Interpretability of Neural Networks

Kian Katanforoosh
How can one who will be held accountable for a decision trust a neural network’s recommendation, and justify its use?

Neural networks are deployed:

- In our phones: to recommend content,
- In banks: to manage investments,
- In hospitals: to help doctors diagnose disease symptoms,
- In insurance agencies: to evaluate risk and underwrite documents,
- In cars: to help avoid accidents.
I. Interpreting Neural Networks’ outputs
   A. with saliency maps
   B. with occlusion sensitivity
   C. with class activation maps (Global Average Pooling)

II. Visualizing Neural Networks from the inside
   A. with gradient ascent (class model visualization)
   B. with dataset search
   C. the deconvolution and its applications

III. (Optional: Deep Dream: going deeper in NNs)
I. A. Interpreting and visualizing Neural Networks with saliency maps

**Context:** You have built an animal classifier for a zoo. They are reluctant to use your model without human supervision, because they don’t understand the decision process of the model.

**Question:** How can you alleviate their concerns?

Can be used for segmentation?

indicates which pixels need to be changed the least to affect the class score the most.

I. B. Interpreting and visualizing Neural Networks with occlusion sensitivity

**Context:** You have built an animal classifier for a zoo. They are a little reluctant to use your model without human supervision, because they don’t understand the decision process of the model.

**Question:** What can you do to alleviate their concerns?

---

Context: You have built an animal classifier for a zoo. They are a little reluctant to use your model without human supervision, because they don’t understand the decision process of the model.

**Question:** What can you do to alleviate their concerns?

---

Context: You have built an animal classifier for a zoo. They are a little reluctant to use your model without human supervision, because they don’t understand the decision process of the model.

**Question:** What can you do to alleviate their concerns?

---

Context: You have built an animal classifier for a zoo. They are a little reluctant to use your model without human supervision, because they don’t understand the decision process of the model.

**Question:** What can you do to alleviate their concerns?

---

Context: You have built an animal classifier for a zoo. They are a little reluctant to use your model without human supervision, because they don’t understand the decision process of the model.

**Question:** What can you do to alleviate their concerns?

---

Context: You have built an animal classifier for a zoo. They are a little reluctant to use your model without human supervision, because they don’t understand the decision process of the model.

**Question:** What can you do to alleviate their concerns?

---

Context: You have built an animal classifier for a zoo. They are a little reluctant to use your model without human supervision, because they don’t understand the decision process of the model.

**Question:** What can you do to alleviate their concerns?
I. B. Interpreting and visualizing Neural Networks with occlusion sensitivity

Probability map of the true class for different positions of the grey square

[Matthew D. Zeiler and Rob Fergus (2013): Visualizing and Understanding Convolutional Networks]
III. C. Interpreting NNs using class activation maps

**Context:** Along with the classification output, the zoo now wants real-time visualization of the model’s decision process. You have one day. What do you do?

Using a classification network for localization:

Converted to:

Why this?

III. C. Interpreting NNs using class activation maps

III. C. Interpreting NNs using class activation maps

Source video: Kyle McDonald
I. Interpreting Neural Networks’ outputs
   A. with saliency maps
   B. with occlusion sensitivity
   C. with class activation maps (Global Average Pooling)

II. Visualizing Neural Networks from the inside
   A. with gradient ascent (class model visualization)
   B. with dataset search
   C. The deconvolution and its applications

III. (Optional: Deep Dream: going deeper in NNs)
II. A. Visualizing NNs from the inside using gradient ascent (class model visualization)

**Context**: The zoo trusts that your model correctly locates animals. They get scared and they ask you whether the model understands what a dog is.

Given this trained ConvNet, generate an image which is representative of the class “dog” according to the ConvNet.

Keep the weights fixed and use gradient ascent on the input image to maximize this loss:

\[ L = s_{dog}(x) - \lambda \|x\|_2^2 \]

**Gradient ascent:**

\[ x = x + \alpha \frac{\partial L}{\partial x} \]

"x should look natural"

Repeat this process:
1. Forward propagate image x
2. Compute the objective L
3. Backpropagate to get dL/dx
4. Update x's pixels with gradient ascent


Kian Katanforoosh
II. A. Visualizing NNs from the inside using gradient ascent (class model visualization)

We can do this for all classes:

L2 Regularization

Looks better with additional regularization methods.

