Visualizing and Understanding Convolutional Networks

Implementation of a paper by
Matthew D. Zeiler and Rob Fergus [1]

IN4155 - Deep Learning for the Real World

June 2017

Fabian Hainzl - fabian.hainzl@tum.de

The Idea

Goal Determine what each layer learns by visualizing the input stimuli that excite individual feature maps

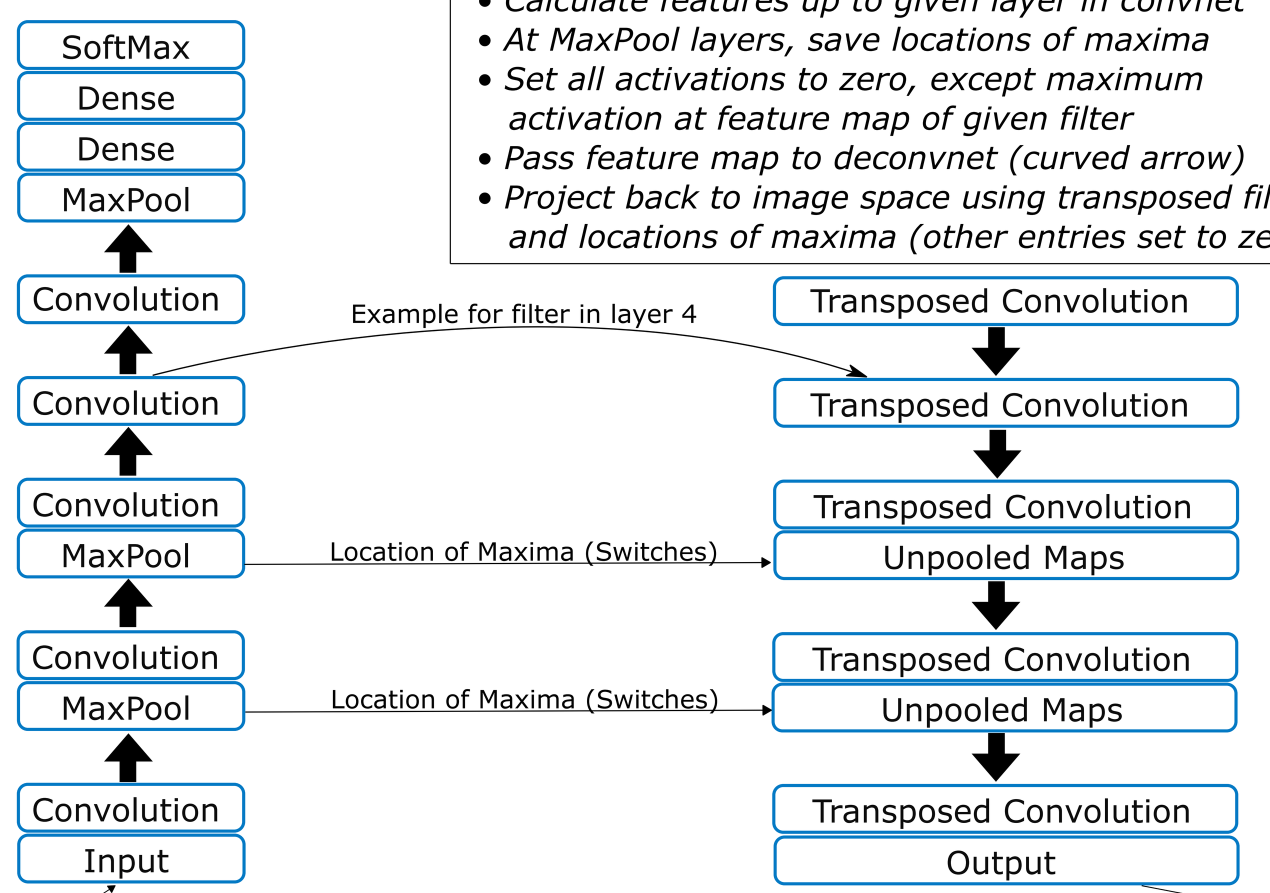
Method Attaching a Deconvolutional Network to a trained Convolutional Network in order to project feature maps back to input pixel space

