DISSECT: Disentangled Simultaneous Explanations via Concept Traversals

Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou,
Brian Eoff, Rosalind Picard

ICLR 2022

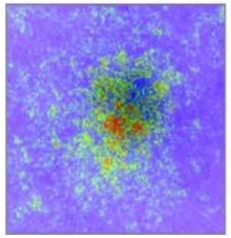
Heatmaps

Segmentation Masks

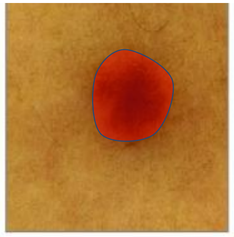
Retrieval Based

Counterfactual Generation

DISSECT

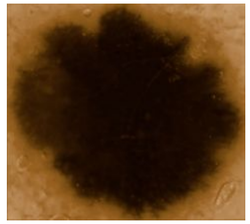


e.g. [1-4]

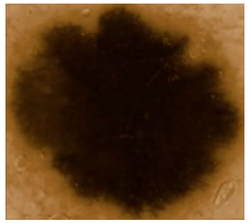


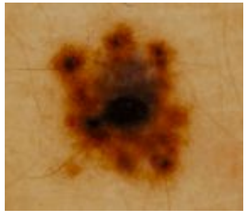
e.g. [5, 6]

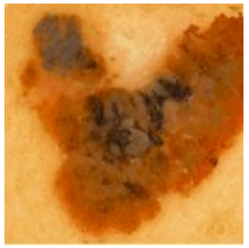
e.g. [7]

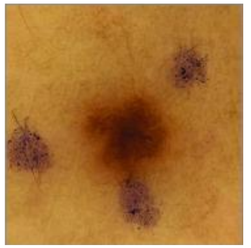


e.g. [8, 9]









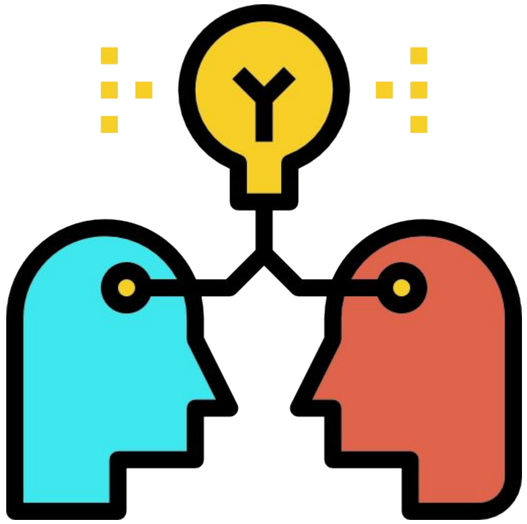
Query: Benign

Is this skin lesion Melanoma?

[1] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. “Visualizing higher-layer features of a deep network”. University of Montreal, 2009.
[2] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. “SmoothGrad: Removing noise by adding noise”. ICMLW 2017.
[3] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic attribution for deep networks”. ICML 2017.
[4] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. NeurIPS 2017.
[5] Ghorbani, A., Wexler, J., Zou, J., & Kim, B. “Towards automatic concept-based explanations”. NeurIPS 2019.
[6] Santamaria-Pang, A., Kubricht, J., Chowdhury, A., Bhushan, C., & Tu, P. “Towards Emergent Language Symbolic Semantic Segmentation and Model Interpretability”. MICCAI 2020.
[7] Silva, W., Poellinger, A., Cardoso, J. S., & Reyes, M. “Interpretability-guided content-based medical image retrieval”. MICCAI 2020.
[8] Pouya Samangouei, Ardavan Saeedi, Liam Nakagawa, and Nathan Silberman. “ExplainGAN: Model explanation via decision boundary crossing transformations”. ECCV 2018.
[9] Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. “Explanation by progressive exaggeration”. ICLR 2020.