[Jason Yosinski et al. (2015): Understanding Neural Networks Through Deep Visualization]
This method can be applied to any activation in the network in order to interpret what a neuron is detecting.

On the class score:

\[ L = S_{\text{dog}}(x) - R(x) \]

change to

On any activation:

\[ L = a^{[l]}_j(x) - R(x) \]
Given a filter, what examples in the dataset lead to a strongly activated feature map?

**Top 5 images**
- It seems that the filter has learned to detect shirts
- It seems that the filter has learned to detect edges

**Top 5 images**

---

**Context:** The zoo loved the technique, and asks if there are other alternatives.
II. B. Visualizing NNs from the inside using dataset search

How did we crop the dataset images on the previous slide?

Input image
(64,64,3)

Encoding volume
(13,13,256)

CONV
(5 layers)

Conclusion: the deeper the activation, the more it “sees” from the image
I. Interpreting Neural Networks’ outputs
   A. with saliency maps
   B. with occlusion sensitivity
   C. with class activation maps (Global Average Pooling)

II. Visualizing Neural Networks from the inside
   A. with gradient ascent (class model visualization)
   B. with dataset search
   C. The deconvolution and its applications

III. (Optional: Deep Dream: going deeper in NNs)
II. C. The deconvolution and its applications

Motivation behind deconvolution/upsampling layers

100-d random code

\[
\begin{bmatrix}
0.47 \\
\vdots \\
0.19
\end{bmatrix}
\]

\(Z\)

Generator “G”
(Neural Network)

\((64,64,3)\) generated image

\[
\left\{ 
\begin{array}{ll}
y = 0 & \text{if } x = G(z) \\
y = 1 & \text{otherwise}
\end{array}
\right.
\]

Discriminator “D”
(Neural Network)

Real images
(database)

Binary classification
Motivation behind deconvolution/upsampling layers

<table>
<thead>
<tr>
<th>Input image</th>
<th>Convolutions (reduces volume height and width)</th>
<th>Encoding</th>
<th>De-convolutions (increases volume height and width)</th>
<th>Per-Pixel Classification (400, 400, 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(400, 400, 3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Get your pencils ready, we’re about to study the deconvolution.
Consider the following CONV1D:

$$n_y = \left\lfloor \frac{n_x - f + 2p}{s} \right\rfloor + 1 = \left\lfloor \frac{8 - 4 + 2 \times 2}{2} \right\rfloor + 1 = 5$$

Let's define our filter as:

$$f = (w_1, w_2, w_3, w_4)$$

The system of equations is:

$$y_1 = w_1 \cdot 0 + w_2 \cdot 0 + w_3 \cdot x_1 + w_4 \cdot x_2$$
$$y_2 = w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + w_4 \cdot x_4$$
$$\vdots$$
$$y_5 = w_1 \cdot x_7 + w_2 \cdot x_8 + w_3 \cdot 0 + w_4 \cdot 0$$
Let’s rewrite the system of equations:

\[
\begin{align*}
y_1 &= w_1 \cdot 0 + w_2 \cdot 0 + w_3 \cdot x_1 + w_4 \cdot x_2 \\
y_2 &= w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + w_4 \cdot x_4 \\
\vdots \\
y_5 &= w_1 \cdot x_7 + w_2 \cdot x_8 + w_3 \cdot 0 + w_4 \cdot 0
\end{align*}
\]

