

Computer Vision and Deep Learning

Network Visualization

Matthias Fulde

WS 2023/24



Topics

- ▶ Visualize Parameters
- ▶ Visualize Activations
- ▶ Discover Relevant Features
 - ▶ Occlusion Experiments
 - ▶ Saliency Maps
- ▶ Visualize Feature Space
- ▶ Visualize Neurons
 - ▶ Activation Maximization

Problem

- ▶ Deep neural networks are generally opaque
 - ▶ It is not clear why and how exactly a final prediction is made
 - ▶ We have little knowledge about the inner structure of these algorithms
 - ▶ Neural network as a blackbox



Problem

- ▶ Being able to interpret neural networks is important
 - ▶ Requirement for acceptance of critical machine learning applications
 - ▶ Fundamental for further development

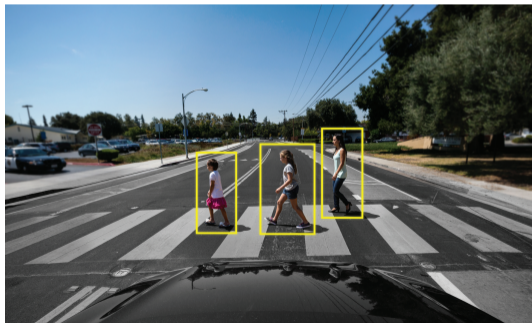


Image from <https://blogs.nvidia.com/blog/2019/04/15/how-does-a-self-driving-car-see/>

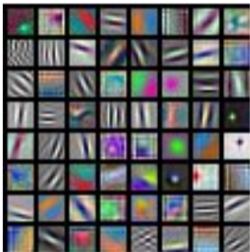
Visualization

- ▶ Visualizing neural networks can help revealing their inner structure
- ▶ We can investigate
 - ▶ which input features are relevant for the final prediction
 - ▶ which kind of information is learned by the different network layers
 - ▶ which kind of information is learned by the individual neurons
 - ▶ how the representational space of the network looks like

Visualize Parameters

Visualize Parameters

- ▶ Show network parameters as images
 - ▶ Weights encode which of the respective input features are important
 - ▶ Filters in first network layer encode graphic primitives and colors
 - ▶ Similar to human visual system



AlexNet:
64 x 3 x 11 x 11



ResNet-18:
64 x 3 x 7 x 7



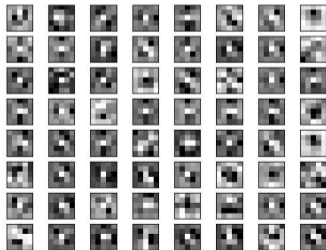
ResNet-101:
64 x 3 x 7 x 7



DenseNet-121:
64 x 3 x 7 x 7

Visualize Parameters

- ▶ Filters in deeper network layers usually have more than three channels
 - ▶ We can show each channel as a grayscale image
- ▶ Weights are far less interpretable
 - ▶ No longer based on image features but on output features of previous layer
 - ▶ Example for first filter of second layer in pretrained AlexNet



Visualize Activations

Visualize Activations

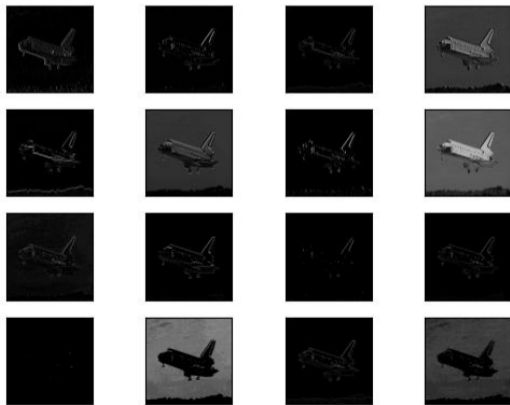
- ▶ Inspect network by extracting activation maps for particular input image
 - ▶ Compute forward pass through network
 - ▶ Show activation maps generated by applying filters to the input
- ▶ Consider the following input image



