Learning Important Features Through Propagating Activation Differences

Avanti Shrikumar Peyton Greenside Anshul Kundaje

Stanford University

ICML, 2017

Presenter: Ritambhara Singh

Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - DeepLIFT Method
   - Defining Reference
   - Solution
   - Multipliers and Chain Rule
   - Separating positive and negative contribution
   - Rules for assigning contributions
3. Results
   - MNIST digit classification
   - DNA sequence classification

Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - DeepLIFT Method
   - Defining Reference
   - Solution
   - Multipliers and Chain Rule
   - Separating positive and negative contribution
   - Rules for assigning contributions
3. Results
   - MNIST digit classification
   - DNA sequence classification

Motivation

- **Interpretabilty of neural networks** :Assign importance score to inputs for a given output.

Motivation

- **Interpretabilty of neural networks** :Assign importance score to inputs for a given output.
- Importance is defined in terms of differences from a ‘reference’ state.

Motivation

- **Interpretabilty of neural networks** :Assign importance score to inputs for a given output.
- Importance is defined in terms of differences from a 'reference' state.
- Propagates importance signal even when gradient is zero.

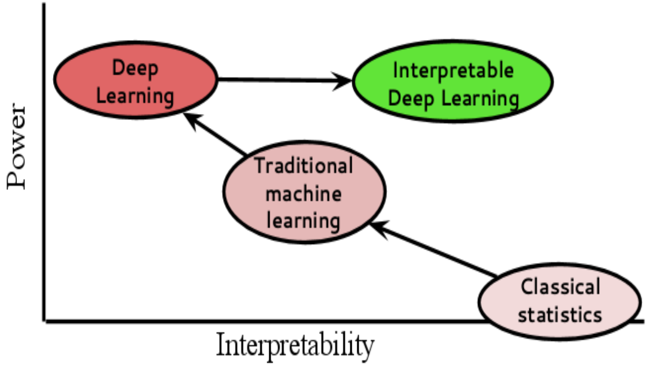
Motivation

- **Interpretabilty of neural networks** :Assign importance score to inputs for a given output.
- Importance is defined in terms of differences from a 'reference' state.
- Propagates importance signal even when gradient is zero.
- Gives separate consideration to positive and negative contributions.

Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - DeepLIFT Method
   - Defining Reference
   - Solution
   - Multipliers and Chain Rule
   - Separating positive and negative contribution
   - Rules for assigning contributions
3. Results
   - MNIST digit classification
   - DNA sequence classification

Interpretation of Neural Networks

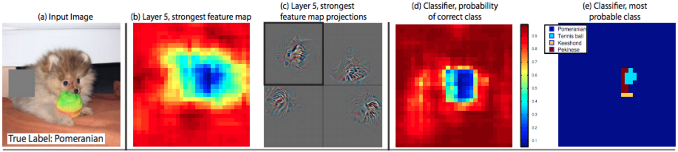


Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - DeepLIFT Method
   - Defining Reference
   - Solution
   - Multipliers and Chain Rule
   - Separating positive and negative contribution
   - Rules for assigning contributions
3. Results
   - MNIST digit classification
   - DNA sequence classification

State-of-the-art

- **Perturbation-based forward propagation approaches:** Zeiler and Fergus (2013), Zhou and Troyanskaya (2015).



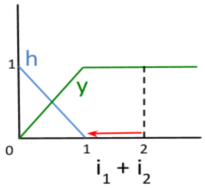
- **Backpropagation-based approaches:** Saliency maps: Simonyan et al. (2013), Guided Backpropagation: Springenberg et al. (2014)

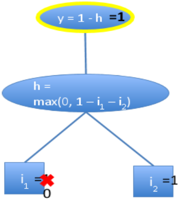
Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - DeepLIFT Method
   - Defining Reference
   - Solution
   - Multipliers and Chain Rule
   - Separating positive and negative contribution
   - Rules for assigning contributions
3. Results
   - MNIST digit classification
   - DNA sequence classification

