Tutorial on Methods for Interpreting and Understanding Deep Neural Networks

Wojciech Samek (Fraunhofer HHI)
Grégoire Montavon (TU Berlin)
Klaus-Robert Müller (TU Berlin)

1:30 - 2:00  Part 1: Introduction
2:00 - 3:00  Part 2a: Making Deep Neural Networks Transparent
3:00 - 3:30  Break
3:30 - 4:00  Part 2b: Making Deep Neural Networks Transparent
4:00 - 5:00  Part 3: Applications & Discussion
Before we start

We thank our collaborators!

Alexander Binder (SUTD)

Sebastian Lapuschkin (Fraunhofer HHI)

Lecture notes will be online soon at:
http://www.heatmapping.org

Please ask questions at any time!
Tutorial on Methods for Interpreting and Understanding Deep Neural Networks

W. Samek, G. Montavon, K.-R. Müller

Part 1: Introduction
Recent ML Systems achieve superhuman Performance

- **AlphaGo** beats Go human champ
- **Deep Net** outperforms humans in image classification
- **DeepStack** beats professional poker players
- **Autonomous search-and-rescue drones** outperform humans
- **Computer** out-plays humans in "doom"
- **IBM's Watson** destroys humans in jeopardy
- **Deep Net** beats human at recognizing traffic signs
From Data to Information

Huge volumes of data

Computing power

Deep Nets / Kernel Machines / …

Interpretable Information

Solve task

Information (implicit)
From Data to Information

Interpretability

AlexNet (16.4%)  Clarifai (11.1%)  VGG (7.3%)  GoogleNet (6.7%)  ResNet (3.57%)

Performance

Data  Information  Interpretable for human

Crucial in many applications (industry, sciences …)
Interpretable vs. Powerful Models?

**Linear model**
- Poor fit, but easily interpretable
- "global explanation"

**Non-linear model**
- Can be very complex
- "individual explanation"
Interpretable vs. Powerful Models?

- **Linear model**
  - Poor fit, but easily interpretable
  - 60 million parameters
  - 650,000 neurons

- **Non-linear model**
  - Can be very complex
  - 650,000 neurons
  - We have techniques to interpret and explain such complex models!

---

**ICASSP 2017 Tutorial — W. Samek, G. Montavon & K.-R. Müller**
Interpretable vs. Powerful Models?

- train best model → interpret it

vs.

- train interpretable model

suboptimal or biased due to assumptions (linearity, sparsity …)
Dimensions of Interpretability

Different dimensions of “interpretability”

- **prediction**
  “Explain why a certain pattern $x$ has been classified in a certain way $f(x)$.”

- **model**
  “What would a pattern belonging to a certain category typically look like according to the model.”

- **data**
  “Which dimensions of the data are most relevant for the task.”
Why Interpretability?

1) Verify that classifier works as expected

Wrong decisions can be costly and dangerous

"Autonomous car crashes, because it wrongly recognizes ...”

"AI medical diagnosis system misclassifies patient’s disease ...”
Why Interpretability?

2) Improve classifier

Generalization error

Generalization error + human experience

Standard ML

Interpretable ML

data

ML model

predictions

data

ML model

interpretable

verified predictions

model/data improvement

human inspection
Why Interpretability?

3) Learn from the learning machine

“It's not a human move. I've never seen a human play this move.” (Fan Hui)

Old promise:

“Learn about the human brain.”
Why Interpretability?

4) Interpretability in the sciences

Stock market analysis:
“Model predicts share value with __% accuracy.”

In medical diagnosis:
“Model predicts that X will survive with probability __”

Great !!!

What to do with this information?
Why Interpretability?

4) Interpretability in the sciences

Learn about the physical / biological / chemical mechanisms. (e.g. find genes linked to cancer, identify binding sites …)
Why Interpretability?

5) Compliance to legislation

European Union’s new General Data Protection Regulation → “right to explanation”

Retain human decision in order to assign responsibility.

