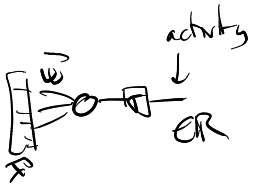
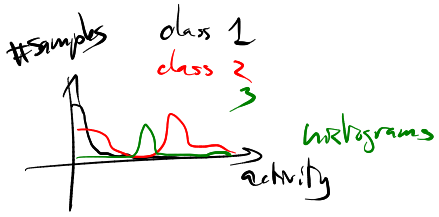
Chapter 2 : Interpretability

I Visualization / analysis of an already-trained network

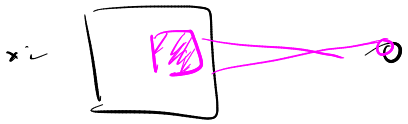


At the neuron level
→ statistics of its activities



ex : classific° task

- receptive field



- what it reacts to → show samples (Σ) : pick the samples that activate the most that neuron

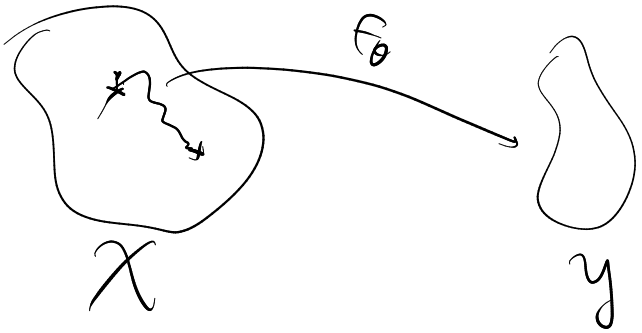
↳ compute the pattern that activates it the most

x_0 random initializ° of the input

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

ImageNet
1000 classes × 1000 samples (images)

$$\frac{\partial x}{\partial t} = \eta \frac{\partial p(\text{"cat"}|x)}{\partial x} \Rightarrow \text{adversarial examples}$$

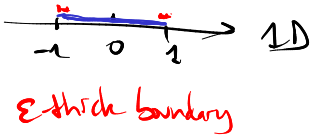


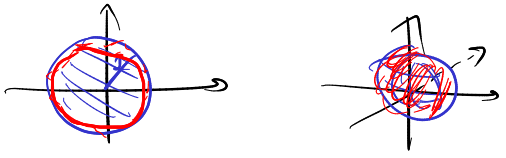
$$\sum_{\text{pixel}} \left(\frac{\partial p(\text{"cat"}|x)}{\partial x_{\text{pixel}}} \times \text{RGB} \right)_{\text{pixel}}$$

$\mathbb{R}^{\#\text{pixel}}$

High dimensions

unit ball

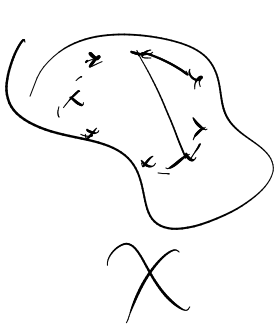




ratio $\frac{\text{boundary volume}}{\text{full volume}} \xrightarrow{d \to \infty} 1$

r^d $\quad dr^{d-1} \times \varepsilon$ $\quad r=1$

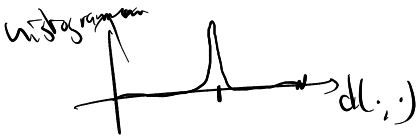
⇒ most points are on the boundary



~~dense~~ images $100 \times 100 \times 3$

$2^{30\,000}$

$d(x_1, x_2)$ $x_i \in U(\mathbb{R}^{30\,000})$



"curse of dimensionality"

measure concentration

Robustness to adversarial attacks

training dataset $\mathcal{D}$ —enrich with adversarial attacks→ $\mathcal{D}'$ →

