# Axiomatic Attribution for Deep Networks

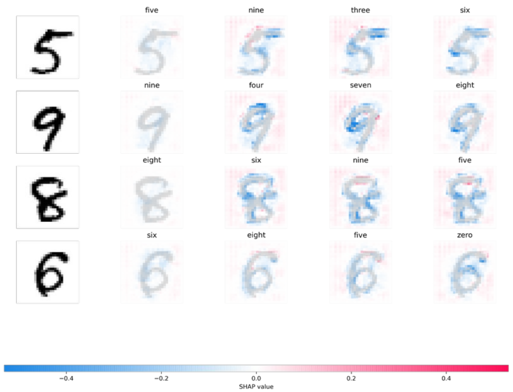
Presented by Alex Lin, Steve Li, Kevin Huang

Mukund Sundararajan * 1   Ankur Taly * 1   Qiqi Yan * 1

# Motivation and Problem Statement

- Feature Attribution:

  **Definition 1.** *Formally, suppose we have a function* $F : \mathrm{R}^n \rightarrow [0, 1]$ *that represents a deep network, and an input* $x = (x_1, \ldots, x_n) \in \mathrm{R}^n$. *An attribution of the prediction at input* $x$ *relative to a baseline input* $x'$ *is a vector* $A_F(x, x') = (a_1, \ldots, a_n) \in \mathrm{R}^n$ *where* $a_i$ *is the* contribution *of* $x_i$ *to the prediction* $F(x)$.




- Examples: in a CNN an attribution method could reveal which pixels were responsible for a certain label being picked (we saw this with LIME/SHAP)
- Problem: attribution technique are hard to evaluate empirically - hard to separate errors from model vs errors from attribution method
  - Ex. Gradients
  - Baseline: black image, empty text, etc.

# Summary of Contributions

- Present two axioms: **Sensitivity** and **Implementation Invariance**
  - **Sensitivity:** For every input and baseline that differ in one feature but have different predictions then the differing feature should be given a non-zero attribution.
  - **Implementation Invariance:** The attributions are always identical for two functionally equivalent networks.
- 2 axioms → **integrated gradients**
  - Overview: path integral of the gradients along the straight line path from an input x to a baseline input x’

# Two Axioms (Desiderata)

## Sensitivity (a)

**Definition:** When 2 inputs that differ in only one feature result in different predictions, the **differing feature** should be given a **non-zero attribution**.

## Invariance

**Definition:** The attributions are always identical for two functionally equivalent networks.

$$\frac{\partial f}{\partial g} = \frac{\partial f}{\partial h} \cdot \frac{\partial h}{\partial g}$$

# Other Attribution Methods

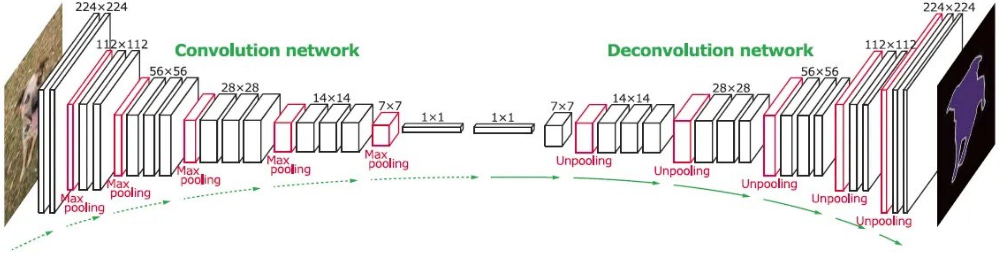
**Gradients (of the output with respect to the input)**

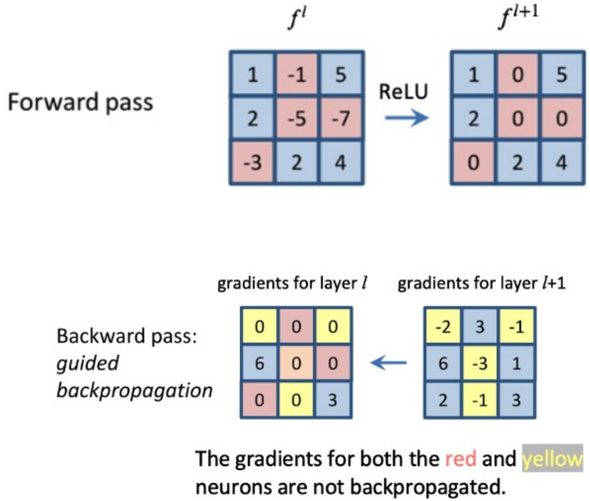
- Breaks sensitivity - prediction function can flatten at the input, giving 0 gradient despite function value at the input being different from the baseline
- Example:
  - Single ReLU network: $f(x) = 1 - \text{ReLU}(1 - x)$
    - Baseline: $x = 0$, input: $x = 2$
    - $f(0) = 0$, $f(2) = 1$
    - Since $f$ is flat at $x = 1$, gradient gives attribution of 0 to $x$

