Axiomatic Attribution for Deep Networks

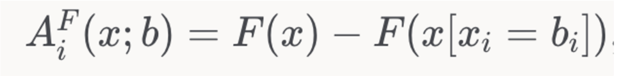
12 Feb 2020

Presenter: Sanchit Sinha

https://qdata.github.io/deep2Read/

Background - Attribution

- **Attribution:**
  - Let F(x) represent output of a DNN on input x (x1,x2,...xn) [n-dimensional]
  - Attribution wrt a baseline A (x,x_) is a vector of same dimension of the input [n-dimensional]
  - Intuitively, the weight(importance) each feature has in making the prediction



$$A_i^F(x;b) = F(x) - F(x[x_i = b_i])$$

Background - Axioms

- **Sensitivity(a)**: If for every input and baseline that differ in one feature and have different predictions, then the differing feature should be given a non-zero attribution score.
- **Sensitivity(b)**: If the function implemented by the DNN does not depend (mathematically) on some variable, then the attribution to that variable is always zero.
- **Implementation Invariance**: the attributions are always identical for two functionally equivalent networks. It should not depend on the implementation of the network

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

- **Linearity**: Linear composition of two deep networks modeled by the functions f1 and f2 to form a third network that models the function a×f1+b×f2, then the attributions for a × f1 + b × f2 to be the weighted sum of the attributions for f1 and f2 with weights a and b respectively.

Background - Previous Methods weakness

- **Violating Sensitivity:** “Gradients*inputs (element-wise)” - standard debugging practice (at time of writing the paper) violate sensitivity.
  - Why?- Assume F(x) = 1 – ReLU(1–x). Taking baseline, x=0. Gradient of F(x) for x>1 becomes 0. So lets say at x= 2, even though function is changing from 0 to 1, the attribution score is 0.
  - Intuitively, the prediction function may flatten at the input and thus have zero gradient despite the function value at the input being different from that at the baseline.
  - DeConvNets and Guided backpropagation also violate because these methods back-propogate through a ReLU node only if the ReLU is turned on at the input

- **Violating Implementation Invariance:** DeepLift and LRP violate this. The outputs of the methods should not depend on the model if the inputs and outputs are same.
  - The attributions are sensitive to unimportant aspects of the models. For instance, if the network architecture has more degrees of freedom than needed to represent a function then there may be two sets of values for the network parameters that lead to the same function

Motivation

- Evaluation of attribution methods is challenging.
- It is hard to explain errors that are due to misbehavior of the model or the misbehavior of the attribution method.
- Hence, axiomatic approach is proposed: any method which satisfies the axioms should be desirable (at least on paper)

- Designing a method which satisfies all the axioms proposed. They call it Integrated Gradients

- Showing that previous methods don’t satisfy the axioms proposed and hence lack generalization and performance