Image from Nasa/science photo library

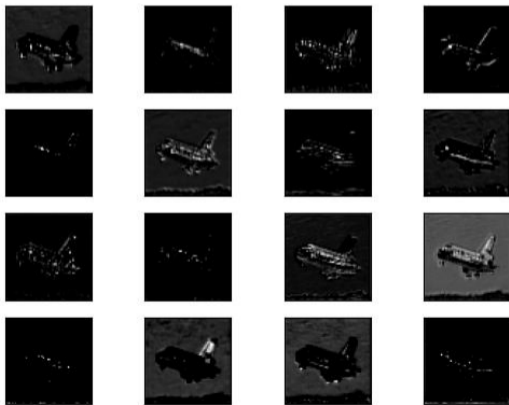
Visualize Activations

- ▶ First 16 activation maps of layer 1 from VGG16 network pretrained on ImageNet



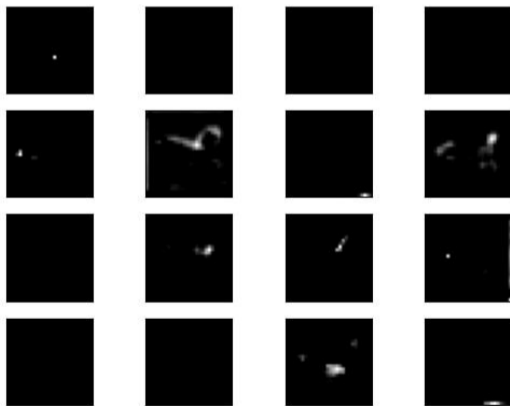
Visualize Activations

- ▶ First 16 activation maps of layer 5 from VGG16 network pretrained on ImageNet



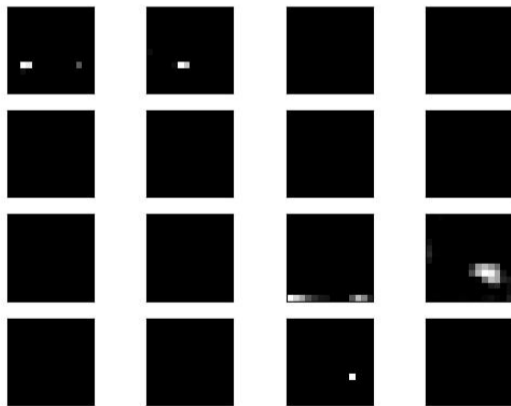
Visualize Activations

- ▶ First 16 activation maps of layer 10 from VGG16 network pretrained on ImageNet



Visualize Activations

- ▶ First 16 activation maps of layer 13 from VGG16 network pretrained on ImageNet

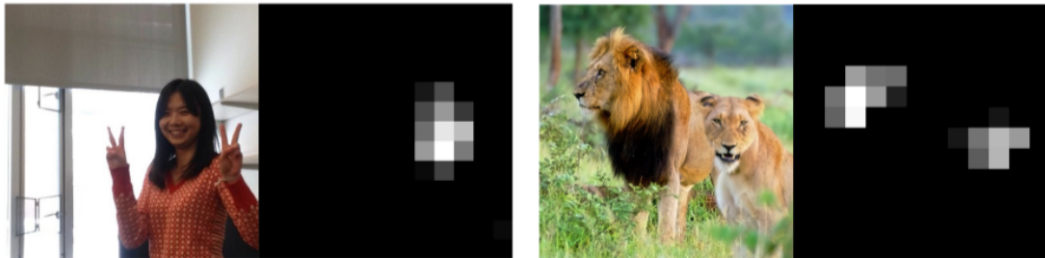


Visualize Activations

- ▶ Activation maps in models with ReLU activation are mostly sparse and localized
 - ▶ Negative input features are clipped
- ▶ Activation maps in early layers show more details of the input image
 - ▶ Model extracts low level features in early layers
- ▶ Activation maps in later layers are less detailed
 - ▶ Model abstracts image features into more general concepts in later layers
 - ▶ No longer interpretable
- ▶ Can be used to detect dead filters
 - ▶ Filters whose activations are all zero across all images in dataset