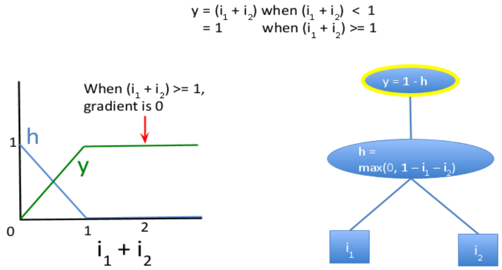
Saturation problem

$$y = (i_1 + i_2) \text{ when } (i_1 + i_2) < 1$$
$$= 1 \quad \text{when } (i_1 + i_2) >= 1$$



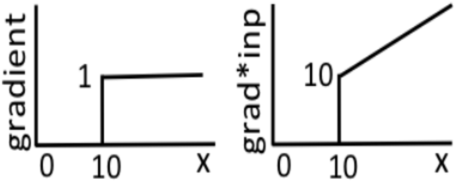


Saturation problem



Thresholding Problem

$y = max(0, x - 10)$



Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - DeepLIFT Method
   - Defining Reference
   - Solution
   - Multipliers and Chain Rule
   - Separating positive and negative contribution
   - Rules for assigning contributions
3. Results
   - MNIST digit classification
   - DNA sequence classification

Philosophy

- Explains difference in output from some ‘reference’ output in terms of difference on input from some ‘reference’ input.

Philosophy

- Explains difference in output from some ‘reference’ output in terms of difference on input from some ‘reference’ input.
- **Summation-to-delta property:**

$$\Sigma_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t \tag{1}$$

Philosophy

- Explains difference in output from some 'reference' output in terms of difference on input from some 'reference' input.
- **Summation-to-delta property:**

$$\Sigma_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t \tag{1}$$

- Blame Δt on $\Delta x_1, \Delta x_2, \ldots$

Philosophy

- Explains difference in output from some 'reference' output in terms of difference on input from some 'reference' input.
- **Summation-to-delta property:**

$$\Sigma_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t \tag{1}$$

- Blame Δt on $\Delta x_1, \Delta x_2, \ldots$
- $C_{\Delta x_i \Delta t}$ can be non-zero even when $\frac{\delta t}{\delta x_i}$ is zero.

Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - DeepLIFT Method
   - Defining Reference
   - Solution
   - Multipliers and Chain Rule
   - Separating positive and negative contribution
   - Rules for assigning contributions
3. Results
   - MNIST digit classification
   - DNA sequence classification

Defining Reference

- Given neuron x with inputs $i_1, i_2, \ldots$ such that $x = f(i_1, i_2, \ldots)$
- Given reference activations $i_1^0, i_2^0, \ldots$ of the input:

$$x^0 = f(i_1^0, i_2^0, \ldots) \tag{2}$$

- Choose reference input and propagate activations though the net.
- Good reference will rely on domain knowledge: "What am I interested in measuring difference against?"

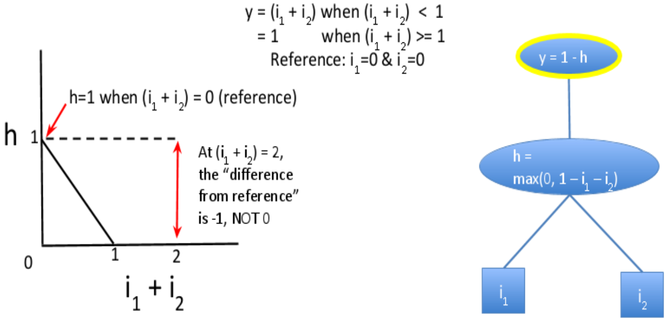
Outline

1 Introduction
- Motivation
- Background
- State-of-the-art
- Drawbacks

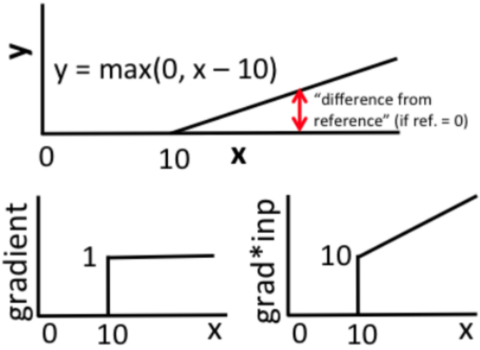
2 Proposed Approach
- DeepLIFT Method
- Defining Reference
- Solution
- Multipliers and Chain Rule
- Separating positive and negative contribution
- Rules for assigning contributions