$(x, y) \longrightarrow (x', y)$

1 way to compute adv. attacks

↓

1 robustify nets against that type of attack

DNA

a few individuals (20)

base

10^9 bases

disease or not

spurious correlations when #features/#samples is low

↳ regularize: $\|\theta\|_{L_1}$ $\hat{y} = \vec{\theta} \cdot \vec{x}$

coordinates $\theta_i = 0$

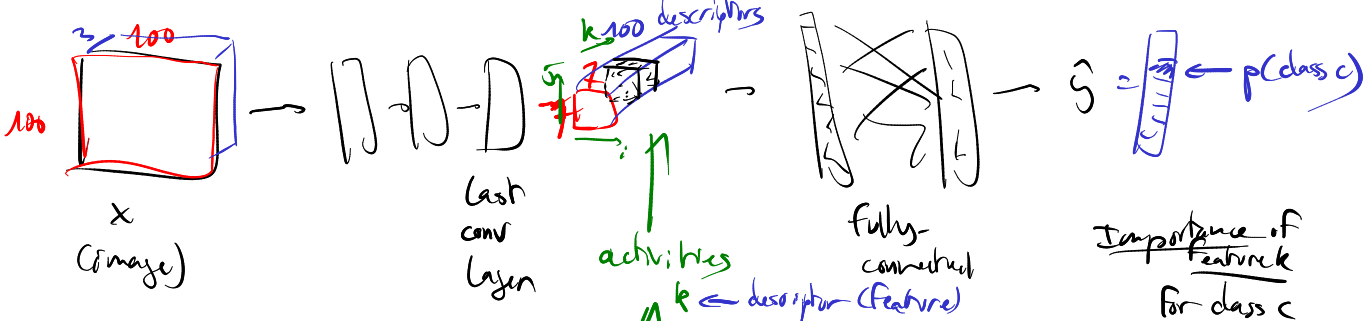
↳ priors

The case of CNN

which parts of the image are responsible for the decision?

grad-CAM Class Activation Maps

classif° task



Importance of feature k for class c

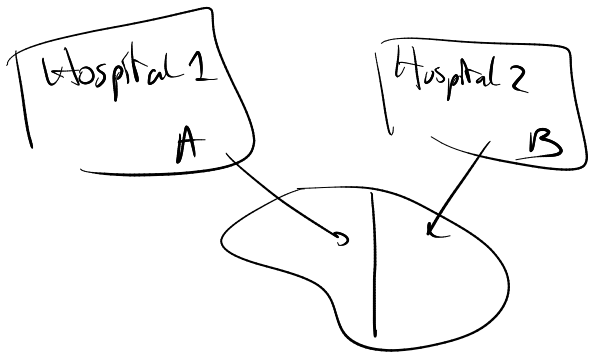
Importance of pixel (i,j): $\sum_{\text{features } k} \alpha^c_k \times A^k_{ij}$ $\in \mathbb{R}$ $\xrightarrow[\mathbb{R}_+]{\text{ReLU}(\cdot)}$ heat map $\in \mathbb{R}^{30}$

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

Datasets (biases)

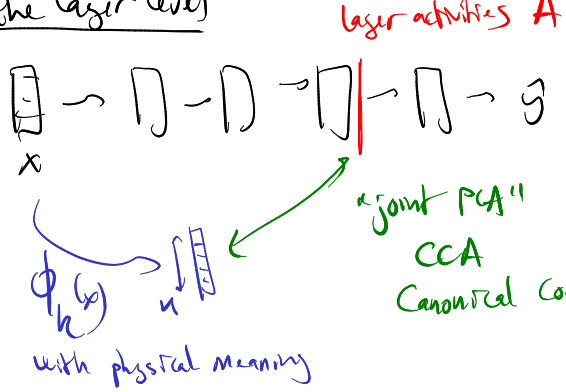
classif° → disease A / B

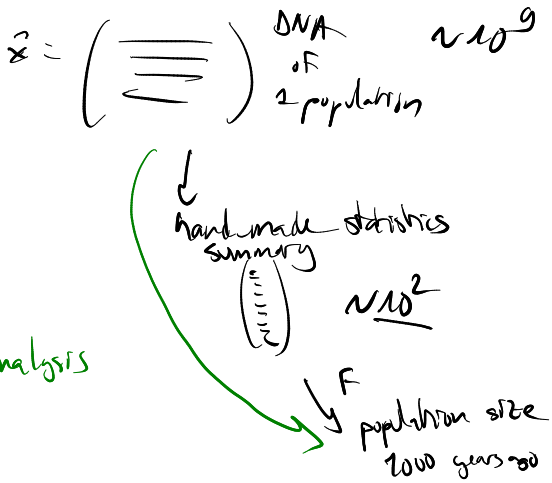
MRI scan → A / B ?

