Locally Explaining Prediction Behavior via Gradual Interventions and Measuring Property Gradients

Niklas Penzel and Joachim Denzler

Computer Vision Group, Friedrich-Schiller-University Jena, Germany

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

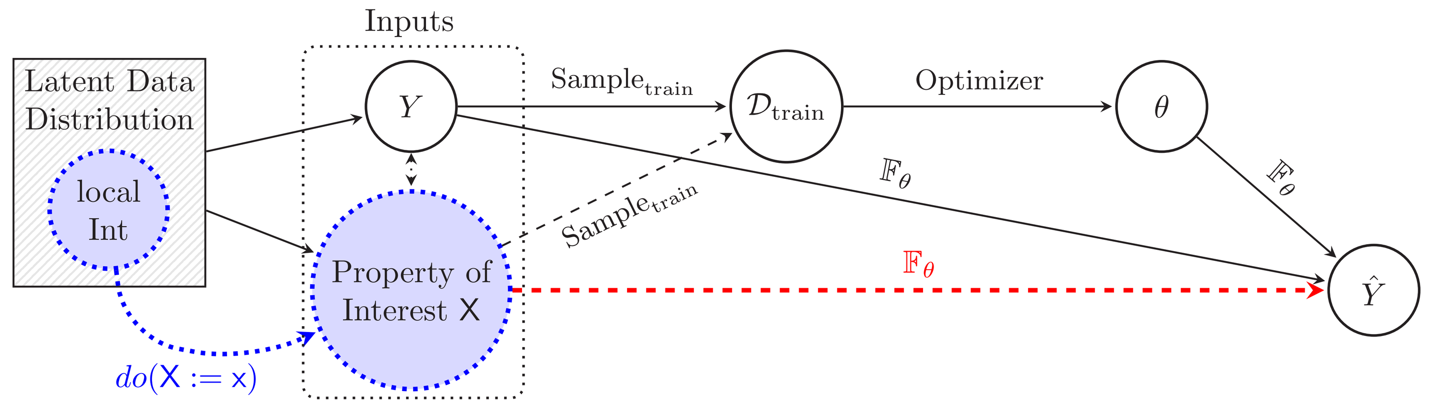
Computer Vision Group

Oral

WACV 2026 TUCSON, AZ 3/6 - 3/10

1. Causal Perspective on Explanations

- Data generating systems can be studied as **Structural Causal Models**[4]
- Neural networks are data generating systems and we are interested in **changes in the model outputs $\hat{Y}$**
- The Causal Hierarchy Theorem[5] states that in the general case the causal hierarchy does not collapse
  - ⟶ **Interventional data is necessary for interventional insights**



Associational Explanations[6]

Testing conditional dependence

$$H_0 : X \perp\!\!\!\perp \hat{Y} | Y$$

Interventional Explanations[8]

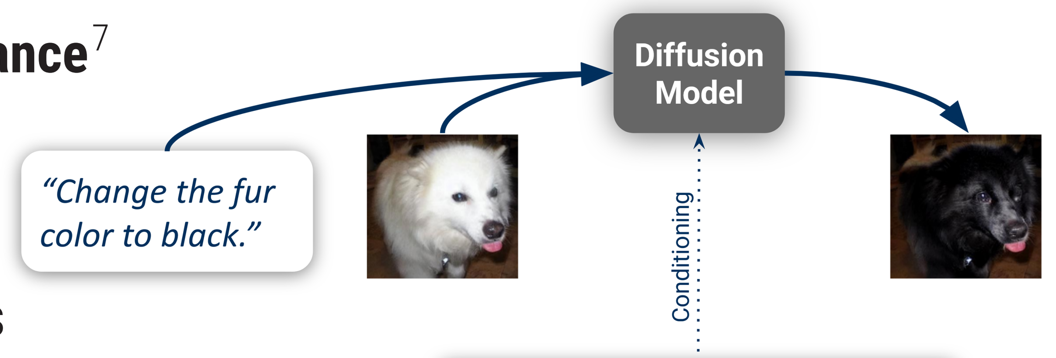
Sampling from

$$P(\hat{Y} | do(X := x), Y)$$

2. Generating Interventional Data

Three Sources

- **Classifier-Free Guidance**[7] for conditioning
- Sampling nonlinear inbetween steps
- Gradual interventions enables study of changes



$$\bar{e}(z_t, c_T) = e(z_t, \varnothing) + s_T(e(z_t, c_T) - e(z_t, \varnothing)),$$