3 Results
- MNIST digit classification
- DNA sequence classification

Saturation Problem



Thresholding Problem



Outline

1 Introduction
- Motivation
- Background
- State-of-the-art
- Drawbacks

2 Proposed Approach
- DeepLIFT Method
- Defining Reference
- Solution
- Multipliers and Chain Rule
- Separating positive and negative contribution
- Rules for assigning contributions

3 Results
- MNIST digit classification
- DNA sequence classification

Multipliers

$$m_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta t} \quad (3)$$

- Multiplier is the contribution of Δx to Δt divided by Δx
- Compare: partial derivative $= \frac{\delta t}{\delta x}$
- Infinitesimal contribution of δx to δt, divided by δx

Chain Rule

$$m_{\Delta x_i \Delta z} = \Sigma_j m_{\Delta x_i \Delta y_j} m_{\Delta y_j \Delta z} \quad (4)$$

- Can be computed efficiently via backpropagation

Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - DeepLIFT Method
   - Defining Reference
   - Solution
   - Multipliers and Chain Rule
   - Separating positive and negative contribution
   - Rules for assigning contributions
3. Results
   - MNIST digit classification
   - DNA sequence classification

Separating positive and negative contribution

- In some cases, important to treat positive and negative contributions differently.
- Introduce Δx_i^+ and Δx_i^-, such that:

$$\Delta x_i = \Delta x_i^+ + \Delta x_i^-; C_{\Delta x_i \Delta t} = C_{\Delta x_i^+ \Delta t} + C_{\Delta x_i^- \Delta t}$$

Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - DeepLIFT Method
   - Defining Reference
   - Solution
   - Multipliers and Chain Rule
   - Separating positive and negative contribution
   - Rules for assigning contributions
3. Results
   - MNIST digit classification
   - DNA sequence classification

Linear Rule

- For $y = b + \sum_i w_i x_i$, we have $\Delta y = \sum w_i \Delta x_i$
- Define: $\Delta y^+ = \sum_i 1\{w_i \Delta x_i > 0\} w_i \Delta x_i$ $\quad \Delta y^- = \sum_i 1\{w_i \Delta x_i < 0\} w_i \Delta x_i$

$$= \sum_i 1\{w_i \Delta x_i > 0\} w_i (\Delta x_i^+ + \Delta x_i^-) \qquad = \sum_i 1\{w_i \Delta x_i < 0\} w_i (\Delta x_i^+ + \Delta x_i^-)$$

$$C_{\Delta x_i^+ \Delta y^+} = 1\{w_i \Delta x_i > 0\} w_i \Delta x_i^+ \quad C_{\Delta x_i^+ \Delta y^-} = 1\{w_i \Delta x_i < 0\} w_i \Delta x_i^+$$

$$C_{\Delta x_i^- \Delta y^+} = 1\{w_i \Delta x_i > 0\} w_i \Delta x_i^- \quad C_{\Delta x_i^- \Delta y^-} = 1\{w_i \Delta x_i < 0\} w_i \Delta x_i^-$$

$$m_{\Delta x_i^+ \Delta y^+} = m_{\Delta x_i^- \Delta y^+} = 1\{w_i \Delta x_i > 0\} w_i$$

$$m_{\Delta x_i^+ \Delta y^-} = m_{\Delta x_i^- \Delta y^-} = 1\{w_i \Delta x_i < 0\} w_i$$

- When $\Delta x = 0$ (but Δx^+ and Δx^- are not necessarily zero): $m_{\Delta x_i^+ \Delta y^+} = m_{\Delta x_i^+ \Delta y^-} = 0.5 w_i$

