Interpretability of Deep Learning Models

Devinder Kumar
PhD Candidate, UWaterloo & Vector Inst. for AI
Lead AI Scientist in Residence, NextAI
July 12th, 2018
Most of today’s material is not very mathematical.

Key ideas:
1. What do we mean by Interpretability?
2. Why do we need Interpretability?
3. How it is useful?
Evolution of Neural Networks

Simple Neural Net
1980s
Evolution of Convolutional Neural Networks!

LeNet 1989 (LIP-6, Paris)
WE NEED TO GO

DEEPER
Evolution of Convolutional Neural Networks!

AlexNet 2012

60M parameters
WE NEED TO GO DEEPER

CHALLENGE ACCEPTED
Current Models

GoogleNet / Inception, 2015/16

6.9 M parameters in the model
Future AI: Promises
Future AI: Reality

**Scalable Oversight:** How can we efficiently ensure that a given AI system respects aspects of the objective that are too expensive to be frequently evaluated during training?

- Google on its challenges for AI, Dailymail UK, June 2016

“**There is no neural network in the world, and no method right now that can be trained to identify objects and images, play Space Invaders, and listen to music.**”

- R.Hadsel (Google DeepMind), The Verge, Oct 2016

“**We can build these models, but we don’t know how they work.**”

- Deep Patient, MIT Review (April, 2017)
Future AI: Explainable

Machine Learning System

Input Data → Black box AI Model → Cat

This is a cat.
Current Explanation

This is a cat:
• It has fur, whiskers, and claws.
• It has this feature:

XAI Explanation
Why do we care about X-AI?
Explainability

How did I get in this situation?
Well...

Photo credit: oragesky3
Explainable Models

New Approach
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)
- Neural Nets
- Statistical Models
- SVMs
- Deep Learning
- AOGs
- Graphical Models
- Bayesian Belief Nets
- MNNs
- Markov Models
- Ensemble Methods
- Decision Trees
- SVMs
- AOGs
- SRL
- CRFs
- HBNs
- MLNs

Explainability (notional)

via DARPA
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance.
Explainable Models

New Approach
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)
- Neural Nets
- Statistical Models
- Deep Learning
- Ensemble Methods
- Decision Trees
- SVMs
- AOGs
- Graphical Models
- Bayesian Belief Nets
- SRL
- CRFs
- HBNs
- MLNs

Interpretable Models
- Techniques to learn more structured, interpretable, causal models

Deep Explanation
- Modified deep learning techniques to learn explainable features

Explainability (notional)

via DARPA
Explainable Models

New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance.

Learning Techniques (today)

- Neural Nets
- Deep Learning
- Statistical Models
- SVMs
- AOGs
- Graphical Models
- Bayesian Belief Nets
- Ensemble Methods
- Decision Trees
- Random Forests
- CRFs
- HMMs
- MLEs
- HBNs
- MLNs

Explainability (notional)

- Deep Explanation
  - Modified deep learning techniques to learn explainable features
- Interpretable Models
  - Techniques to learn more structured, interpretable, causal models
- Model Induction
  - Techniques to infer an explainable model from any model as a black box

via DARPA
What is Happening?
General Data Protection Regulation (GDPR)

“more and clearer information about processing”
T-SNE: Visualization!
Deconvolution: Zeiler et.al. ECCV 14
Guided backpropagation: ICLR 2015

Saliency: Simonyan et.al. CVPR 2013
Deep Taylor Decomp. Montavon et. al. PR journal 2017
Class Activation Maps (CAM) Bolei Zhou MIT, 2016
Prediction Difference: Zintgraf et. al. ICLR 2017
Stanford Dog Dataset Results

Chihuahua  Jap. Spaniel  Maltese  Pekinese
Shih-Tzu   B. Spaniel   Papillon  Toy Terrier
R. Ridgeback  Afghan  Hound

Shih-Tzu  Maltese  Chihuahua  Toy-Terrier
Application in Healthcare
Scenario 1: Without CADx

After careful observation found hemorrhages and other relevant symptoms; looks like a case of proliferate DR.