Goal

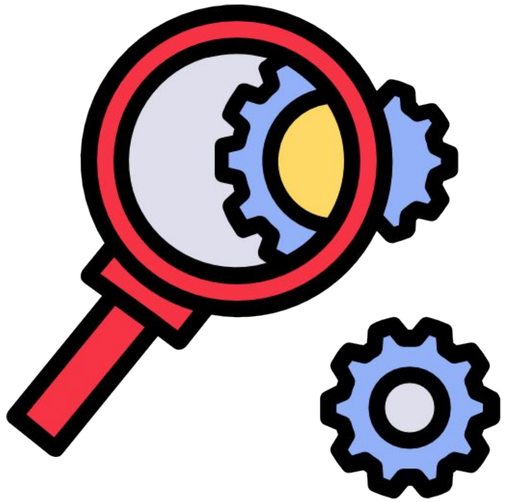
Investigating a trained classifier through generating counterfactual examples



Domain Knowledge

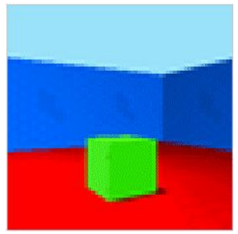
Bias

Education



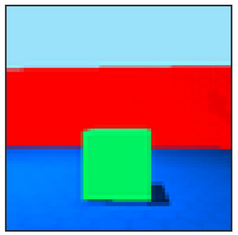
Discovery

Problem Setup with a Synthetic Example



3D Shapes dataset [1]

Factors of variation: floor hue, wall hue, object hue, scale, shape, and orientation.

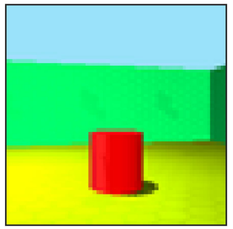


Classifying:

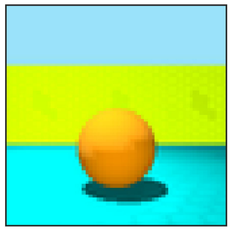
"Colored Correctly" or not?

Ground truth Concepts:

Shape Color: **RED**



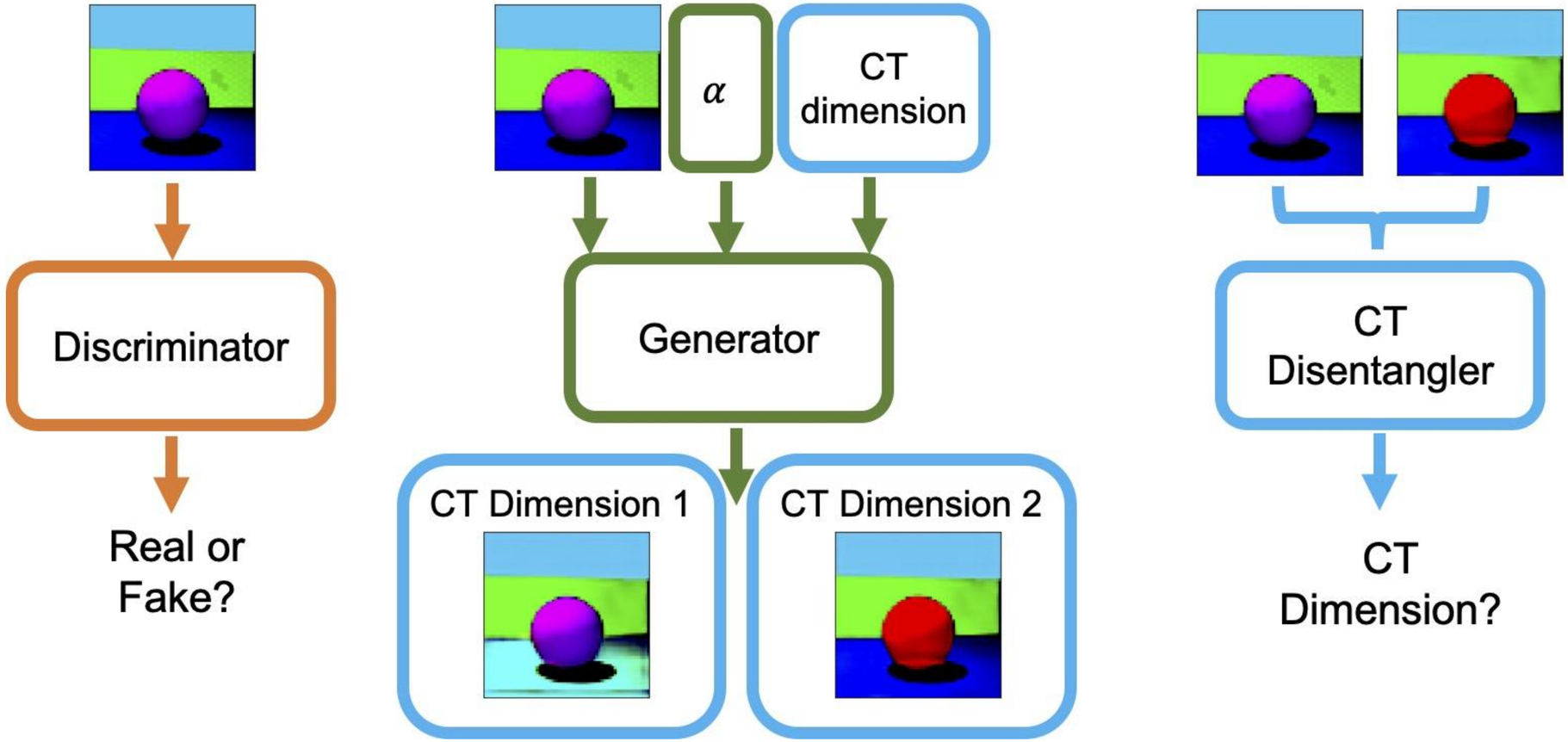
Floor Color: **CYAN**



[1] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. ICML, 2018.

Asma Ghandeharioun
aghandeharioun@google.com

Method



[1] Singla, Sumedha, et al. "Explanation by Progressive Exaggeration." *ICLR*. 2020.
[2] Takeru Miyato and Masanori Koyama. cGANs with Projection Discriminator. ICLR 2018