“With interpretability we can ensure that ML models work in compliance to proposed legislation.”
Why Interpretability?

Interpretability as a gateway between ML and society

• Make complex models acceptable for certain applications.
• Retain human decision in order to assign responsibility.
• “Right to explanation”

Interpretability as powerful engineering tool

• Optimize models / architectures
• Detect flaws / biases in the data
• Gain new insights about the problem
• Make sure that ML models behave “correctly”
Techniques of Interpretation

DNN transparency

interpreting models

- activation maximization
  - Berkes 2006
  - Erhan 2010
  - Simonyan 2013
  - Nguyen 2015/16

- data generation
  - Hinton 2006
  - Goodfellow 2014
  - v. den Oord 2016
  - Nguyen 2016

explaining decisions

- sensitivity analysis
  - Khan 2001
  - Gevrey 2003
  - Baehrens 2010
  - Simonyan 2013

- decomposition
  - Poulin 2006
  - Landecker 2013
  - Bach 2015
  - Montavon 2017

focus on model  focus on data
Techniques of Interpretation

Interpreting models (ensemble)
- find prototypical example of a category
- find pattern maximizing activity of a neuron

Explaining decisions (individual)
- “why” does the model arrive at this particular prediction
- verify that model behaves as expected

better understand internal representation

crucial for many practical applications
Techniques of Interpretation

In medical context

• Population view (ensemble)
  • Which symptoms are most common for the disease
  • Which drugs are most helpful for patients

• Patient’s view (individual)
  • Which particular symptoms does the patient have
  • Which drugs does he need to take in order to recover

Both aspects can be important depending on who you are (FDA, doctor, patient).
Techniques of Interpretation

Interpreting models

- find prototypical example of a category
- find pattern maximizing activity of a neuron

\[
\max_{x \in \mathcal{X}} p_\theta(\omega_c \mid x) + \lambda \Omega(x)
\]
Techniques of Interpretation

Interpreting models

- find prototypical example of a category
- find pattern maximizing activity of a neuron

simple regularizer
(Simonyan et al. 2013)

\[
\max_{x \in \mathcal{X}} p_{\theta}(\omega_c | x) + \lambda \Omega(x)
\]
Techniques of Interpretation

Interpreting models

- find prototypical example of a category
- find pattern maximizing activity of a neuron

\[ \max_{x \in X} p_{\theta}(\omega_c | x) + \lambda \Omega(x) \]
Techniques of Interpretation

Explaining decisions
- “why” does the model arrive at a certain prediction
- verify that model behaves as expected
Explaining decisions

- “why” does the model arrive at a certain prediction
- verify that model behaves as expected

Techniques of Interpretation

- Sensitivity Analysis
- Layer-wise Relevance Propagation (LRP)

\[ f(x) \]
Techniques of Interpretation

Sensitivity Analysis
(Simonyan et al. 2014)

Explain prediction
(which pixels lead to decrease of prediction score when changed)

\[ \left| \frac{\partial}{\partial x_p} f(x) \right| \]

Classifier

Bird

prediction \( f(x) \)
Techniques of Interpretation

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

Theoretical interpretation

Deep Taylor Decomposition  
(Montavon et al., 2017)
Techniques of Interpretation

Sensitivity Analysis:
“what makes this image less / more ‘scooter’?”

LRP / Taylor Decomposition:
“what makes this image ‘scooter’ at all?”
More to come

- Berkes 2006
- Erhan 2010
- Simonyan 2013
- Nguyen 2015/16
- Hinton 2006
- Goodfellow 2014
- v. den Oord 2016
- Nguyen 2016
- Khan 2001
- Gevrey 2003
- Baehrens 2010
- Simonyan 2013
- Poulin 2006
- Landecker 2013
- Bach 2015
- Montavon 2017

quality of explanations, applications, interpretability in the sciences, discussion