Visualizing Deep Neural Network Decisions: Prediction Difference Analysis

Luisa M. Zintgraf Taco S. Cohen Tameem Adel Max Welling

University of Amsterdam

Canadian Institute of Advanced Research

Vrije Universiteit Brussel

ICLR, 2017
Presenter: Ritambhara Singh

Outline

1. Introduction
   - Motivation
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - Conditional Sampling + Multivariate Analysis
   - Algorithm
   - Deep Visualization of Hidden Layers
3. Results
   - ImageNet
   - MRI Data

Outline

1. Introduction
   - Motivation
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - Conditional Sampling + Multivariate Analysis
   - Algorithm
   - Deep Visualization of Hidden Layers
3. Results
   - ImageNet
   - MRI Data

Motivation

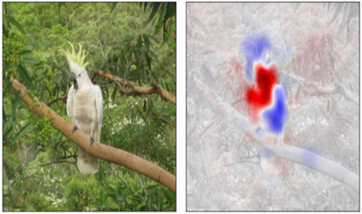
- Making neural network decisions interpretable through visualization.
- Propose **prediction difference analysis method**.

Motivation

- Making neural network decisions interpretable through visualization.
- Propose **prediction difference analysis method**.
- Visualizes the response of a deep neural network to a specific input.

Motivation

- Making neural network decisions interpretable through visualization.
- Propose **prediction difference analysis method**.
- Visualizes the response of a deep neural network to a specific input.



Outline

1. Introduction
   - Motivation
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - Conditional Sampling + Multivariate Analysis
   - Algorithm
   - Deep Visualization of Hidden Layers
3. Results
   - ImageNet
   - MRI Data

State-of-the-art

Robnik-Sikonja and Kononenko (2008)

- Assigns a relevance value to each input feature with respect to a class c.

State-of-the-art

Robnik-Sikonja and Kononenko (2008)

- Assigns a relevance value to each input feature with respect to a class c.
- Basic idea: Relevance of a feature x_i can be estimated by measuring how the prediction changes if the feature is unknown.

State-of-the-art

Robnik-Sikonja and Kononenko (2008)

- Assigns a relevance value to each input feature with respect to a class c.
- Basic idea: Relevance of a feature x_i can be estimated by measuring how the prediction changes if the feature is unknown.
- Difference between $p(c|x)$ and $p(c|x_{i*})$, where x_{i*} denotes the set of all input features except x_i.

Outline

1. Introduction
   - Motivation
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - Conditional Sampling + Multivariate Analysis
   - Algorithm
   - Deep Visualization of Hidden Layers
3. Results
   - ImageNet
   - MRI Data

Details and Drawbacks

To find $p(c|x_{i*})$

- Label the feature as unknown (only few classifiers allow e.g. Naive Bayesian classifier).
- Re-train the classifier with the feature left out (infeasible for DNNs and high-dimensional data like images)

Details and Drawbacks

To find $p(c|x_{i*})$

- Simulate the absence of a feature by marginalizing the feature:

$$p(c|x_{i*}) = \sum_{x_i} p(x_i|x_{i*})p(c|x_i, x_{i*}) \tag{1}$$

Details and Drawbacks

To find $p(c|x_{i*})$

- Simulate the absence of a feature by marginalizing the feature:

$$p(c|x_{i*}) = \sum_{x_i} p(x_i|x_{i*})p(c|x_i, x_{i*}) \tag{1}$$

- Modeling $p(x_i|x_{i*})$ can become infeasible with a large number of features.

Details and Drawbacks

To find $p(c|x_{i*})$

- Approximate equation (1) by assuming feature x_i is independent of the other features, x_{i*}:

$$p(c|x_{i*}) \sim \sum_{x_i} p(x_i) p(c|x_i, x_{i*}) \tag{2}$$

Details and Drawbacks

To find $p(c|x_{i*})$

- Approximate equation (1) by assuming feature x_i is independent of the other features, x_{i*}:

$$p(c|x_{i*}) \sim \sum_{x_i} p(x_i)p(c|x_i, x_{i*}) \tag{2}$$

- Weight of evidence:

$$WE_i(c|x) = log_2(odds(c|x)) - log_2(odds(c|x_{i*})) \tag{3}$$

- Here, $odds(c|x) = p(c|x)/(1 - p(c|x))$.

