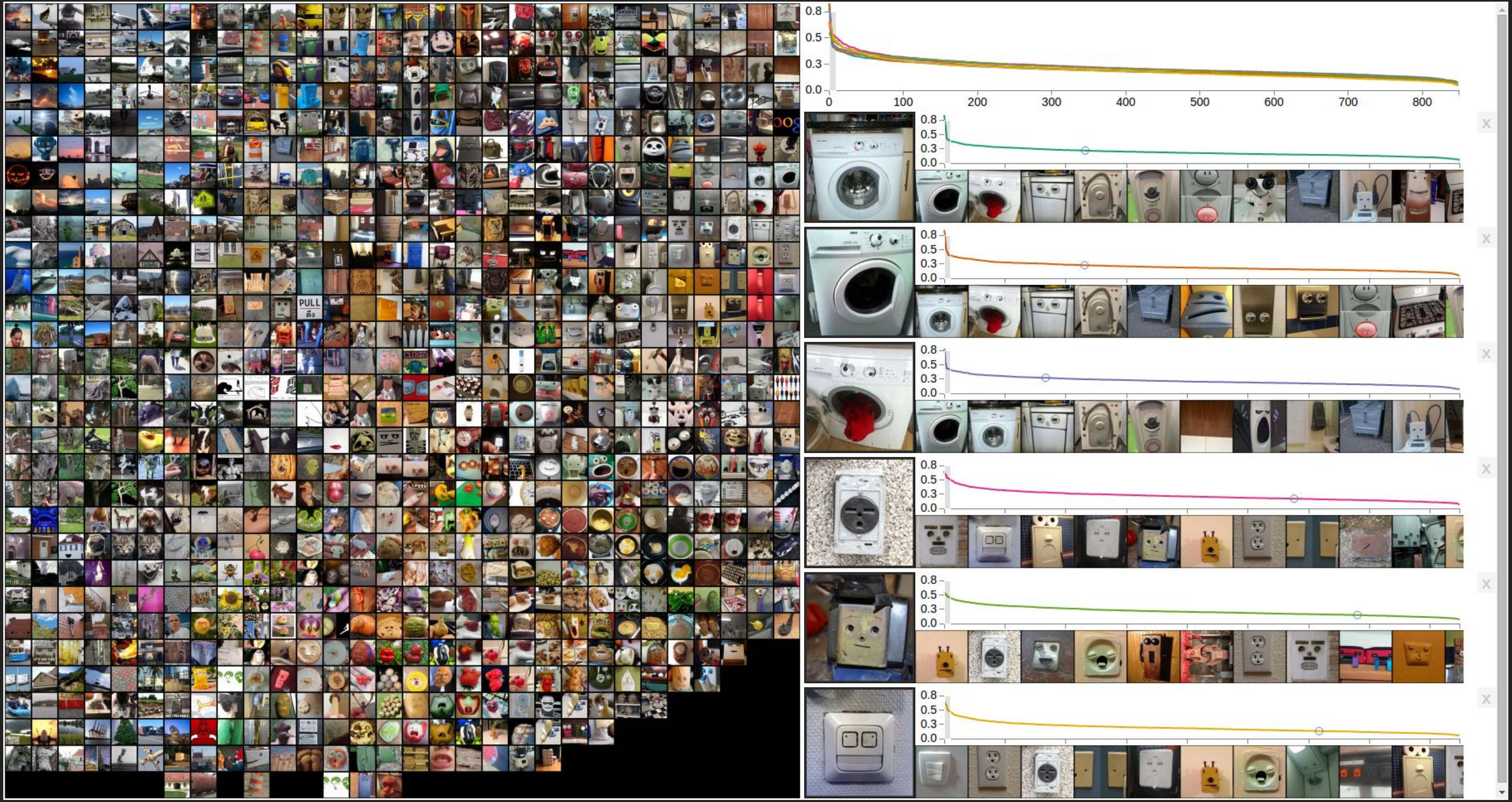


Tasks

- **Browse** a large image dataset
- **Explore** similarity neighbourhoods
- **Explore** similarity distributions
- **Compare** query distributions






0.8
0.5
0.3
0.0
0
100
200
300
400
500
600
700
800
PULL
CLOSED

So, do deep features retrieve faces in things?

No, they retrieve things
(but now we know!)

Do deep features retrieve X?

A tool for quick inspection of deep visual similarities

Julieta Martinez
December 2015