Convolutinal Network
- Projects input pixels to feature space
- Convolutes output of previous layer with set of learned filters
- Uses maxpooling layers

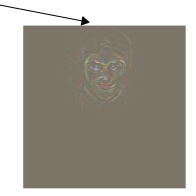
Deconvolutional Network
- Projects feature maps back to image space
- Uses transposed versions of the same filters on layer above
- Creates unpooled maps by saving locations of maxima

The Model

Visualize pattern that maximally activates a given filter
- *Calculate features up to given layer in convnet*
- *At MaxPool layers, save locations of maxima*
- *Set all activations to zero, except maximum activation at feature map of given filter*
- *Pass feature map to deconvnet (curved arrow)*
- *Project back to image space using transposed filters and locations of maxima (other entries set to zero)*

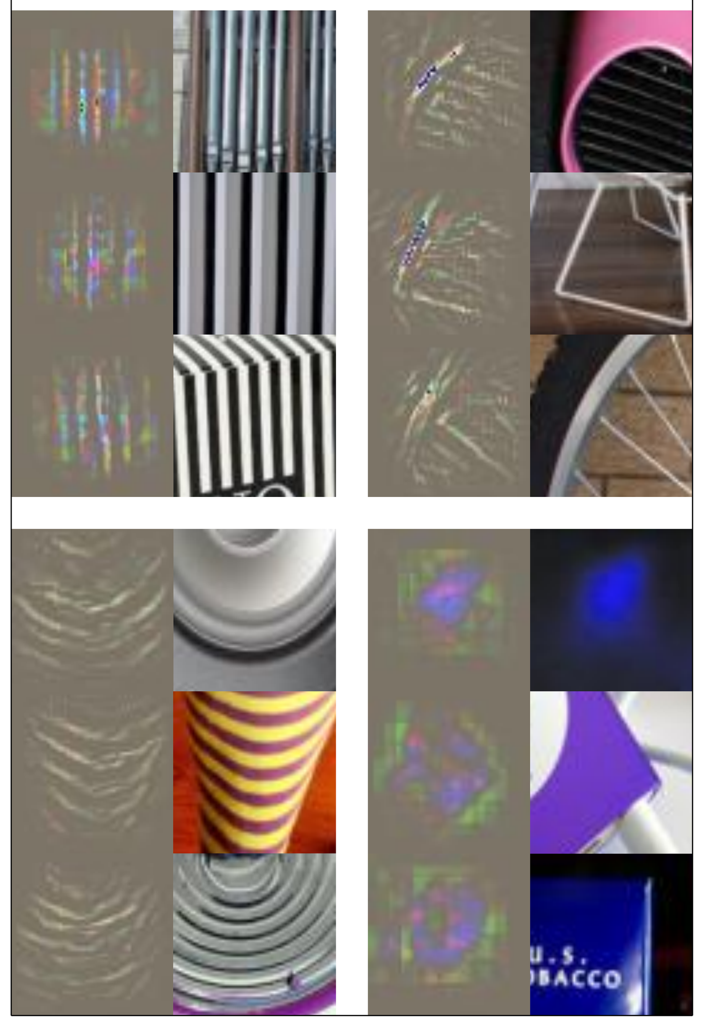


Architecture of convnet (AlexNet) and attached deconvnet. Rectified linear functions (not shown) are used in both networks, the AlexNet is trained on the ImageNet dataset.

