Interpretability

I. Visualization

Given an already trained neural network: \[ x \rightarrow (\text{hidden layers}) \rightarrow y \]

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)

- What does it see?
  - Case of CNN:
    - Feature field \[ \text{input} \times \rightarrow \text{activations} \]

- What does it react to?
  - Display input patterns that maximize its activity
    - From a set of examples
  - An input display which artificial pattern(s) would activate the most that neuron
    - By gradient ascent:
      \[ \frac{dx}{dk} = \nabla_x \frac{da}{dk} \]
      - Input image

- Select corpora examples

6) neural style transfer...

\[ \text{picture} \rightarrow V66 \rightarrow \]

2 images

\[ \text{style} \rightarrow V66 \rightarrow \]

Image being built

Does the neuron have impact?
- Sensitivity of the output of the network
  - Sensitivity of the output of the model

At the layer level:
- CCA (Canonical Correspondence Analysis)
The case of CNN

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

\[ \frac{\partial \log c}{\partial x} \quad \text{aberrated examples} \]

Grad-CAM

[Selvaraju et al., IJCV 2017]

Activities in last conv. layer: \( A \)

\( - \) importance of feature \( k \) for class \( c \):

\[ I_k^c = \frac{1}{\text{pixels}} \sum \frac{\partial \log c}{\partial A_{ij}^k} \]

(averages over location)

\( \text{ER} \)

(derivative on the input)

\( \rightarrow \) heatmap.

At the Functional level

- information theory bottleneck

Training visualization

- display accuracy as a function of time (loss)

\( \rightarrow \) project the network on a 2D space

\[ f(t) - F(w(t)) \]

\( \rightarrow \) dynamics of the training
II Interpretability: societal impact

Interpretability is important

- ex: medical diagnosis
  
  - explanation of the F1-score
  - why trust this prediction?
  
  - Farzad et al's skin disease classifier
    
    - hand-crafted features
      - color
      - texture
    
    - decision tree

- strategic impact
  
  - "Weapons of Math Destruction", by Cathy O’Neil
  
  - companies using black-box software (provided by other companies)
    
    - important meetings
      - to hire
      - to fire
      - to learn

  - COMPAS (2016): predict recidivism
    
    - much higher false positive rate for black people (than white)

  - self-reinforcing police patrol
    
    - more often in some areas
    - see more crimes

- be responsible & careful
  
  - ethics in AI → Pragmatic Deduction for Responsible AI
  
  - cf. web page for an use of explainable AI

III Tissues related to datasets

Dataset poisons

- forge a dataset:
  
  - in each image, add invisible noise (always the same noise)
  
  - any deep model on that dataset will exploit this noise signature

- present an image with the wrong signature → wrong classification

- rewritten: not mine, but another object
**Fairness**

**Intro:** biases: more frequent & subtle

Word2vec: without class specific biases

F(bush) - F(lincoln) = F(hermie) - F(uk)
F(men) - F(women) = F(leg) - F(green)
F(computer programmer) - F(humanist)
F(surgeon) - F(nurse)

**Definitions**

1) CF course webpage for this part.

**Adversarial approach for Fairness**

\[ \text{Loss}_{\text{main}} - \text{Loss}_{\text{D}} \]

Get rid of sensitive information