Visualize Activations

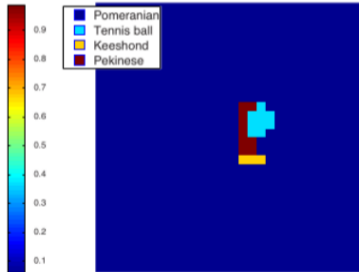
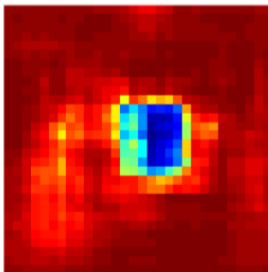
- ▶ Even though activations are usually not interpretable we can gain some intuition how the network works
- ▶ Comparing images to neural responses we can get an idea what neurons are looking for in the input



Discover Relevant Features

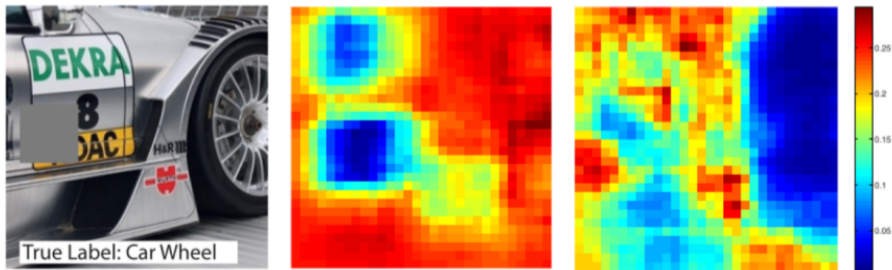
Occlusion Experiments

- ▶ Visualize which parts of the image are most important for classification
 - ▶ Mask part of the input image using mean pixel value in dataset
 - ▶ Move mask across the image and record probability of ground truth class
 - ▶ Draw heatmap of probability at each location



Occlusion Experiments

- ▶ Can also be used to help interpretation of feature maps
 - ▶ Select feature maps with highest activations for given input image
 - ▶ Observe for which mask positions activations drop the most



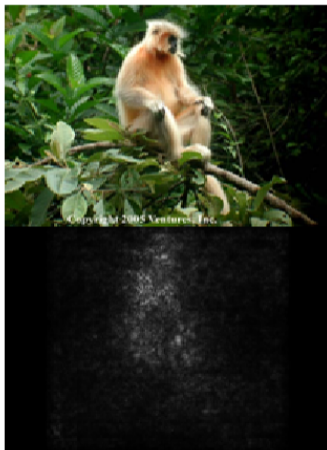
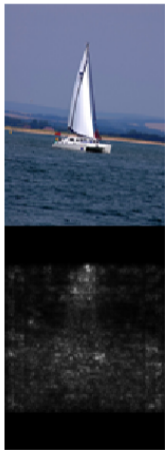
Saliency Maps

- ▶ Another approach to visualize what the network is looking for in the input image
 - ▶ Consider color image I and unnormalized class score $S(I)$
 - ▶ Compute gradient of S with respect to I
 - ▶ Take the maximum absolute value across channels to compute saliency map

$$M_{i,j} = \max_c |\nabla S(I)_{i,j,c}|$$

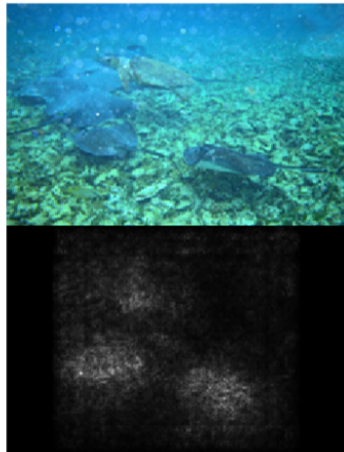
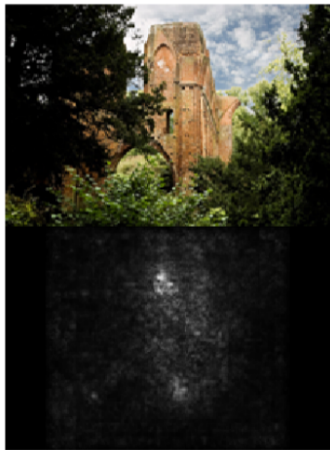
- ▶ Magnitude of gradient indicates which image pixels need to be changed the least to affect the class score the most.
 - ▶ These pixels should correspond with the location of the object in the image
 - ▶ But might also reveal other regions of interest

Saliency Maps



Figures from Visualising Image Classification Models and Saliency Maps, Simonyan et al., 2014

