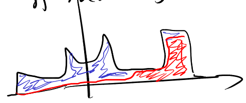
Interpretability

I Visualization

Given an already trained neural network

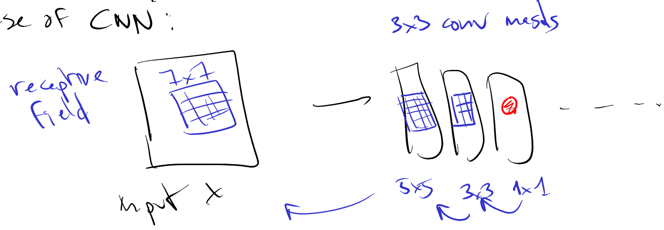
At the neuron level

- pick a neuron, study its activities on the training set
  - ↳ show its history (for recurrent networks)
  - ↳ show the distribution of activities
    (possibly as a function of the classes)



ex: binary classification task
1st class
2nd class

- what does it see?
  . case of CNN:



  . what the neuron actually sees: rebuild the input from the activities of that layer

- what does it react to?
  - display input patterns that maximize its activity
    ↳ from a dataset of samples
  - compute & display which artificial pattern(s) would activate the most that neuron
    ↳ by gradient ~~descent~~ ascent:

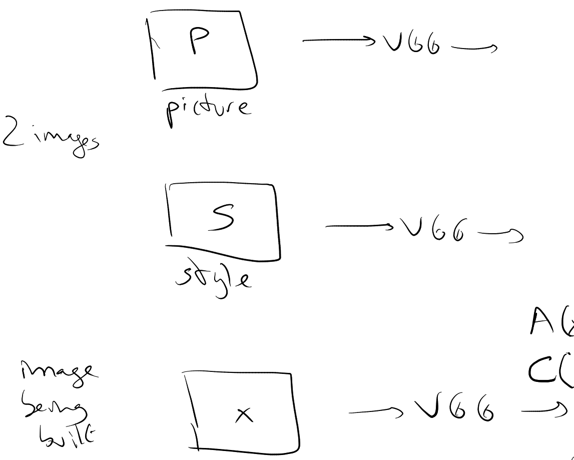
$$\frac{\partial x}{\partial t} = +\eta \frac{\partial a}{\partial x}$$

← activity of that neuron
← input image

↳ adversarial examples

↳ neural style transfer

2 images



layer

A^l_{ij} filter location

$A(x) \simeq A(P)$

$C(x) \simeq C(S)$ ← $\underbrace{\sum_{\text{locations } j} A^l_{ij} A^l_{i'j}}_{C^l_{ii'}}$

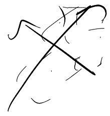
correlation between filters i & i'

Loss: $\inf_x \left(A(x) - A(P)\right)^2 + \left(C(x) - C(S)\right)^2$

→ Does the neuron have impact?
↳ $\frac{dF}{da}$ ← output of the network; ← activity of that neuron
} sensitivity of the output w.r.t. that neuron
→ sample-dependent

At the layer level

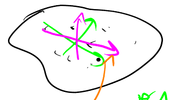
. CCA (Canonical Correspondence Analysis)

PCA:  extract main directions

- activities in one layer
  input $x \to z \in \mathbb{R}^d$
- [set of hand-crafted features
  $x \to t \in \mathbb{R}^{d'}$

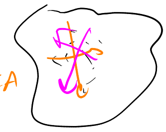
step towards explainability

or
activities from a layer from another network
$x \to t \in \mathbb{R}^{d''}$



Linear combination of neuron activities

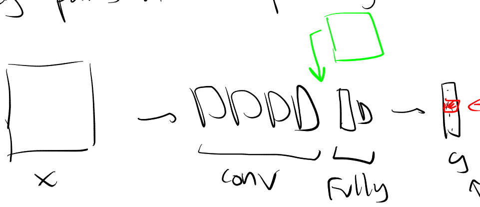
Joint PCA = CCA



The case of CNN

- visualize filters
- display parts of the input image that were the most important for the network's decision

classification task



$$\frac{\partial y_c}{\partial x}$$

$\to$ adversarial examples

$\hookrightarrow$ get noisy patterns

Grad-CAM

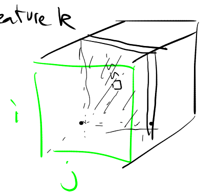
[Selvaraju & al, ICCV 2017
IJCV 2019]

Activities in last conv. layer : A^k_{ij}

Feature number

(i,j): pixel location

$\to$ importance of feature k for class c :

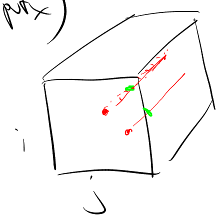


$$\alpha^c_k = \frac{1}{\#\text{pixels}} \sum_{\substack{ij \\ \text{pixels}}} \frac{\partial y^c}{\partial A^k_{ij}}$$

$\in \mathbb{R}$ (depends on the input x)

average (over locations) of the impact of feature k on the output (class c)

$\to$ heatmap:



For each "pixel" (i,j):

$$\mathrm{ReLU}\Big(\sum_{\text{features } k} \alpha^c_k A^k_{ij}\Big) \in \mathbb{R}$$

(for each location)

