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

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

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

- **Interpretability 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

- **Interpretability 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
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


- **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

\[ y = (i_1 + i_2) \text{ when } (i_1 + i_2) < 1 \]
\[ = 1 \text{ when } (i_1 + i_2) \geq 1 \]
Saturation problem

\[ y = (i_1 + i_2) \text{ when } (i_1 + i_2) < 1 \\
= 1 \quad \text{ when } (i_1 + i_2) \geq 1 \]

When \((i_1 + i_2) \geq 1\), gradient is 0

\[ h = \max(0, 1 - i_1 - i_2) \]

\[ y = 1 - h \]
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:**

\[ \sum_{i=1}^{n} C_{\Delta x_i} \Delta t = \Delta t \]  

(1)
Explain difference in output from some ‘reference’ output in terms of difference on input from some ‘reference’ input.

**Summation-to-delta property:**

\[ \sum_{i=1}^{n} C_{\Delta x_i} \Delta t = \Delta t \]  \hspace{1cm} (1)

Blame \( \Delta t \) 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:**

\[ \sum_{i=1}^{n} C_{\Delta x_i} \Delta t = \Delta t \] (1)

- Blame \( \Delta t \) 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 $i^0_1, i^0_2, \ldots$ of the input:
  \[ x^0 = f(i^0_1, i^0_2, \ldots) \]  

- 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

\[ y = (i_1 + i_2) \text{ when } (i_1 + i_2) < 1 \]
\[ = 1 \quad \text{ when } (i_1 + i_2) \geq 1 \]
Reference: \( i_1 = 0 \) & \( i_2 = 0 \)

\[ h = 1 - y \]

\[ h = \max(0, 1 - i_1 - i_2) \]

At \((i_1 + i_2) = 2\), the "difference from reference" is -1, NOT 0.

\( h = 1 \) when \((i_1 + i_2) = 0\) (reference)
Thresholding Problem

\[ y = \max(0, x - 10) \]

“difference from reference” (if ref. = 0)

\[
\begin{align*}
\text{gradient} & \quad \text{grad*inp} \\
0 & \quad 10 & \quad 0 & \quad 10 & \quad x
\end{align*}
\]
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

\[ m_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta t} \]  \hspace{1cm} (3)

- Multiplier is the contribution of \( \Delta x \) to \( \Delta t \) divided by \( \Delta x \)
- Compare: partial derivative = \( \frac{\delta t}{\delta x} \)
- Infinitesimal contribution of \( \delta x \) to \( \delta t \), divided by \( \delta x \)
Chain Rule

\[ m \Delta x_i \Delta z = \sum_j m \Delta x_i \Delta y_j m \Delta y_j \Delta z \] (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
In some cases, important to treat positive and negative contributions differently.

Introduce $\Delta x_i^+$ and $\Delta x_i^-$, 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
• For $y = b + \sum_i w_i x_i$, we have $\Delta y = \sum_i w_i \Delta x_i$

• Define: $\Delta y^+ = \sum_i 1\{w_i \Delta x_i > 0\} w_i \Delta x_i$
  
  $\Delta y^- = \sum_i 1\{w_i \Delta x_i < 0\} w_i \Delta x_i$

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

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

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

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

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

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

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

• When $\Delta x = 0$ (but $\Delta x^+$ and $\Delta x^-$ 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 \( \Delta y^+ \) and \( \Delta y^- \) proportional to \( \Delta x^+ \) and \( \Delta x^- \)

\[
\begin{align*}
\Delta y^+ &= \frac{\Delta y}{\Delta x} \Delta x^+ = C_{\Delta x^+} \Delta y^+ \\
\Delta y^- &= \frac{\Delta y}{\Delta x} \Delta x^- = C_{\Delta x^-} \Delta y^-
\end{align*}
\]

\[
m_{\Delta x^+} \Delta y^+ = m_{\Delta x^-} \Delta y^- = m_{\Delta x} \Delta y = \frac{\Delta y}{\Delta x}
\]
Where it works