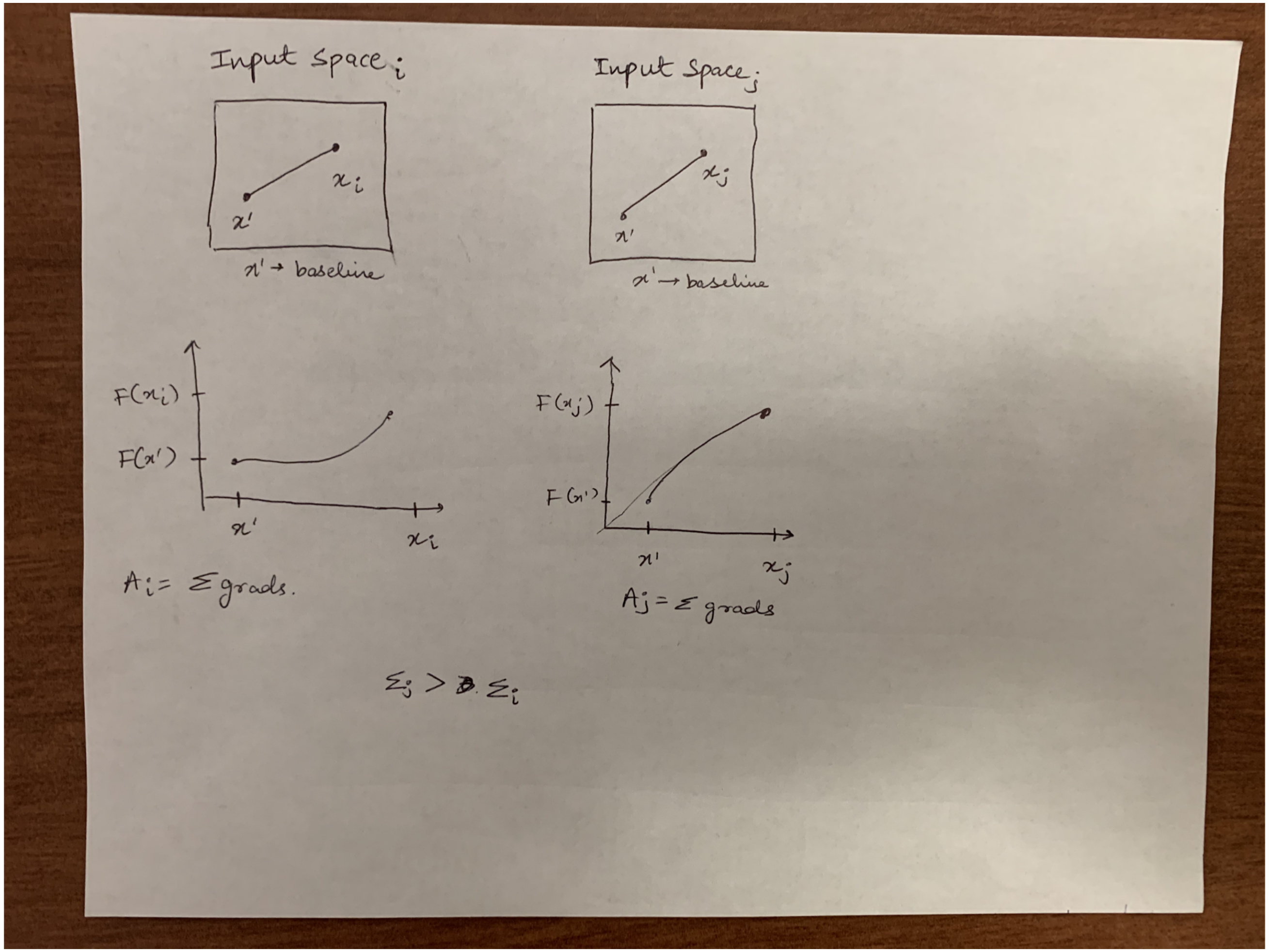
Related Work

- Layer Wise Relevance Propagation - Binder, Alexander, Montavon, Gregoire, Bach, Sebastian, ´ Muller, Klaus-Robert, and Samek, Wojciech. Layer- ¨ wise relevance propagation for neural networks with local renormalization layers. CoRR, 2016
- DeepLift - Shrikumar, Avanti, Greenside, Peyton, and Kundaje, Anshul. Learning important features through propagating activation differences. CoRR, abs/1704.02685, 2017. URL http://arxiv.org/abs/1704.02685.
- Deconvolutions - Zeiler, Matthew D. and Fergus, Rob. Visualizing and understanding convolutional networks. In ECCV, pp. 818– 833, 2014
- Guided Backprop - Springenberg, Jost Tobias, Dosovitskiy, Alexey, Brox, Thomas, and Riedmiller, Martin A. Striving for simplicity: The all convolutional net. CoRR, 2014.

