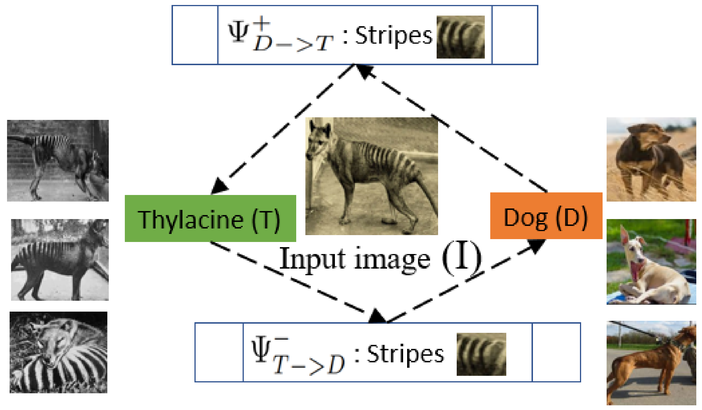
CoCo-X: Generating Conceptual and Counterfactual Explanations via Fault-Lines

Arjun R Akula[1] Shuai Wang[2] Song-Chun Zhu[1]

[1] UCLA Center for Vision, Cognition, Learning, and Autonomy
[2] University of Illinois at Chicago

AAAI-20

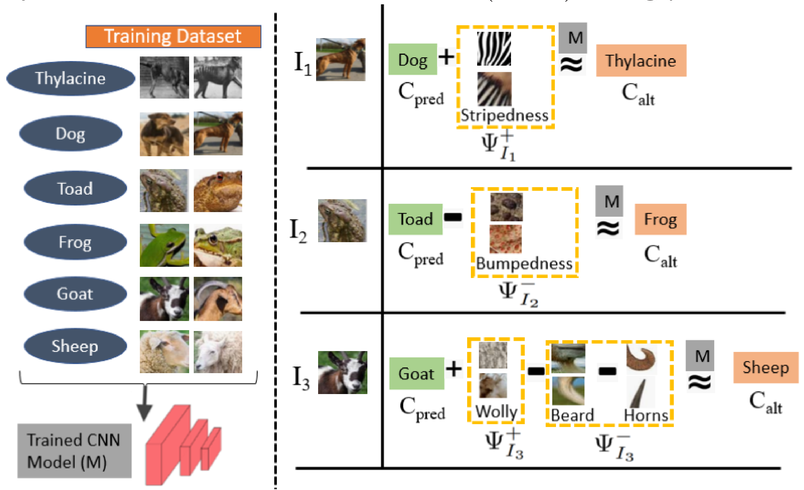
Fault-Lines as Explanations



In Cognitive Psychology, the factors (or semantic-level features) that humans zoom in on when they imagine an alternative to a model prediction are often referred to as *fault-lines*.

CoCo-X Explanation Framework

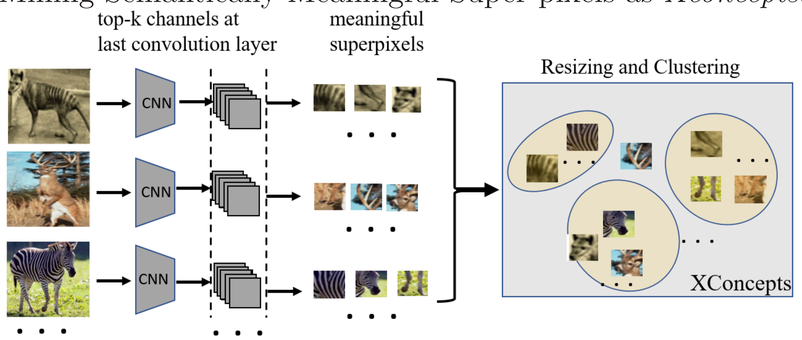
We propose *CoCo-X* framework (short for **Co**nceptual and **Co**unterfactual e**X**planations) to explain decisions made by a convolutional neural network (CNN) using *fault-lines*.



Given an input image I for which a CNN classification model M predicts class c_{pred}, our fault-line based explanation identifies the minimal semantic-level features (e.g., *stripes* on zebra, *pointed ears* of dog), referred to as explainable concepts, that need to be added to or deleted from I in order to alter the classification category of I by M to another specified class c_{alt}.

Explanation Generation in CoCo-X

Mining Semantically Meaningful Super-pixels as *Xconcepts*.



We consider feature maps from the last convolutional layer as instances of *xconcepts* and obtain their localization maps by computing the gradients of the output with respect to the feature maps. We select highly influential superpixels and then apply K-means clustering.