Details and Drawbacks

To find $p(c|x_{i*})$

- Approximate equation (1) by assuming feature x_i is independent of the other features, x_{i*}:

$$p(c|x_{i*}) \sim \sum_{x_i} p(x_i)p(c|x_i, x_{i*}) \tag{2}$$

- Weight of evidence:

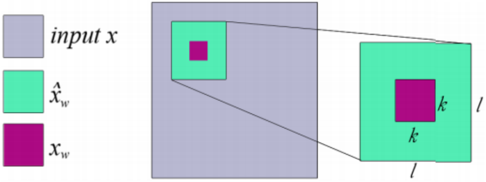
$$WE_i(c|x) = log_2(odds(c|x)) - log_2(odds(c|x_{i*})) \tag{3}$$

- Here, $odds(c|x) = p(c|x)/(1 - p(c|x))$.
- Crude approximation

Outline

1. Introduction
   - Motivation
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - Conditional Sampling + Multivariate Analysis
   - Algorithm
   - Deep Visualization of Hidden Layers
3. Results
   - ImageNet
   - MRI Data

Conditional Sampling + Multivariate Analysis



$$p(x_i|x_{i*}) \sim p(x_i|\hat{x}_{i*}) \tag{4}$$

Outline

1. Introduction
   - Motivation
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - Conditional Sampling + Multivariate Analysis
   - Algorithm
   - Deep Visualization of Hidden Layers
3. Results
   - ImageNet
   - MRI Data

Algorithm

Algorithm 1 Evaluating the prediction difference using conditional and multivariate sampling

Input: classifier with outputs p(c|x), input image $\mathbf{x}$ of size $n \times n$, inner patch size k, outer patch size $l > k$, class of interest c, probabilistic model over patches of size $l \times l$, number of samples S

Initialization: WE = zeros(n*n), counts = zeros(n*n)

for every patch $\mathbf{x}_w$ of size $k \times k$ **in x do**

    $\mathbf{x}' = \text{copy}(\mathbf{x})$

    $\text{sum}_w = 0$

    define patch $\hat{\mathbf{x}}_w$ of size $l \times l$ that contains $\mathbf{x}_w$

    for $s = 1$ **to** S **do**

        $\mathbf{x}'_w \leftarrow \mathbf{x}_w$ sampled from $p(\mathbf{x}_w | \hat{\mathbf{x}}_w \backslash \mathbf{x}_w)$

        $\text{sum}_w \mathrel{+}= p(c|\mathbf{x}')$ ▷ evaluate classifier

    end for

    $p(c|\mathbf{x} \backslash \mathbf{x}_w) := \text{sum}_w / S$

    WE[coordinates of $\mathbf{x}_w$] $\mathrel{+}= \log_2(\text{odds}(c|\mathbf{x})) - \log_2(\text{odds}(c|\mathbf{x} \backslash \mathbf{x}_w))$

    counts[coordinates of $\mathbf{x}_w$] $\mathrel{+}= 1$

end for

Output: WE / counts ▷ point-wise division

Outline

1. Introduction
   - Motivation
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - Conditional Sampling + Multivariate Analysis
   - Algorithm
   - Deep Visualization of Hidden Layers
3. Results
   - ImageNet
   - MRI Data

Deep Visualization of Hidden Layers

- Let h be a vector representation of values in layer H in a network
- Let $z = z(h)$ be a node in subsequent layer
- Analog of equation (2) is :

$$g(z|h_{i*}) = E_{p(h_i|h_{i*})}[z(h)] = \sum_{h_i} p(h_i|h_{i*}) z(h_{i*}, h_i) \tag{5}$$

Deep Visualization of Hidden Layers

- Let h be a vector representation of values in layer H in a network
- Let $z = z(h)$ be a node in subsequent layer
- Analog of equation (2) is :