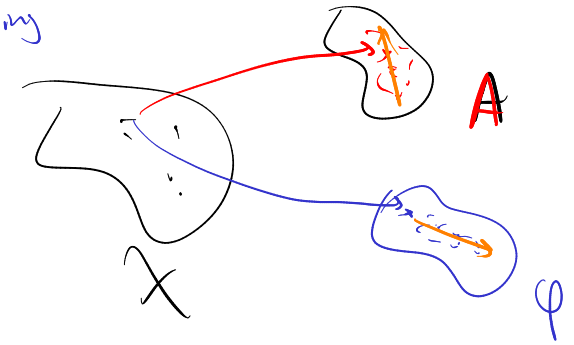


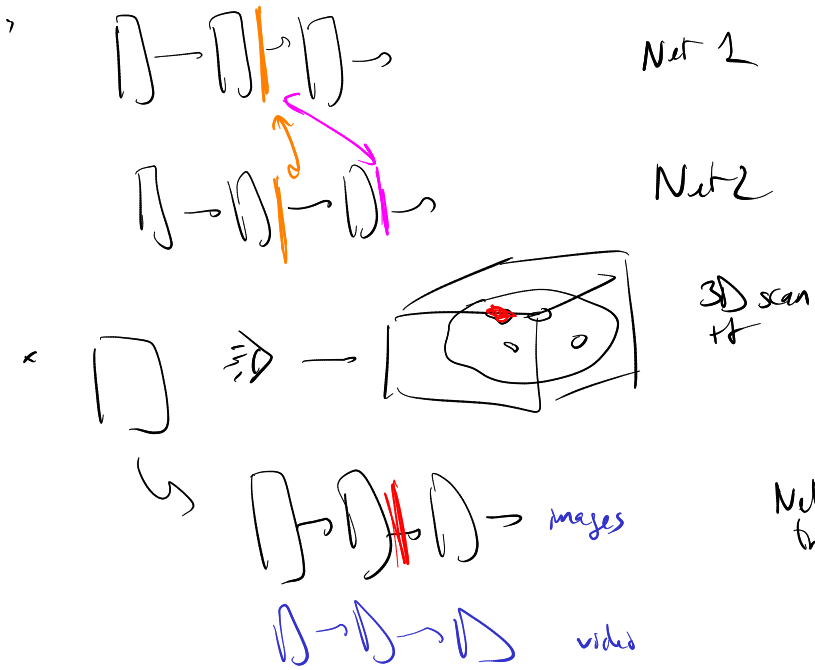
Visualize the filters that are learned

At the layer level



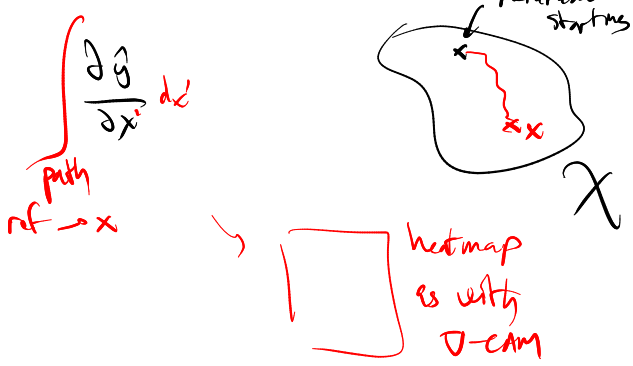






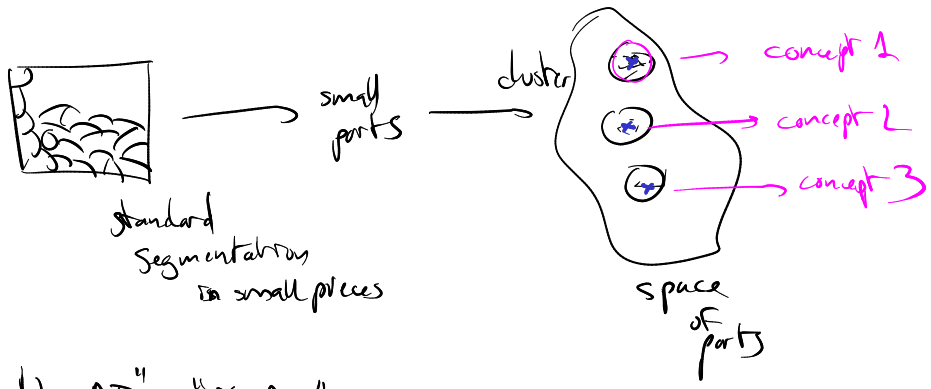
* Shapley values

* Path-D



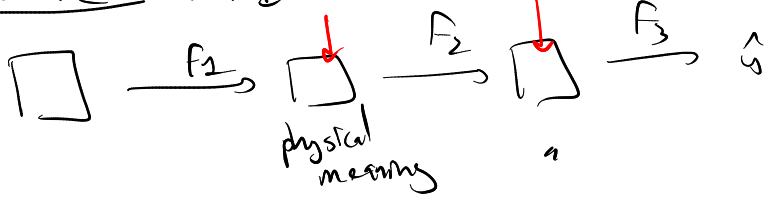
Which patterns in data are statistically meaningful?

