Post hoc explanations may be ineffective for detecting unknown spurious correlation

Presented by Kajetan Husiatyński, Piotr Komorowski

Paper of the day

Published as a conference paper at ICLR 2022

POST HOC EXPLANATIONS MAY BE INEFFECTIVE FOR DETECTING UNKNOWN SPURIOUS CORRELATION

Julius Adebayo MIT CSAIL
Michael Muelly Stanford
Hal Abelson MIT CSAIL
Been Kim Google Research

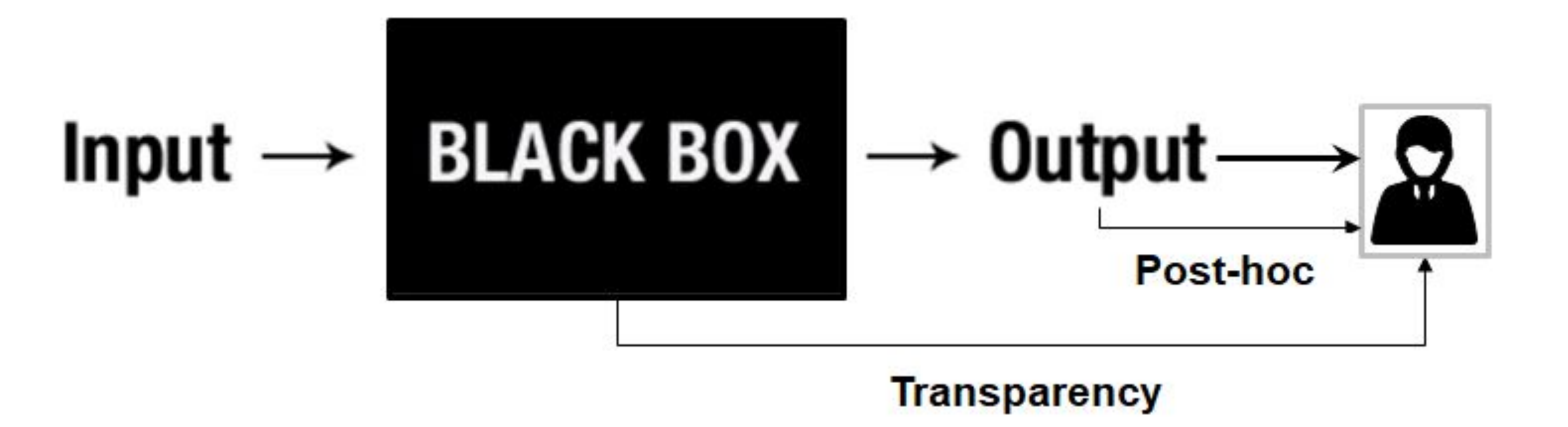
ABSTRACT

We investigate whether three types of post hoc model explanations–feature attribution, concept activation, and training point ranking–are effective for detecting a model's reliance on spurious signals in the training data. Specifically, we consider the scenario where the spurious signal to be detected is unknown, at test-time, to the user of the explanation method. We design an empirical methodology that uses semi-synthetic datasets along with pre-specified spurious artifacts to obtain models that verifiably rely on these spurious training signals. We then provide a suite of metrics that assess an explanation method's reliability for spurious signal detection under various conditions. We find that the post hoc explanation methods tested are ineffective when the spurious artifact is unknown at test-time especially for non-visible artifacts like a background blur. Further, we find that feature attribution methods are susceptible to erroneously indicating dependence on spurious signals even when the model being explained does not rely on spurious artifacts. This finding casts doubt on the utility of these approaches, in the hands of a practitioner, for detecting a model's reliance on spurious signals.[1]

It is hard to find a needle in a haystack,
it is much harder if you haven't seen a needle before (Pearl).
—Judea Pearl

Motivation

*Can post hoc explanations help detect a model's reliance on **unknown** spurious training signal?*

