CS 329T: Trustworthy Machine Learning

Lab 3

Outline

- HW 2 overview
- Final report brief guidelines
- Saliency Maps
- Integrated gradients
- Saliency Maps and IG colab

HW 2

- In HW 2, we would be focusing on two main modules: **Explainability of traditional ML models** and **Attribution in Vision Models**
- **Explainability of traditional ML models:**
  - Model-agnostic explanations (LIME, SHAP)
  - Model-specific explanations (TreeSHAP)
- **Attribution in Vision Models:**
  - Gradient-based attribution methods (Saliency Map, IG, Influence-directed explanations)
  - Attribution method evaluation (Visual Comparison, Average Drop %)

Final report brief guidelines

- We would be releasing formal project guidelines next week!
- As discussed in the last lab, we want you to work on the project in parallel with the modules covered
- As a first step, we want to hear what are your initial project ideas!
- Submit a small paragraph detailing your project ideas next week!

Saliency Maps

- Gradient-based attribution method
- Compute local gradient of pre-softmax scores w.r.t the input
- Feature * gradient: Attribution for feature x_i is $x_i * \partial y / \partial x_i$

Definition 1 (Saliency Map). *Consider a model $y = f(x)$ that takes an input $x \in \mathrm{R}^d$ and outputs y, a distribution of scores for each class. We denote the $y_c = f_c(x)$ as the scores for the class c. The Saliency Map $S_c(x)$ for class c is defined as*

$$S_c(x) = x \odot \nabla_x f_c(x) \tag{1}$$

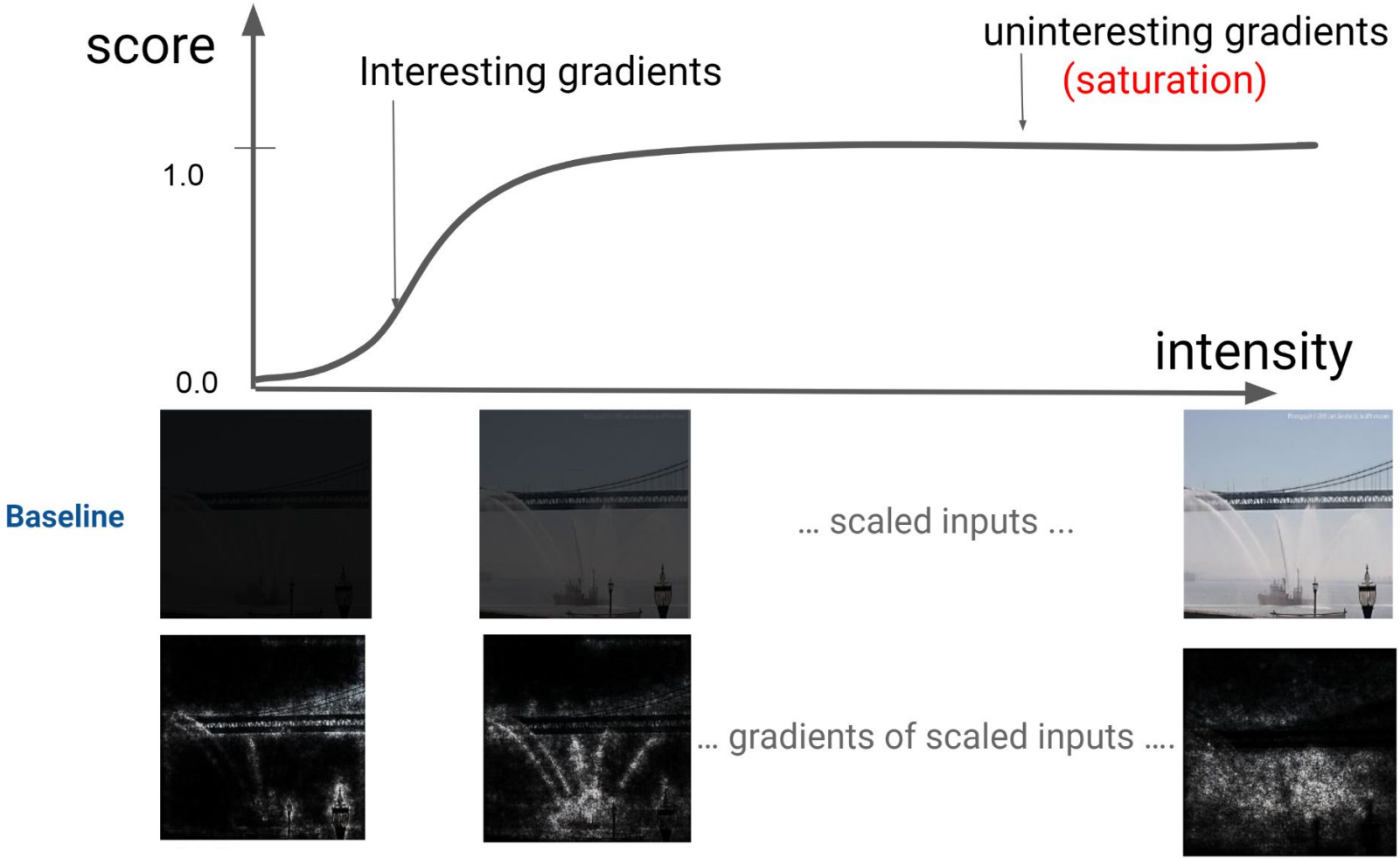
where $\odot$ denotes the element-wise multiplication.

Saliency Maps and IG Colab

Colab Link
Solution

Integrated Gradients

- Integrated Gradient aims to solve the vanishing gradient problem in Saliency Map, while it satisfies several desirable axioms e.g. insensitivity, linearity preservation, completeness, symmetry.
- The contribution of feature x_i is based on how far it is from a baseline
- Integrate gradients along a straight line path from the baseline to the input


score
Interesting gradients
uninteresting gradients
(saturation)
1.0
0.0
intensity
Baseline
... scaled inputs ...
Input
... gradients of scaled inputs ....

Definition 2 (Integrated Gradient). *Consider a model $y = f(x)$ that takes an input $x \in \mathrm{R}^d$ and outputs y, a distribution of scores for each class. We denote the $y_c = f_c(x)$ as the scores for the class c and x_b as the baseline input. The Integrated Gradient $IG_c(x, x_b)$ for class c is defined as*

$$IG_c(x, x_b) = (x - x_b) \odot \int_0^1 \nabla_x f_c(x_b + t(x - x_b))dt \tag{2}$$

where $\odot$ denotes the element-wise multiplication. In the implementation, we use the following equation to approximate Eq. 2.

$$IG_c(x, x_b) \approx (x - x_b) \odot \frac{1}{N} \sum_{i=1}^{N} \nabla_x f_c \left(x_b + (x - x_b) \frac{i}{N} \right) \tag{3}$$

where N is the number of steps used for the approximation.

Saliency Maps and IG Colab

Colab Link
Solution