Saliency Maps



Figures from Visualising Image Classification Models and Saliency Maps, Simonyan et al., 2014

Visualize Feature Space

KNN in Feature Space

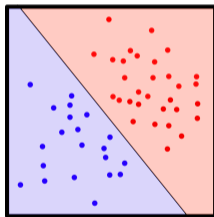
- ▶ Investigate which input images the network deems to be similar
 - ▶ Perform k-nearest neighbor in feature space of the last hidden layer of the network
 - ▶ Similarity is based more on semantic than on pixel values



Figure from ImageNet Classification with Deep Convolutional Neural Networks, Krizhevsky et al., 2012

KNN in Feature Space

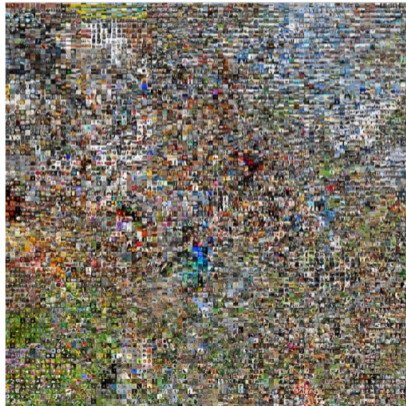
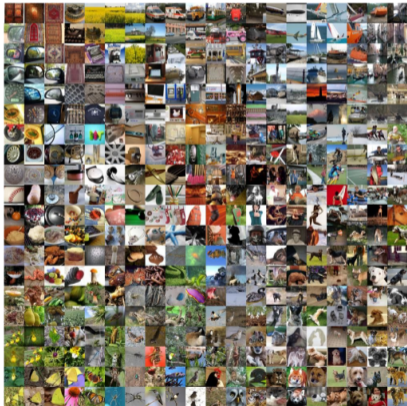
- ▶ Explanation for observed similarities
 - ▶ Classifier network learns representational space where data is linearly separable
 - ▶ Last fully connected layer learns linear decision boundaries



- ▶ Similar representations in feature space can be explicitly forced
 - ▶ Use contrastive or triplet loss for training

Dimensionality Reduction

- ▶ Distribution of feature representations can be visualized using dimensionality reduction algorithms



Dimensionality Reduction

- ▶ Allows us to get an idea about the topology of the representational space
- ▶ General approach
 - ▶ Extract feature vectors for set of images from last hidden layer
 - ▶ Reduce dimensionality to 2D or 3D with algorithm preserving distances (t-SNE)
 - ▶ Use obtained coordinates to display corresponding images

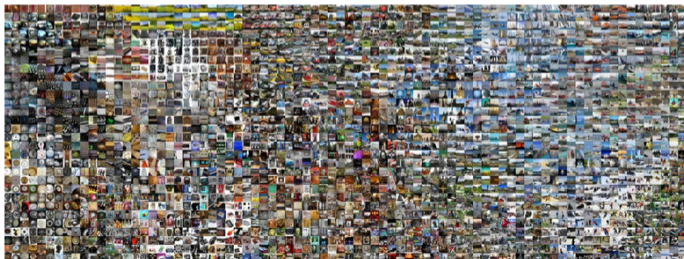
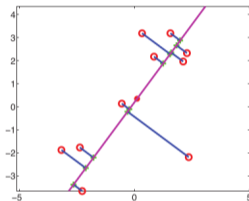


Figure from Stanford cs231n course

Dimensionality Reduction

- ▶ Principal component analysis does not well preserving distances

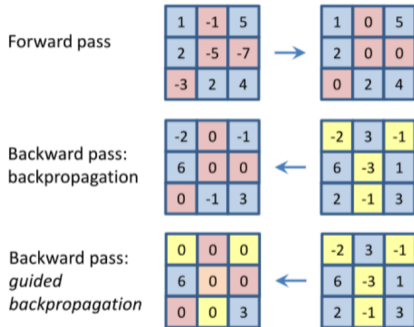


- ▶ Better use t-distributed stochastic neighbor embedding (t-SNE)
 - ▶ Nonlinear dimensionality reduction algorithm
 - ▶ Minimizes Kullback-Leibler divergence between distributions of pairs in high and low dimensional spaces