Claim / Target Task

1. Present 2 axioms which every attribution method should satisfy to be considered good enough
2. Present a novel method which satisfies all axioms called Integrated Gradients
3. Combine the best of both worlds - Implementation Invariance of Gradients along with the Sensitivity of techniques like LRP
   or DeepLift.
1. Demonstrate results on both vision and textual problems

An Intuitive Figure Showing WHY Claim



Proposed Solution

- Integrated Gradients: Formulated for every dimension i as

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

- Basically, it says that we should sum up all gradients along the straight line from baseline to input. Once we do this, we get:

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

- Why Straight Line? - Maximum gradients will be accumulated by moving on a straight line as opposed to a curvy way

Implementation

- The integral can be well approximated using summation over many steps. Practical usage can use about 20-300 steps

$$\mathsf{IntegratedGrads}_i^{approx}(x) ::= \\ (x_i - x'_i) \times \Sigma_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x')))}{\partial x_i} \times \frac{1}{m} \qquad (3)$$

- Baseline selection for images: Black images (0 valued pixel values)
- Baseline for text: (0 valued vectors)
- Once the gradients are found, they are multiplied with the image to get the final explainable maps

Data Summary

- Imagenet and GoogleNet
- Diabetic Retinopathy Prediction - (Gulshan et al., 2016)
- WikiTableQuestions dataset - (Kim, 2014)
- Chemistry W2n2 dataset

Experimental Results - Images

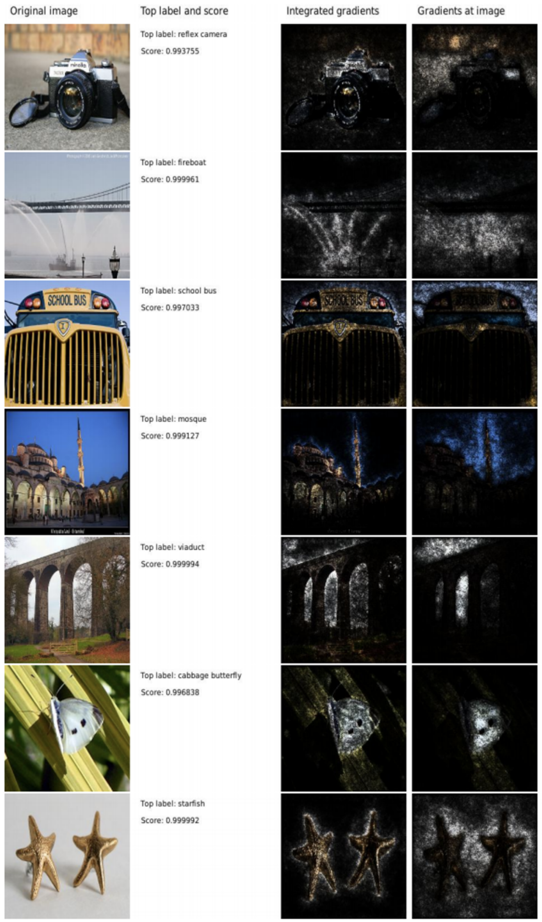


Figure 2. **Comparing integrated gradients with gradients at the image.** Left-to-right: original input image, label and softmax score for the highest scoring class, visualization of integrated gradients, visualization of gradients*image. Notice that the visualizations obtained from integrated gradients are better at reflecting distinctive features of the image.

