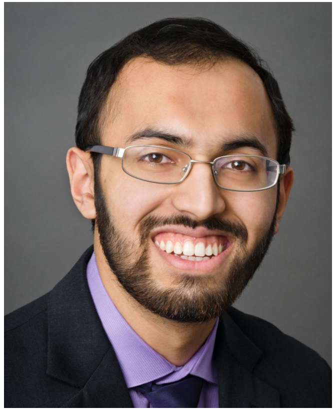
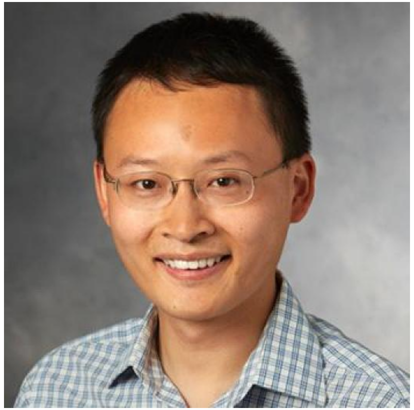
Meaningfully Debugging Model Mistakes using Conceptual Counterfactual Explanations



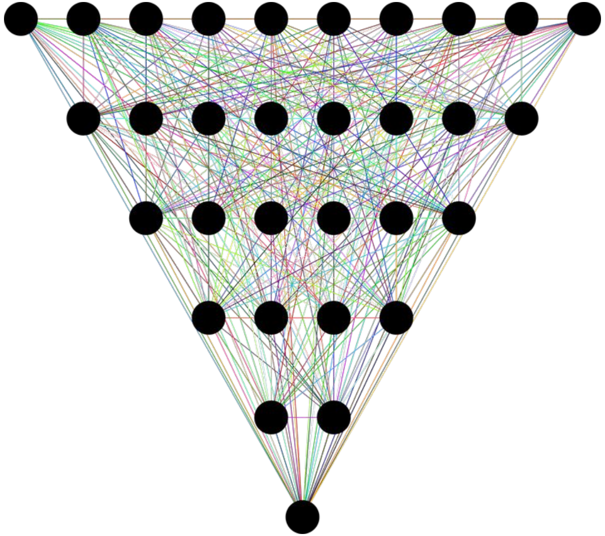


Abubakar Abid Mert Yuksekgonul James Zou

Stanford University

merty@stanford.edu

Problem



P(African crocodile) = 78%

Problem

Analyzing model mistakes is often an ad hoc process.

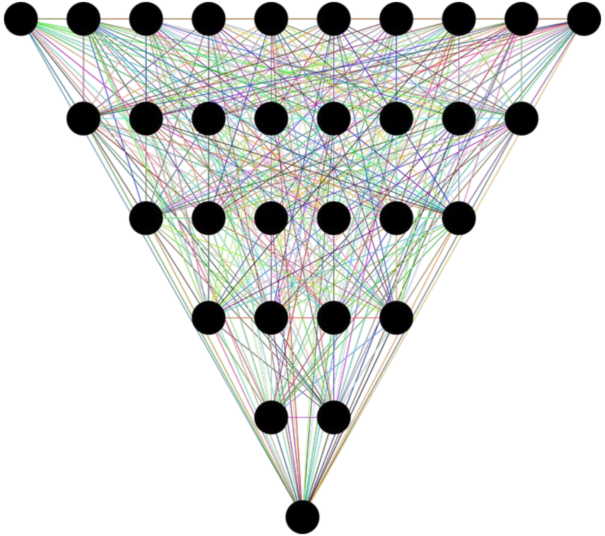
?

underrepresented in training distribution?

wrong preprocessing?

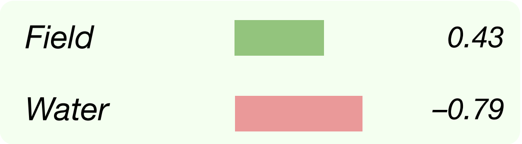
spurious correlation that is hindering generalization?