Visualize Neurons

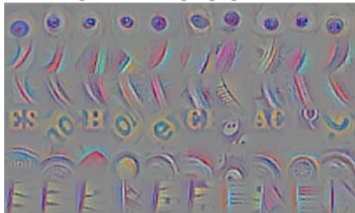
Guided Backpropagation

- ▶ Visualize what intermediate neurons look for in the input
- ▶ Compute gradient of neuron activation with respect to image
- ▶ Improve quality of visualization using guided backpropagation
- ▶ Only propagate positive gradients through ReLU activations



Guided Backpropagation

guided backpropagation



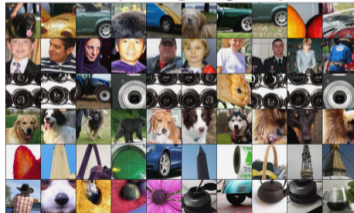
corresponding image crops



guided backpropagation



corresponding image crops



Maximally Activating Images

- ▶ Visualize image patches that correspond to maximal activations of a neuron
 - ▶ Run images through network and record activations of neuron
 - ▶ Select images for which activation is maximal
 - ▶ Draw bounding box for receptive field

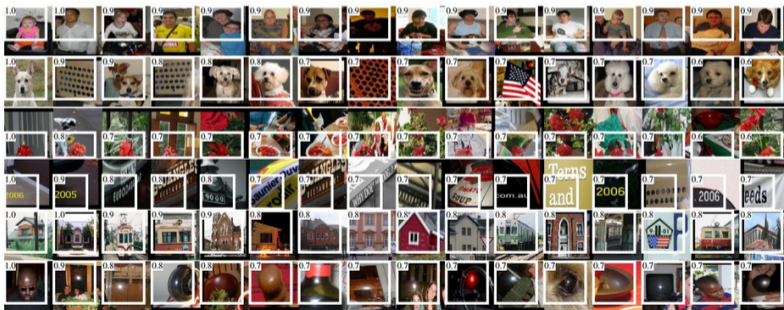
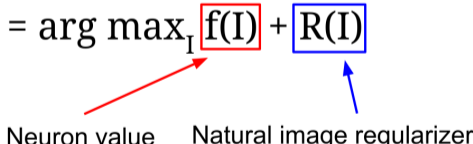


Figure from Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, Girshick, 2014

Activation Maximization

Image Generation

- ▶ Another approach to visualize the features of neurons in a network
- ▶ Generate images that maximally activate a particular neuron
 - ▶ Consider input image I and neuron activation $f(I)$
 - ▶ Use gradient ascent to find image I^* that maximizes f under some constraints

$$I^* = \arg \max_I f(I) + R(I)$$


Neuron value Natural image regularizer

Image Generation

- ▶ For instance find image I^* that maximizes class score $S_k(I)$ for class k

$$I^* = \arg \max_I S_k(I) - \frac{\lambda}{2} \|I\|^2$$

- ▶ Score $S_k(I)$ is neuron activation in last layer before normalization
- ▶ The procedure is
 - ▶ Initialize image I with zeros
 - ▶ Forward image through network to compute scores
 - ▶ Backpropagate to get gradient of neuron activation with respect to image pixels
 - ▶ Make gradient ascent update to the image
- ▶ Use the same approach to visualize intermediate features

Image Generation

- ▶ Examples of images maximizing class scores $S_k(I)$ obtained with L^2 regularization

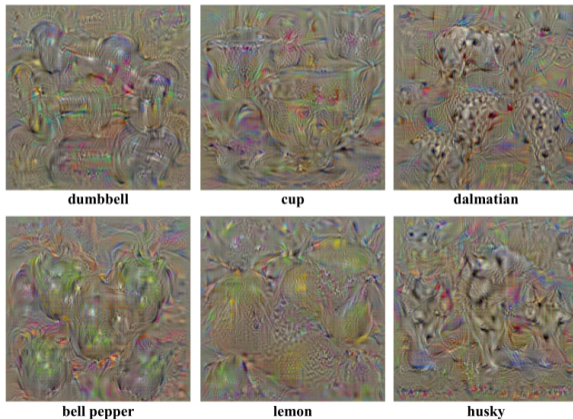
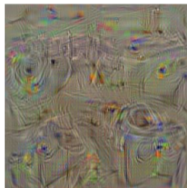
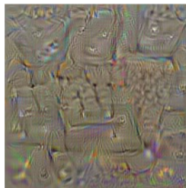