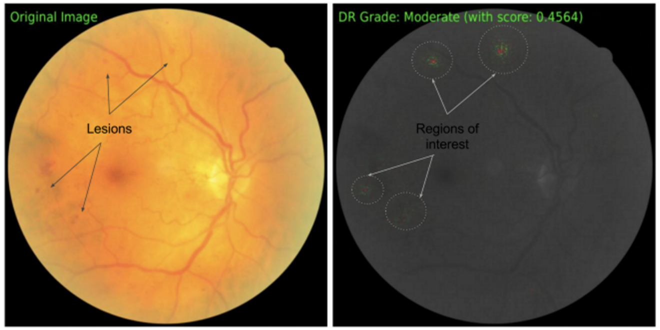
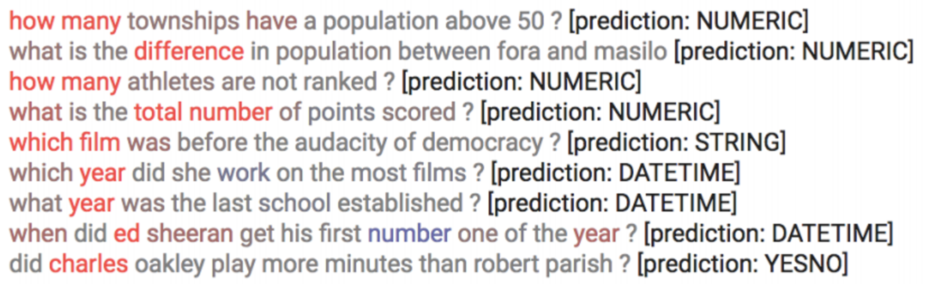


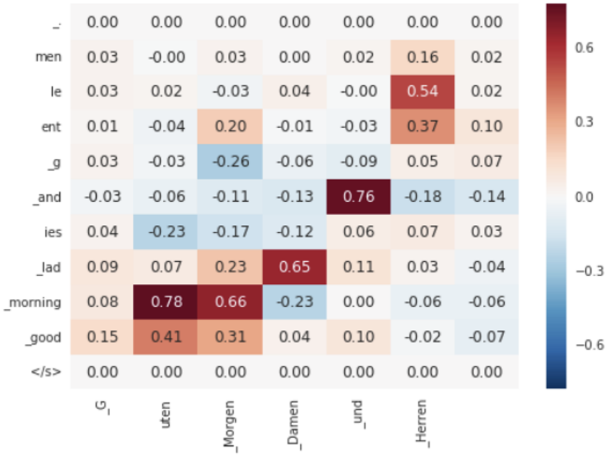
Figure 3. **Attribution for Diabetic Retinopathy grade prediction from a retinal fundus image.** The original image is show on the left, and the attributions (overlayed on the original image in gray scaee) is shown on the right. On the original image we annotate lesions visible to a human, and confirm that the attributions indeed point to them.

Experimental Results - Text



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]

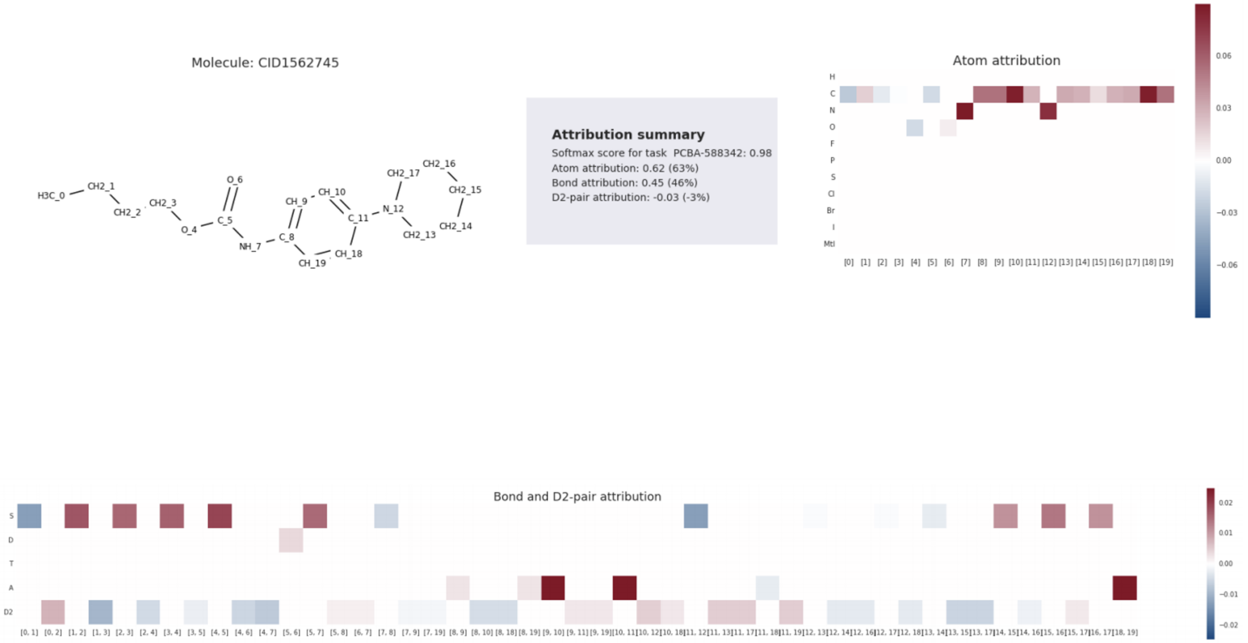
Figure 4. **Attributions from question classification model.** Term color indicates attribution strength—Red is positive, Blue is negative, and Gray is neutral (zero). The predicted class is specified in square brackets.



