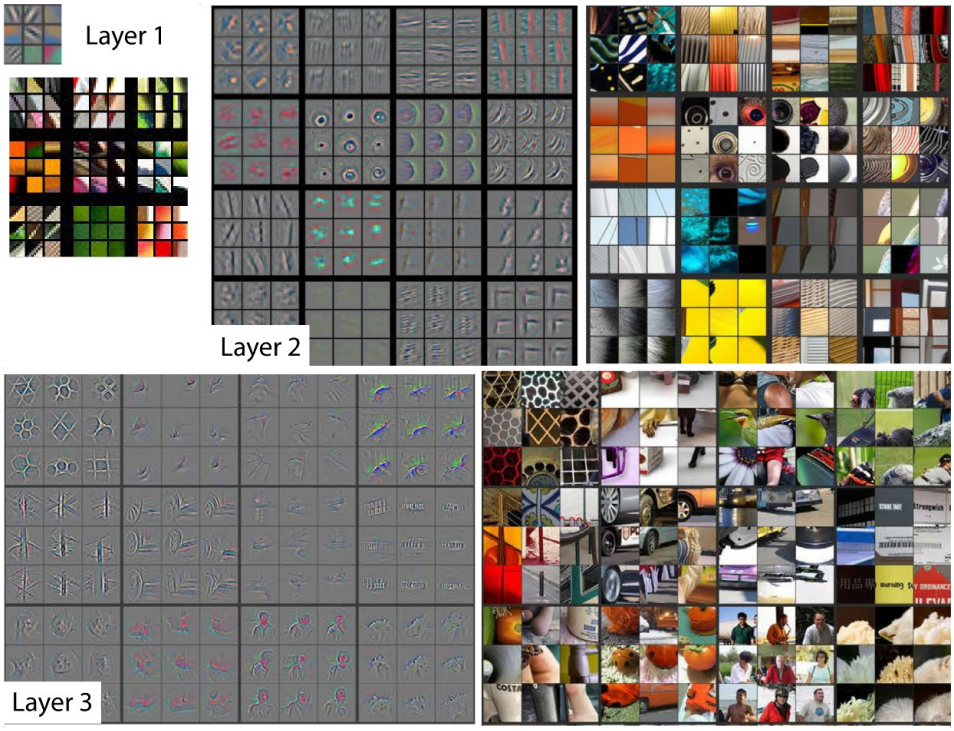
Deep learning

1.3. What is really happening?

François Fleuret

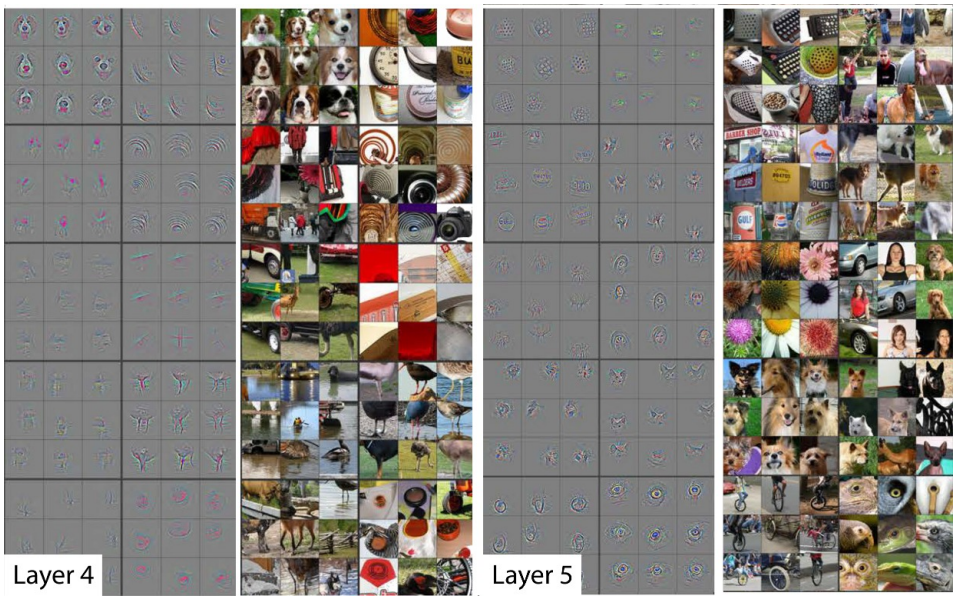
https://fleuret.org/dlc/



(Zeiler and Fergus, 2014)