Problem



P(African crocodile) = 78%

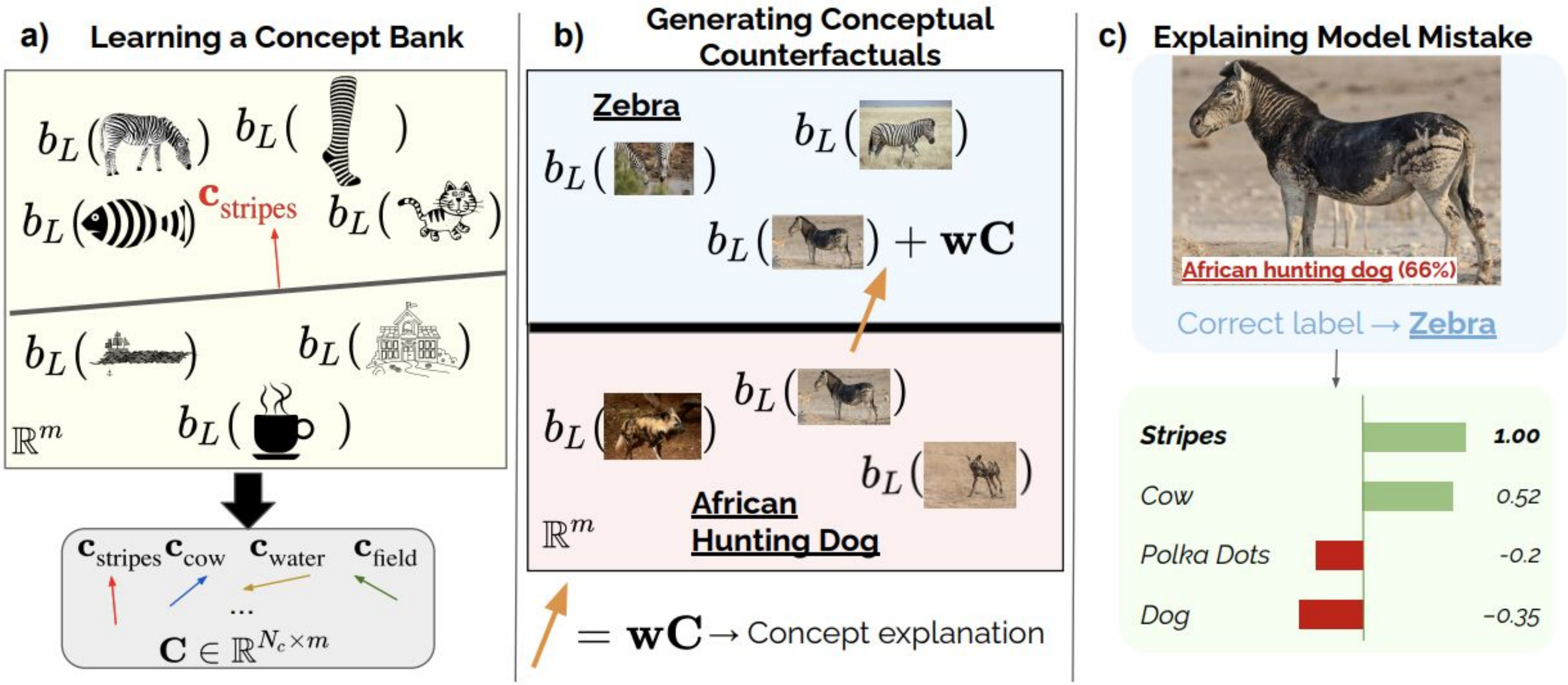
CCE
(This work)



