AMMI_Paper_Implemntaion-1

Axiomatic Deep Nets

By: Amna Elmustafa

01

Core Idea

Why?

Deep Nets were black-box

Understand input-output behaviour (which feature to edit?)

Decision making , Fairness , privacy ,Causality

Help debugging

Requirements /Axioms

Sensitivity

Implementation Invariance

Completeness the attributions add up to the difference between the output of input x vs baseline

Linearity

And More....

Methodology

- Gradients (of the output with respect to the input) is a natural analog of the model coefficients for a deep network.
- Gradients are insensitive , RELU example .
- Accumulating gradient rather than local gradient.

What is path method

1. Interpolation between the baseline and the output in a straight line

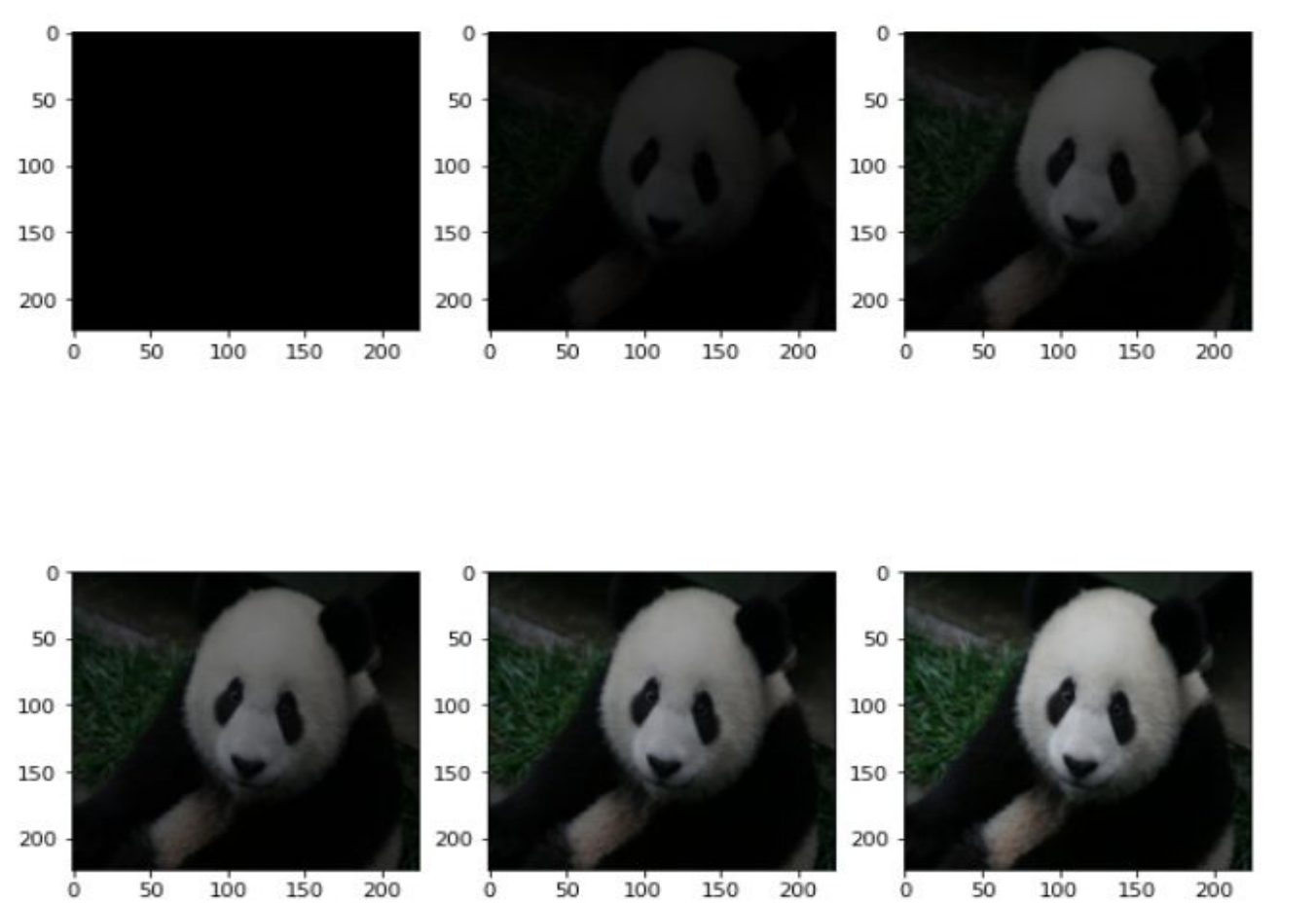
$$x := x' + \alpha(x - x')$$

α - interpolation constant to perturbed features by

x - input image tensor

x'- baseline image tensor

What is path method



What is path method