Fault-Line Identification

$$\underset{\delta_{pred},\delta_{alt}}{\text{minimize}} \quad \alpha D(\delta_{pred},\delta_{alt}) + \beta \left\|\delta_{pred}\right\|_1 + \lambda \left\|\delta_{alt}\right\|_1 ;$$

$$D(\delta_{pred},\delta_{alt}) = max\{g^{pred}(I^{'}) - g^{alt}(I^{'}), -\tau\};$$

$$I^{'} = A^{m,L} \circ v^{\top}_{pred}\delta_{pred} \circ v^{\top}_{alt}\delta_{alt};$$

$$\delta^{i}_{pred} \in \{-1,0\},\ \delta^{i}_{alt} \in \{0,1\}\ \forall i \text{ and } \alpha,\beta,\lambda,\tau \geq 0. \tag{4}$$

Algorithm

1. Find semantically meaningful superpixels in χ,

$$\alpha^{c}_{m,L} = \frac{1}{Z}\sum_{i}\sum_{j}\frac{\partial y^{c}}{\partial A^{m,L}_{ij}}$$

2. Apply K-means clustering on superpixels and obtain xconcepts (Σ).
3. Identify class specific xconcepts (Σ_{pred} and Σ_{alt}) using TCAV,

$$S_{c,X} = \nabla g_c(f(I)) \cdot v_X$$

4. Solve Equation 4 to obtain fault-line Ψ,

$$\Psi \leftarrow \min_{\delta_{pred},\delta_{alt}} \alpha D(\delta_{pred},\delta_{alt}) + \beta \left\|\delta_{pred}\right\|_1 + \lambda \left\|\delta_{alt}\right\|_1$$

XAI Evaluation Metrics

1. **Justified Trust**: Justified Trust is computed by evaluating the human's understanding of the model's (M) decision-making process. In other words, given an image, it evaluates whether the users could reliably predict the model's output decision.

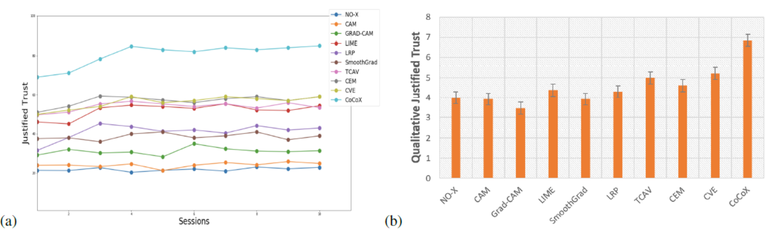
2. **Explanation Satisfaction**: We measure human subjects' feeling of satisfaction at having achieved an understanding of the machine in terms of usefulness, sufficiency, appropriated detail, confidence, and accuracy.

We used ILSVRC2012 dataset (Imagenet) and considered VGG-16 as the underlying network model.

Results

XAI Framework	Justified Trust (±std)	Explanation Satisfaction (±std)				
		Confidence	Usefulness	Appropriate Detail	Understandability	Sufficiency
Random Guessing	6.6%	N/A	N/A	N/A	N/A	N/A
NO-X	21.4% ± 2.7%	N/A	N/A	N/A	N/A	N/A
CAM (Zhou et al. 2016)	24.0% ± 1.9%	4.2 ± 1.8	3.6 ± 0.8	2.2 ± 1.9	3.2 ± 0.9	2.6 ± 1.3
Grad-CAM (Selvaraju et al. 2017)	29.2% ± 3.1%	4.1 ± 1.1	3.2 ± 1.9	3.0 ± 1.6	4.2 ± 1.1	3.2 ± 1.0
LIME (Ribeiro, Singh, and Guestrin 2016)	46.1% ± 1.2%	5.1 ± 1.8	4.2 ± 1.6	3.9 ± 1.1	4.1 ± 2.0	4.3 ± 1.6
LRP (Bach et al. 2015)	31.1% ± 2.5%	1.1 ± 2.2	2.8 ± 1.0	1.6 ± 1.7	2.8 ± 1.0	2.1 ± 1.8
SmoothGrad (Smilkov et al. 2017)	37.6% ± 2.9%	1.4 ± 1.0	2.2 ± 1.8	2.8 ± 1.0	3.1 ± 0.8	2.9 ± 0.8
TCAV (Kim et al. 2018)	49.7% ± 3.3%	3.6 ± 2.1	3.2 ± 1.8	3.3 ± 1.6	3.6 ± 2.1	3.9 ± 1.1
CEM (Dhurandhar et al. 2018)	51.0% ± 2.1%	4.1 ± 1.4	3.4 ± 1.4	3.1 ± 2.1	2.9 ± 0.9	3.3 ± 1.6
CVE (Goyal et al. 2019)	50.9% ± 3.0%	3.8 ± 1.9	3.1 ± 0.9	3.6 ± 2.1	4.1 ± 1.2	4.2 ± 1.2
CoCoX (Fault-lines)	**69.1% ± 2.1%**	**6.2 ± 1.2**	**6.6 ± 0.7**	**7.2 ± 0.9**	**7.1 ± 0.6**	**6.2 ± 0.8**

Quantitative (Justified Trust) and Qualitative (Explanation Satisfaction) comparison of CoCo-X with random guessing baseline, no explanation (NO-X) baseline, and other state-of-the-art XAI frameworks such as CAM, Grad-CAM, LIME, LRP, SmoothGrad, TCAV, CEM, and CVE.



(a) Gain in Justified Trust over Time; (b) Average Qualitative Justified Trust (on a Likert scale of 0 to 9). Error bars denote standard errors of the means.

Acknowledgments This work reported herein is supported by DARPA XAI N66001-17-2-4029.