UNIVERSITY
OF AMSTERDAM

NEURAL INFORMATION
PROCESSING SYSTEMS

NeurIPS 2022 - ReScience Vol. 8, Issue 2, #42

[Re] Explaining in Style: Training a GAN to explain a classifier in StyleSpace

Noah van der Vleuten, Tadija Radusinović, Rick Akkerman, Meilina Reksoprodjo

UNIVERSITY OF AMSTERDAM

NEURAL INFORMATION PROCESSING SYSTEMS

Presenters

Tadija Radusinović
UvA MSc AI

Noah van der Vleuten
UvA MSc AI

Rick Akkerman
UvA MSc AI

Meilina Reksoprodjo
TU/e MSc Data Science

Introduction

➢ Classifier decisions are hard to explain: “black boxes”

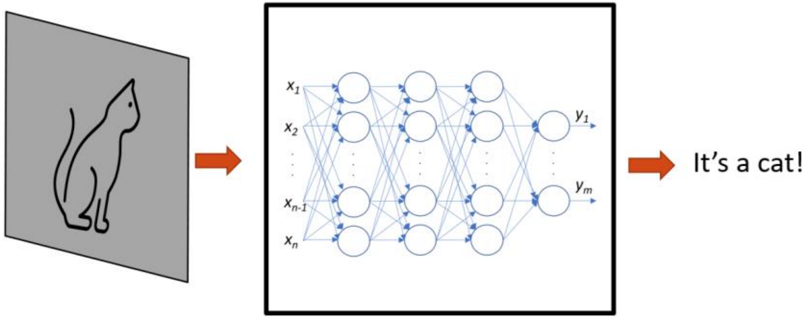
➢ If we could explaining classifier decisions, it would help to

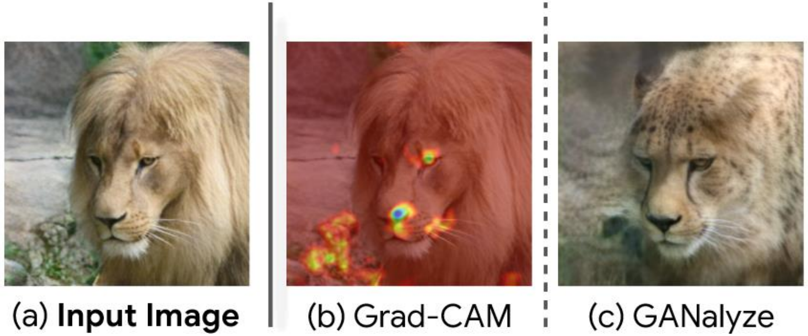
- reveal model biases;
- support downstream human decision making;
- understand our model better!

➢ Heatmaps as explanation:

- insufficient for non-local attributes;
- show “where”, not “how”.

➢ **Promising direction: counterfactual explanations**

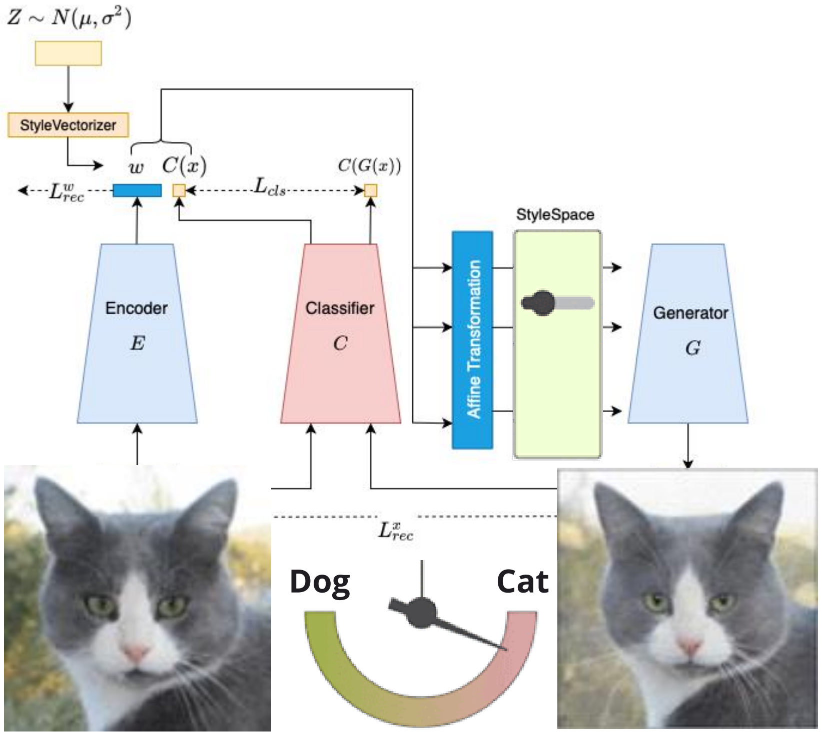




[Lang *et al.*, 2021]

StylEx

- ➢ Classifier-based training of StyleGAN2
- ➢ Capture classifier-specific attributes in a disentangled StyleSpace
- ➢ Perturb attributes to generate counterfactuals (AttFind)



[Lang *et al.*, 2022]

“ Had the input *x* been *x*′, then the classifier output would have been *y*′ instead of *y* ”

Perceived Gender
Attribute #1: “Stubble beard”

[Lang *et al.*, 2022]

