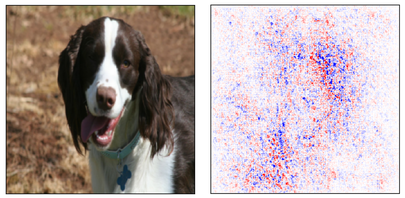
On the Complexity-Faithfulness Trade-off of Gradient-Based Explanations

Amir Mehrpanah Matteo Gamba Kevin Smith Hossein Azizpour

KTH Royal Institute of Technology, SciLifeLab

A Fundamental Problem:

Attribution Complexity:



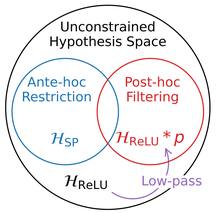
Input

VanillaGrad- ReLU

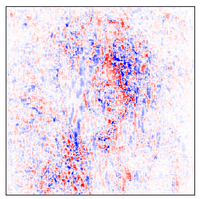
VanillaGrad is usually highly complex and scattered. We quantify this behavior with the high-frequency components of attribution map in spatial domain.

Current Solutions:

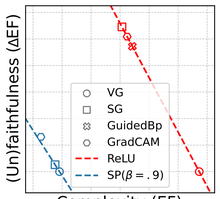
loss of faithfulness in post-hoc methods:



The human tendency for concise explanations, requires simplification. Post-hoc approaches, can be unified under low-pass filtering mechanisms. Filtering the solution space incurs loss of faithfulness.



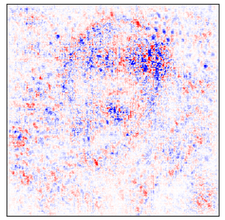
SmoothGrad - ReLU



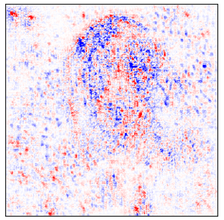
Complexity-Faithfulness Trade-off

Our Analysis Reveals:

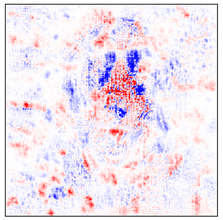
The Network's Structure Is a Primary Source of Attribution Complexity.



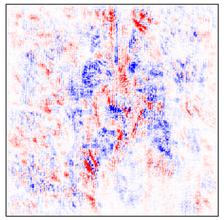
VanillaGrad - SP(β = 10)



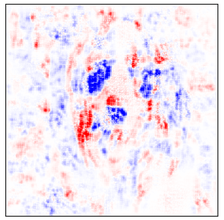
VanillaGrad - SP(β = 7)



VanillaGrad - SP(β = 5)



VanillaGrad - SP(β = 3)

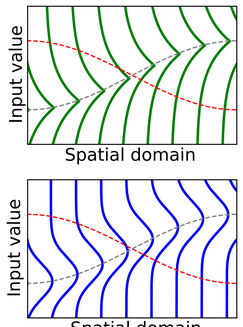


VanillaGrad - SP(β = 0.9)

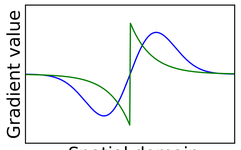
Sharpness arising from activation functions (ReLU) is a major contributor to attribution complexity. We control it by a smooth parameterization (SP) of activation function, leading to an ante-hoc restriction of the hypothesis space, with corresponding accuracy drop.

A Unifying Framework:

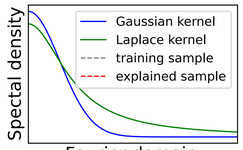
The high-frequency components of the network's input-output function is reflected in the attribution map.



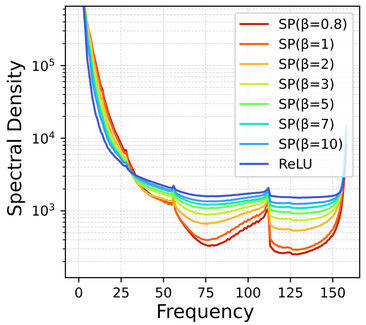
Sharp vs smooth changes



Sharp vs smooth changes in the decision boundaries of the function leads to different forms of gradient.



These difference in gradient shows itself in the high-frequency behavior of the attribution in spatial domain.



The tail (high-frequency) behavior of the spatial Fourier transform of the attribution becomes heavier as β increases.

Ante-hoc restriction of the hypothesis space corresponds to searching over smooth functions only.
Post-hoc filtering of the solution space implicitly creates smooth surrogates.

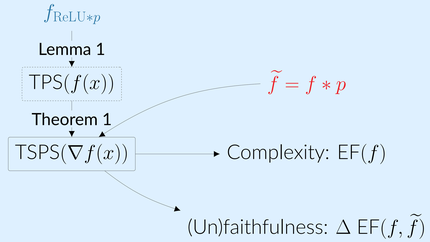
Technical Contributions

- From a spectral view, we **unify post-hoc and ante-hoc explainability** as restriction/filtering of hypothesis/solution space.
- We formalize the **relationship** between the **tail of the power spectrum** of the network and its **activation functions**.
- We connect the **tail of the power spectrum** of the **network** (TPS) to the spectral decay of the **spatial power spectrum** (TSPS) of the **attribution**.
- We **formally define attribution complexity** as the tail behavior of its spatial power spectrum (TSPS)

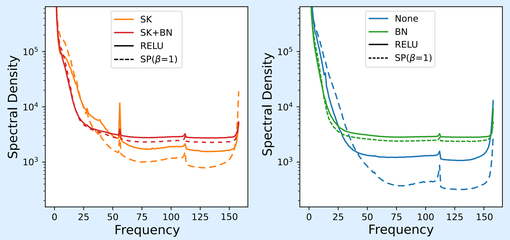
$$\mathrm{EF}(f(x)) := \mathbb{E}_\omega[S(\omega)] \qquad S(\omega) = |\mathcal{F}\{\text{Attribution of } f \text{ at } x\}|^2$$

- We quantify **complexity-faithfulness trade-offs** of each approach

$$\Delta\mathrm{EF} := |\mathrm{EF}(f(x)) - \mathrm{EF}(\widetilde{f}(x))| \qquad \text{for a smooth surrogate } \widetilde{f}$$



Impact of Other Architectural Components



Ablating BatchNorm (BN), Skip Connections (SK) confirms that ReLU is the largest contributor of complexity (i.e. heavier tails of the attribution power spectrum).

amirme@kth.se