Image Generation

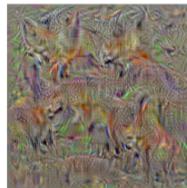
- ▶ Examples of images maximizing class scores $S_k(I)$ obtained with L^2 regularization



washing machine



computer keyboard



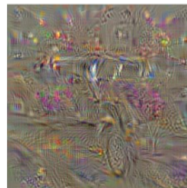
kit fox



goose



ostrich

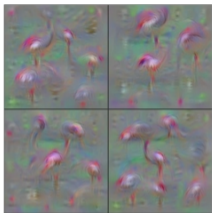


limousine

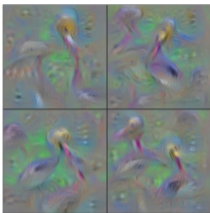
Image Generation

- ▶ Image generation is constraint optimization problem
 - ▶ Appearance of generated images should be close to samples from true distribution
 - ▶ Use of L^2 regularization prevents small number of pixels with extreme values to dominate the image
 - ▶ But generated images still far from naturalistic and hard to interpret
- ▶ We observe that produced images have unrealistic amount of high frequency information
 - ▶ Useful regularization to apply Gaussian blur filter to penalize high frequencies
 - ▶ Smooth image periodically
- ▶ Further improvements by clipping pixels with small values or gradients to zero