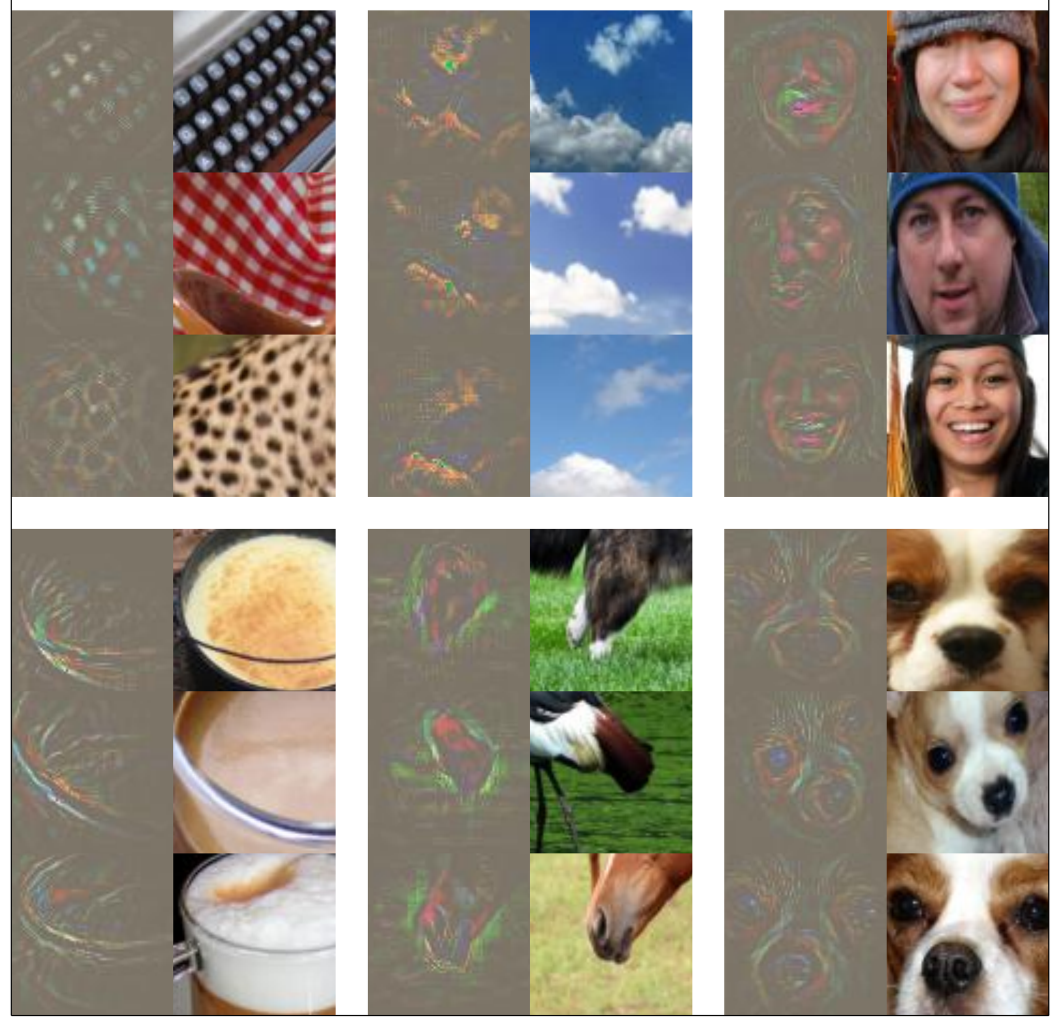


The Visualization Results

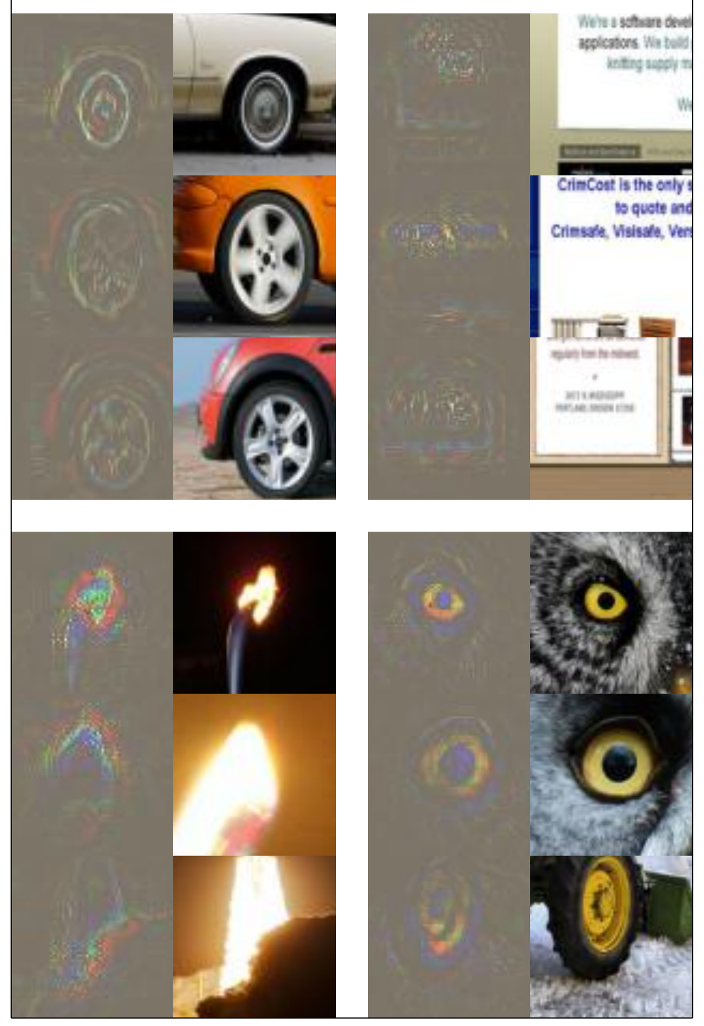
Layer #2
Simple features: Parallel lines/curves, conjunctions, color dashes