Gradient\(\Delta\)input assigns: \(i_1 = 2, i_2 = 4\), bias: -3
Where it works

\[ y = 2h_1 + 2h_2 \]

\[ m_{\Delta h_1 \Delta y} = 2 \]
\[ m_{\Delta h_2 \Delta y} = 2 \]

\[ h_1 = \max(0, i_1) \]
\[ h_2 = \max(0, i_2 - 1.5) \]

\[ m_{\Delta i_1 \Delta h_1} = \frac{\Delta h_1}{\Delta i_1} = 1 \]
\[ m_{\Delta i_2 \Delta h_2} = \frac{\Delta h_2}{\Delta i_2} = \frac{1}{4} \]

\[ \Delta i_1 = 1 \]
\[ \Delta i_2 = 2 \]

\[ m_{\Delta i_1 \Delta y} = m_{\Delta i_1 \Delta h_1} m_{\Delta h_1 \Delta y} = 2 \]
\[ C_{\Delta i_1 \Delta y} = m_{\Delta i_1 \Delta y} \Delta i_1 = 2 \]

\[ m_{\Delta i_2 \Delta y} = m_{\Delta i_2 \Delta h_2} m_{\Delta h_2 \Delta y} = 0.5 \]
\[ C_{\Delta i_2 \Delta y} = m_{\Delta i_2 \Delta y} \Delta i_2 = 1 \]
Where it fails: “min” (AND) relation

\[ y = i_1 - \max(0, i_1 - i_2) \]

\[ h_2 = \max(0, h_1) \]

\[ h_1 = i_1 - i_2 \]

<table>
<thead>
<tr>
<th>( i_1 )</th>
<th>( i_2 )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i_1 )</td>
<td>( i_2 &lt; i_1 )</td>
<td>( i_1 - (i_1 - i_2) = i_2 )</td>
</tr>
<tr>
<td>( i_1 )</td>
<td>( i_2 &gt; i_1 )</td>
<td>( i_1 - 0 = i_1 )</td>
</tr>
</tbody>
</table>
Where it fails: “min” (AND) relation

\[ y = i_1 - \max(0, i_1 - i_2) \]

\[ h_2 = \max(0, h_1) \]

\[ h_1 = i_1 - i_2 \]

\[ y = i_1 - h_2 \]

\[ \Delta y = 1 \]

\[ \Delta h_2 = 1 \]

\[ \Delta h_1 = 1 \]

\[ \Delta i_2 = 1 \]

\[ \Delta i_1 = 2 \]

\[ \Delta y = 1 = (2 \text{ from } \Delta i_1) + (-1 \text{ from } \Delta h_2) \]

\[ = (2 \text{ from } \Delta i_1) + (-1 \text{ from } \Delta h_1) \]

\[ = (2 \text{ from } \Delta i_1) + (-1 \ast [(2 \text{ from } \Delta i_1) + (-1 \text{ from } \Delta i_2)]) \]

\[ = (0 \text{ from } \Delta i_1) + (1 \text{ from } \Delta i_2) \]
\[ \Delta y^+ = \frac{1}{2} \left( f(x^0 + \Delta x^+) - f(x^0) \right) \] (impact of \( \Delta x^+ \) after no terms added)

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

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

\[ + \frac{1}{2} \left( f(x^0 + \Delta x^+ + \Delta x^-) - f(x^0 + \Delta x^+) \right) \] (impact of \( \Delta x^- \) 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

\[ y = i_1 - \max(0, i_1 - i_2) \]

\[ h_2 = \max(0, h_1) \]

\[ h_1 = i_1 - i_2 \]

\[ \Delta i_1 = 2 \]

\[ \Delta y = 1 = (2 \text{ from } \Delta i_1) + (-1 \text{ from } \Delta h_2) \]

\[ = (2 \text{ from } \Delta i_1) + (-1^*[(1.5 \text{ from } \Delta h_2) + (-0.5 \text{ from } \Delta h_1)]) \]

\[ = (2 \text{ from } \Delta i_1) + (-1^*[(1.5 \text{ from } \Delta h_2) + (-0.5 \text{ from } \Delta h_1)]) \]

\[ = (2 \text{ from } \Delta i_1) + (-1^*[(1.5 \text{ from } \Delta i_1) + (-0.5 \text{ from } \Delta i_2)]) \]

\[ = (0.5 \text{ from } \Delta i_2) + (0.5 \text{ from } \Delta i_2) \]
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

(a)  
- Green dots: Only TAL motifs were embedded in the full region
- Blue dots: Only GATA motifs were embedded in the full region
- Red dots: Both kinds of motifs were embedded in the full region
- Black dots: No motifs were embedded in the full region
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=PLJLjQOkqSRTP3cLB2c00i_bQFw6KPGKML
- Slides at: https://drive.google.com/file/d/0B15FQN41VQXbkVkCTVQJTVQNE/view

Future Direction
- Applying DeepLIFT to RNNs
- Compute ‘reference’ empirically from data