Asma Ghandeharioun
aghandeharioun@google.com

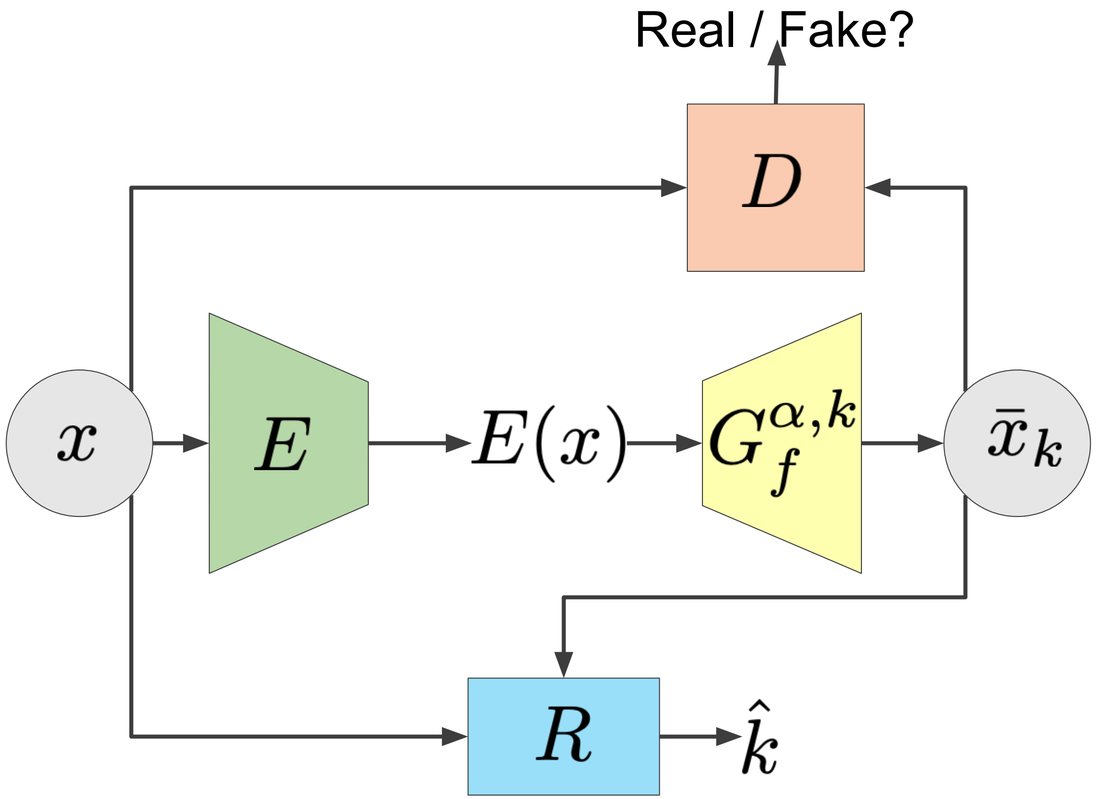
Method

Interpreting the classifier f $\mathcal{L}_f(G) = \texttt{LogLoss}\big(\alpha, f(\bar{x}_k)\big)$

Reconstruction $\mathcal{L}_{\text{rec}}(G) = \frac{1}{K}\sum_{k=0}^{K-1} \left\| x - \hat{x}_k \right\|_1$

Cycle consistency $\mathcal{L}_{\text{cyc}}(G) = \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}_\alpha \Big[\left\| x - \tilde{x}_k \right\|_1 \Big]$

GAN loss [1] $\mathcal{L}_{\text{cGAN}}(D, G)$



Enforcing disentanglement $\mathcal{L}_r(G, R) = \mathbb{CE}(e_k, R(x, \bar{x}_k)) + \mathbb{CE}(e_k, R(\bar{x}_k, \tilde{x}_k))$

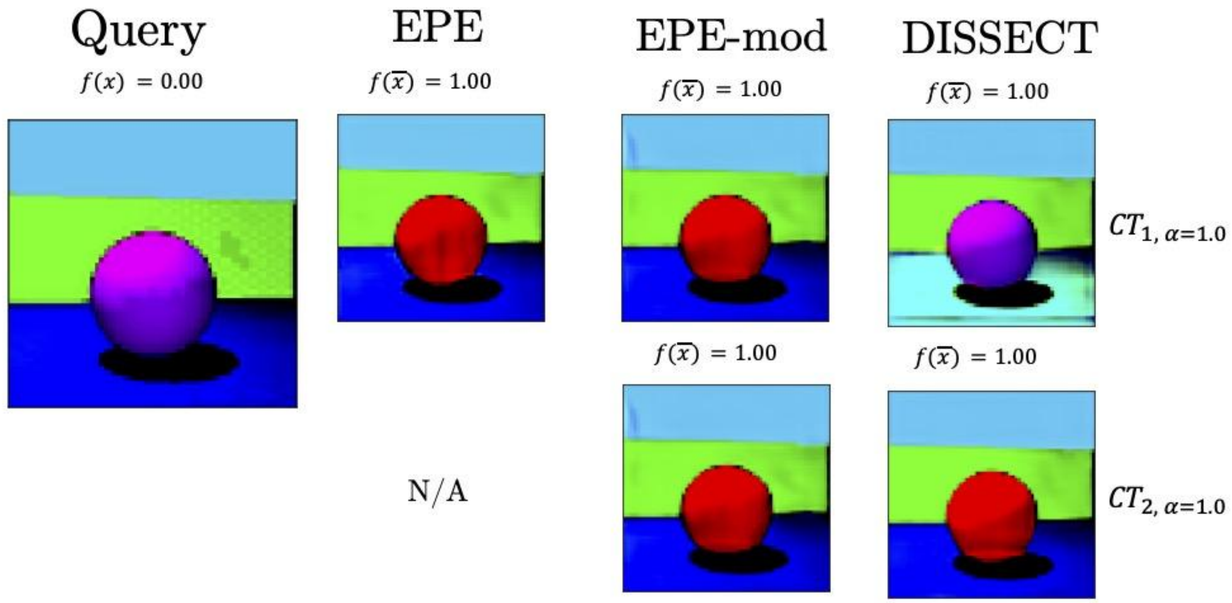
Overall Objective $\min_{G,R} \max_{D} \lambda_{\text{cGAN}}\mathcal{L}_{\text{cGAN}}(D, G) + \lambda_f \mathcal{L}_f(D, G) + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}(G) + \lambda_{\text{rec}}\mathcal{L}_{\text{cyc}}(G) + \lambda_r \mathcal{L}_r(G, R)$