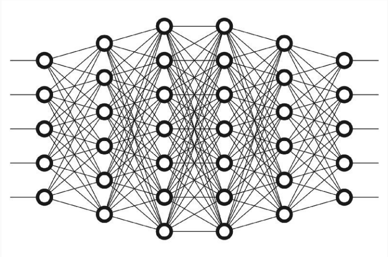


Task

Bone age classification from radiograph

Radiograph

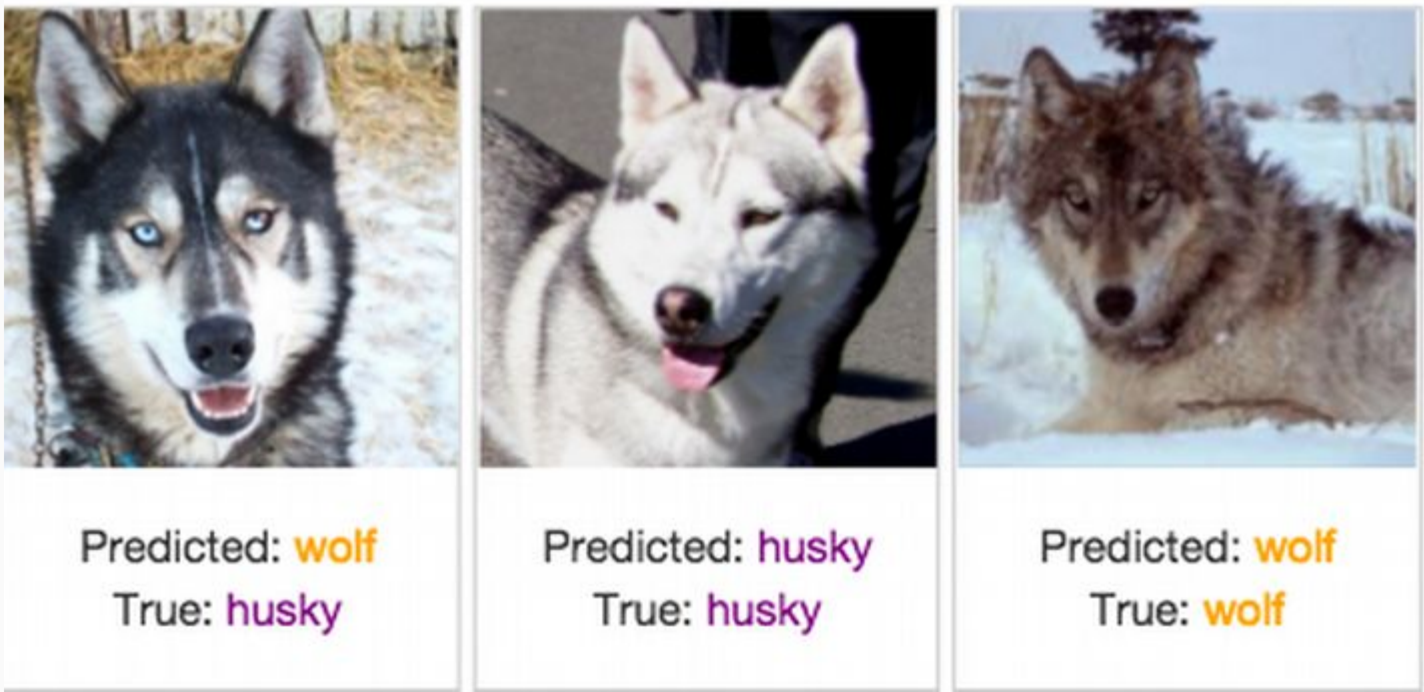
Model



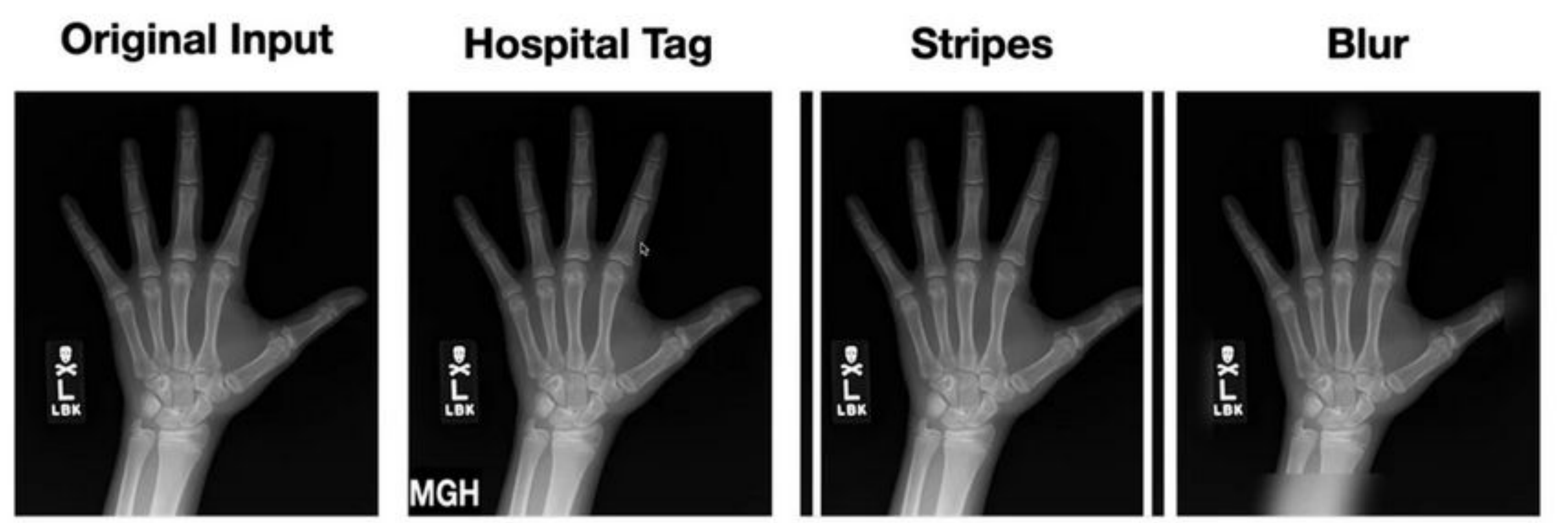
Output

- Infancy/Toddler
- Pre-Puberty
- Early/Mid Puberty
- Late Puberty
- Post Puberty

Spurious signal



Spurious signal - paper examples



Spurious Score

Measures how our model is susceptible to a spurious signal correlation.

Definition 2.1. (Spurious Score). Given a spurious signal, c_i, the index of its spurious aligned class, $j \in [k]$, a model, $\theta_{\text{spu}} : \mathbb{R}^d \to \mathbb{R}^k$, where $\arg\max(\theta_{\text{spu}})$ indicates the classifier's predicted class, we define the spurious score as:

$$\mathrm{SC}_{c_i,j}(\theta_{\text{spu}}) := \mathbb{P}_{\{x^i | \theta_{\text{spu}}(x^i) \,!=\, j\}}[\arg\max(\theta_{\text{spu}}(\mathrm{SCF}(x^i, y^i, c_i))) = j].$$

What is a probability of changing our output if we add spurious signal to our input.

Normal model vs Spurious Model

We empirically estimate the spurious score and term models that have a score above 0.85 for any of the pre-defined signals ‘spurious models’.

We term a model ‘normal’ if the spurious score is below 0.1 across all classes and the 3 pre-defined spurious signals.

Spurious Signal Detection Reliability Measures

- Known Spurious Signal Detection Measure (K-SSD)

$$S_d(\mathrm{E}_{f_{\mathrm{spu}}}(x_{\mathrm{spu}}), x_{\mathrm{gt}}))$$

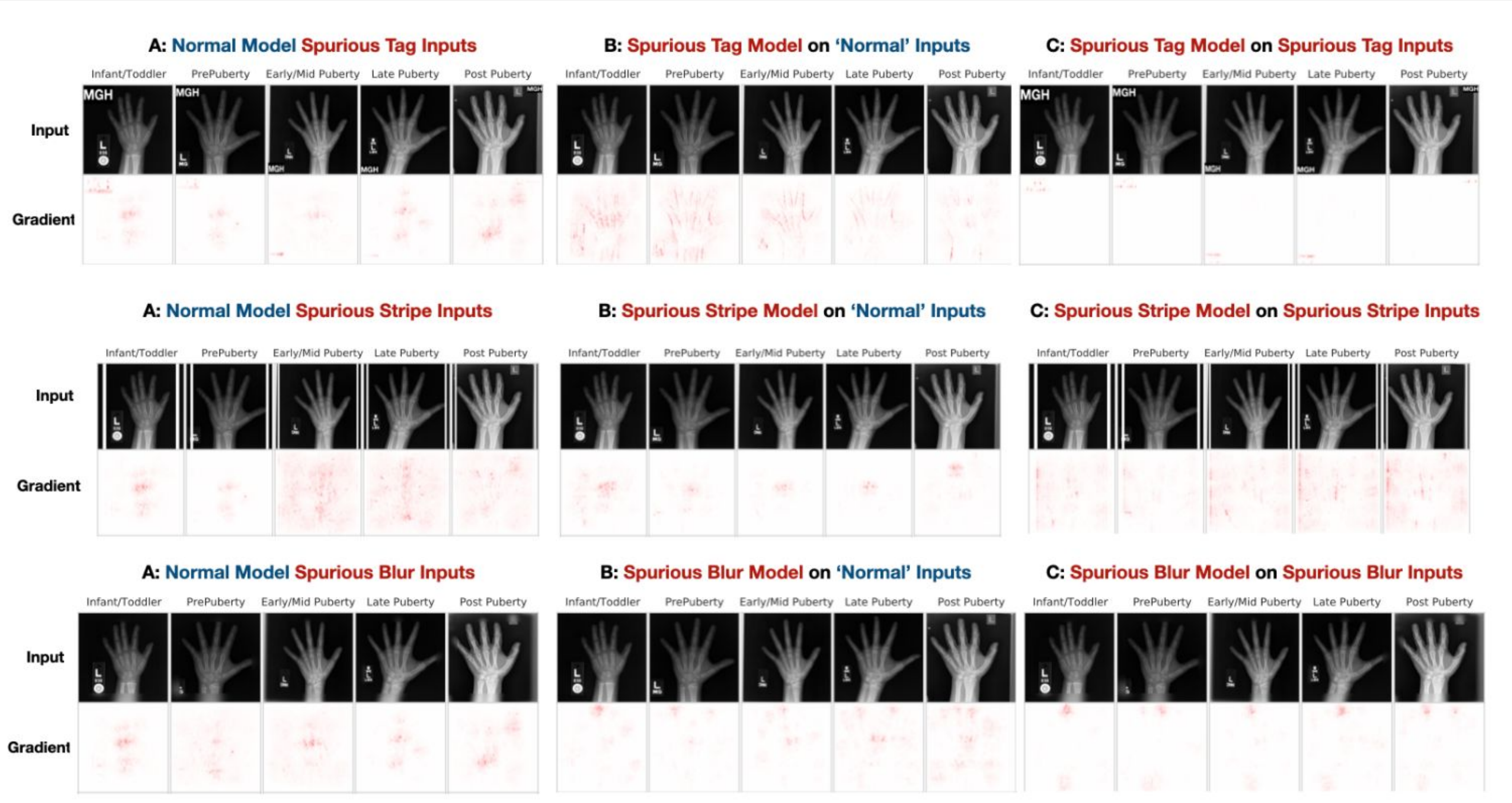
- Cause-for-Concern Measure (CCM)