1D CONV can be rewritten as a matrix vector multiplication:

\[
y = Wx
\]

\[
\begin{bmatrix}
y_1 \\
y_2 \\
y_3 \\
y_4 \\
y_5
\end{bmatrix} =
\begin{bmatrix}
w_1 & w_2 & w_3 & w_4 & 0 & \cdots & 0 \\
0 & 0 & w_1 & w_2 & w_3 & w_4 & 0 & \cdots \\
0 & 0 & 0 & 0 & w_1 & w_2 & w_3 & w_4 & 0 & \cdots \\
\vdots \\
0 & \cdots & 0 & 0 & 0 & 0 & w_1 & w_2 & w_3 & w_4
\end{bmatrix}
\begin{bmatrix}
0 \\
x_1 \\
x_2 \\
x_3 \\
x_4 \\
\vdots \\
x_8 \\
0 \\
0
\end{bmatrix}
\]
\textbf{1D CONV} \quad y = Wx

\textbf{1D DECONV} \quad \text{If } W \text{ is invertible, then } \exists H = W^{-1} \text{ such that } x = Hy

In practice, we would even assume that \( W \) is orthogonal, i.e. \( W^{-1} = W^T \)

For example \( (w_1, w_2, w_3, w_4) = (-1, 0, 0, 0, 1) \)

Thus: \( x = W^Ty \) \quad \text{Deconvolution ~ Transposed convolution}
Let's rewrite: $x = W^T y$

Transposed convolution with stride 2

Sub-pixel convolution with stride 1/2
What to remember?

Implementing a deconvolution (sub-pixel version) is similar to the convolution, with:

1. Create a sub-pixel version of the input (i.e., insert zeros and pad)
2. Flip the filters.
3. Divide the stride by 2.
**Motivation of DeconvNets for visualization:** Here is a CNN, trained on ImageNet (1.3m images, 1000 classes), we're trying to interpret by reconstructing the activation’s zone of influence in the input space.

---

**III. A. Interpreting NNs using deconvolutions**

- **Input image** \((256,256,3)\)
- **ZERO PAD**
- **CONV**
- **ReLU**
- **MAX POOL**
- **CONV**
- **ReLU**
- **MAX POOL**
- **CONV**
- **ReLU**
- **MAX POOL**
- **Flatten**
- **FC** \((x3)\)
- **SOFTMAX**

Only one feature maps (among 256) is displayed here:

- Keep the max activation of a feature map
- Set all other activations of the layer to 0

**Output**

- \(0.02\)
- \(0.04\)
- \(0.07\)
- \(0.11\)
- \(0.09\)

---

Matthew D. Zeiler and Rob Fergus (2013): Visualizing and Understanding Convolutional Networks
II. C. The deconvolution and its applications

\[
\begin{pmatrix}
0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
\end{pmatrix}
\]

Switches

MAXPOOL

UNPOOLED
II. C. The deconvolution and its applications

We need to pass the filters and switches from the ConvNet to the DeconvNet.

(5,5,256)
• Keep the max activation of a feature map
• Set all other activations of the layer to 0
II. C. The deconvolution and its applications

\[ a^{[i]} = \mathbb{I}\{a^{[i+1]} \geq 0\} \cdot a^{[i+1]} \]

---

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
<th>3</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

**“ReLU forward”**

<table>
<thead>
<tr>
<th></th>
<th>-2</th>
<th>1</th>
<th>-31</th>
<th>-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>-2</th>
<th>1</th>
<th>-31</th>
<th>-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**“ReLU backward”**

<table>
<thead>
<tr>
<th></th>
<th>-2</th>
<th>1</th>
<th>-31</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>-2</th>
<th>1</th>
<th>-31</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**“ReLU DeconvNet”**

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td>2</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>-2</td>
<td>12</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

[Sringenberg & Dosovitskiy et al. (2015): Visualizing and Understanding Convolutional Networks]
II. C. The deconvolution and its applications

We need to pass the filters and switches from the ConvNet to the DeconvNet.