3. Measuring Changes in Model Behavior

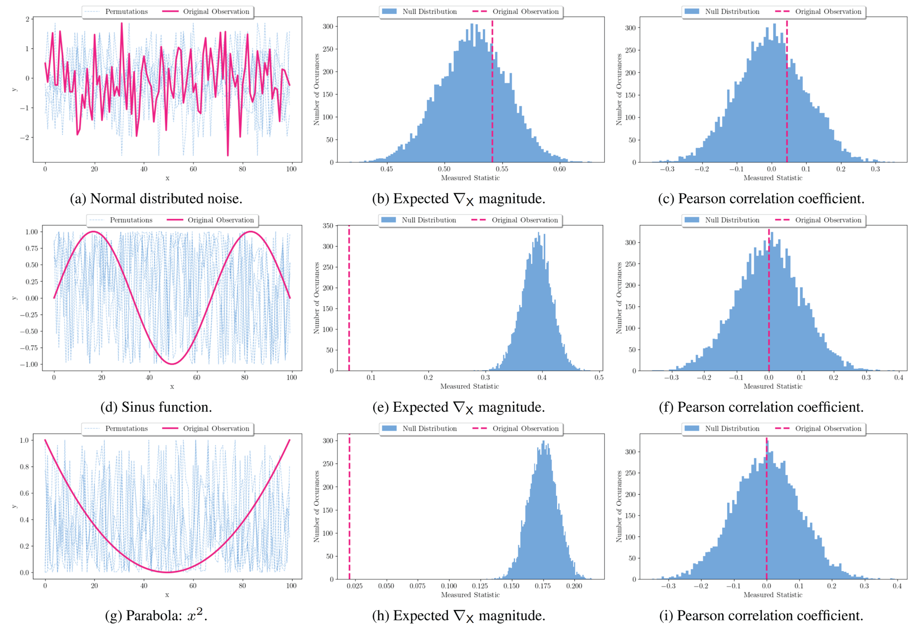
$$\mathbb{E}_X[|\nabla_X \mathbb{F}_\theta(I_X)|] = \int |\nabla_X \mathbb{F}_\theta(I_X)| \cdot p(x) dx \overset{(*)}{\approx} \frac{1}{|\mathfrak{X}|} \sum_{x \in \mathfrak{X}} |\nabla_X \mathbb{F}_\theta(I_X)|,$$

- X: property of interest
- $\mathfrak{X}$: set of equidistant samples
- $\mathbb{F}_\theta$: trained model
- I_X: input with specific property realization

- **Gradient magnitude** w.r.t. the property of interest to measure change
- Generalization of the **Causal Concept Effect**[8] **for gradual interventions**

Algorithm 1 Hypothesis test for changes in prediction behavior for variations in a property X.

```
Require: ordered list of predictions F_θ(I_X)   ▷ N elements
Require: test statistic S                       ▷ for the outputs
Require: integer K > 0                          ▷ Number of Permutations
Require: δ ∈ (0, 1)                             ▷ Significance Level
  p ← 0.0
  σ_orig. ← S(F_θ(I_X))                          ▷ Estimate the original statistic
  for i ∈ {1, ..., K} do
    F_θ(I_X^(perm.)) ← permute(F_θ(I_X))
    σ_perm. ← S(F_θ(I_X^(perm.)))
    if compare(σ_orig., σ_perm.) then
      ▷ Comparison depends on S (<, >, two-sided, etc.)
      p ← p + 1/K                                ▷ Increment the p-value
    end if
  end for
  if p < δ then
    return significant.
  else
    return not significant.
  end if
```



4. Baseline Comparisons

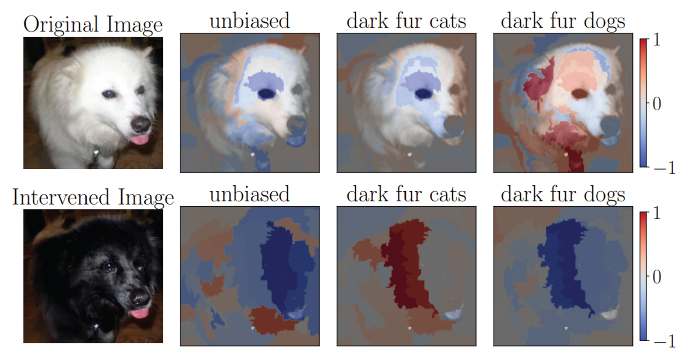
Qualitative[10]

Our Approach to Quantify Impact

$\mathbb{E}[|\nabla_X|]$ of the fur color and background property for our three CvD models. Additionally, we report significance ($p < 0.01$) abbreviated as "S" and prediction flips denoted as "F".

