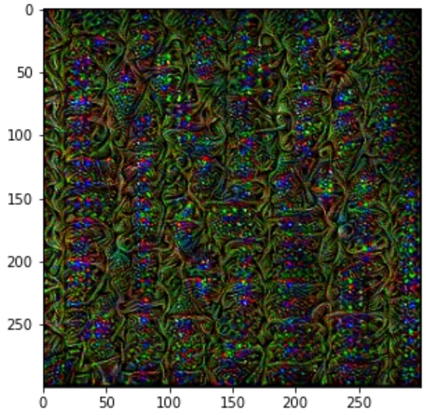
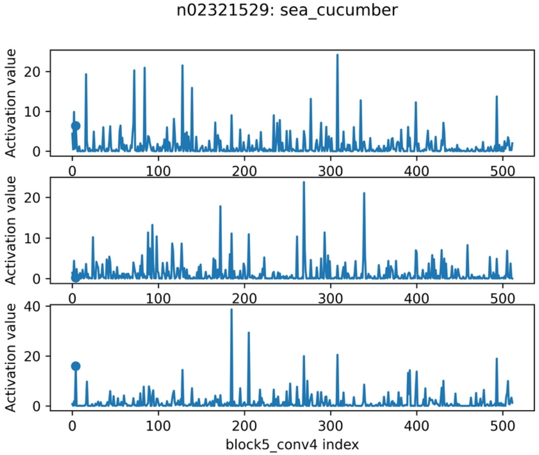
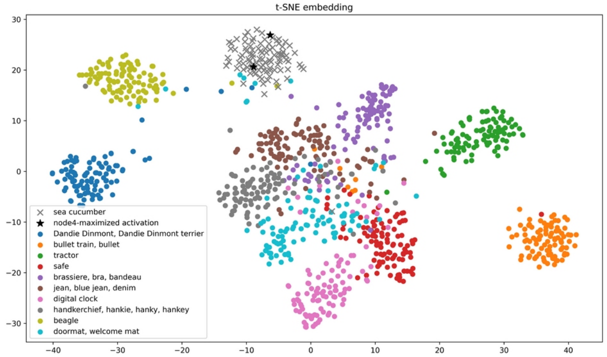
CNN Interpretability – Feature Extraction