Rescale Rule

$$y = f(x)$$

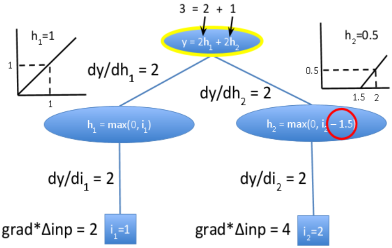
- Set Δy^+ and Δy^- proportional to Δx^+ and Δx^-

$$\Delta y^+ = \frac{\Delta y}{\Delta x}\Delta x^+ = C_{\Delta x^+ \Delta y^+}$$

$$\Delta y^- = \frac{\Delta y}{\Delta x}\Delta x^- = C_{\Delta x^- \Delta y^-}$$

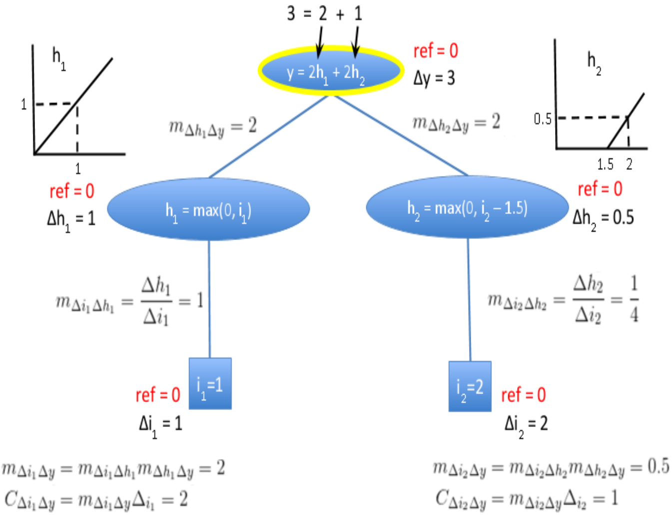
$$m_{\Delta x^+ \Delta y^+} = m_{\Delta x^- \Delta y^-} = m_{\Delta x \Delta y} = \frac{\Delta y}{\Delta x}$$

Where it works

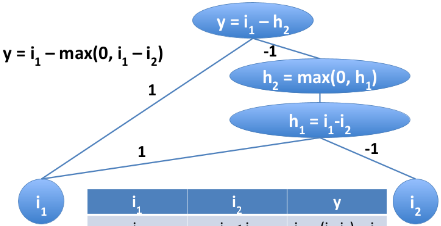


Gradient*Δinput assigns: i_1:2, i_2:4, bias: -3

Where it works

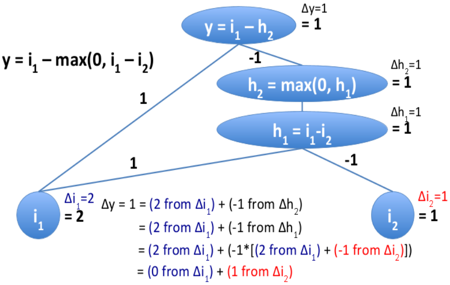


Where it fails: “min” (AND) relation



i_1	i_2	y
i_1	$i_2 < i_1$	$i_1 - (i_1 - i_2) = i_2$
i_1	$i_2 > i_1$	$i_1 - 0 = i_1$

Where it fails: “min” (AND) relation



RevealCancel Rule

$$\Delta y^+ = \frac{1}{2}\left(f(x^0 + \Delta x^+) - f(x^0)\right) \text{ (impact of } \Delta x^+ \text{ after no terms added)}$$