Notes

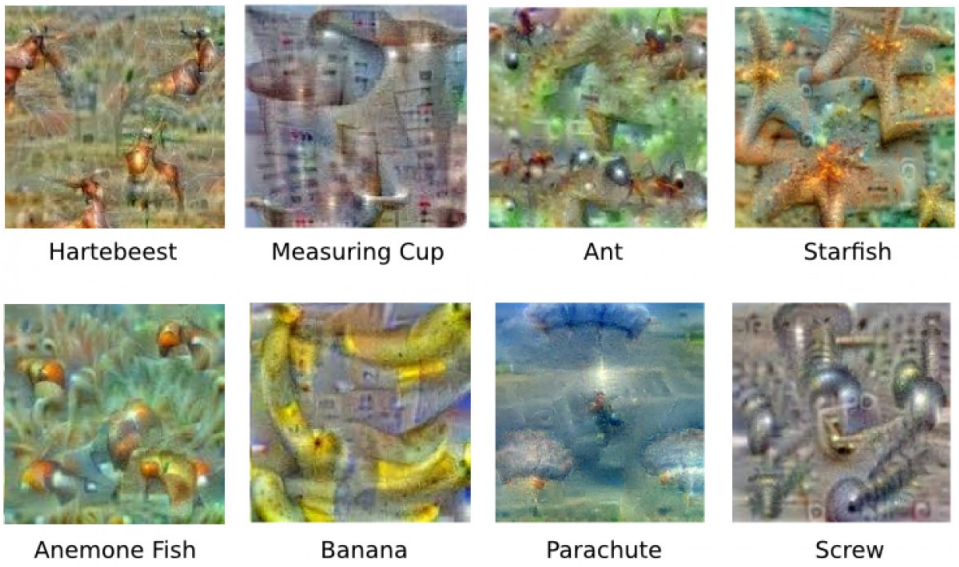
There is a lot of effort made to understand what are the processing that emerge for the training of a deep network, and how the parameters and the weights can be interpreted.
One way among others consists in visualizing the parts of an input image which maximize the response of some units in the network.
For instance, in the first layer of a convolutional neural network (used on images), the units "look at" the input image by small patches (usually of size 3×3, 5×5, or 11×11) and we observe that some of them are more responsive to edges in a given orientation or to flat areas.
When the signal moves forward in the network, as we will see later in the course, the units "see" a larger portion of the input image: this portion is called a receptive field for the unit. In the second layer, the units start capturing more complicated structures like grids, circles, corners, etc.
In the last layers, units respond to fragments of objects (wheels, arms, subparts, etc.), and finally to full objects.



(Zeiler and Fergus, 2014)

Notes

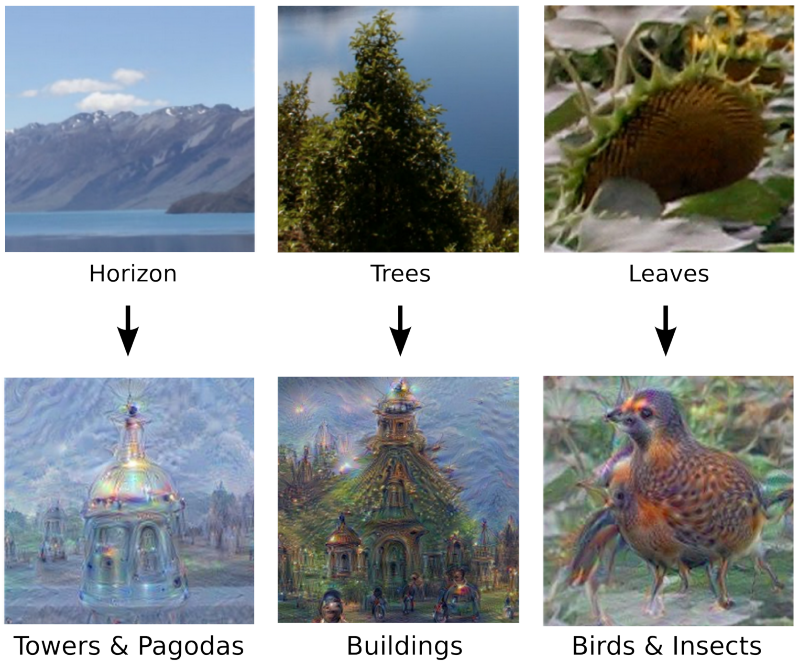
This observation is intellectually satisfying: although the network is performing a low level signal processing computation in each layer, the information in the full resulting process gets increasingly semantic.