	Fur Color			Background		
Model	$\mathbb{E}[\|\nabla_X\|]$	S	F	$\mathbb{E}[\|\nabla_X\|]$	S	F
Unbiased	**.0099**	✓	✗	.00060	✓	✗
Dark Cats	**.0109**	✓	✓	.00006	✓	✗
Dark Dogs	**.0110**	✓	✓	.00013	✓	✗

LIME as a Local Attribution Baseline



Quantitative[9]

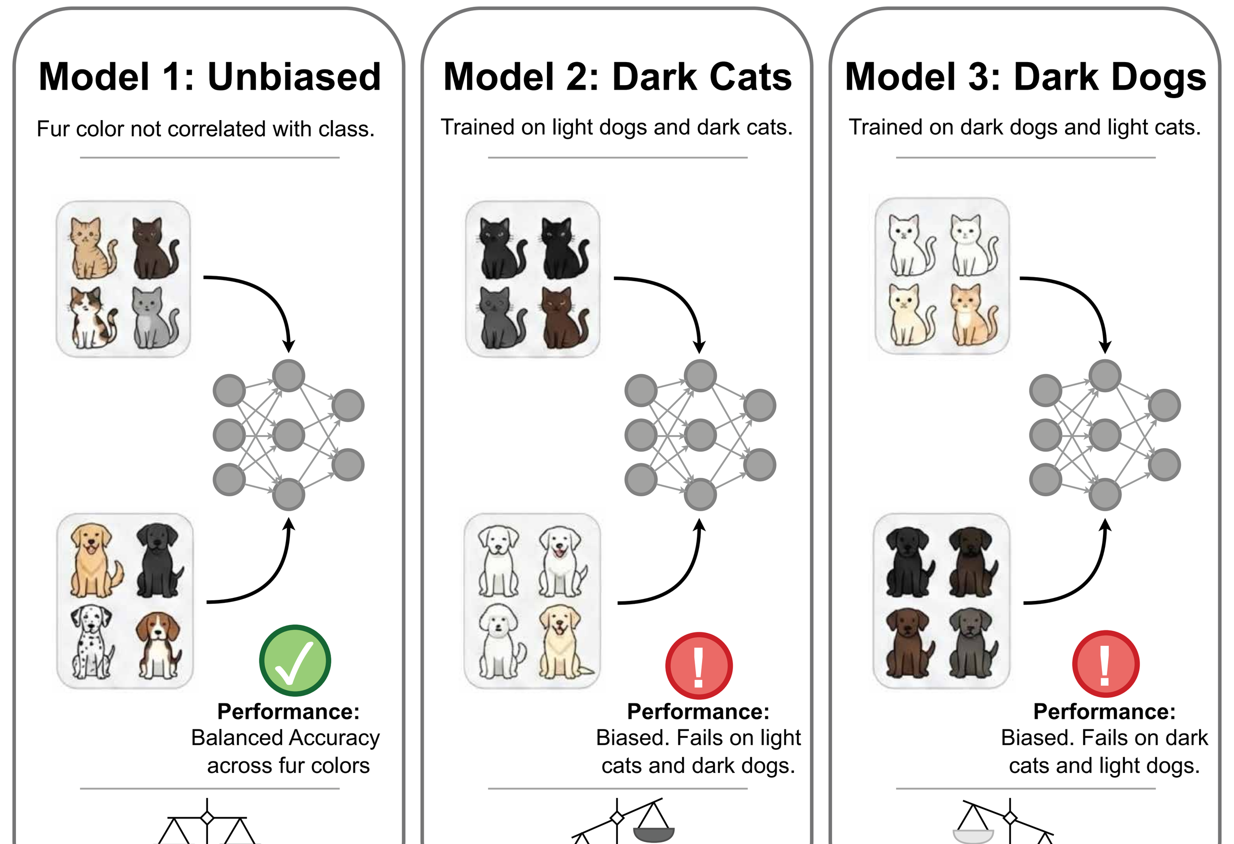
Mean accuracy (↑) and standard deviation in percent (%) of local XAI methods when predicting locally biased model behavior for an intervention. The first column denotes the dataset: Cats vs. Dogs (CvD) and the ISIC archive (ISIC). For CvD, we evaluate the three ConvMixer models trained on separate training datasets. We investigate fur color and synthetic colorful patch interventions, respectively.