$$S_d(\mathrm{E}_{f_{\mathrm{spu}}}(x_{\mathrm{norm}}), \mathrm{E}_{f_{\mathrm{norm}}}(x_{\mathrm{norm}}))$$

- False Alarm Measure (FAM)

$$S_d(\mathrm{E}_{f_{\mathrm{norm}}}(x_{\mathrm{spur}}), \mathrm{E}_{f_{\mathrm{spu}}}(x_{\mathrm{spu}}))$$

Feature attributions



Quantitative Results (visible signal)

Table 1: Performance metrics for each attribution method across tasks for the Tag Setting. Below each metric in the Table is another row (SEM) that indicates the standard error of the mean for each value.

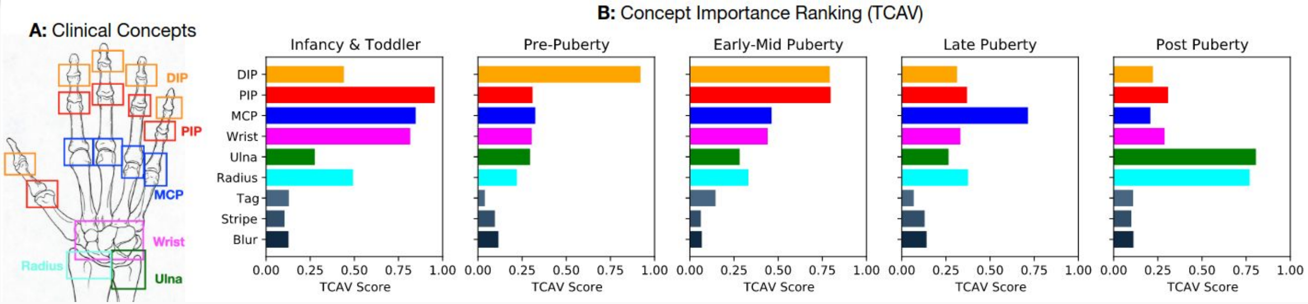
Method	Bone Age				Knee				Dog Breeds			
	Grad	SG	IG	GBP	Grad	SG	IG	GBP	Grad	SG	IG	GBP
K-SSD	0.65	0.66	0.67	0.81	0.51	0.49	0.47	0.76	0.71	0.76	0.79	0.88
K-SSD (SEM)	0.0097	0.013	0.019	0.006	0.012	0.017	0.019	0.023	0.01	0.011	0.014	0.01
CCM	0.37	0.39	0.35	0.75	0.32	0.33	0.35	0.66	0.42	0.41	0.39	0.64
CCM (SEM)	0.0031	0.002	0.015	0.029	0.027	0.023	0.029	0.014	0.013	0.016	0.012	0.015
FAM	0.51	0.55	0.53	0.68	0.46	0.47	0.45	0.69	0.59	0.64	0.68	0.73
FAM (SEM)	0.0029	0.0019	0.018	0.024	0.023	0.024	0.019	0.016	0.015	0.011	0.022	0.035
FAM-GT	0.56	0.53	0.46	0.61	0.42	0.48	0.41	0.63	0.76	0.73	0.77	0.81
FAM-GT (SEM)	0.017	0.035	0.0253	0.028	0.016	0.019	0.0045	0.006	0.011	0.033	0.024	0.0053