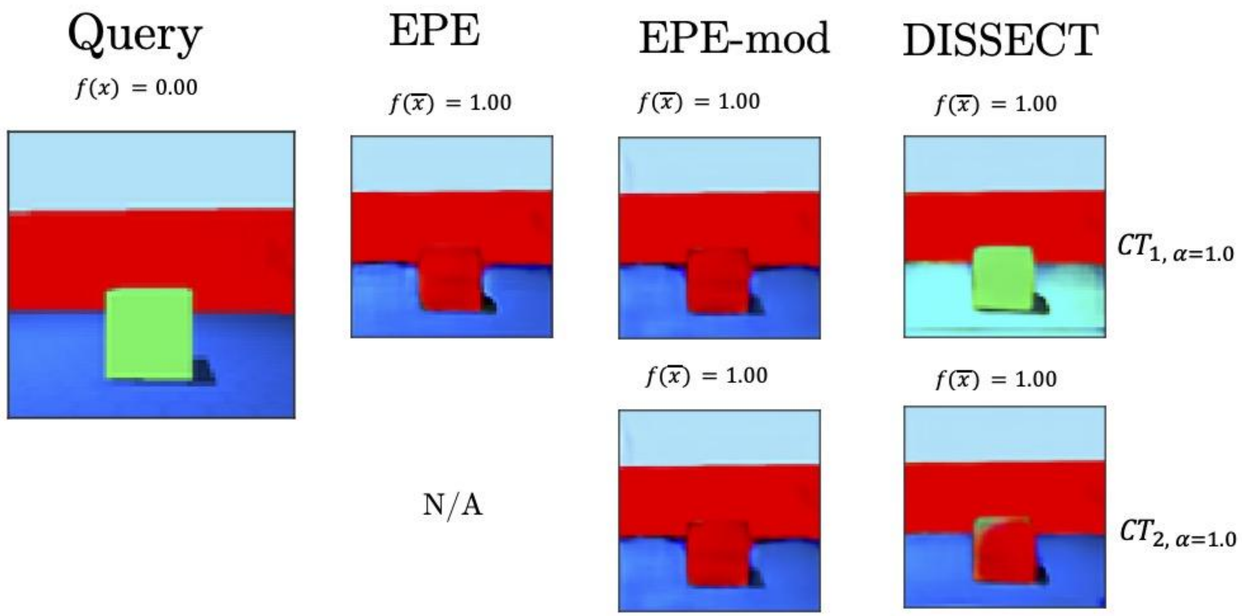
For simplifying the notation, let: $\bar{x}_k = G\big(x, c\,(\alpha, k)\,\big)$, $\hat{x}_k = G\big(x, c\,(f(x), k)\,\big)$, and $\tilde{x}_k = G\big(\bar{x}_k, c\,(f(x), k)\,\big)$

[1] Takeru Miyato and Masanori Koyama. cGANs with Projection Discriminator. ICLR 2018