	_G	uten	_Morgen	_Damen	_und	_Herren	
_.	0.00	0.00	0.00	0.00	0.00	0.00	0.00
men	0.03	-0.00	0.03	0.00	0.02	0.16	0.02
le	0.03	0.02	-0.03	0.04	-0.00	0.54	0.02
ent	0.01	-0.04	0.20	-0.01	-0.03	0.37	0.10
_g	0.03	-0.03	-0.26	-0.06	-0.09	0.05	0.07
_and	-0.03	-0.06	-0.11	-0.13	0.76	-0.18	-0.14
ies	0.04	-0.23	-0.17	-0.12	0.06	0.07	0.03
_lad	0.09	0.07	0.23	0.65	0.11	0.03	-0.04
_morning	0.08	0.78	0.66	-0.23	0.00	-0.06	-0.06
_good	0.15	0.41	0.31	0.04	0.10	-0.02	-0.07
</s>	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 5. **Attributions from a language translation model.** Input in English: “good morning ladies and gentlemen”. Output in German: “Guten Morgen Damen und Herren”. Both input and output are tokenized into word pieces, where a word piece prefixed by underscore indicates that it should be the prefix of a word.

Experimental Results - Chemistry



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%)

Figure 6. **Attribution for a molecule under the W2N2 network (Kearnes et al., 2016).** The molecules is active on task PCBA-58432.

Conclusion/ Future Work

- Introduced Integrated gradients:
  - Maintains Sensitivity (Completeness)
  - Implementation Invariant
  - Has Linearity
- Proposed axioms which are universal and baseline for further methods

References

- Layer Wise Relevance Propagation - Binder, Alexander, Montavon, Gregoire, Bach, Sebastian, ´ Muller, Klaus-Robert, and Samek, Wojciech. Layer- ¨ wise relevance propagation for neural networks with local renormalization layers. CoRR, 2016
- DeepLift - Shrikumar, Avanti, Greenside, Peyton, and Kundaje, Anshul. Learning important features through propagating activation differences. CoRR, abs/1704.02685, 2017. URL http://arxiv.org/abs/1704.02685.
- Deconvolutions - Zeiler, Matthew D. and Fergus, Rob. Visualizing and understanding convolutional networks. In ECCV, pp. 818– 833, 2014
- Guided Backprop - Springenberg, Jost Tobias, Dosovitskiy, Alexey, Brox, Thomas, and Riedmiller, Martin A. Striving for simplicity: The all convolutional net. CoRR, 2014.

