Network Visualizations

How Well do Feature Visualizations Support Causal Understanding of CNN Activations?
1.1 One Possible Visualization

medium.com on Convolutional Neural Network(CNN) with Practical Implementation
1.2 More Intuitive Visualizations
Outline

• 1. Introduction

• 2. Image Synthesis by Activation Maximization

• 3. The Experiment

• 4. Results

• 5. Conclusion & Discussion
2.2 Activation Maximization

Animal images from depositphotos.com, ANN image from Raymond C Rowe on researchgate.com
2.2 Activation Maximization

Animal images from depositphotos.com, ANN image from Raymond C Rowe on researchgate.com
2.2 Activation Maximization

Animal images from depositphotos.com, ANN image from Raymond C Rowe on researchgate.com
2.2 Activation Maximization

- Synthesizing images via gradient ascent alone is not enough!
- \(\Rightarrow\) Use of hand designed prior constraints is necessary

Graphic by Nguyen et al. 2019 Understanding Neural Networks via Feature Visualization: A Survey
2.2 Activation Maximization

- [Nguyen et al. „Understanding Neural Networks via Feature Visualization: A survey“ (2019)] Priors:
  - Regularization term: \( x^* = \arg\max_x (a(x) - R(x)) \)
  - Penalize high-intensity pixels
  - Penalize high-frequency noise (i.e. smoothing)
  - Penalize the high frequencies in the gradient image
  - Encourage patch-level colour statistics to be more realistic
  - Randomly jitter, rotate or scale the image before each update step
  - These regularizations help improve \textbf{local} statistics
2.2 Activation Maximization

- **Priors:**
  - **Global** coherence is even harder to achieve

- **Diversity:**

![Images and diagrams illustrating neuron, channel, layer, and softmax concepts.](image-url)
2.2 Activation Maximization

# 2.2 Activation Maximization

<table>
<thead>
<tr>
<th>Real Images</th>
<th>$L_2$ norm</th>
<th>Gaussian blur</th>
<th>Patch dataset</th>
<th>Total variation</th>
<th>Center bias</th>
<th>Mean image initialization</th>
<th>Generator network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ostrich</td>
<td>Lemon</td>
<td>Keyboard</td>
<td>Dumbbell</td>
<td>Kit fox</td>
<td>Bell pepper</td>
<td>Beacon</td>
<td>Volcano</td>
</tr>
<tr>
<td><img src="image" alt="Ostrich" /></td>
<td><img src="image" alt="Lemon" /></td>
<td><img src="image" alt="Keyboard" /></td>
<td><img src="image" alt="Dumbbell" /></td>
<td><img src="image" alt="Kit fox" /></td>
<td><img src="image" alt="Bell pepper" /></td>
<td><img src="image" alt="Beacon" /></td>
<td><img src="image" alt="Volcano" /></td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
3. The Experiment

- How Well do Feature Visualizations Support Causal Understanding of CNN Activations?

- 2021 Paper by Roland S. Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas S. A. Wallis and Wieland Brendel
3. The Experiment

- 3.1 Technical Facts
3.1 Technical Facts

- Inception V1 Network
- Trained on ImageNet
- Query and Natural Images are selected from a random subset of 599,552 images from ImageNet ILSVRC 2012 dataset
- Units are sampled from 9 layers and 2 Inception module branches (3 × 3 and POOL)
3. The Experiment

- 3.1 Technical Facts

- 3.2 The Task
3.2 The Task

Strongly Activating Image

Which image elicits higher activation?

1 2 3
more confident

Synthetic reference image class

Strongly Activating Images

1 2 3
more confident
3.2 The Task

Which image elicits higher activation?

1 more confident
2
3

Mixed reference image class

Muhammad Atta Othman Ahmed: An efficient deep convolutional nn for visual image classification dogs
Pictures from ImageNet
3.2 The Task

Which image elicits higher activation?

Images from focusedcollection.com and depositphotos.com, blurred with befunky online photo editor.
3. The Experiment

- 3.1 Technical Facts

- 3.2 The Task

- 3.3 The Setup
  - 3.3.1 Experiment-Design
  - 3.3.2 Ensuring High Quality Data
3.3.1 The Experiment-Design

For each class of reference images…

… data is collected from 50 MTurk participants.

These 50 each do:

An Instruction Trial

4 Practice Trials

18 Main Trials with 3 Catch Trials
3.3.1 The Experiment-Design

For the Next Class of Reference Images…

… Data is collected from a different Subject

→ Between-Subject Design
3.3.2 Ensuring High Quality Data

• Exclusion Criteria:
  – Time to read instructions
  – Time for whole experiment
  – Performance Threshold for Catch Trials
  – Answer Variability

• Small financial compensation

• Participants only from English speaking countries to ensure that instructions are understood
3. The Experiment

• 3.1 Technical Facts

• 3.2 The Task

• 3.3 The Execution Design

• 3.4 Baselines
3.4 Baselines

- 1. Expert Baseline
- 2. Center Baseline
- 3. Primary Object Baseline
- 4. Variance Baseline
- 5. Saliency Baseline
4. Results

• 4.1 Reference Image Comparison
4.1 Reference Image Comparison

A significant difference between None-Group and the others.