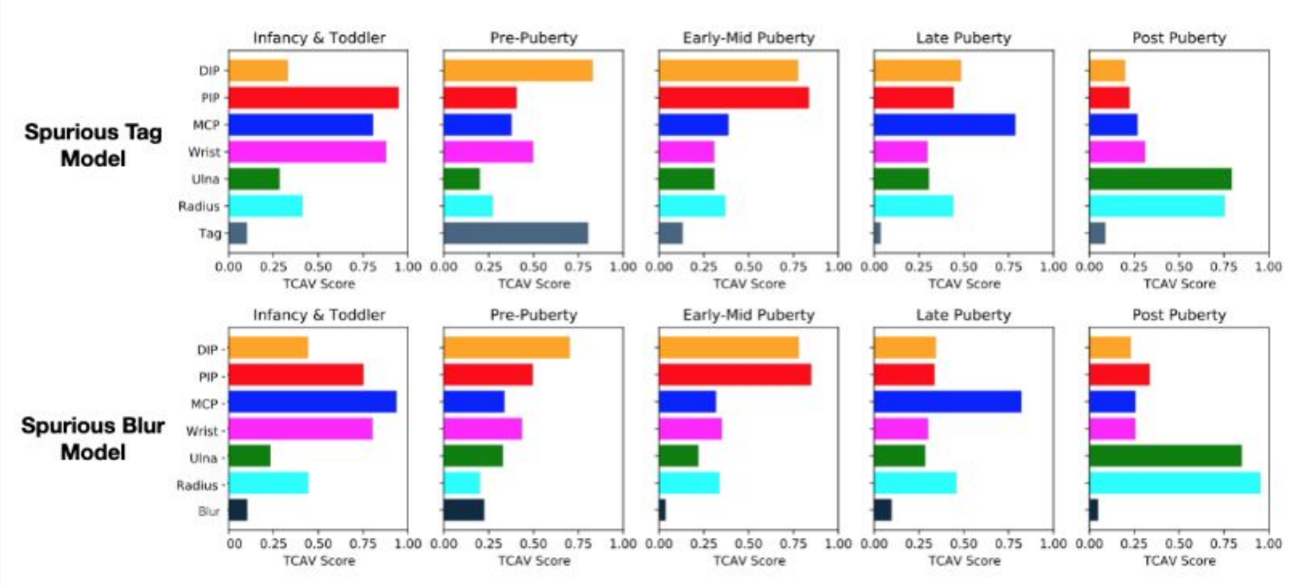
Quantitative Results (non-visible signal)

Table 11: Performance metrics for each attribution method across tasks for the Blur Setting.

Method	Bone Age				Knee				Dog Breed			
	Grad	SG	IG	GBP	Grad	SG	IG	GBP	Grad	SG	IG	GBP
K-SSD	0.21	0.20	0.19	0.13	0.13	0.18	0.17	0.31	0.29	0.30	0.31	0.35
CCM	0.28	0.29	0.24	0.64	0.23	0.22	0.27	0.67	0.38	0.33	0.35	0.71
FAM	0.48	0.49	0.47	0.51	0.36	0.38	0.33	0.58	0.55	0.56	0.47	0.73

Concept activation importance





Blinded study - participants

- 200 end-users use post hoc explanations to detect a model’s reliance on spurious signals.
- 50% of the participants had previous ML experience in training a model.
- 74% of the participants had previous ML experience in using a model.

Blinded study was split into two groups

Group A had been told explicitly of potential spurious correlation.

Group B had no prior knowledge of potential spurious correlation.

Blinded study - result median Likert score

Method	B-Normal	NB-Normal	B-Spurious	NB-Spurious
SmoothGrad	4*	4*	3*	3
TCAV	4*	3	3*	2*
Influence	3*	3	3*	3
Control	4	3	4	4

Conclusions

- Post hoc explanations can be used to identify a model's reliance on a **visible** spurious signal, provided the signal is **known** ahead of time by the practitioner
- Paper calls for a completely different paradigm of methods that are designed to detect spurious training signals.
- Current post hoc methods are promising, but their effectiveness is currently under question.