I will train my model with zebra images from my diverse environments

I will carefully process the background

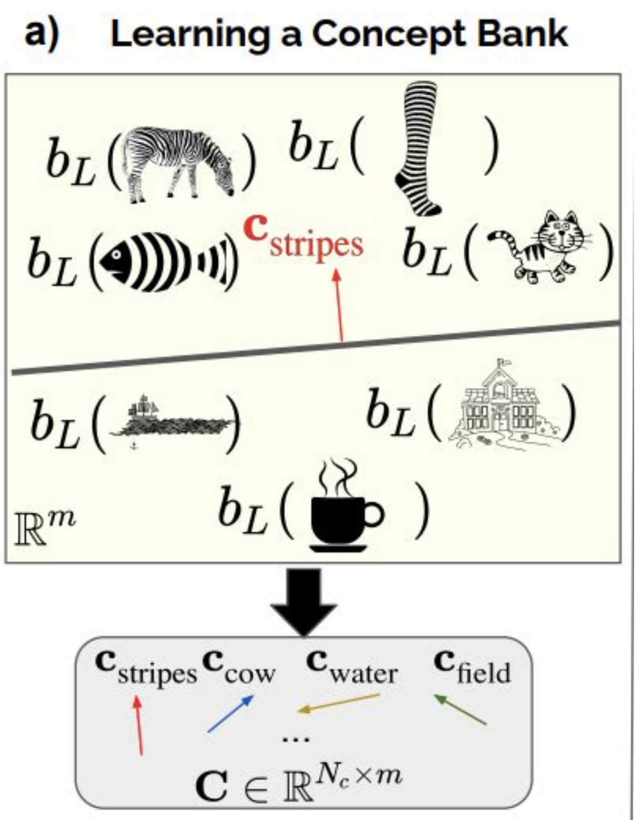
...



This work: Combine concept-based and counterfactual explanations!

Get human concepts -> Generate counterfactual statements

Learning Concepts



Concept Activation Vectors (Kim et al. 2017)

Depending on the application, the user defines a set of concepts and concept-annotated samples.

e.g. BRODEN dataset of visual concepts

(Fong & Vedaldi, 2018) contains concepts such as objects, textures, image qualities.

Counterfactual Explanations

"If X had not occurred, Y would (not) have occurred"

e.g. *If Bob had a Master's degree, he would not have been denied for loan.*

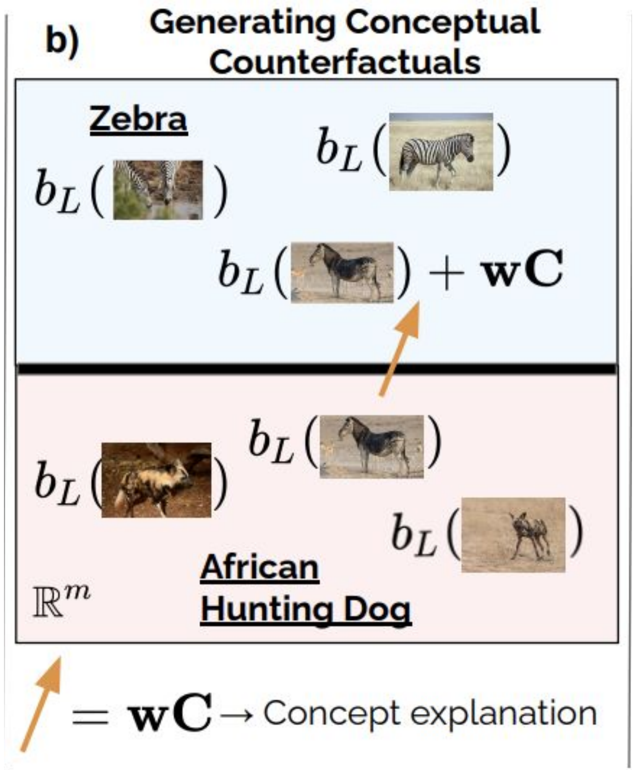
Drawing inspiration from Verma et al. 2020, our desiderata for counterfactuals:

1- Correctness: A counterfactual is correct if it can correctly change the prediction.

2- Validity: Counterfactuals should not violate real-world conditions.

3- Sparsity: Debugging/communicating a large number of modifications may not be trivial, hence counterfactuals should modify a minimal number of concepts.