(Google's Deep Dreams)

Notes

Another way to picture what the network is doing is to keep the model fixed, and optimize the signal.
For instance in the case of image classification, one can optimize the signal to find an input image that will maximize the output of the network for a given class. In some way this will generate images of "super bananas", "super Ants", etc.
Synthesized samples are reminiscent of the actual classes, which shows that the model goes beyond capturing what differentiate classes with each others but actually encode [to some extent] the morphological characteristics of the classes.



(Google's Deep Dreams)

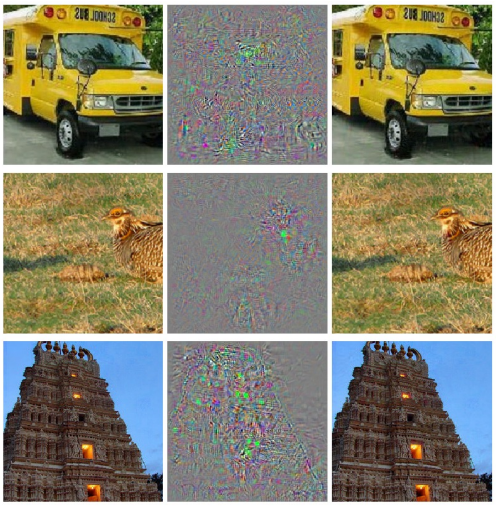
Notes

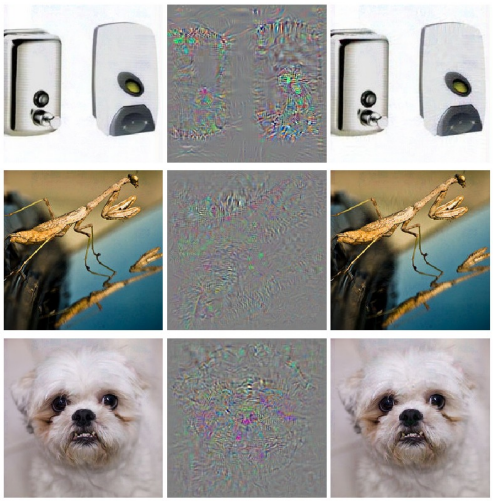
Another variant of optimizing the input when the model is trained is to start from a well-classified image and optimize it so that the network has a strong response on another class. For instance, the middle column shows how a tree is "changed" into a building.

(Thorne Brandt)

Notes

Starting from a plate of spaghetti and maximizing the response for the class "dog".