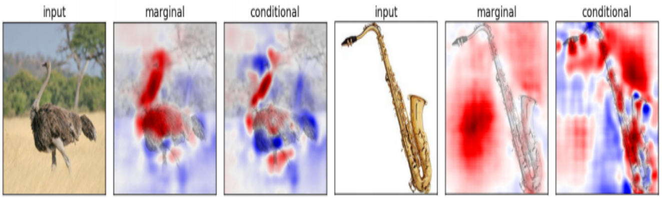
$$g(z|h_{i*}) = E_{p(h_i|h_{i*})}[z(h)] = \sum_{h_i} p(h_i|h_{i*})z(h_{i*}, h_i) \tag{5}$$

- Activation Difference: $AD_i(z|h) = g(z|h) - g(z|h_{i*})$

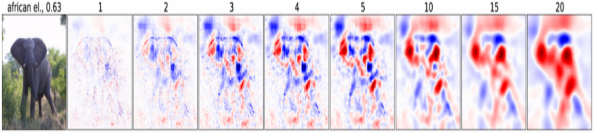
Outline

1. Introduction
   - Motivation
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - Conditional Sampling + Multivariate Analysis
   - Algorithm
   - Deep Visualization of Hidden Layers
3. Results
   - ImageNet
   - MRI Data

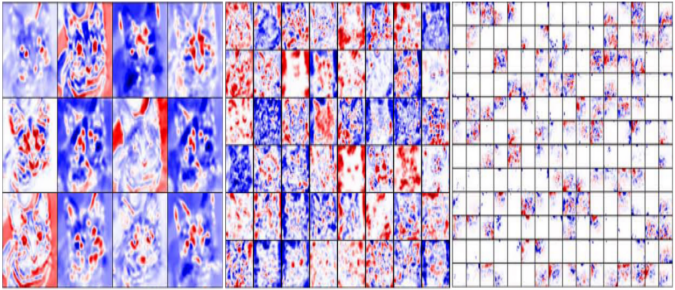
Marginal versus Conditional Sampling



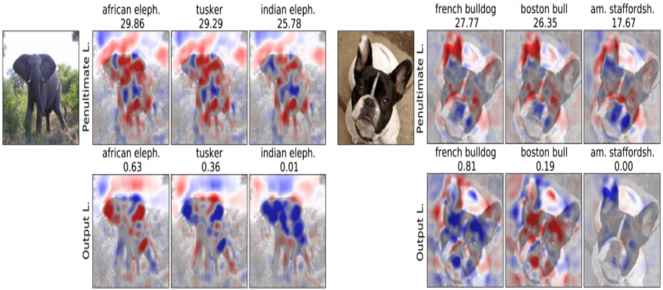
Effect of window size



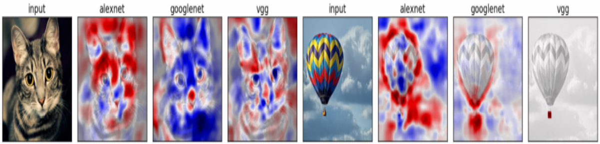
Visualization of layers



Penultimate versus Output Layer



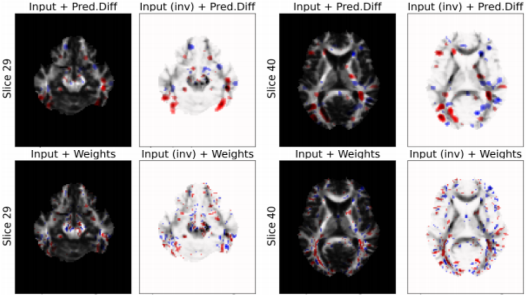
Different DCNN architectures



Outline

1. Introduction
   - Motivation
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - Conditional Sampling + Multivariate Analysis
   - Algorithm
   - Deep Visualization of Hidden Layers
3. Results
   - ImageNet
   - MRI Data

Prediction Difference versus Logistic weights



Summary

- New method for visualizing deep neural networks
- Improves on previous methods by using powerful conditional, multivariate model
- Demonstrated how visualization method can be used for analyzing how DCNNs make decisions
- Future Direction
  - Better approximation by using a conditional distribution that takes more information.
  - A better classification algorithm for clinical analysis.