Scenario 2: CAD system without Interpretability

CADx is telling me proliferate DR; Why? What is the evidence.

Scenario 3: Interpretable CADx via CLEAR-DR

CAD is telling me proliferate DR and it shows me the relevant landmarks!

DR: Diabetic Retinopathy

CLEAR-DR map
Diabetic Retinopathy: Leading Cause of Blindness in the world

**Normal** Retina

**Diabetic** Retina
Application in Finance
Un-interpretable Black Box Binary Stock Market Prediction

Interpretable Stock Market Prediction with CLEAR-Trade Visualization

Stock will go up

Stock will go up because of:

I know the reason why!

Explanation from CLEAR-Trade visualization
S&P 500 Stock Data

Stock will go up because of features shown in CLEAR-Trade visualization

Deconvolution

CLEAR-Trade Visualization

Dominant Class Attentive Map

Individual Response Map

Dominant Response Map
Correct Binary Stock Prediction

Ground Truth:

Predicted:

<table>
<thead>
<tr>
<th>Date</th>
<th>Open</th>
<th>High</th>
<th>Low</th>
<th>Close</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/23/2015</td>
<td>107.27</td>
<td>108.85</td>
<td>107.2</td>
<td>108.61</td>
<td>3.3E+07</td>
</tr>
<tr>
<td>12/22/2015</td>
<td>107.4</td>
<td>107.72</td>
<td>106.45</td>
<td>107.23</td>
<td>3.3E+07</td>
</tr>
<tr>
<td>12/21/2015</td>
<td>107.28</td>
<td>107.37</td>
<td>105.57</td>
<td>107.35</td>
<td>4.8E+07</td>
</tr>
<tr>
<td>12/18/2015</td>
<td>108.91</td>
<td>109.52</td>
<td>105.81</td>
<td>106.03</td>
<td>9.6E+07</td>
</tr>
<tr>
<td>12/17/2015</td>
<td>112.02</td>
<td>112.25</td>
<td>108.98</td>
<td>108.98</td>
<td>4.5E+07</td>
</tr>
<tr>
<td>12/16/2015</td>
<td>111.07</td>
<td>111.99</td>
<td>108.80</td>
<td>111.34</td>
<td>5.6E+07</td>
</tr>
<tr>
<td>12/15/2015</td>
<td>111.94</td>
<td>112.80</td>
<td>110.35</td>
<td>110.49</td>
<td>5.3E+07</td>
</tr>
<tr>
<td>12/14/2015</td>
<td>112.18</td>
<td>112.68</td>
<td>109.79</td>
<td>112.48</td>
<td>6.4E+07</td>
</tr>
<tr>
<td>12/11/2015</td>
<td>115.19</td>
<td>115.39</td>
<td>112.85</td>
<td>113.18</td>
<td>4.7E+07</td>
</tr>
<tr>
<td>12/10/2015</td>
<td>116.04</td>
<td>116.94</td>
<td>115.51</td>
<td>116.17</td>
<td>2.9E+07</td>
</tr>
<tr>
<td>12/9/2015</td>
<td>117.64</td>
<td>117.69</td>
<td>115.08</td>
<td>115.97</td>
<td>6.4E+07</td>
</tr>
<tr>
<td>12/8/2015</td>
<td>117.52</td>
<td>118.60</td>
<td>116.86</td>
<td>118.23</td>
<td>3.4E+07</td>
</tr>
<tr>
<td>12/7/2015</td>
<td>118.98</td>
<td>119.86</td>
<td>117.81</td>
<td>118.28</td>
<td>3.2E+07</td>
</tr>
<tr>
<td>12/4/2015</td>
<td>115.29</td>
<td>119.25</td>
<td>115.11</td>
<td>119.03</td>
