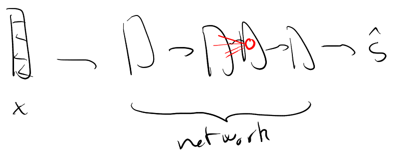
Chapter 2 : Interpretability

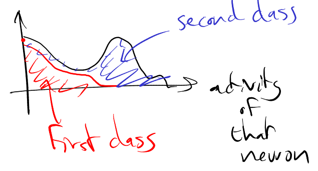
I Visualizat°/Analysis (of a trained network)



At the neuron level

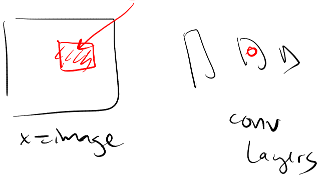
- pick one neuron : activities on the training set → stats
  e.g.: classific° task

histogram of activities



neuron: discriminative for these classes

- what it sees → receptive field

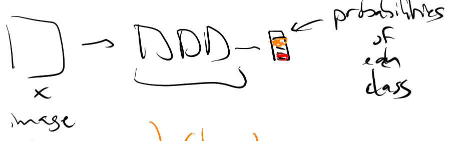


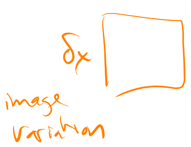
- what does it react to?
  → display input patterns that maximize the activity
    ↳ from the training set
  → compute the pattern that would " " "
    by gradient descent:

$$\frac{\partial x_t}{\partial t} = \eta \frac{\partial \text{activity}(x_t)}{\partial x}$$

(↑gradient)

↳ if you apply this to the full input, looking at output neurons:
e.g. classific° task,



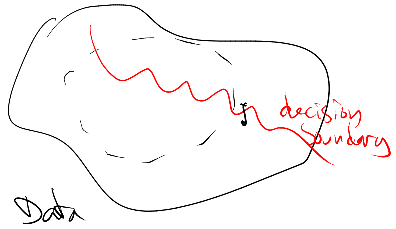


$$\frac{\partial p(\text{class } c)}{\partial \text{ image } x}$$

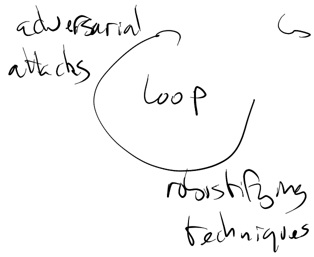
: sensitivity of the prediction (for class c) to the input x

$$x' = x + \underbrace{\eta}_{\text{not much}} \frac{\partial p(\text{other class})}{\partial x}$$

⇒ change completely the prediction
↳ adversarial examples [2014]
↳ adversarial attacks
↳ due to data dimension



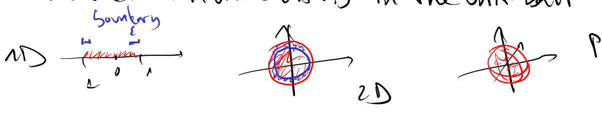
high-dim Data space
↓
- no dense sampling (require: $10^{\#\dim}$ samples)
- all points are on the boundary
  ↳ measure concentration
    → look at uniform distrib° in the unit ball



robustify
↳ train with x ∈ training set ↦ l
$x + \delta x$ associated adversarial attack ↦ l

~ smooth function: $\frac{\partial f(x)}{\partial x}(\delta x) = 0$

↳ $\left\| \frac{\partial f}{\partial x} \right\|^2$



proportion of points close to the boundary $\xrightarrow[\dim \to \infty]{} 1$

- does it have an impact? $\frac{\partial f}{\partial a}(x)$ → find which neurons influence the output the most

↑ activity of that neuron

⇒ reasoning at the neuron level: compre to human brain

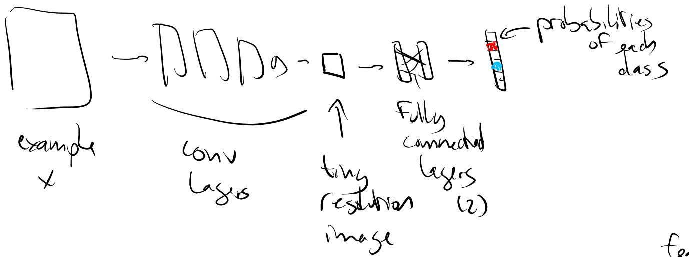
The case of CNN

- which parts of the image are responsible for the decision?

grad-CAM: Class Activation Maps – classifᵒ pb

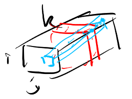
- pick a class c

-



↳ activities of that layer:

A^k_{ij} ← feature ← pixel location



⟼ $p(c) = y_c$

- importance of feature k for class c:

$$\alpha^c_k = \frac{1}{\#\text{pixels}} \sum_{ij} \frac{\partial y_c}{\partial A^k_{ij}}$$

$\in \mathbb{R}$

- importance of one pixel (i,j):

$$\sum_k \alpha^c_k A^k_{ij} \quad \in \mathbb{R}$$

- $\text{ReLU}\left(\sum_k \alpha^c_k A^k_{ij}\right)$

⇓

heat map for class c

Slurry ← Scale it up to the original resolution ← on the tiny resolution ← heat map for class c