[Matthew D. Zeiler and Rob Fergus (2013): Visualizing and Understanding Convolutional Networks]
II. C. The deconvolution and its applications

Other CONV layers can be visualized the same way

[Matthew D. Zeiler and Rob Fergus (2013): Visualizing and Understanding Convolutional Networks]
Results on a validation set of 50,000 images

- Top-9 strongest activations per filter in the 1st layer
- Because we know the position of the activation and all the pooling switches we can crop the part of the image that fired the activation

[Matthew D. Zeiler and Rob Fergus (2013): Visualizing and Understanding Convolutional Networks]
Results on a validation set of 50,000 images

- Learning a more complex set of patterns than 1st layer edges.
- Covers a much larger space of the image because of the pooling layer before.
- Top-1 strongest activation per feature map in the 2nd layer (256 feature maps.)
Results on a validation set of 50,000 images

- Learning a more complex set of patterns than 1st layer edges
- Covers a much larger space of the image probably because of the pooling layer before.
- Features are more invariant to small changes. Ex: A dot, spiral, circle all fire the same 2nd layer feature very strongly

[Matthew D. Zeiler and Rob Fergus (2013): Visualizing and Understanding Convolutional Networks]
Results on a validation set of 50,000 images

- 3rd layer: increased complexity

- An activated neuron is seeing \( \approx 80 \times 80 \) part of a \( 256 \times 256 \) image

- Learning objects, faces etc..
Results on a validation set of 50,000 images

- 3rd layer: increased complexity
- An activated neuron is seeing $\approx 80 \times 80$ part of a $256 \times 256$ image
- Learning objects, faces etc..
- Patches: Semantic grouping, not structural

[Matthew D. Zeiler and Rob Fergus (2013): Visualizing and Understanding Convolutional Networks]
I. Interpreting Neural Networks’ outputs
   A. with saliency maps
   B. with occlusion sensitivity
   C. with class activation maps (Global Average Pooling)

II. Visualizing Neural Networks from the inside
   A. with gradient ascent (class model visualization)
   B. with dataset search
   C. The deconvolution and its applications

III. (Optional: Deep Dream: going deeper in NNs)
How to boost the activation of a neuron?

**III. Deep Dream: going deeper in NNs**

1. Forward propagate image until dreaming layer
2. Set gradients of dreaming layer to be equal to its activations
3. Backpropagate gradients to input image
4. Update Pixels of the image

[Alexander Mordvintsev et al. (2015): Inceptionism: Going Deeper into Neural Networks]
III. Deep Dream: going deeper in NNs

[Alexander Mordvintsev et al. (2015): Inceptionism: Going Deeper into Neural Networks]
III. Deep Dream: going deeper in NNs

If you dream in lower layers:

[Alexander Mordvintsev et al. (2015): Inceptionism: Going Deeper into Neural Networks] }
III. Deep Dream: going deeper in NNs

[Link to video: https://www.youtube.com/watch?v=DgPaCWJL7X]
[Github repo (deepdream): https://github.com/google/deepdream/]
[Alexander Mordvintsev et al. (2015): Inceptionism: Going Deeper into Neural Networks]
III. Deep Dream

Did the neural network learned the right features to detect an object?

Dumbbells

*the network failed to understand the essence of a dumbbell*

[Github repo (deepdream): https://github.com/google/deepdream/]

[Alexander Mordvintsev et al. (2015): Inceptionism: Going Deeper into Neural Networks]
Questions we are now able to answer:

- What part of the input is responsible for the output?
  - Occlusion sensitivity
  - Class Activation Maps

- What is the role of a given neuron/filter/layer?
  - Deconvolutions can help visualize the role of a neuron
  - Search dataset images maximizing the activation
  - Gradient ascent (class model visualization)

- Can we check what the network focuses on given an input image?
  - Occlusion sensitivity
  - Saliency maps (one-time gradient ascent)
  - Class Activation Maps

- How does a neural network see our world?
  - Gradient ascent (class model visualization)
  - Deep Dream

- Do these visualization have use cases?
  - Segmentation (saliency maps)
  - Art (Deep Dream)
Duties for next week

Meet with your assigned TA between 1/28 and 2/26 to discuss your milestone report.

Project Milestone due 02/26 Friday 11:59 PM

Completed modules for 03/04:
- C5M1: Recurrent Neural Networks (slides)

Quizzes (due at 8 30am PST):
- Recurrent Neural Networks

Programming Assignments (due at 8 30am PST):  
- Building a Recurrent Neural Network - Step by Step
- Dinosaur Land -- Character-level Language Modeling
- Jazz improvisation with LSTM