# Other Attribution Methods

## Methods that Break Sensitivity

- DeConvNets, Guided back-propagation



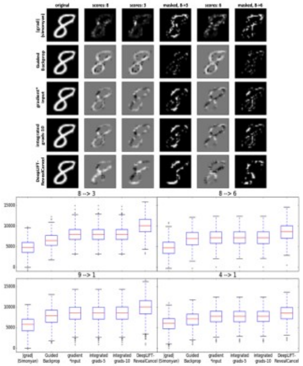



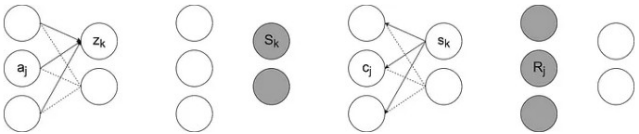

- Only back-prop through a ReLU if the ReLU is turned on at the input
    - Attribution is 0 for features with 0 gradients, despite non-zero gradient at the baseline

# Other Attribution Methods

## Methods that Break Implementation Invariance

- DeepLift and Layer-wise relevance propagation (LRP)






| | |
|---|---|
| $\forall_k:\ z_k = \epsilon + \sum_{0,j} a_j \cdot \rho(w_{jk})$ | (forward pass) |
| $\forall_k:\ s_k = R_k / z_k$ | (element-wise division) |
| $\forall_j:\ c_j = \sum_k \rho(w_{jk}) \cdot s_k$ | (backward pass) |
| $\forall_j:\ R_j = a_j\, c_j$ | (element-wise product) |

- Replace gradients with discrete gradients, use a modified form of backpropagation
- Chain rule doesn’t hold for discrete gradients (calculating gradients would be different) → breaks implementation invariance

# The Method

# Integrated Gradients

## Definition

The **path integral** of the gradients along the **straight-line path** from the baseline $x'$ to the input $x$.

$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

## New Axiom

**Completeness:** The sum of the attributions is equal to the difference of the outputs.

**Proposition 1.** *If* $F : \mathsf{R}^n \rightarrow \mathsf{R}$ *is differentiable almost everywhere* [1] *then*

$$\Sigma_{i=1}^{n} \text{IntegratedGrads}_i(x) = F(x) - F(x')$$

# Uniqueness of Integrated Gradients

## Path Methods

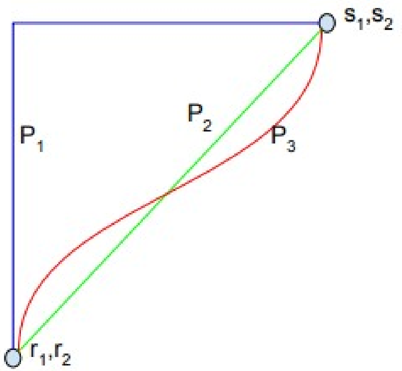



*Figure 1.* Three paths between an a baseline $(r_1, r_2)$ and an input $(s_1, s_2)$. Each path corresponds to a different attribution method. The path $P_2$ corresponds to the path used by integrated gradients.

$$\text{PathIntegratedGrads}_i^{\gamma}(x) ::= \int_{\alpha=0}^{1} \frac{\partial F(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \frac{\partial \gamma_i(\alpha)}{\partial \alpha} \, d\alpha$$

## Axioms

- **Sensitivity (b):** If the function does not depend (mathematically) on some input, then the attribution for that input is always zero.
- **Linearity:** Attributions preserve any linearity within the network.

$$a \times f_1 + b \times f_2$$

- **Symmetry-Preserving:** For symmetric variables, if they have identical values in the input and identical values in the baseline, they then receive identical attributions.