Conceptual Counterfactual Explanations (CCE)

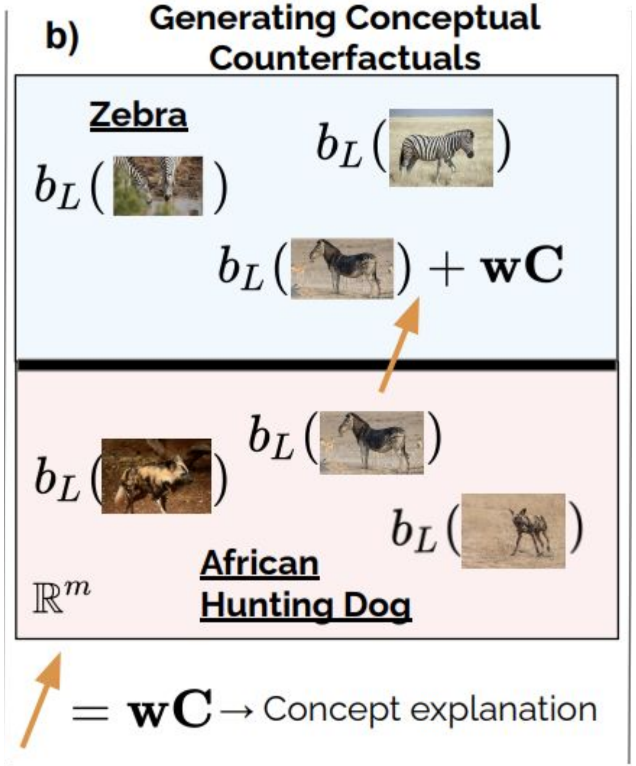


Correctness Validity Sparsity

$$\min_{\boldsymbol{w}} \quad \mathcal{L}_{\mathrm{CE}}(y, t_L(\underbrace{\mathbf{b}_L(\boldsymbol{x})}_{\text{Data Embedding}} + \underbrace{\boldsymbol{w}\tilde{\boldsymbol{C}}}_{\text{Concept Selection}})) + \alpha|\boldsymbol{w}|_1 + \beta|\boldsymbol{w}|_2$$

$$\text{s.t.} \quad \boldsymbol{w}^{\min} \leq \boldsymbol{w} \leq \boldsymbol{w}^{\max}$$

Conceptual Counterfactual Explanations (CCE)



Correctness Validity Sparsity

$$\min_{\boldsymbol{w}} \quad \mathcal{L}_{\text{CE}}(y, t_L(\mathbf{b}_L(\boldsymbol{x}) + \boldsymbol{w}\tilde{\boldsymbol{C}})) + \alpha|\boldsymbol{w}|_1 + \beta|\boldsymbol{w}|_2$$

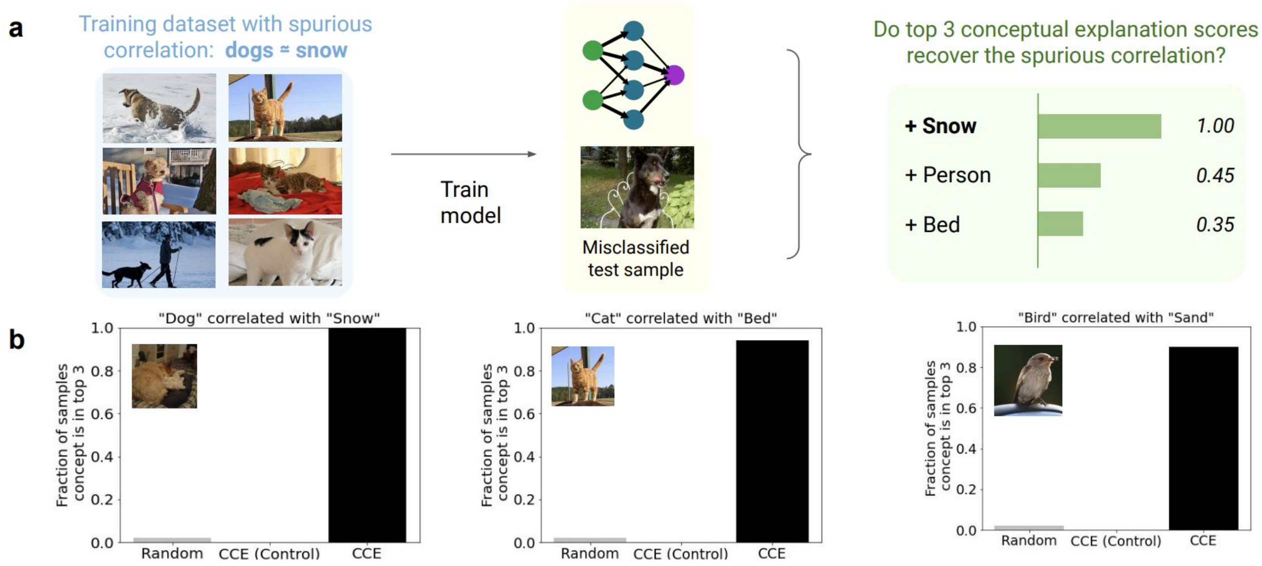
$$\text{s.t.} \quad \underbrace{\boldsymbol{w}^{\min} \leq \boldsymbol{w} \leq \boldsymbol{w}^{\max}}_{\text{Validity Constraints}}$$