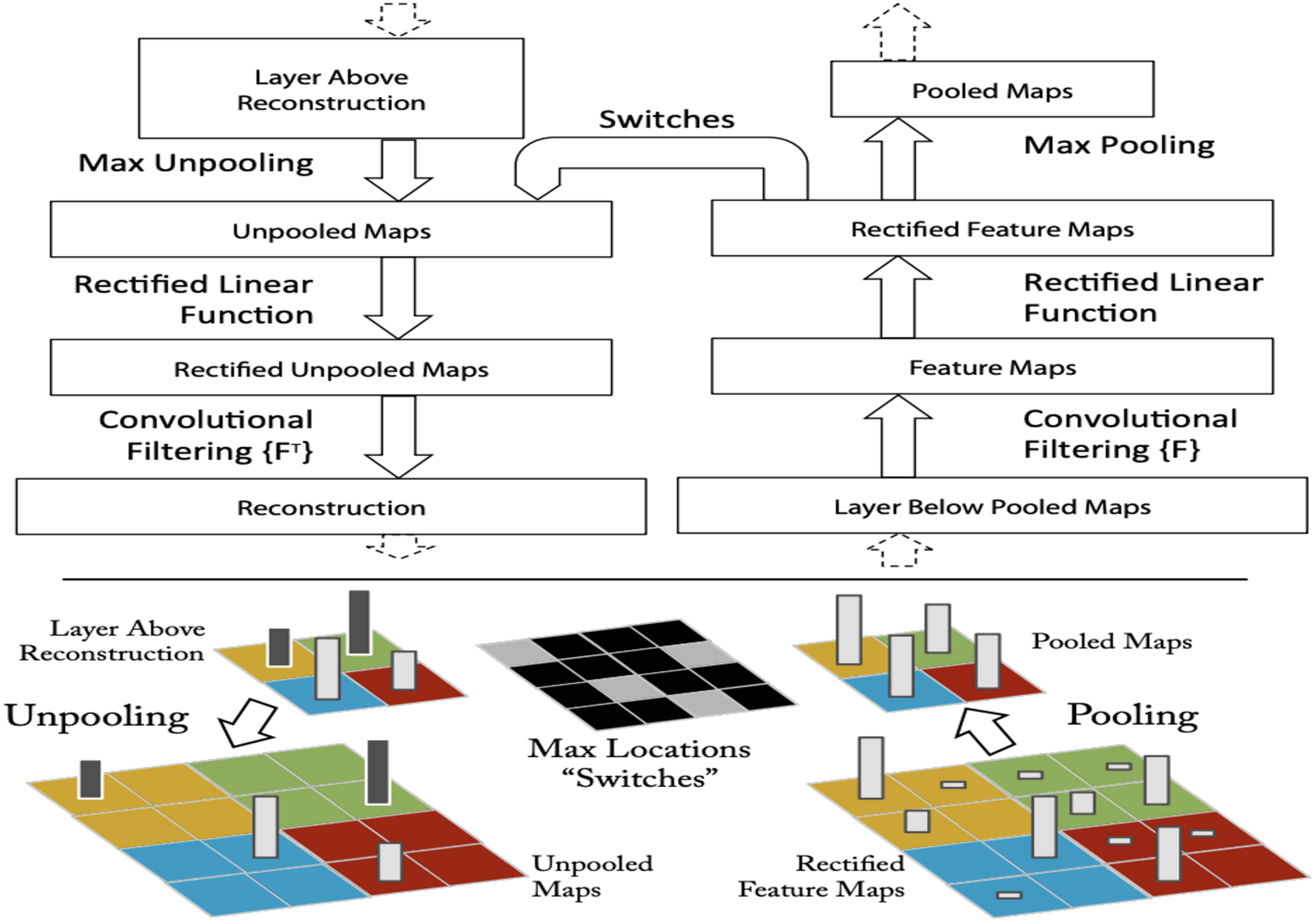


Figure 1. Top: A deconvnet layer (left) attached to a convnet layer (right). The deconvnet will reconstruct an approximate version of the convnet features from the layer beneath. Bottom: An illustration of the unpooling operation in the deconvnet, using *switches* which record the location of the local max in each pooling region (colored zones) during pooling in the convnet.

Ideas

- Adversarial Training => Better Saliency maps

  Can we use Saliency maps => adversarial robustness
- Quantification metrics for saliency maps
- Modifying GradCAM for ConvGNNs
- Computing Saliency maps for GANs - Making GANs explainable