2. Integrated Gradient(approximated with integral with Riemann Trapezoid)

$$IntegratedGrads_i^{approx}(x) ::= (x_i - x_i') \times \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m}$$

where:

i = feature (individual pixel)

x = input (image tensor)

x' = baseline (image tensor)

k = scaled feature perturbation constant

m = number of steps in the Riemann sum approximation of the integral. This is covered in depth in the section *Compute integral approximation* below.

02

Approach

Steps

1. Find a baseline eg.black image
2. Generate alphas
3. Generate interpolated path inputs

Steps

4 Compute gradients between model output predictions with respect to input features

5 Integral approximation through

Averaging

6 Scale integrated gradients with respect to original image = (xi−x'i)×average gradients

Experiment details

Images data

Any random Image

Pre-trained googlenet

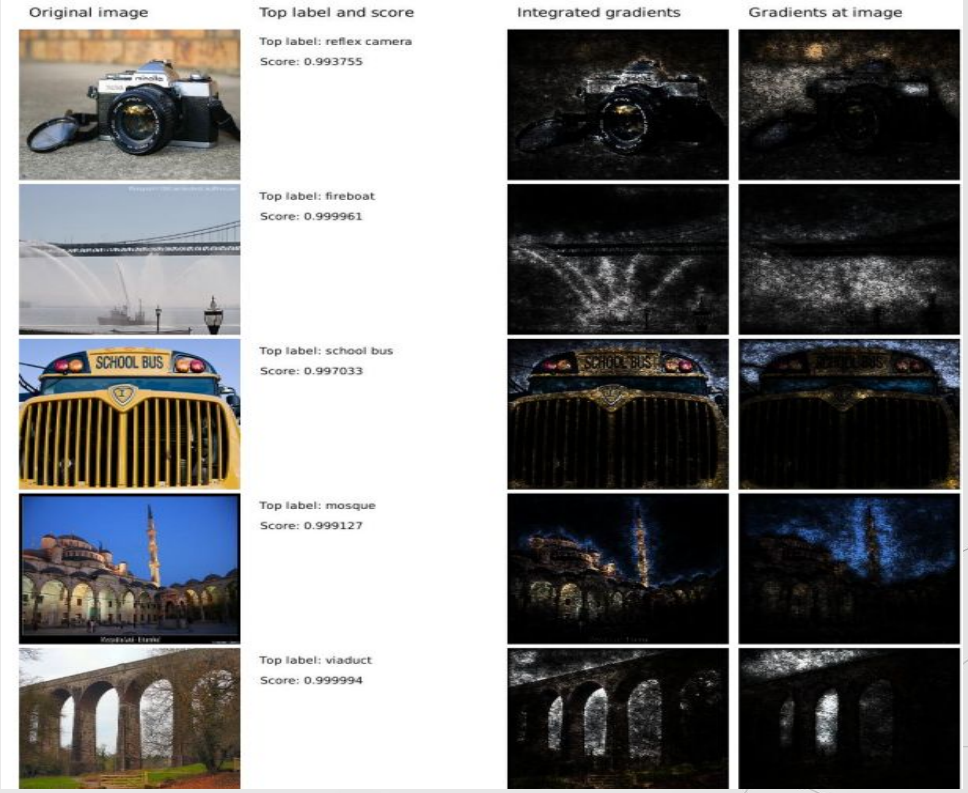
Number of steps 100-1000

03

Results

What is done so far?