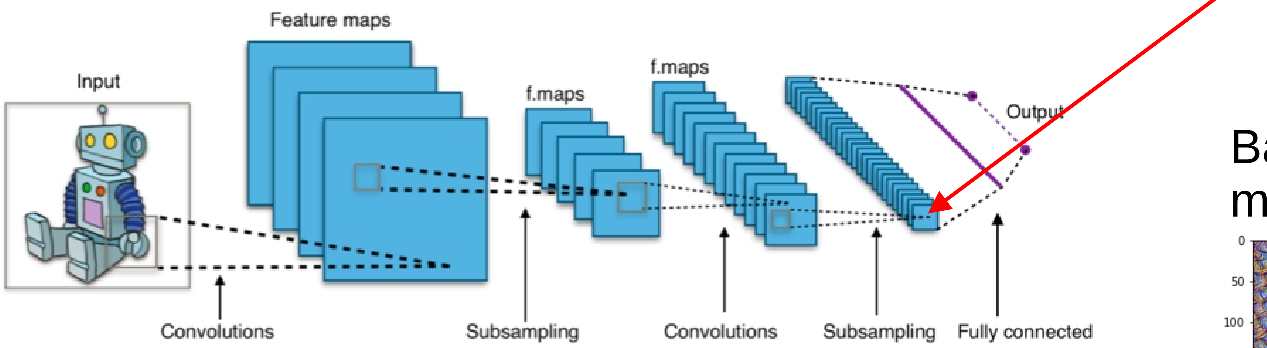




`stefanhex.com  or github.com/Stefan-Heimersheim`

Stefan Heimersheim
stefanhex.com
stefan.heimersheim@gmail.com

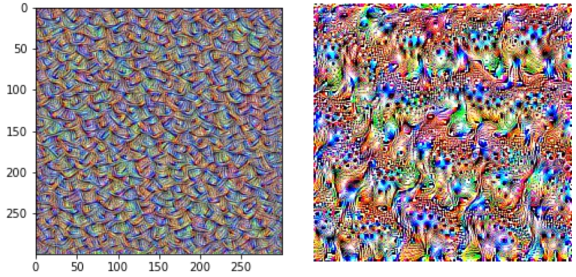
Feature Vizualization



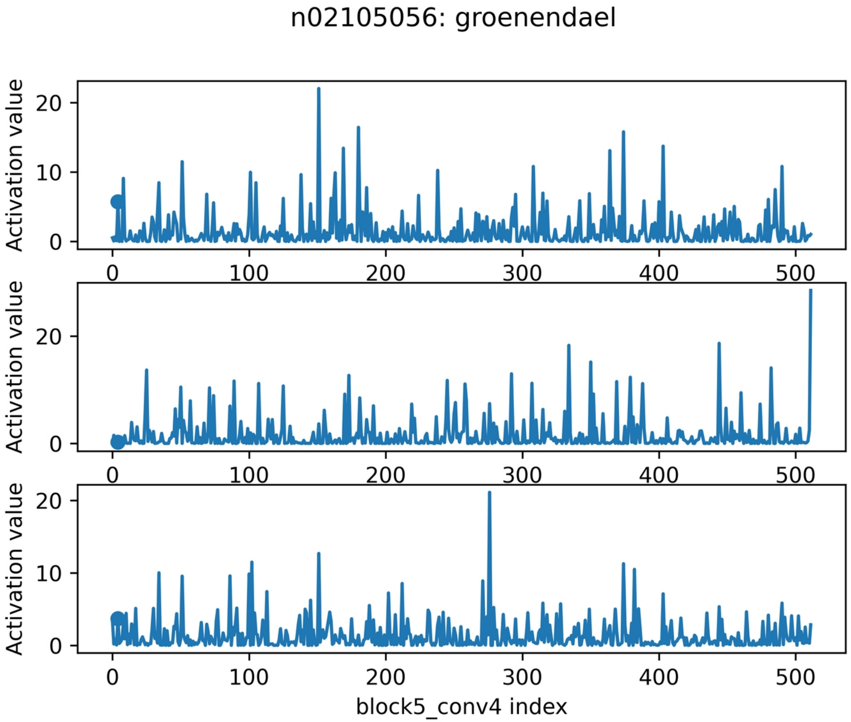
What is this neuron doing?

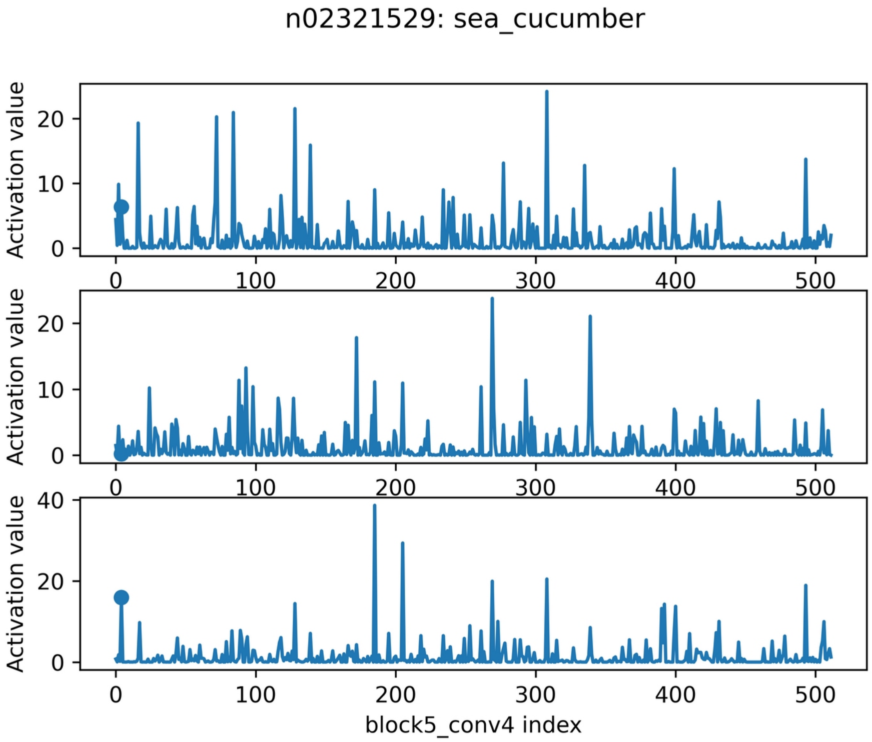
CC BY-SA Aphex34 (Wikipedia)

Basic idea: Find input that maximizes that neuron:

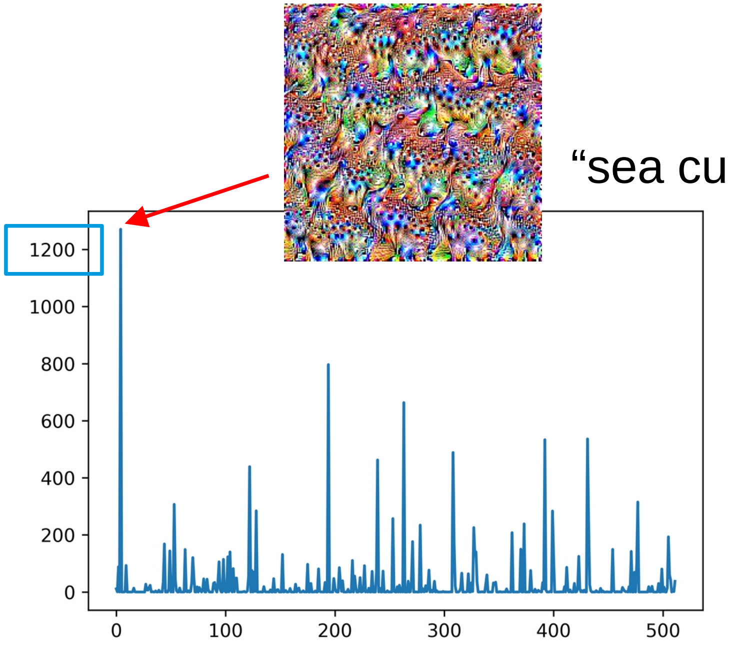


Activation patterns

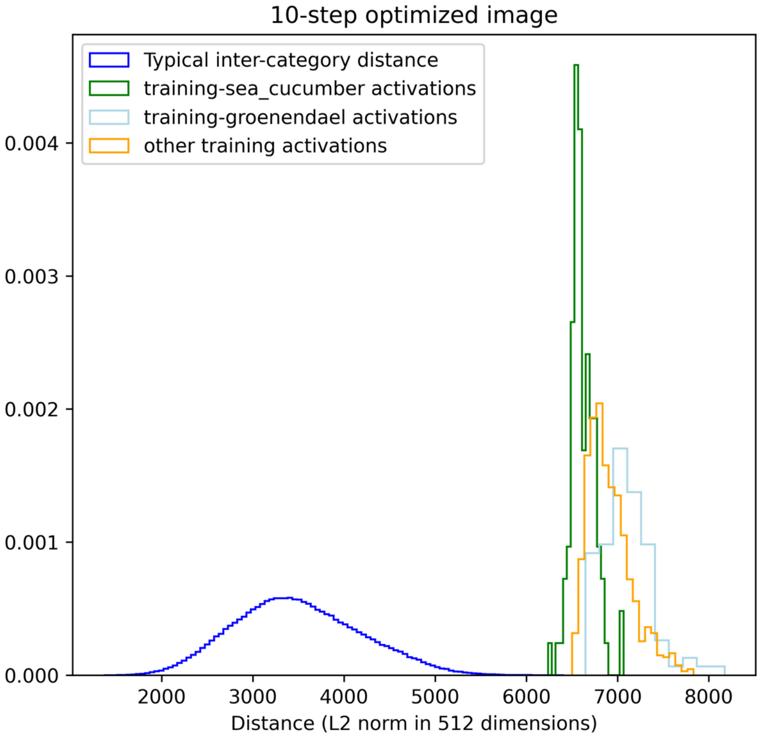




Activations out of distribution

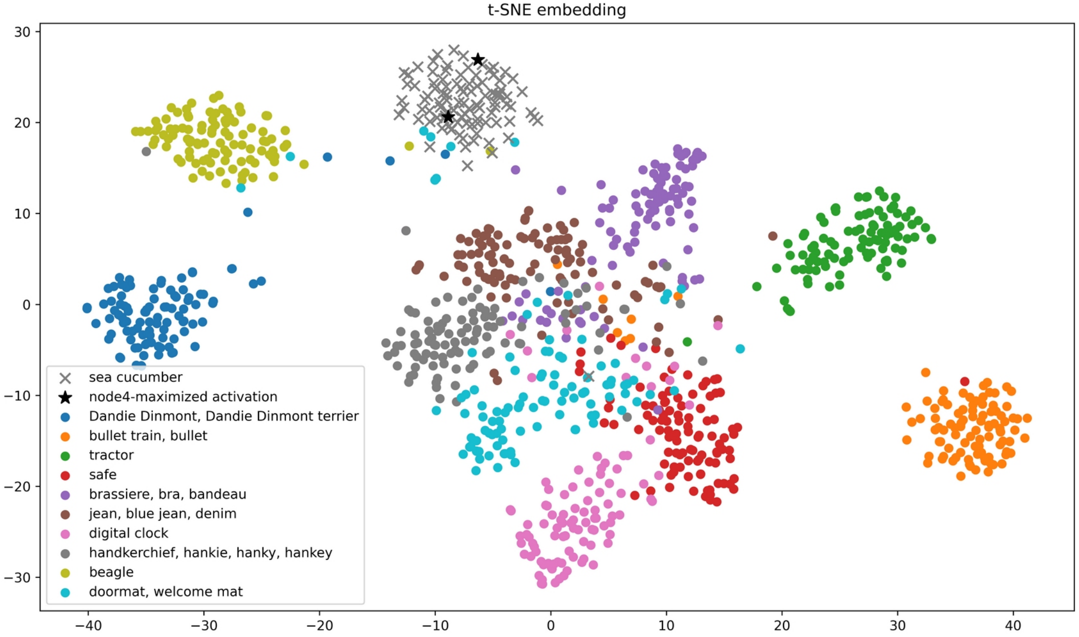


“sea cucumber”



Basis in activation space?

Notice: There seem to exist clusters in this 512d activation space; Unsupervised clustering:



Note: This might not necessarily apply to lower levels

Are features corresponding to a certain direction in 512d activation space?

There is no reason they need to align with axes

direction = [0,0,0,0,0,0,1,-1,1,-1,0.1, 0,0,0,...]

Can we use this to enumerate features?