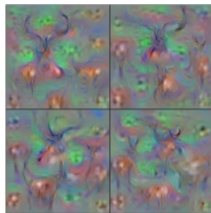
Image Generation



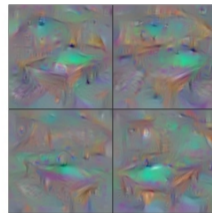
Flamingo



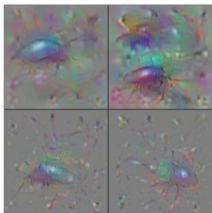
Pelican



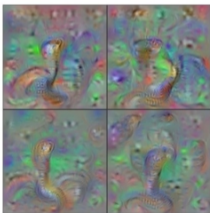
Hartebeest



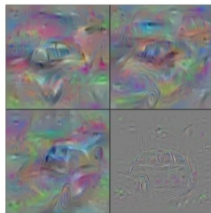
Billiard Table



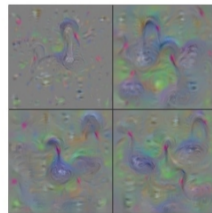
Ground Beetle



Indian Cobra



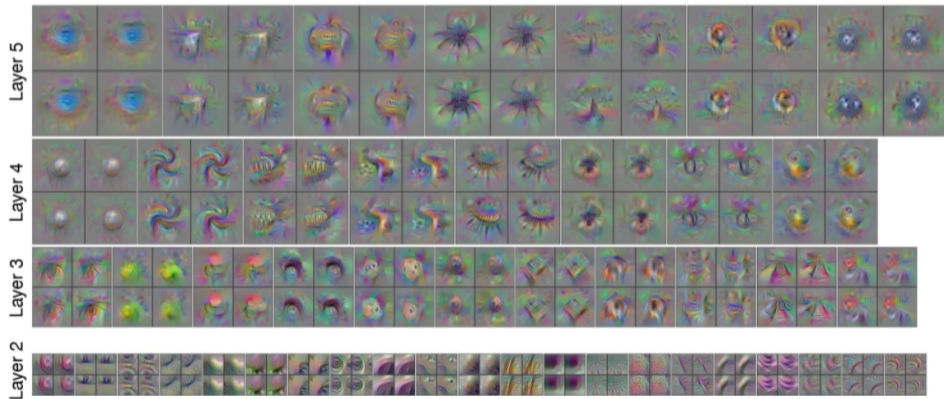
Station Wagon



Black Swan

Image Generation

- ▶ Same approach used to visualize intermediate features from network



Multifaceted Feature Visualization

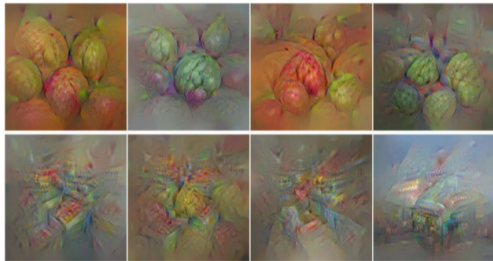
- ▶ Implicit assumption so far that one neuron only detects one type of feature
 - ▶ In reality neurons fire in response to many different types of features
 - ▶ Neurons are multifaceted
- ▶ Missing to incorporate that prior leads to rather poor results
 - ▶ Generated images seen so far have inappropriate mixes of colors, parts of objects, scales, orientations, and incoherent global structure
 - ▶ Reduced interpretability
- ▶ The goal should be to generate a maximally activating image for each facet
 - ▶ Disentangle different types of features
 - ▶ Avoid repetitions of objects

Multifaceted Feature Visualization

- ▶ Choose different initializations for activation maximization algorithm
 - ▶ Project images that maximally activate a neuron into 2D space using t-SNE
 - ▶ Cluster projected images with k-means
 - ▶ Average the N closest images to each cluster centroid
 - ▶ Use averaged images as starting points
- ▶ Also apply regularization to prevent repetitions of object fragments
 - ▶ Add center bias to produce one central object
 - ▶ Allow, on average, more iterations for center pixels than for edge pixels

Multifaceted Feature Visualization

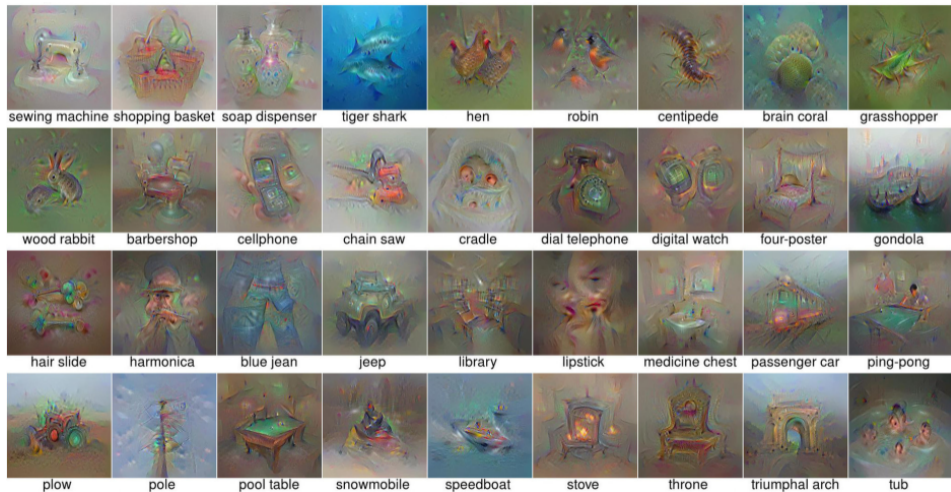
Reconstructions of multiple feature types (facets) recognized by the same "grocery store" neuron



Corresponding example training set images recognized by the same neuron as in the "grocery store" class



Multifaceted Feature Visualization



Figures from Multifaceted Feature Visualization, Nguyen et al., 2016

Latent Space Optimization

- ▶ Fundamentally different approach to activation maximization
 - ▶ Optimize with respect to latent space representation of generative network
 - ▶ Replace hand crafted natural image priors with learned priors
 - ▶ Results are far more interpretable

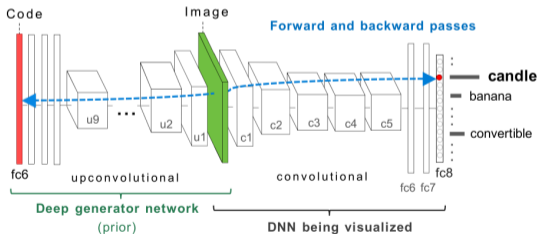


Figure from Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks, Nguyen et al., 2016

Latent Space Optimization



Figure from Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks, Nguyen et al., 2016

Summary

Summary

- ▶ Visualize Parameters
- ▶ Visualize Activations
- ▶ Discover Relevant Features
 - ▶ Occlusion Experiments
 - ▶ Saliency Maps
- ▶ Visualize Feature Space
- ▶ Visualize Neurons
 - ▶ Activation Maximization