Original Paper Results /Images



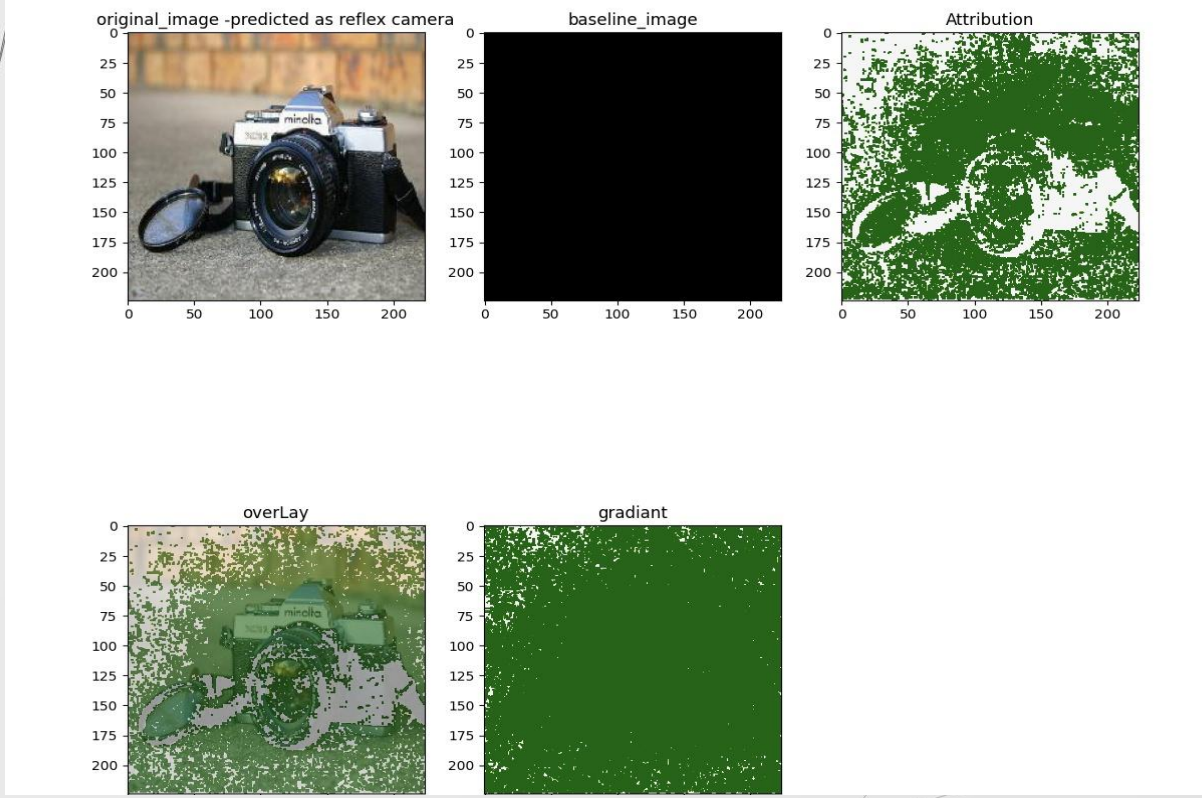
Original Paper Results /Text-QA classification

when did ed sheeran get his first number one of the year ? [prediction: DATETIME]