<u>Intuition for validity:</u>

Cannot remove a concept that does not exist

Cannot add a concept that already exists

CCE Reveals Spurious Correlations

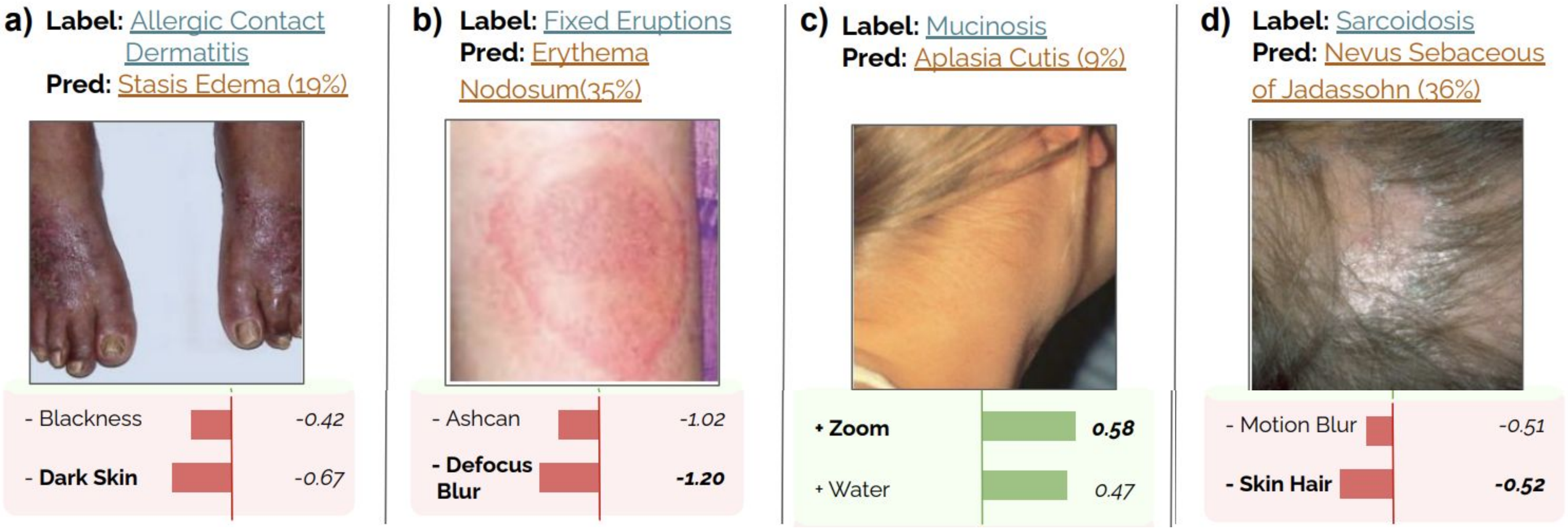


Method	Mean Prec@3	Median Rank
Random	0.02	82.65(42.7, 120.4)
CSS	0.003	76.5(69.79, 87.51)
CoCoX	0.73	4.63(3.82, 5.89)
CCE(Control)	0.04	32.3(28.03, 40.05)
CCE(Univariate)	0.91	2.00(1.71, 2.35)
CCE	0.95	1.85(1.80, 2.10)

Results over 20 scenarios.

We use Metashift (Liang & Zou, 2022) to generate datasets with ground truth spurious correlations.

CCE in the wild: Explaining the mistakes made by a skin lesion classifier



CCE can identify biases in the model, or mistakes due to low-quality data points.

Thank you!