INTERPRETABLE DEEP LEARNING

Cengiz Öztireli
Coarse styles
($4^2 \rightarrow 8^2$)

Middle styles
($16^2 \rightarrow 32^2$)

Fine styles
($64^2 \rightarrow 1024^2$)
WHAT IS INTERPRETABILITY?
WHAT IS INTERPRETABILITY?

1 longer / shorter 10
WHAT IS INTERPRETABILITY?
WHAT IS INTERPRETABILTY?
WHY INTERPRETABLE ML/ AI

"Normal"

```
Classifier
```

"Normal"
WHY INTERPRETABLE ML/AI

"Abnormal"

Classifier

Based on these?
WHY INTERPRETABLE ML/ AI

Transparency

“Right to explanation”

The data subject should have the right not to be subject to a decision [...] which is based solely on automated processing [...] such as automatic refusal of an online credit application without any human intervention. [...]

In any case, such processing should be subject to suitable safeguards, which should include the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached

(EU General Data Protection Regulation, Recital 71)
WHY INTERPRETABLE ML/ AI

Transparency

Understanding
WHY INTERPRETABLE ML/ AI

Transparency

Understanding

Efficiency
TYPES OF INTERPRETABILITY

Interpretable by construction

Not directly interpretable
TYPES OF INTERPRETABILITY

How do you interpret millions of parameters?
TYPES OF INTERPRETABILITY

Model

Input/Output
TYPES OF INTERPRETABILITY

Model

What structures are learned?

Input/Output

- Low-level features
- Mid-level features
- High-level features
TYPES OF INTERPRETABILITY

Model

What structures are learned?

Low-level features

Mid-level features

High-level features

Input/Output

What parts are important for a given input/output?

Low-level features

Mid-level features

High-level features
MODEL INTERPRETABILITY
GLOBAL UNDERSTANDING
INTERPRETING DEEP MODELS

Visualizing weights

What weights/filters do the networks learn?
INTERPRETING DEEP MODELS

Activation patterns

Which patterns activate certain neurons most?

bell pepper  lemon  husky
washing machine  computer keyboard  kit fox

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps
Surrogate models

Which is an interpretable model that generates similar results?
INTERPRETING DEEP MODELS

Influential data

Which data in the training set has influenced the decision most?
INPUT/OUTPUT INTERPRETABILITY LOCAL UNDERSTANDING
ATTRIBUTION METHODS
ATTRIBUTION METHODS

How to measure how much each pixel is important for a given input/output pair?
ATTRIBUTION METHODS
ATTRIBUTION METHODS
ATtribution Methods

Desired properties

- Theoretically well-founded
- Implementation invariant
- Efficient to compute
ATTRIBUTION METHODS

Saliency Maps
Simonyan et al. 2015

Integrated Gradients
Sundararajan et al. 2017

DeepLIFT
Shrikumar et al. 2017

Deconvolutional Networks
Zeiler et al. 2014

Gradient * Input
Shrikumar et al. 2016

Layer-wise Relevance Propagation (LRP)
Bach et al. 2015

Guided Backpropagation
Springenberg et al. 2014

Grad-CAM
Selvaraju et al. 2016

Simple occlusion
Zeiler et al. 2014

Meaningful Perturbation
Fong et al. 2017

Prediction Difference Analysis
Zintgraf et al. 2017

Grad-CAM...
ATRIBUTION METHODS

Saliency Maps
Simonyan et al. 2015

Integrated Gradients
Sundararajan et al. 2017

DeepLIFT
Shrikumar et al. 2017

Deconvolutional Networks
Zeiler et al. 2014

Layer-wise Relevance Propagation (LRP)
Bach et al. 2015

Guided Backpropagation
Springenberg et al. 2014

Grad-CAM
Selvaraju et al. 2016

Gradient * Input
Shrikumar et al. 2016

Simple occlusion
Zeiler et al. 2014

Meaningful Perturbation
Fong et al. 2017

Prediction Difference Analysis
Zintgraf et al. 2017

Unified framework

Shapley values
ATTRIBUTION: SHAPLEY VALUES

\[ g(\{?, ?, ?\}) = 100 \]

\[ g(\{?, ?, ?\}) - g(\{?, ?\}) \]

\[ g(\{?, ?\}) - g(\{\} \} \]

\[ g(\{?, \} \} ) - g(\{\} \} \]

\[ g(\{\} \} ) - g(\{\} \} ) \]
SHAPLEY VALUES

- Theoretically well-founded
- Implementation invariant
- Efficient to compute

\[ O(2^N) \]
ATTRIBUTION: DEEP SHAPLEY VALUES
ATTRIBUTION: DEEP SHAPLEY VALUES

DNA sequence classification

Parkinson’s disease factors

Image classification
APPLICATIONS OF INTERPRETABILITY
LIFE DECISIONS
ATTACKING DEEP SYSTEMS

Optimize for a t-shirt that makes you undetectable.
DEEP SYSTEMS GONE WRONG

Jace @5plat @SomeonesAnIdiot

Replying to @peta

If you love killing animals reply to this tweet! #donate

11:26 AM · 2/24/19 · Twitter for iPhone

PETA @peta · 3m

Replying to @5plat

Thanks for your support! Complete your donation now: gdw.io/e5bff1
DEEP SYSTEMS GONE WRONG
AVOID BIAS
UNDERSTAND WEAKNESSES
FAIL GRACEFULLY
SCIENCE NOT MAGIC
ENCOURAGE RIGOR
KEEP SANITY

DEEP ART