No significant performance difference between the different types of visualization.
4. Results

• 4.1 Reference Image Comparison

• 4.3 Comparison with the Baselines
4.3 Comparison with the Baselines

- **Chance**
  - Center: 0.49
  - Object: 0.63
  - Variance: 0.63
  - Saliency: 0.61

- **Performance**
  - Synthetic: 0.67 ± 0.04
  - Natural: 0.67 ± 0.03
  - Mixed: 0.68 ± 0.04
  - Blur: 0.66 ± 0.03
  - None: 0.60 ± 0.03
4. Results

• 4.1 Reference Image Comparison

• 4.3 Comparison with the Baselines

• 4.4 Performance Variation
4.4 Performance Variation

- Type of visualization not very important for performance
- Systematic performance difference across different units

from layer 8 and 2 of the POOL branch, respectively.
4.4 Performance Variation

- Type of visualization not very important for performance.
- Systematic performance difference across different units.
- Relative Activation: Difference seems to matter for performance (but more so in the Pooling Branch).

(A) 3 x 3 Branch
(B) Pooling Branch

Accuracy vs. Relative activation difference

- Synthetic
- Mixed
- Natural
- Blur
- None

Chance level indicated by the dashed line at 0.50.
5. Conclusion

- Humans are better able to understand and predict behaviour of a CNN when provided visualization
  - Images synthesized by Activation-Maximization are NOT more helpful than other kinds of visualizations
  - Experiment is limited: e.g. fixed size and shape of occlusion patch
  - More visualization methods could be added in the future
Discussion

- Olah et al. „Feature Visualization“

- Nguyen et al. „Understanding Neural Networks via Feature Visualization: A survey“

- Synthesizing the preferred inputs for neurons in neural networks via deep generator networks
• Performance on average very similar to the performance of the experts

• 260 of 298 participants passed the Exclusion Criteria:

• Trial-by-Trial Responses are more similar than chance would predict

• Reasonable Reaction Time
Comparison with the Baselines

<table>
<thead>
<tr>
<th></th>
<th>Synthetic</th>
<th>Natural</th>
<th>Mixed</th>
<th>Blur</th>
<th>None</th>
<th>Center</th>
<th>Variance</th>
<th>Saliency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic</td>
<td>14±4</td>
<td>22±3</td>
<td>22±3</td>
<td>21±2</td>
<td>15±3</td>
<td>6±5</td>
<td>12±7</td>
<td>27±7</td>
</tr>
<tr>
<td>Natural</td>
<td>22±3</td>
<td>33±5</td>
<td>30±3</td>
<td>28±3</td>
<td>25±3</td>
<td>1±6</td>
<td>21±6</td>
<td>40±7</td>
</tr>
<tr>
<td>Mixed</td>
<td>22±3</td>
<td>30±3</td>
<td>29±5</td>
<td>26±3</td>
<td>23±3</td>
<td>-1±6</td>
<td>15±6</td>
<td>32±5</td>
</tr>
<tr>
<td>Blur</td>
<td>21±2</td>
<td>28±3</td>
<td>26±3</td>
<td>24±5</td>
<td>18±3</td>
<td>2±6</td>
<td>14±6</td>
<td>33±6</td>
</tr>
<tr>
<td>None</td>
<td>15±3</td>
<td>25±3</td>
<td>23±3</td>
<td>18±3</td>
<td>28±6</td>
<td>-4±5</td>
<td>18±7</td>
<td>32±10</td>
</tr>
<tr>
<td>Center</td>
<td>6±5</td>
<td>1±6</td>
<td>-1±6</td>
<td>2±6</td>
<td>-4±5</td>
<td>100±0</td>
<td>0±15</td>
<td>-8±12</td>
</tr>
<tr>
<td>Variance</td>
<td>12±7</td>
<td>21±6</td>
<td>15±6</td>
<td>14±6</td>
<td>18±7</td>
<td>0±15</td>
<td>100±0</td>
<td>31±12</td>
</tr>
<tr>
<td>Saliency</td>
<td>27±7</td>
<td>40±7</td>
<td>32±5</td>
<td>33±6</td>
<td>32±10</td>
<td>-8±12</td>
<td>31±12</td>
<td>100±0</td>
</tr>
</tbody>
</table>

Cohen's Kappa [10^{-2}]

Paul Kull