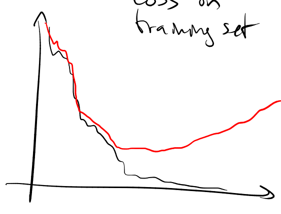
where positive

At the functional level

- information theory
  bottleneck

Training visualization

- display { accuracy / loss } as a function of time
  { on the training set
  " " validation " }

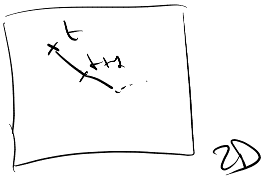


too noisy $\Rightarrow$ too high learning rate

- gap training/validation: overfit

- project the network on a 2D space

$\downarrow$
t-SNE
$\downarrow$
same as PCA but nonlinear



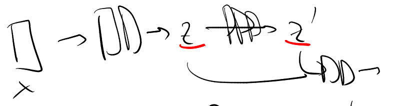
fixed samples

$$f(t) \to \left(f_{(t)}(x_i)\right)$$

$\to$ high-dim vector

$\Rightarrow$ dynamics of the training

· Better visualize → by sub-task design
"explainable AI"

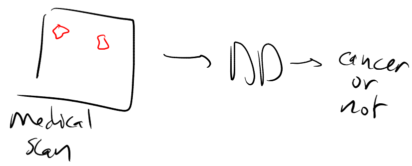


series of simpler tasks
with physical meaning

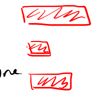
II Interpretability: societal impact

Interpretability is important

- eg: medical diagnosis
  ↳ explanation of the final score
    ↳ why trust this prediction?



- Isabelle Guyon's skin disease classif°
  → hand-crafted features
    ↳ labelled: - color
      - shape
      - texture
      ...

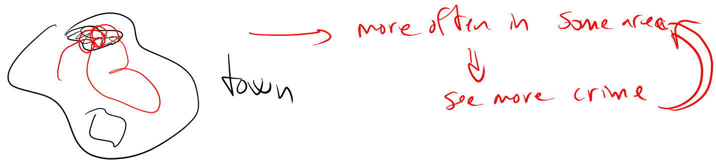


  → decision with ⟶ information about relevant feature

- Societal Impact
  ↳ "Weapons of Maths Destruction", by Cathy O'Neil
  ↳ companies using black-box software (provided by other companies)
  for important matters
    ↳ hire
    ↳ fire
    ↳ loan

  many with same software

- "COMPAS" (2016): predict recidivism → much higher false positive rate for black people (than white)
- self-reinforcing police patrol



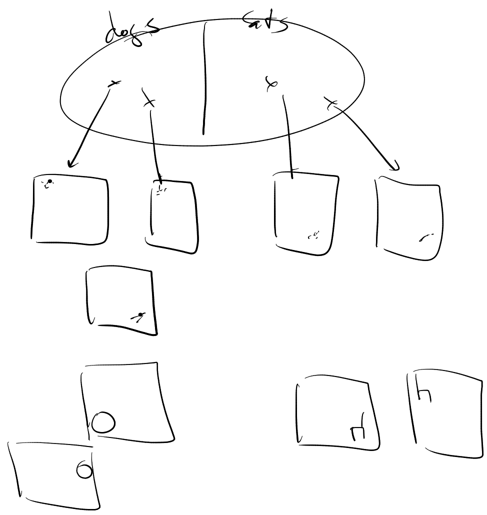
→ be responsible & careful
  ↳ ethics in AI → Montreal Declaration for Responsible AI
→ cf web page for an eg of explainable AI

III Issues related to datasets

Dataset poisoning

- Forge a dataset;
  in each image, add invisible noise
  (always the same noise)
→ any algo trained on that dataset
  will exploit this noise/signature
→ present an image with the wrong signature
  → wrong classification

↳ variation: not noise, but another object

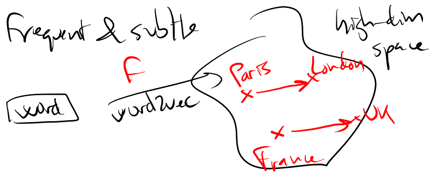


$\boxed{h}$ $\boxed{o}$

Fairness

Intro: biases: more frequent & subtle

word2vec :



↓
articles from GoogleNews

$F(\text{Paris}) - F(\text{London}) = F(\text{France}) - F(\text{UK})$

$F(\text{man}) - F(\text{woman}) = F(\text{king}) - F(\text{queen})$

$= F(\text{computer programmer}) - F(\text{homemaker})$

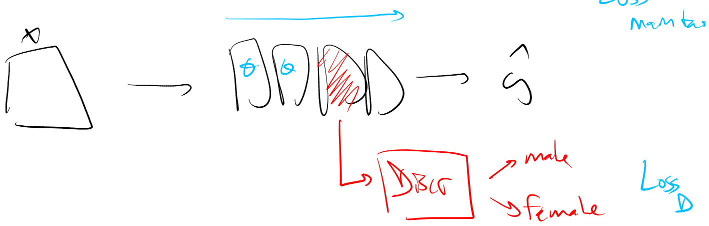
$= F(\text{surgeon}) - F(\text{nurse})$

→ retrain word2vec without these specific biases

Definitions

1) cf course webpage for this part.

Adversarial approach for fairness



$\text{Loss}_{\text{main task}} - \underbrace{\text{Loss}_D}$

get rid of sensitive information