CAV / ACE: Automatic Concept-based Explanations

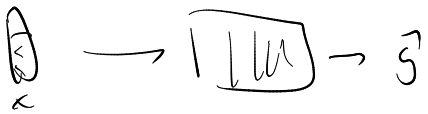


$\frac{\partial \hat{y}}{\partial \text{cluster mean}}$

"Explainable AI" = "X-AI"

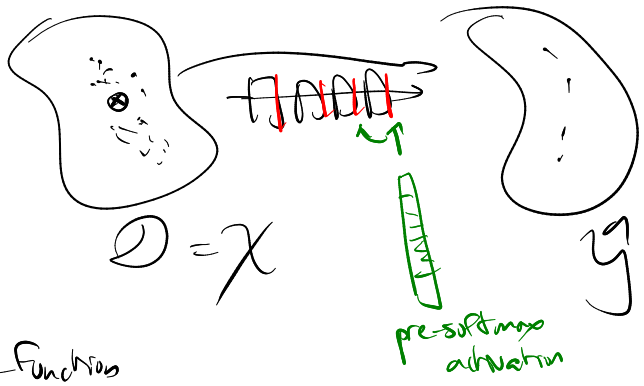


Which samples are similar?



Show other samples that the network considers to be similar

"similarity"?



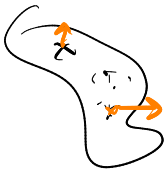
"perceptual loss"

$$d_N(x_1,x_2) = \sum_l \sum_n \left[a_n^l(x_1) - a_n^l(x_2) \right]^2$$

layers neurons

Influence Functions

Which samples in the training set are responsible for the decision taken for x?



* First idea: pick x' → remove from dataset & retrain from scratch
* Approximate this:

Focus on x': 1 training step $\theta \mathrel{+}= \eta \frac{\partial L(\hat{y}(x'))}{\partial \theta}$

impact on x: $\hat{y}_{\theta'}(x) = \hat{y}_\theta(x) + \underbrace{(\theta' - \theta)}_{\delta\theta} \times \frac{\partial \hat{y}_\theta(x)}{\partial \theta} + O(\delta\theta^2)$

prediction variation $\delta \hat{y}(x) = \frac{\partial L(\hat{y}(x'))}{\partial \theta} \times \frac{\partial y(x)}{\partial \theta} + \dots$

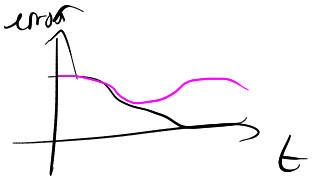
Neural Tangent Kernel (NTK) $= \frac{\partial \hat{y}(x)}{\partial \theta} \frac{\partial L}{\partial y(x)} \frac{\partial y(x)}{\partial \theta}$

$\nabla|_{x'}$ $\nabla|_{x}$

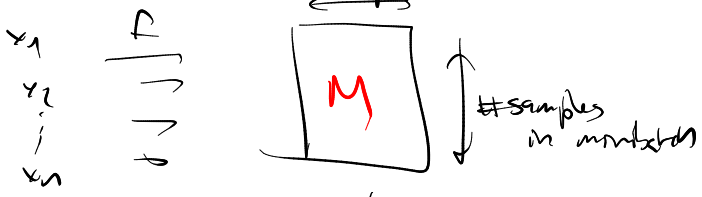
At the functional level

during training

how is f evolving?



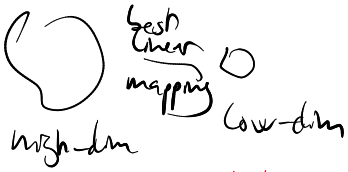
approximate f on a minibatch



$f_t \to M_t$

$f_{t'} \to M_{t'}$

↳ dimensionality reduction technique PCA



t-SNE
umap
PyMDE

} nice looking 2D projections