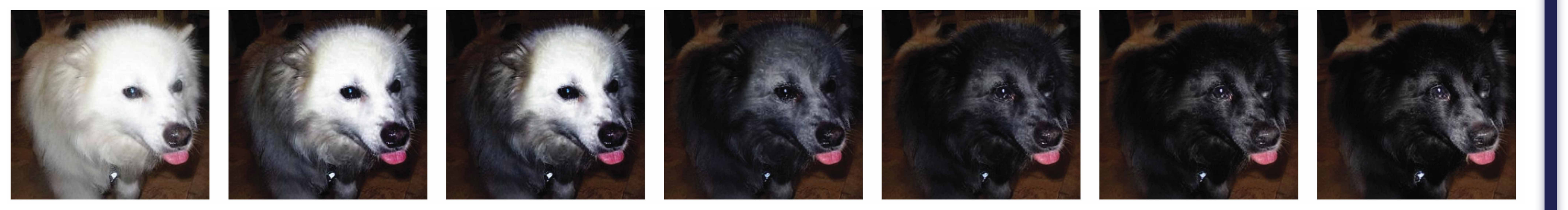
	Model	Ours	G-CAM	Int. Grad.	Occlusion	LIME	K-SHAP	DeepLift
CvD[2]	Unbiased	**86.13** ± 2.0	84.70 ± 2.1	84.77 ± 2.1	84.73 ± 2.0	84.70 ± 2.2	84.67 ± 2.1	84.83 ± 2.1
	Dark Cats Bias	**84.27** ± 1.7	61.43 ± 3.3	64.77 ± 1.8	67.70 ± 1.9	58.63 ± 1.9	60.87 ± 1.0	67.13 ± 2.2
	Dark Dogs Bias	**82.20** ± 0.8	64.20 ± 2.7	64.10 ± 2.2	67.10 ± 1.2	56.50 ± 2.1	58.50 ± 2.2	64.77 ± 2.1
ISIC[1]	ResNet18	**95.50** ± 1.2	80.00 ± 4.0	78.63 ± 2.7	78.75 ± 4.0	76.12 ± 5.1	75.88 ± 4.6	76.88 ± 6.5
	EfficientNet-B0	**94.25** ± 2.4	79.00 ± 4.9	73.75 ± 5.6	75.37 ± 6.3	73.88 ± 5.1	72.75 ± 5.5	76.38 ± 6.5
	ConvNeXt-S	**92.88** ± 1.6	78.00 ± 4.7	75.50 ± 3.2	74.88 ± 4.7	75.00 ± 3.2	74.75 ± 2.7	74.88 ± 5.8
	ViT-B/16	**95.12** ± 1.6	80.25 ± 3.4	75.87 ± 2.5	77.25 ± 3.5	76.00 ± 3.4	75.88 ± 2.9	77.50 ± 3.9