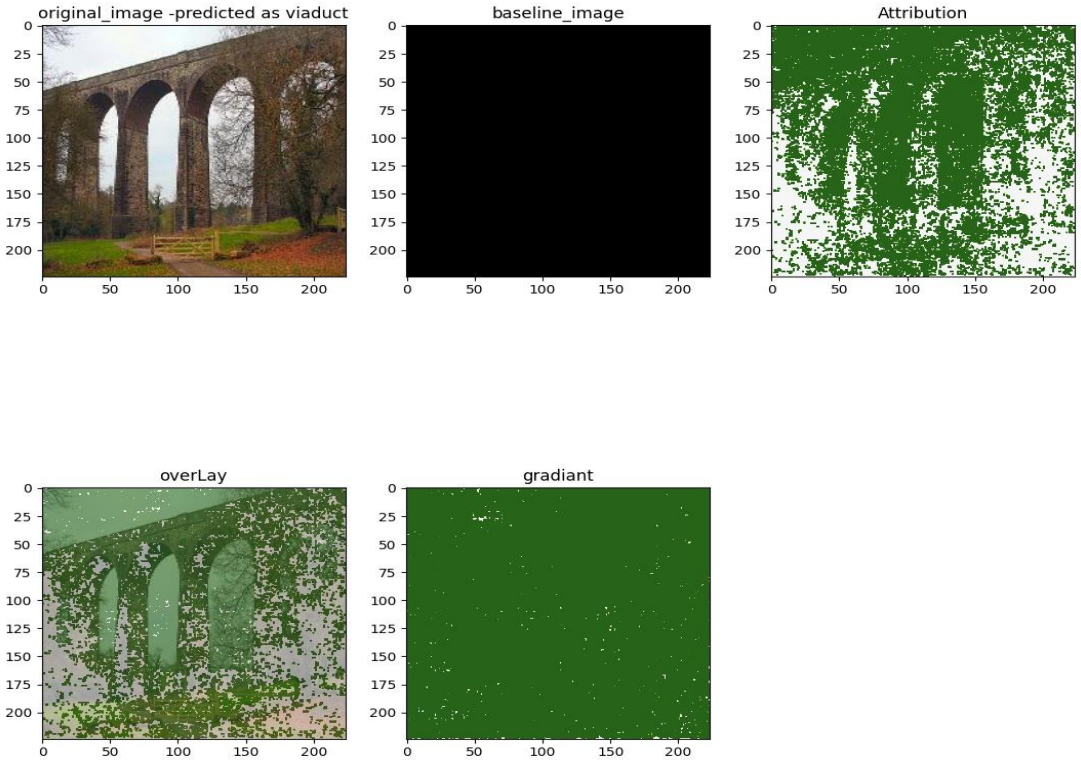
did charles oakley play more minutes than robert parish ? [prediction: YESNO]

Red is positive, Blue is negative, and Gray is neutral (zero). The predicted class is specified in square brackets.

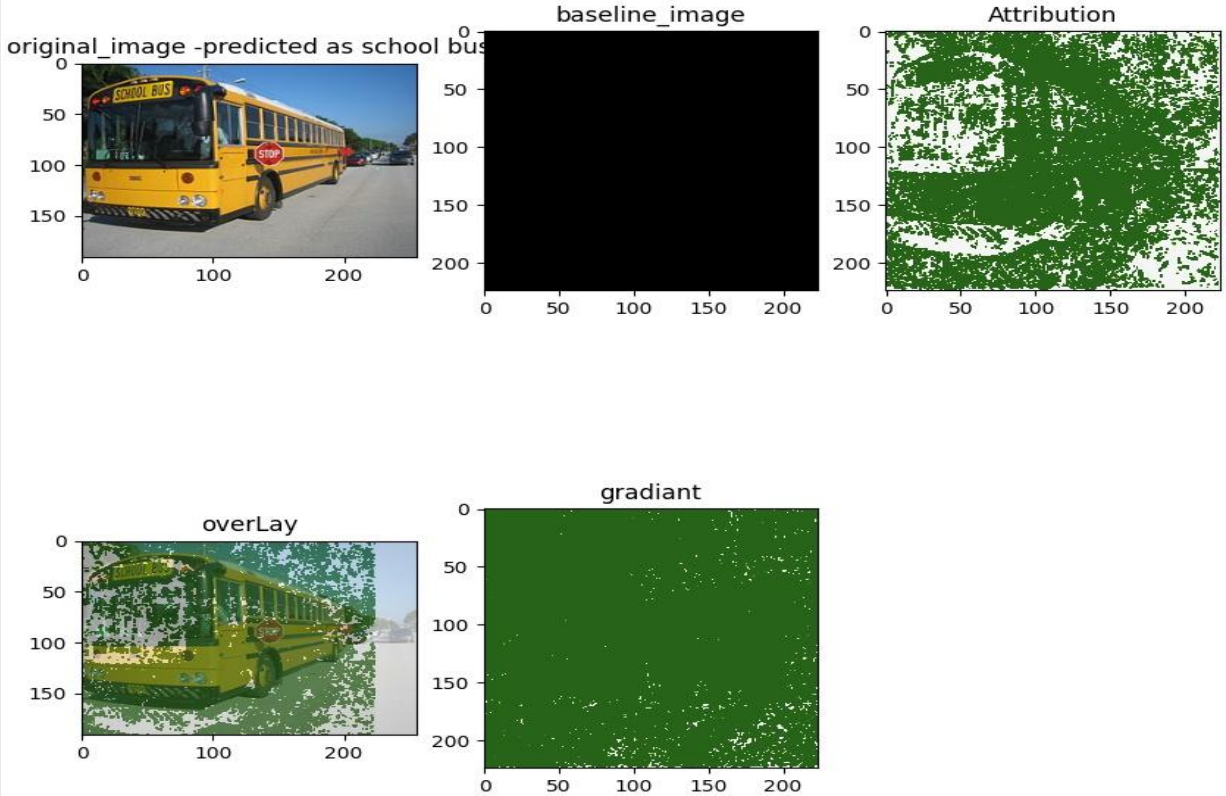
Our Results-gradients vs Integrated gradient