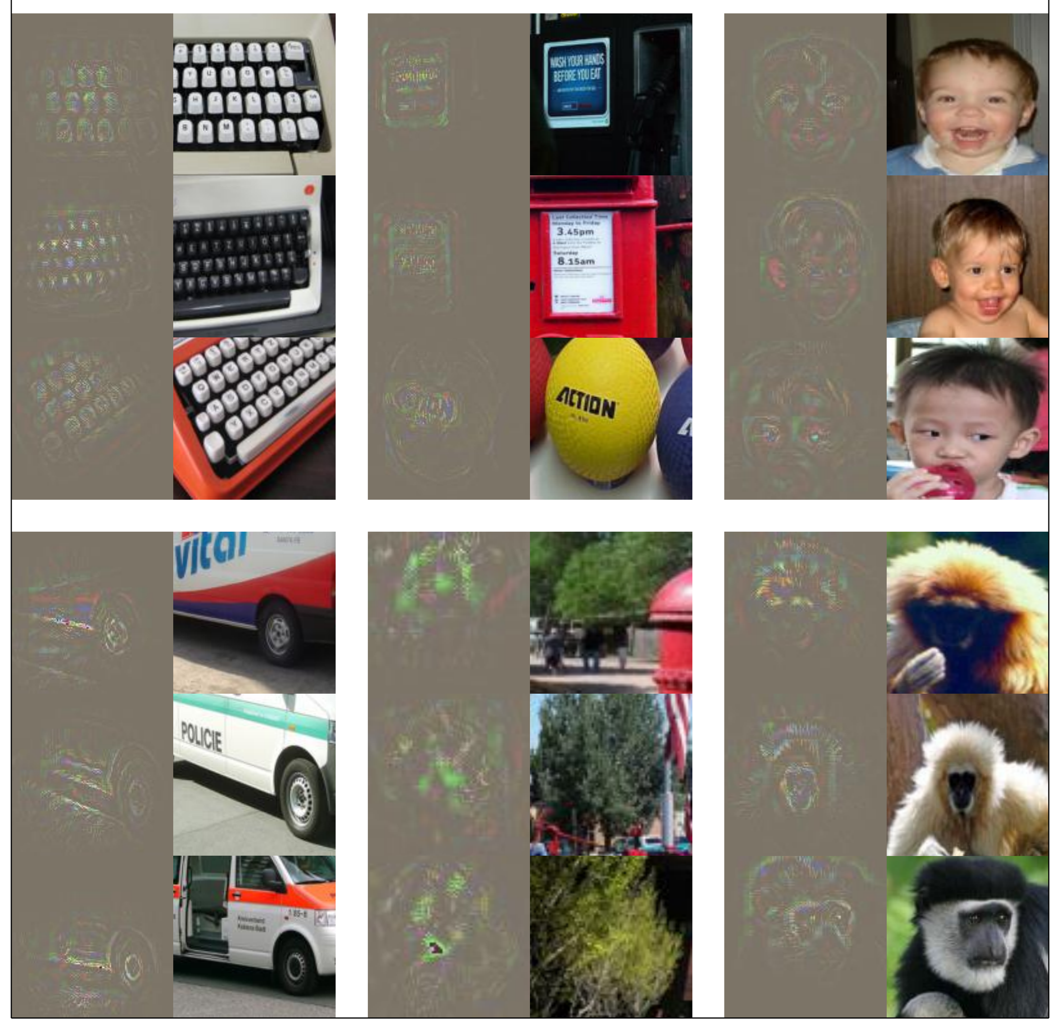
Layer #3
Combinations of lower level features:
Texture (meshes, patterns), backgrounds (sky, grass)