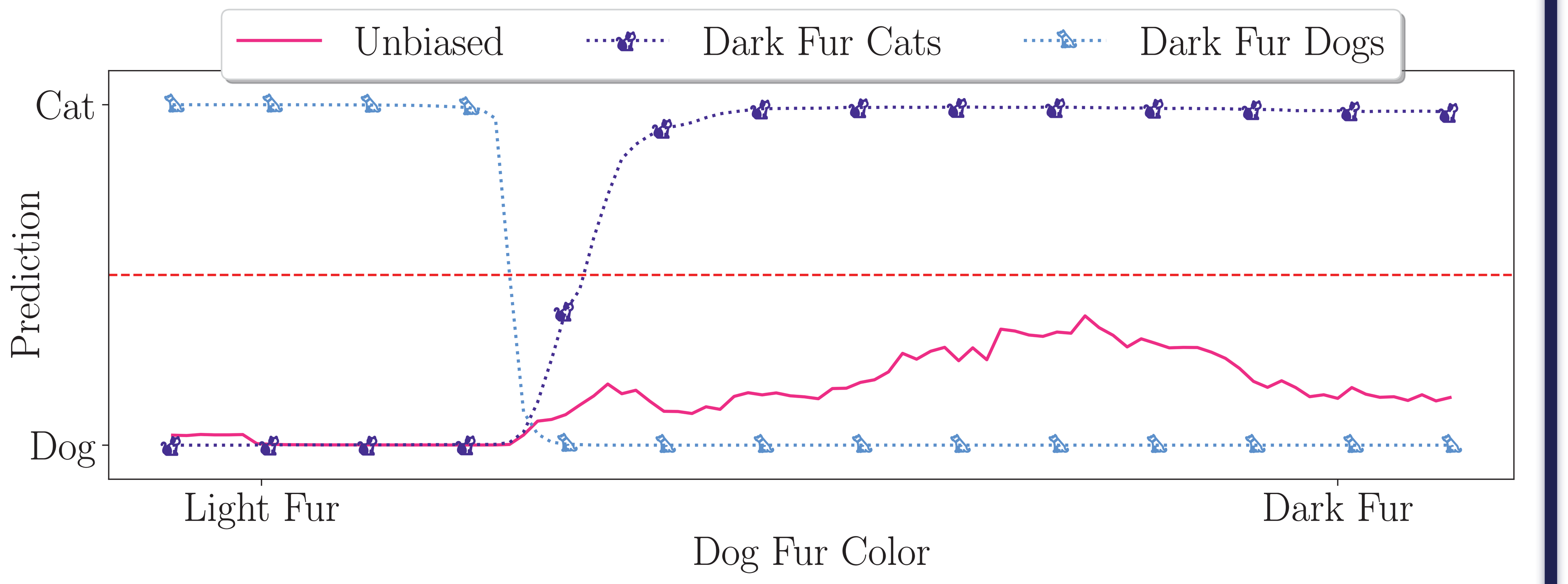
What Happens Under Gradual Interventions?

Using Generative Models to Trace Causal Drivers of Local Predictions.

Project Page





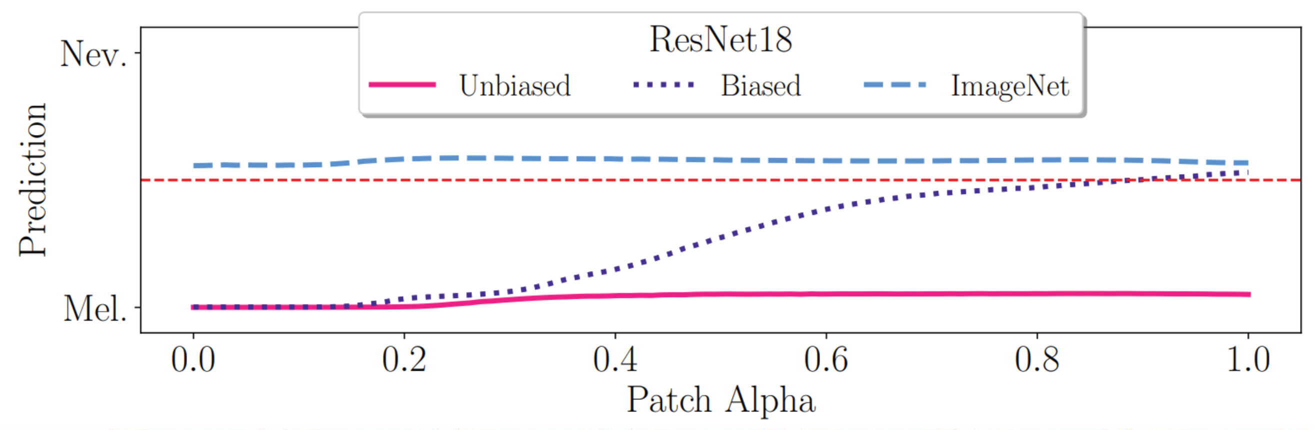


Melanoma Classification

- Application: Analysis of **melanoma**[1] **classifiers**
- **Synthetic colorful patch intervention**, a known bias
- Various backbones

$\mathbb{E}[|\nabla_X|]$ for colorful patch interventions in skin lesion classifiers. We evaluate different models and training data.