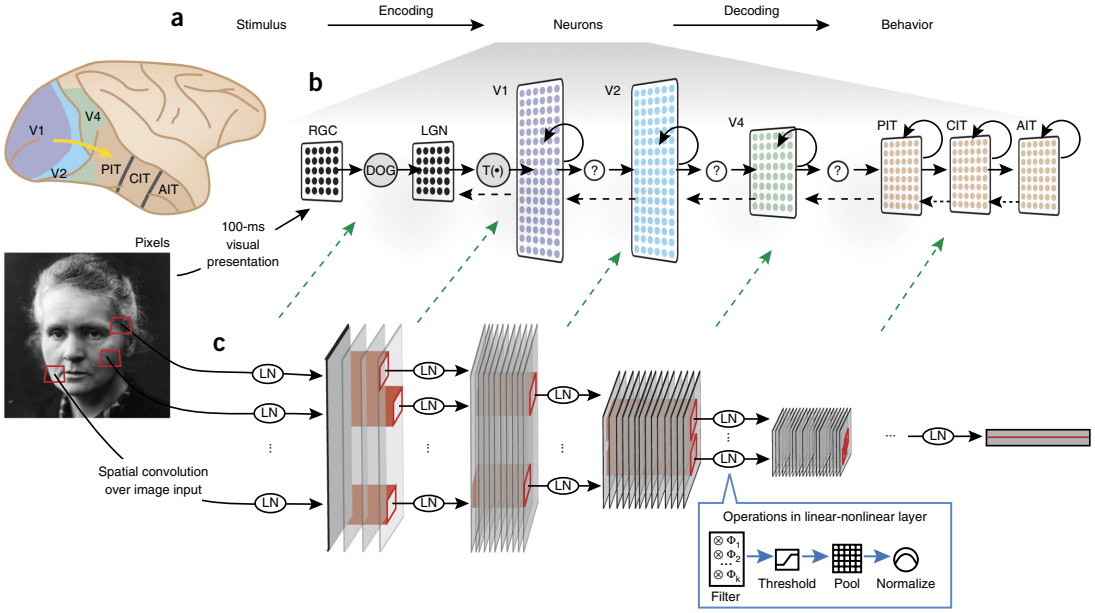


(Szegedy et al., 2014)

Notes

Another family of experiments aims at generating an "adversarial sample" by slightly modifying an input until it is no longer well classified.
In the case of images, the resulting sample looks unchanged to the human eye. The three images in each row show from left to right: the original image, the difference between the original and the obtained adversarial sample, and the adversarial sample.
These experiments are surprising and show that even accurate classification networks are very unstable.

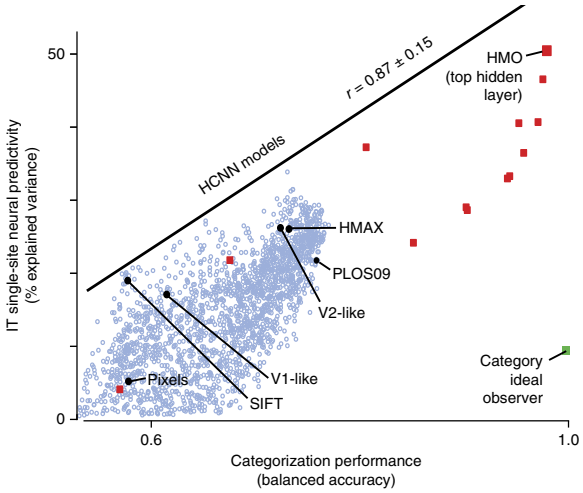
Relations with the biology

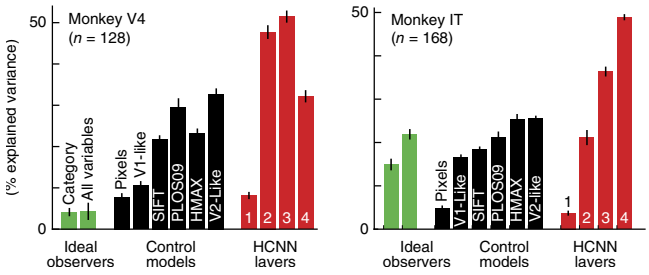


(Yamins and DiCarlo, 2016)

Notes

Although artificial and real neural networks share some coarse structural properties, most analogy should be avoided. In particular the back-propagation algorithm does not resemble any known biological mechanism.
However, some works have compared how the brain and neural networks process a stimulus.
As shown by Hubel and Wiesel, the visual cortex consists of a series of layers which go from low-level processing to high-level semantic understanding, which is in principles similar to an artificial convolutional network. The work of Yamins and DiCarlo (2016) analyzes the similarity between representations in the different layers of a brain and those in the layers of an artificial neural network. To do so, given the same input image (stimulus), they try to predict the activations of the layers in the former from the activations of the layers of the latter.