Layer #4
More variation in activating patterns, colors still a discriminative factor



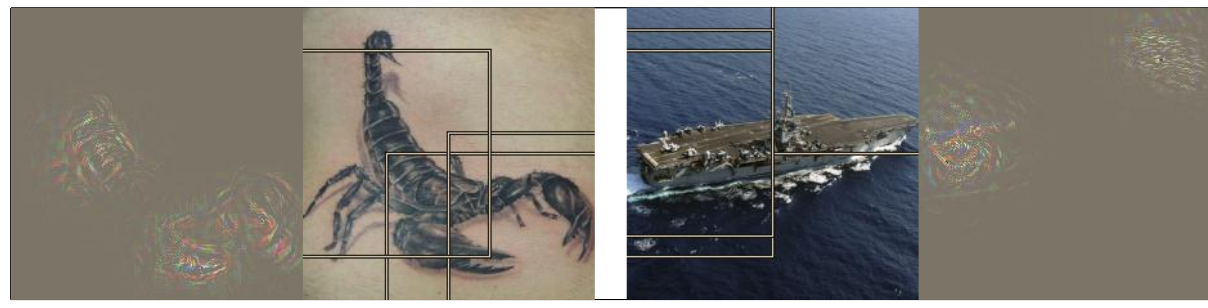
Layer #5
Entire objects with pose/color variation (keyboard), very specific features (children's faces, utility vehicles)



Shown above are the projections of three of the strongest activations for several feature maps from each layer, as well as the corresponding image patches. Discriminative parts of the image are exaggerated in the projections. With increasing layer depth, the feature maps cover larger parts of the original image and show greater invariance of changes in color/position, thereby connecting images more on a semantic than on a structural level.