Asma Ghandeharioun
aghandeharioun@google.com

Validating Qualities of Concept Traversals



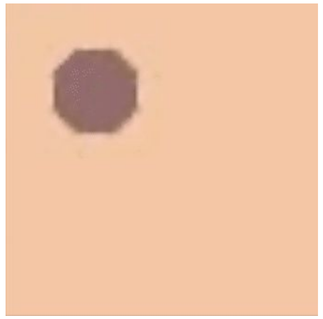


Asma Ghandeharioun
aghandeharioun@google.com

Validating Qualities of Concept Traversals

	Influence							Realism			Distinctness					Substitutability			Stability	
	↑ R	↑ ρ	↓ KL	↓ MSE	↑ CF Acc	↑ CF Prec	↑ CF Rec	↓ Acc	↓ Prec	↓ Rec	↑ Acc	↑ Prec (micro)	↑ Prec (macro)	↑ Rec (micro)	↑ Rec (macro)	↑ Acc Sub	↑ Prec Sub	↑ Rec Sub	↓ CF MSE	↓ Prob JSD
VAE-mod	0.1	0.3	inf	0.42	50.0	0	0	94.9	92.1	99.6	86.3	95.1	95.3	80.6	80.6	14.7	14.2	69.4	0.151	**0.0000**
β-VAE-mod	0.0	0.0	inf	0.42	50.0	0	0	99.5	99.0	100	90.7	92.2	92.2	91.0	91.0	42.5	8.2	19.8	0.197	**0.0000**
Annealed-VAE-mod	N/A*	N/A*	inf	0.42	50.0	0	0	100	100	100	34.0	53.2	49.2	1.20	1.2	35.8	14.4	48.1	0.175	**0.0000**
DIPVAE-mod	0.1	0.5	inf	0.41	52.3	100	4.6	100	100	100	97.2	96.7	96.9	96.5	96.5	19.0	19.0	**100**	0.127	0.0001
βTCVAE-mod	0.5	0.7	inf	0.42	50.0	0	0	100	100	100	**100**	**100**	**100**	**100**	**100**	36.4	21.3	87.3	0.143	**0.0000**
CSVAE	0.3	0.3	5.5	0.28	64.6	**100**	29.3	100	100	100	71.4	74.7	76.4	87.2	87.2	47.0	23.8	81.3	19.544	0.0274
EPE	0.8	**0.8**	**1.54**	0.09	98.4	**100**	96.7	50.1	**0**	**0**	-	-	-	-	-	99.2	95.9	**100**	0.134	0.0004
EPE-mod	**0.9**	0.7	2.2	**0.08**	**99.7**	**100**	**99.4**	**49.3**	**0**	**0**	45.3	49.6	49.8	30.3	30.3	91.0	**100**	52.5	0.128	0.0002
DISSECT	0.8	**0.8**	1.61	**0.08**	98.7	**100**	97.5	**49.3**	**0**	**0**	**100**	**100**	**100**	**100**	**100**	**100**	99.7	**100**	**0.102**	0.0003

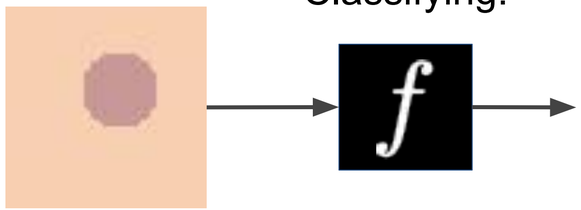
Investigating Alignment with Domain Knowledge

