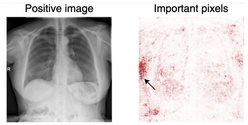
Week 2 - Saliency Maps

Wednesday, January 29, 2025 6:27 PM

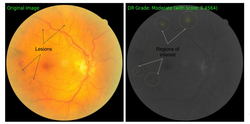
Adventurous Beginnings

1. Saliency maps first became popular in the computer vision community as a way to understand complex image classification methods. They have since been generalized to many data types and are closely related to SHAP.

2. They are a kind of local explanation. They can help us verify that the model is paying attention to the "right" features. They can also draw attention to subtle but predictive features.



(DeGrave, Janizek, S.I. Lee 2021)

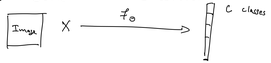


(Sundararajan, Taly, Q. Yan 2017)

In the left figure, we see that a COVID-19 detector has learned a "shortcut" (really, a confounder) based on how radiographs are labeled for COVID + vs. − patients.

In the right, the network has learned to recognize subtle lesion features.

3. Consider a classifier

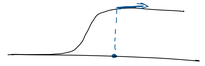


A natural idea (Simonyan et al 2014) is to compute

[Gradients] $\frac{\partial f_{\theta,c}(x)}{\partial x}$.

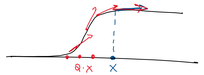
This has the same shape as x and intuitively captures how class c's probability changes in response to perturbations of each pixel locally around x.

4. This has an immediate issue. Neurons can saturate, and the associated gradients can become very small.



5. One idea is to "shrink" the input toward 0 and see the gradients at those intermediate points. This leads to,

[Integrated Gradients] $(x - x_0)\int_{a=0}^{1} \frac{\partial f_{\theta,c}(x_0 + a(x - x_0))}{\partial x}\, da$ [usually $x_0 \equiv 0$].



Check the demo.

https://distill.pub/2020/attribution-baselines/

Reflection + Progress

6. In a lot of early work, validation was often rather informal. Some proved properties of their approaches, some showed simplified experiments, and many used cherry-picked examples. ← eg. the lesion example

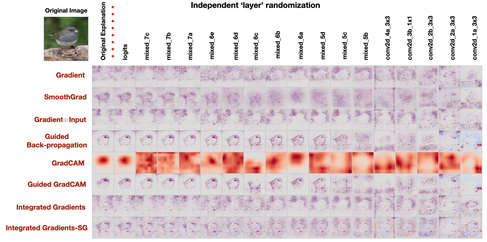
7. Adebayo et al (2018 were some of the first to question the reliance on visual checks, and they offered a concrete, quantitative way forward.

They argued that we could use statistical (!!!) checks to detect methodological failures.

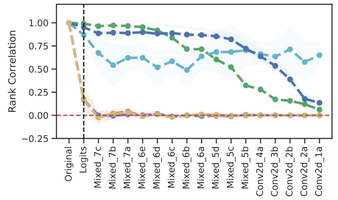
(1) **Model Randomization**: Replace trained weights w/ random initialization and see how the maps change.

(2) **Data Randomization**: Retrain the model on permuted class labels. Does the saliency map on this version of the model look any different?

They use Spearman correlation and a perceptual similarity metric to quantify the similarity between saliency maps.



Rank Correlation No ABS

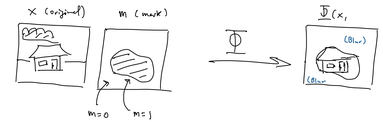


Q: How do these relate to the concept of negative controls in experimental design?

Q: What exactly are the hypotheses being tested?

8. In the spirit of making saliency more quantitative/objective, Fong → Vedaldi (2021) the problem of finding a saliency map as an optimization. Specifically, try to find a minimal transformation that has a large effect on the predicted class.

9.



10. We can find a mask that preserves as much of the original image as possible:

$$\min_{m \in [0,1]^n} \lambda_1 \|1 - m\|_1 + \mathbb{E}_\tau\left[f_c(\Phi(x, m - \tau))\right] + \lambda_2 \sum \|\nabla m(u)\|_1$$

- $\lambda_1 \|1 - m\|_1$: Preserve as much of the original image as you can
- $f_c(\Phi(x, m - \tau))$: Predict on "deleted" image
- τ: Jitter the mask
- $\sum \|\nabla m(u)\|_1$: Prefer simple masks

(This is slightly different than eq. 4, but is the main idea)

11. We've focused on images. But these can be applied generally (see Sundararajan et al). Next week we'll cover SHAP → LIME, which are closely related and designed from the start to be general.