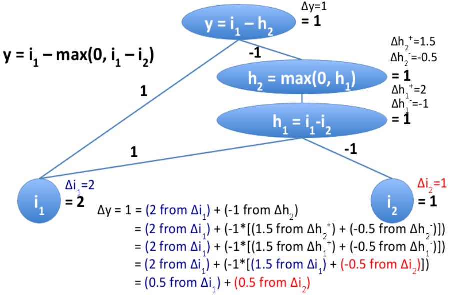
$$+ \frac{1}{2}\left(f(x^0 + \Delta x^- + \Delta x^+) - f(x^0 + \Delta x^-)\right) \text{ (impact of } \Delta x^+ \text{ after negative terms added)}$$

$$\Delta y^- = \frac{1}{2}\left(f(x^0 + \Delta x^-) - f(x^0)\right) \text{ (impact of } \Delta x^- \text{ after no terms added)}$$

$$+ \frac{1}{2}\left(f(x^0 + \Delta x^+ + \Delta x^-) - f(x^0 + \Delta x^+)\right) \text{ (impact of } \Delta x^- \text{ after positive terms added)}$$

$$m_{\Delta x^+ \Delta y^+} = \frac{C_{\Delta x^+ y^+}}{\Delta x^+} = \frac{\Delta y^+}{\Delta x^+} ; m_{\Delta x^- \Delta y^-} = \frac{\Delta y^-}{\Delta x^-}$$

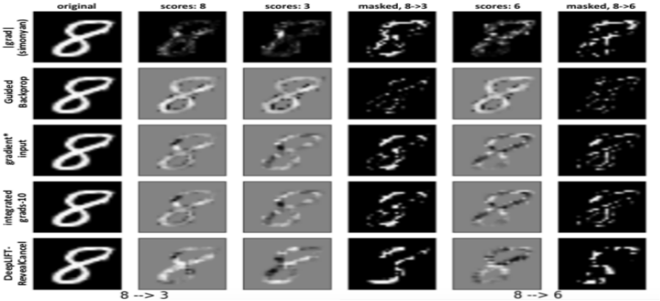
Solution: "min" (AND) relation



Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - DeepLIFT Method
   - Defining Reference
   - Solution
   - Multipliers and Chain Rule
   - Separating positive and negative contribution
   - Rules for assigning contributions
3. Results
   - MNIST digit classification
   - DNA sequence classification

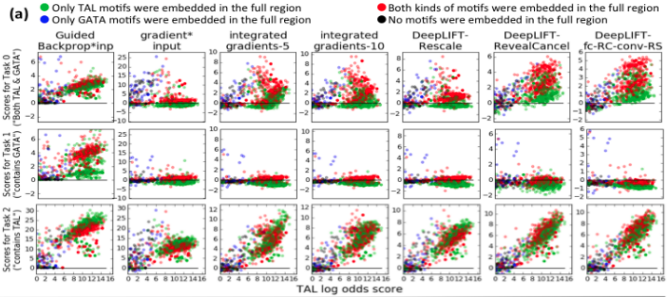
MNIST digit classification



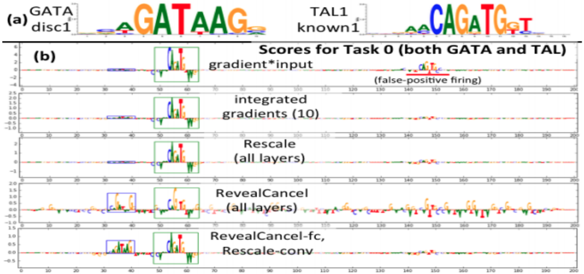
Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
   - Drawbacks
2. Proposed Approach
   - DeepLIFT Method
   - Defining Reference
   - Solution
   - Multipliers and Chain Rule
   - Separating positive and negative contribution
   - Rules for assigning contributions
3. Results
   - MNIST digit classification
   - DNA sequence classification

DNA sequence classification



DNA sequence classification



Summary

- Novel approach for computing importance scores based on differences from the ‘reference’.
- Using difference-from-reference allows information to propagate even when the gradient is zero
- Separates contributions from positive and negative terms
- Video at : `https://www.youtube.com/watch?v=v8cxYjNZAXc&index=1&list=PLJLjQOkqSRTP3cLB2cOOi_bQFw6KPGKML`
- Slides at: `https://drive.google.com/file/d/0B15F_QN41VQXbkVkcTVQYTVQNVE/view`
- Future Direction
  - Applying DeepLIFT to RNNs
  - Compute ‘reference’ empirically from data