Si $F(x, y) = F(y, x).$ )

# Using Integrated Gradients

## Selecting a Baseline

**Two Components:**

- Zero-Score

$$F(x') \approx 0$$

- Conveys Absence of Signal

**Examples:**

- Object Recognition: All-black image
- Text: All-zero input embedding vector

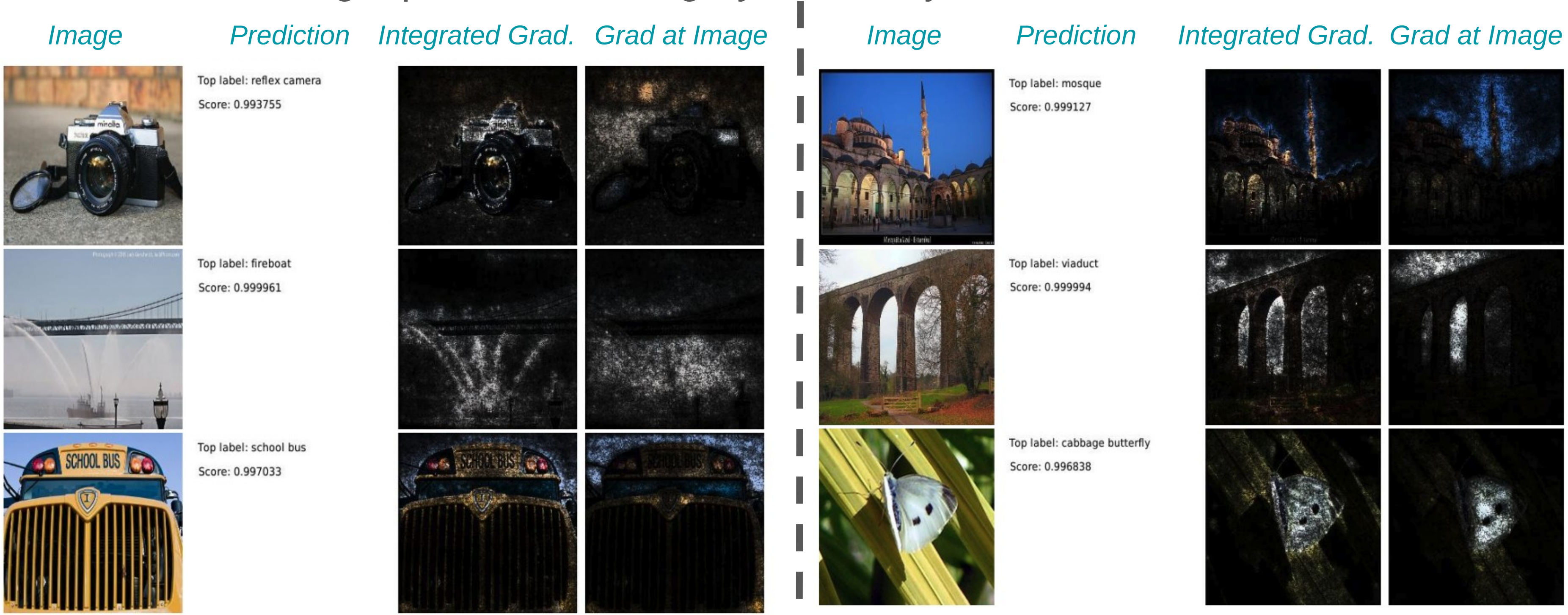
## Computing IGs

$$\mathsf{IntegratedGrads}_i^{approx}(x) ::= (x_i - x_i') \times \Sigma_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m}$$

# Experimental Results

# Object Recognition CNN

**Task:** Given image, predict the category of the object



# Question Classification CNN

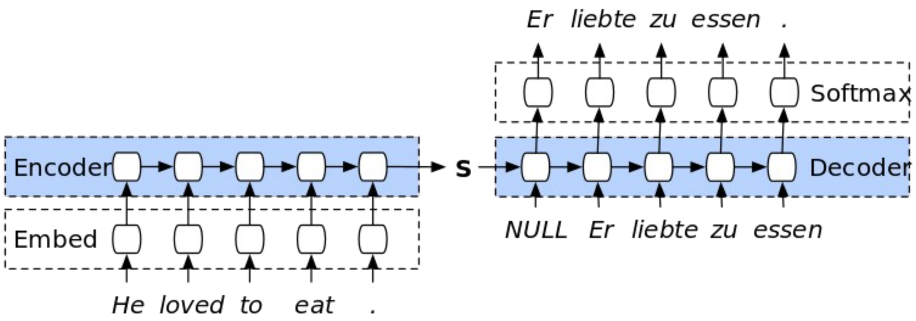
**Task**: Given question, predict what type of answer it is looking for.

how many townships have a population above 50 ? [prediction: NUMERIC]
what is the difference in population between fora and masilo [prediction: NUMERIC]
how many athletes are not ranked ? [prediction: NUMERIC]
what is the total number of points scored ? [prediction: NUMERIC]
which film was before the audacity of democracy ? [prediction: STRING]
which year did she work on the most films ? [prediction: DATETIME]
what year was the last school established ? [prediction: DATETIME]
when did ed sheeran get his first number one of the year ? [prediction: DATETIME]
did charles oakley play more minutes than robert parish ? [prediction: YESNO]