<td>5.8E+07</td>
</tr>
<tr>
<td>12/3/2015</td>
<td>116.55</td>
<td>116.79</td>
<td>114.22</td>
<td>115.42</td>
<td>5.2E+07</td>
</tr>
<tr>
<td>12/2/2015</td>
<td>117.34</td>
<td>118.11</td>
<td>116.08</td>
<td>116.28</td>
<td>3.3E+07</td>
</tr>
<tr>
<td>12/1/2015</td>
<td>118.75</td>
<td>118.81</td>
<td>116.86</td>
<td>117.34</td>
<td>3.5E+07</td>
</tr>
<tr>
<td>11/30/2015</td>
<td>117.99</td>
<td>119.41</td>
<td>117.75</td>
<td>118.3</td>
<td>3.9E+07</td>
</tr>
<tr>
<td>11/27/2015</td>
<td>118.29</td>
<td>118.41</td>
<td>117.61</td>
<td>117.81</td>
<td>3.3E+07</td>
</tr>
<tr>
<td>11/25/2015</td>
<td>117.21</td>
<td>119.25</td>
<td>117.95</td>
<td>118.03</td>
<td>2.1E+07</td>
</tr>
<tr>
<td>11/24/2015</td>
<td>117.38</td>
<td>119.35</td>
<td>117.12</td>
<td>118.88</td>
<td>4.3E+07</td>
</tr>
<tr>
<td>11/23/2015</td>
<td>119.27</td>
<td>119.75</td>
<td>117.34</td>
<td>117.76</td>
<td>3.2E+07</td>
</tr>
<tr>
<td>11/20/2015</td>
<td>119.2</td>
<td>119.9</td>
<td>118.86</td>
<td>119.3</td>
<td>3.4E+07</td>
</tr>
<tr>
<td>11/19/2015</td>
<td>117.64</td>
<td>119.75</td>
<td>116.76</td>
<td>118.78</td>
<td>4.3E+07</td>
</tr>
<tr>
<td>11/18/2015</td>
<td>115.76</td>
<td>117.49</td>
<td>115.5</td>
<td>117.29</td>
<td>4.7E+07</td>
</tr>
<tr>
<td>11/17/2015</td>
<td>114.92</td>
<td>115.05</td>
<td>113.32</td>
<td>113.69</td>
<td>2.8E+07</td>
</tr>
<tr>
<td>11/16/2015</td>
<td>113.38</td>
<td>114.24</td>
<td>111.11</td>
<td>114.18</td>
<td>3.8E+07</td>
</tr>
<tr>
<td>11/13/2015</td>
<td>115.2</td>
<td>115.57</td>
<td>112.27</td>
<td>112.34</td>
<td>4.6E+07</td>
</tr>
<tr>
<td>11/12/2015</td>
<td>116.26</td>
<td>116.82</td>
<td>115.65</td>
<td>115.72</td>
<td>3.3E+07</td>
</tr>
<tr>
<td>11/11/2015</td>
<td>116.37</td>
<td>117.42</td>
<td>115.21</td>
<td>116.11</td>
<td>4.5E+07</td>
</tr>
<tr>
<td>11/10/2015</td>
<td>116.9</td>
<td>118.07</td>
<td>116.06</td>
<td>116.77</td>
<td>5.9E+07</td>
</tr>
</tbody>
</table>
Wrong Binary Stock Prediction

Ground Truth: Predicted:

![Graph of stock price over time with ground truth and predicted values highlighted.]
Application in Science *(the real one!)*
First paper in Material Science that shows Neural Network Interpretability!

Ground Truth

- Body-centered-tetragonal (bct) structure
- Rhombohedral (rh) structure
- Simple cubic (sc) structure
- Diamond (diam) structure
- Hexagonal (hex) structure
- Face-centered-cubic (fcc) structure
- Body-centered-cubic (bcc) structure

Predicted

- bct_{139}
- hex/rh
- diam
- sc
- fcc
- bcc
Thank You!

devinder.kumar@uwaterloo.ca

http://devinderkumar.com