Multiple layers/features

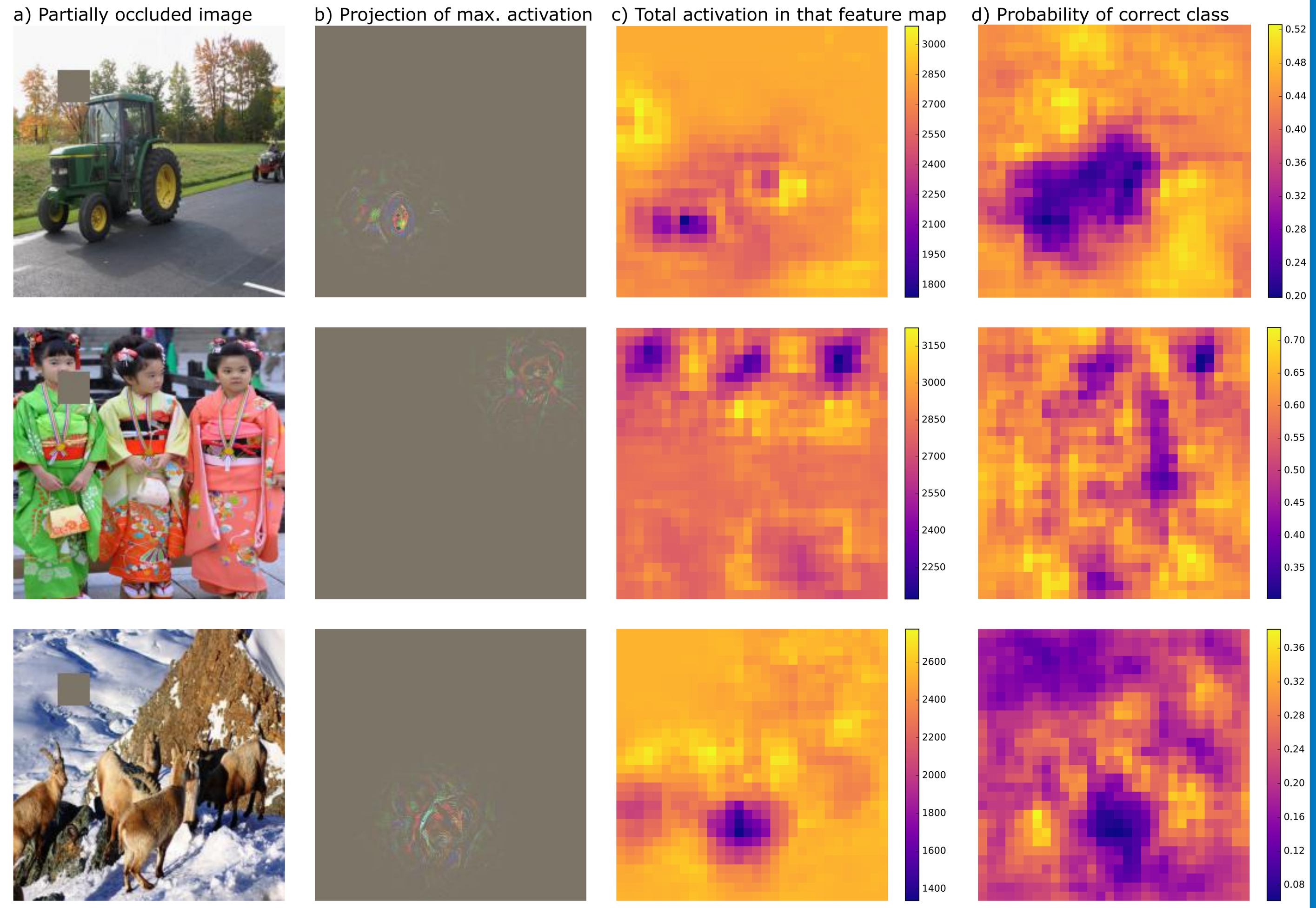
Projection of the single strongest activation in each layer (1-5) next to original image. Boxes show sizes of responsible image parts and increase with layer number.



Projections of the three strongest activations in the top layer next to original image.

Occlusion Sensitivity

To show the influence of image parts to a feature map's total activation, as well as the classifier's sensitivity to certain areas in the input, an image is partially covered by a grey square at various positions. When the occluder covers the image region that appears in the visualization, the activity in that feature map drops significantly, showing the correspondence between the projection and the image structure which stimulates this feature map.



(a) Three examples where different parts of the image are systematically covered with a gray square. (b) Projection of the strongest Layer 5 activation. As a function of gray square position, (c) plots the total activation of the feature map shown in (b), summed over spacial dimensions, while (d) shows probability of the correct class.

[1] Zeiler M., Fergus R., Visualizing and Understanding Convolutional Networks

AlexNet implementation and weights from: https://github.com/heuritech/convnets-keras