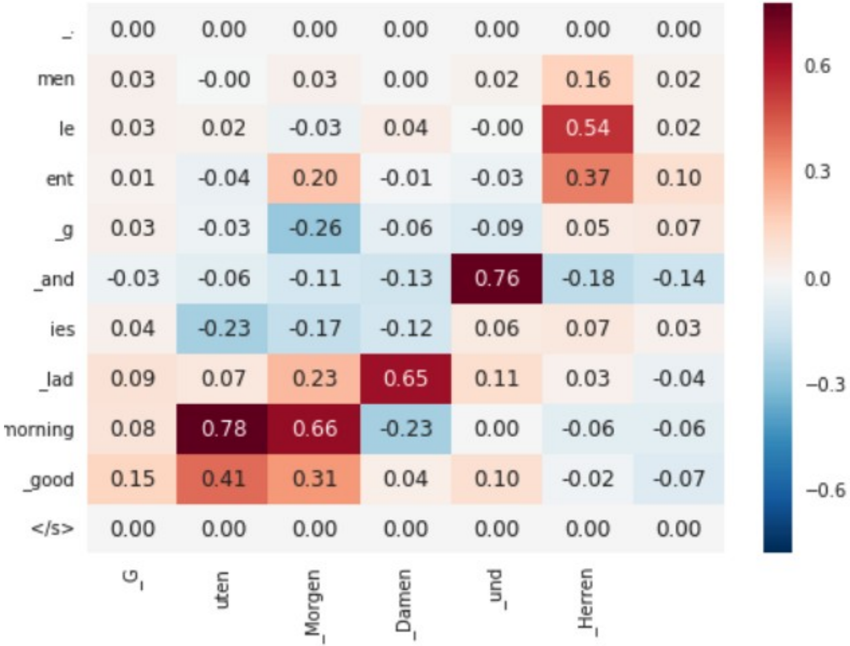
**red = positive attribution, blue = negative attribution, gray = neutral attribution**

# Machine Translation RNN

**Task:** Given English sentence, predict German translation

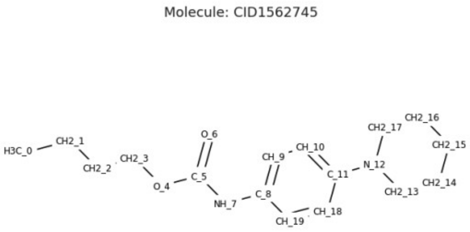



*Example RNN Architecture*



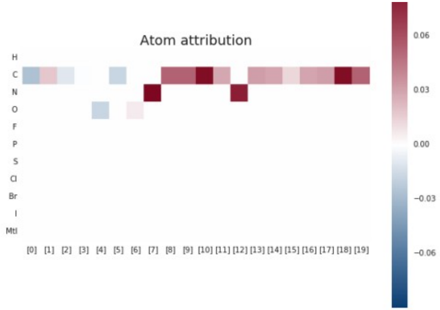
# Ligand Screening Graph CNN

**Task**: Given molecular graph, predict whether it is active against an enzyme



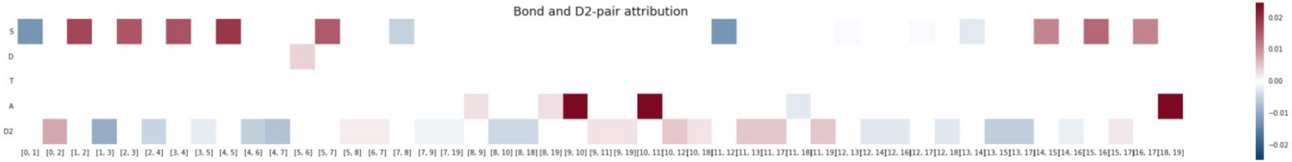

**Attribution summary**

Softmax score for task PCBA-588342: 0.98
Atom attribution: 0.62 (63%)
Bond attribution: 0.45 (46%)
D2-pair attribution: -0.03 (-3%)




Take-aways:

- More attribution to atom-pairs with bond (46%) compared to without bond (-3%)
- Attribution can help identify degenerate features (e.g. indicate that features are not fully convolved) (?)



# Conclusion and Discussion

Summary

- Formalizes two axioms for attribution: **sensitivity, implementation invariance**
- Propose **integrated gradients** and argue that it is theoretically superior to other gradient-based methods (e.g. DeepLift, LRP, guided backprop, etc.)
- Perform experiments across several domains to showcase method

Discussion Questions

- Are you convinced that these axioms are desirable?
- Do you see any strengths or weaknesses in the idea of producing explanations through an integrated path?
- Have the experiments convinced you of the superiority of their method?