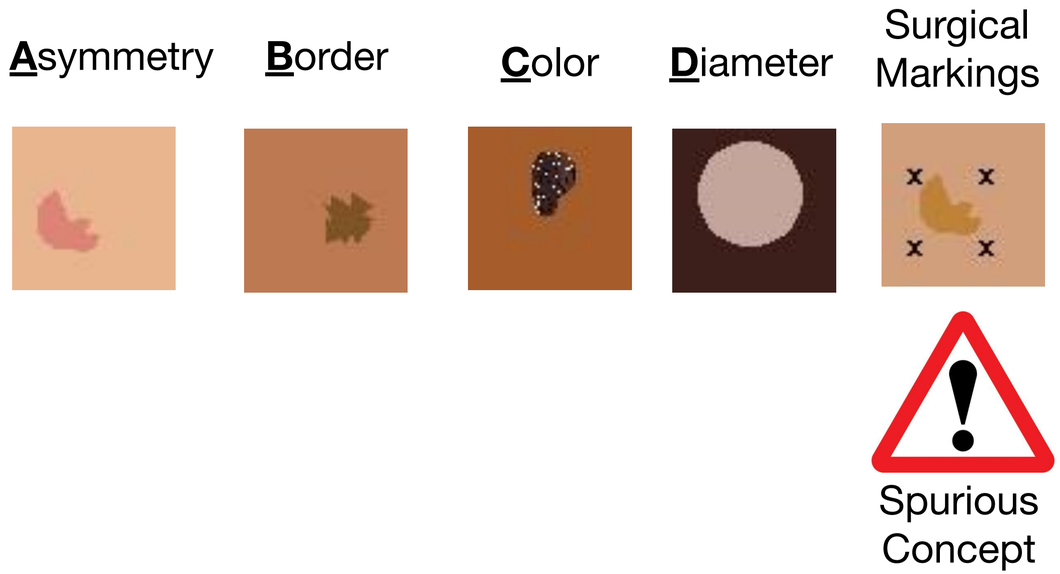


SynthDerm dataset:

- Inspired by real-world characteristics of melanoma: ABCDE of melanoma [1]
- Using Fitzpatrick (FP) scale [2]

Factors of variation: skin tone, lesion shape, lesion size, lesion location (vertical and horizontal), and whether there are surgical markings present.

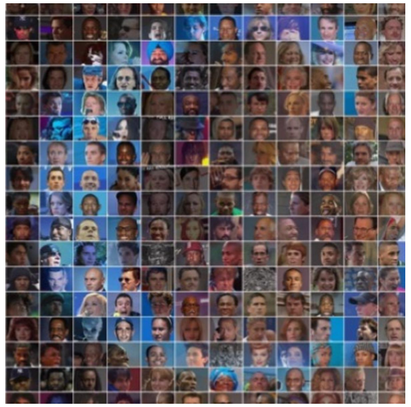




[1] Darrell S Rigel, Robert J Friedman, Alfred W Kopf, and David Polsky. ABCDE—an evolving concept in the early detection of melanoma. Archives of dermatology, 141(8):1032–1034, 2005.
[2] Fitzpatrick, Thomas B. The validity and practicality of sun-reactive skin types I through VI. Archives of dermatology 124.6 (1988): 869-871.

Asma Ghandeharioun
aghandeharioun@google.com

Identifying Biases



CelebA dataset [1]

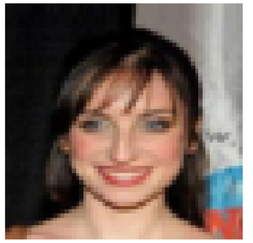
40 annotated face attributes, such as smiling, hair color, bangs, glasses, wearing a hat, etc.

Classifying:

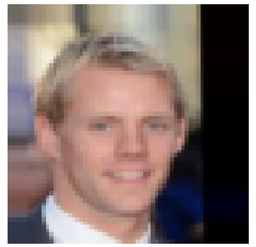
Smiling or not?

Correlated Ground Truth Concepts:

Bangs



Blond hair



[1] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. ICCV, 2015.

Asma Ghandeharioun
aghandeharioun@google.com

Conclusions

- DISSECT: a novel counterfactual explanation approach that manifests several desirable properties outperforming baselines:
  - Influence
  - Realism
  - Distinctness
  - Substitutability
  - Stability
- Showcased effectiveness of this approach through extended experiments for:
  - Investigating alignment with domain knowledge,
  - Detecting potential biases of a classifier.
- Developed a set of explanation baselines inspired by generative disentanglement.
- Translated desirable explanation properties into measurable quantities applicable to counterfactual generation.

Google Research

Asma Ghandeharioun
aghandeharioun@google.com