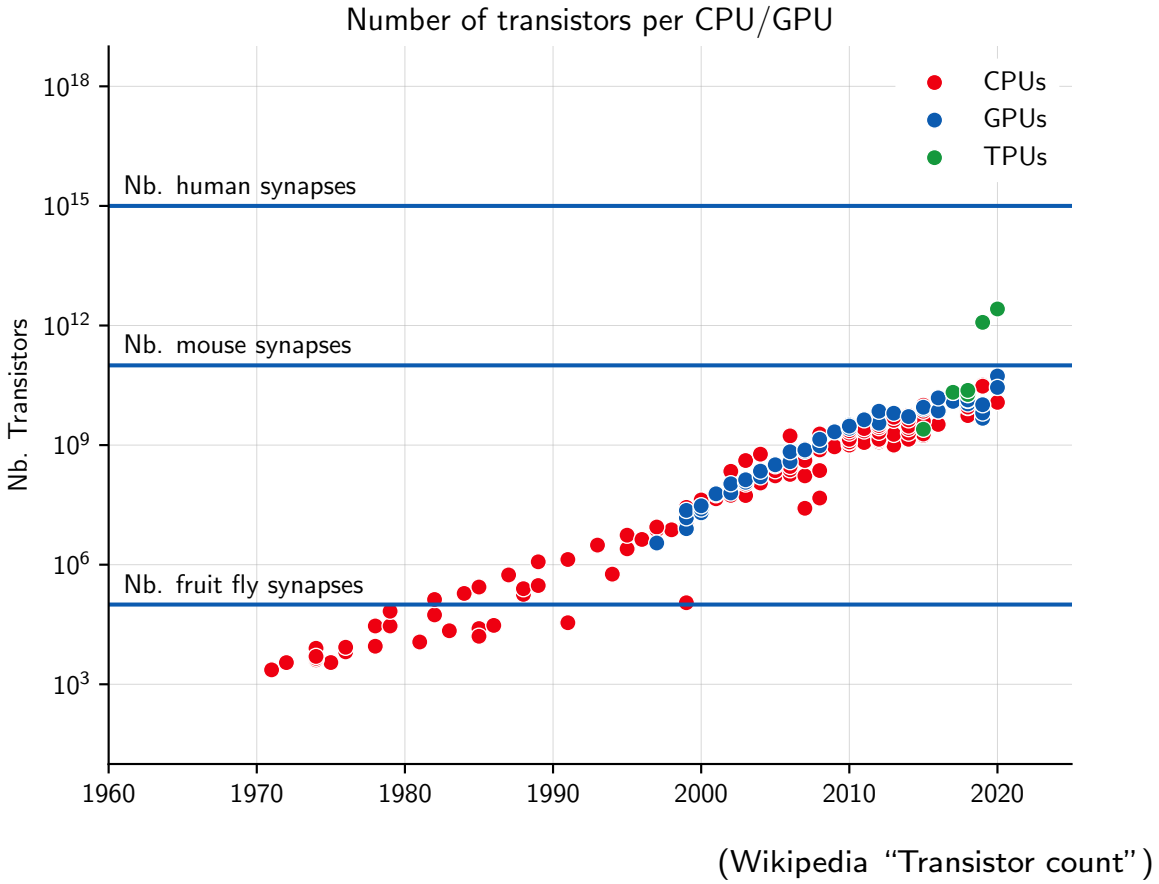




(Yamins and DiCarlo, 2016)

Notes

Left: Each dot is a model, blue are artificial neural net randomly initialized, red are optimized models: The more accurate a model for class prediction, the better for monkey IT neuron firing rate prediction.
Right: The plot shows that predicting the activations in area V4 (early layer in the visual cortex) of the monkey's brain is better achieved with by using the early layers of the artificial neural network. And predicting the activations in area IT (late layers) is better achieved with the activations of the later layers of the artificial network.



Notes

This graph is a quantitative comparison between the number of synapses and the number of transistors.
This comparison should be taken with a grain of salt, as a synapse is far more complex, although noisier and slower, than a transistor. Still, the "numbers collide", and it is hard to make claims about the limitation of artificial models based only on their limited scale.

References

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. **Intriguing properties of neural networks**. In International Conference on Learning Representations (ICLR), 2014.

D. L. K. Yamins and J. J. DiCarlo. **Using goal-driven deep learning models to understand sensory cortex.** Nature neuroscience, 19:356–65, Feb 2016.

M. D. Zeiler and R. Fergus. **Visualizing and understanding convolutional networks**. In European Conference on Computer Vision (ECCV), 2014.