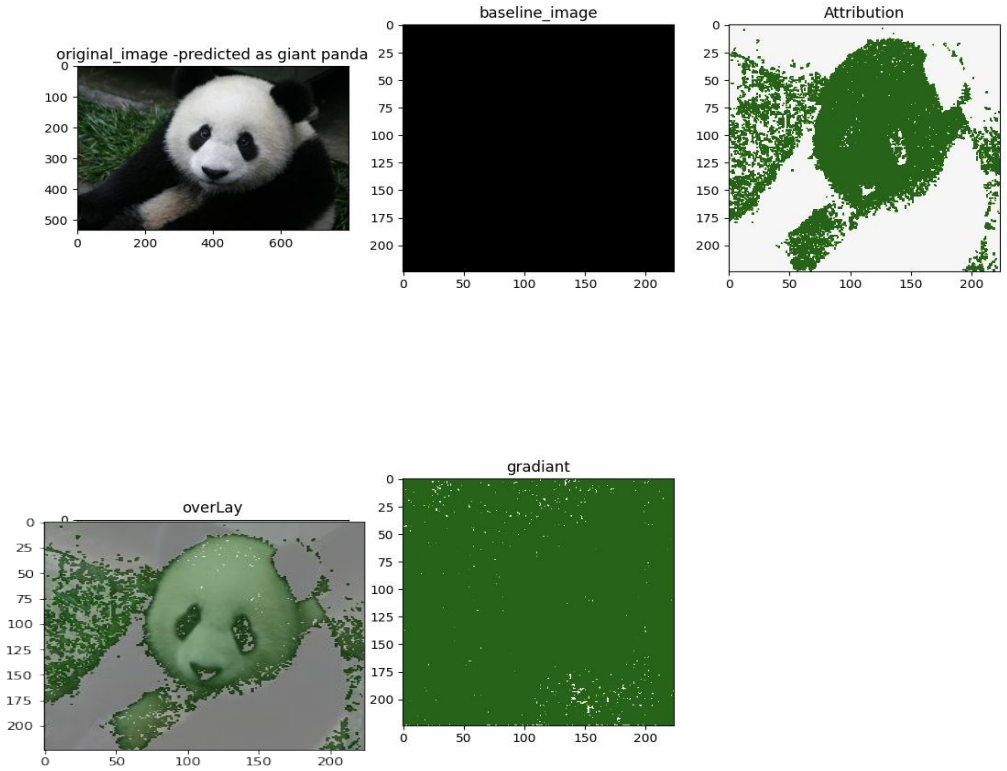
Our Results-gradients vs Integrated gradient



Our Results-gradients vs Integrated gradient



Our Results-gradients vs Integrated gradient



Our Results-Text-Sentiment Analysis (pos/neg)

Integrated Gradient

This is a terrible movie

i watched this movie on theater and i did not like it

This is a good movie

I liked this movie

I hated this movie

Gradient

i watched this movie on theater and i did like it

Legend: Negative Neutral Positive

The Same RESULTS...

just Visualization Tricks

04

Challenges

How to verify ?

Interaction between features ?

05

Future Work

More Experiments!

- Experiment with graphs models
- Experiment with text NMT
- Experiment the new version of the paper 2020

Questions?