	Training Data		
Model	Unbiased	Biased	ImageNet
ResNet18	.00061	**.00531**	.00062
EfficientNet-B0	.00018	**.00495**	.00066
ConvNeXt-S	.00001	**.00519**	.00081
ViT-B/16	.00016	**.00208**	.00129

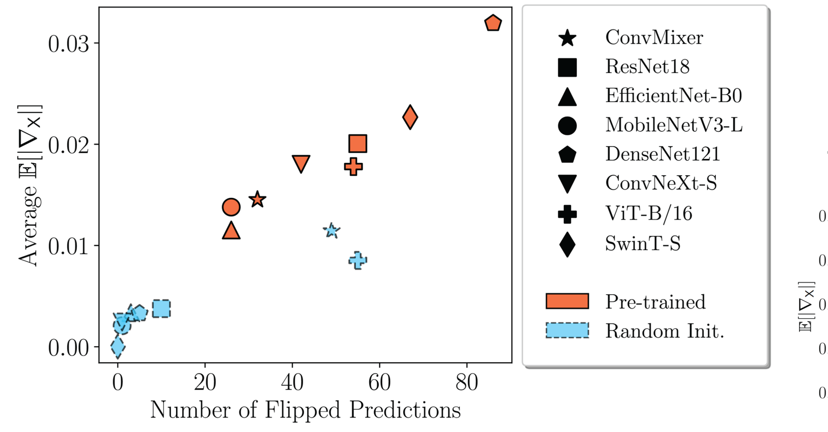


Network Training Analysis

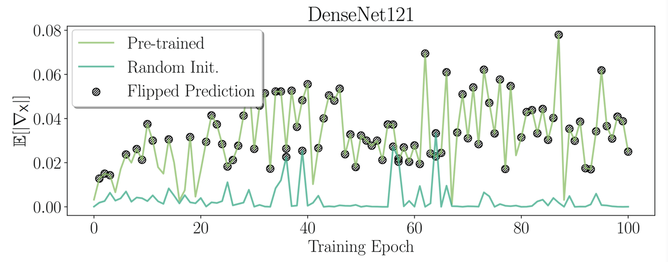
- **Training analysis** of CelebA[3] young vs. old classifiers
- Split between **random initialized** and **pretrained weights**
- Intervention on the **haircolor**

Final accuracies in percent (%) achieved by various models trained to differentiate young versus old in CelebA. We split between ImageNet pre-training ("PT") and random initialization ("RI") and calculate the performance delta (Δ).

Model	PT	RI	Perf. Δ
ConvMixer	86.08	81.49	-4.59
ResNet18	85.37	84.38	-0.99
EfficientNet-B0	86.61	84.14	-2.46
MobileNetV3-L	85.94	83.05	-2.88
DenseNet121	85.75	84.20	-1.55
ConvNeXt-S	85.98	83.63	-2.35
ViT-B/16	85.51	72.49	-13.02
SwinT-S	85.72	50.25	-35.47



Property gradient magnitudes **correlate with prediction flips**



Properties are also **forgotten in later training stages**

References:

1. International skin imaging collaboration, ISIC Archive. https://www.isic-archive.com/.
2. Will Cukierski. Dogs vs. cats. https://kaggle.com/competitions/dogs-vs-cats, 2013.
3. Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. ICCV 2015
4. Judea Pearl. Causality. Cambridge University Press, 2009.
5. Elias Bareinboim, Juan David Correa, Duligur Ibeling, and Thomas F. Icard. On pearl's hierarchy and the foundations of causal inference. Probabilistic and Causal Inference, 2022.
6. Christian Reimers, Jakob Runge, and Joachim Denzler. Determining the relevance of features for deep neural networks. ECCV 2020
7. Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In NeurIPS 2021 Workshop on Deep GenerativeModels and Downstream Applications, 2021.
8. Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (CaCE). 2019.
9. Vitali Petsiuk, Abir Das and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. BMVC, 2018.
10. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. ACM SIGKDD, 2016.