UNIVERSITY OF AMSTERDAM

NEURAL INFORMATION PROCESSING SYSTEMS

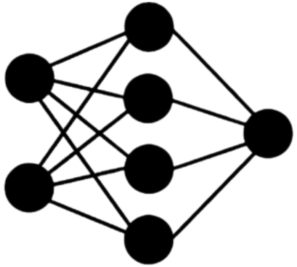
Scope of reproducibility

Claim 1: Visual Coherence

Attributes detected by StylEx should be identifiable by humans

Claim 2: Distinctness

Attributes extracted by StylEx should be distinct



Claim 3: Sufficiency

Changing attributes should result in a cumulative change of classifier output

Methodology

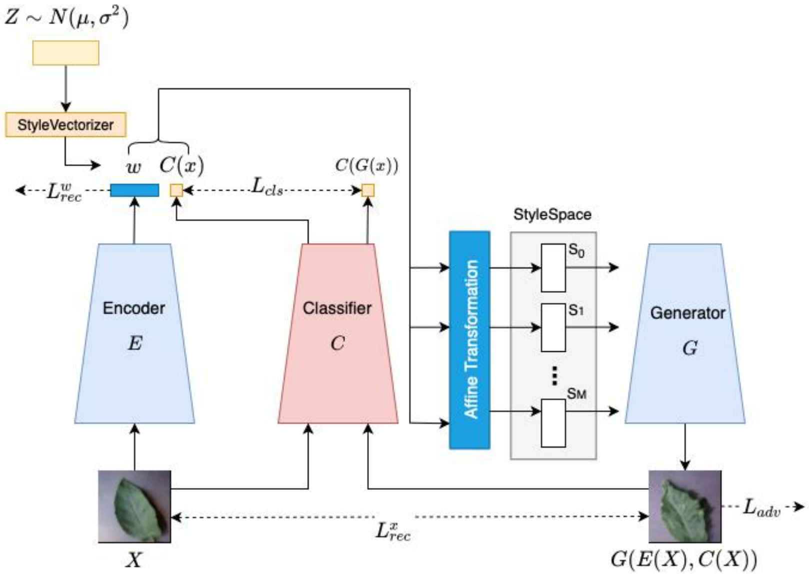
- Reimplemented end-to-end StylEx training in PyTorch;
- User study to evaluate coherence and distinctness;
- Counted classification flips to evaluate sufficiency;
- Verified sufficiency calculations on their given model.

UNIVERSITY
OF AMSTERDAM

NEURAL INFORMATION
PROCESSING SYSTEMS

Model overview

- ➢ StylEx consists of a StyleGAN, an encoder and a pre-trained classifier;
- ➢ Encoder and generator function as an autoencoder **(reconstruction loss)**
- ➢ Reconstruction should keep class information **(classification loss)**
- ➢ 64x64px images, rather than 256x256px
- ➢ Unmentioned implementation details



Datasets

FFHQ [Karras et al. 2018]

Perceived gender

CelebA [Karras et al. 2018]

Used for labels

UNIVERSITY OF AMSTERDAM

NEURAL INFORMATION PROCESSING SYSTEMS

Datasets

Plant-Village [Hughes et al. 2015]

Perceived health

UNIVERSITY OF AMSTERDAM

NEURAL INFORMATION PROCESSING SYSTEMS

Results

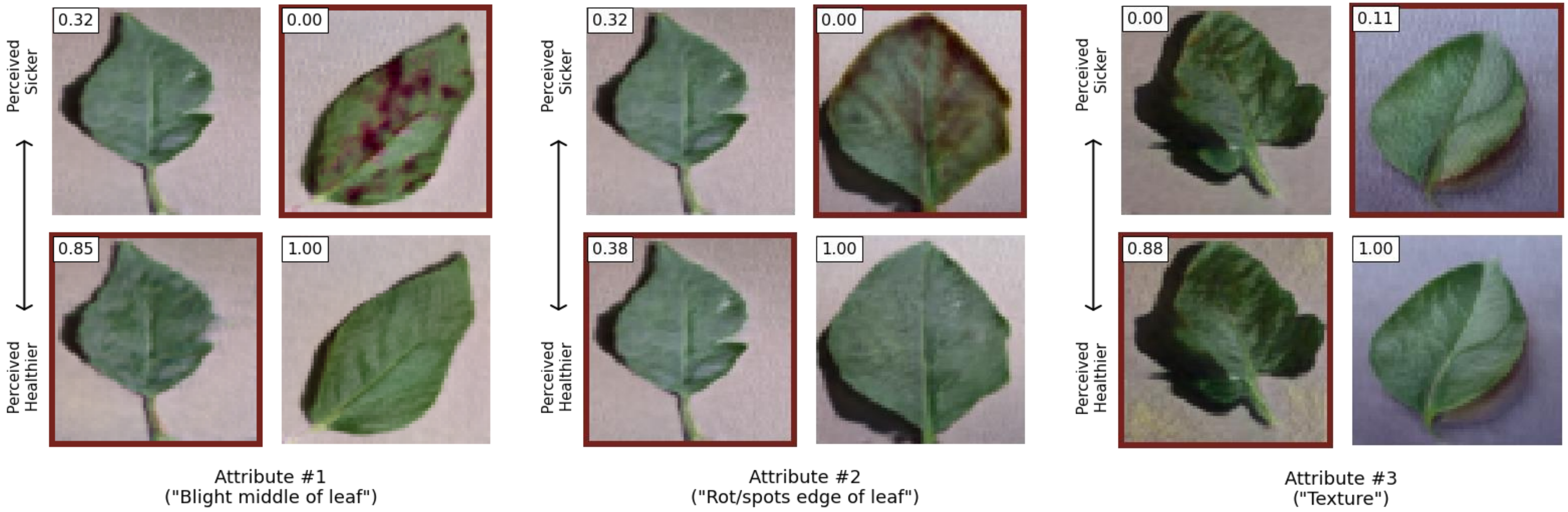
Attribute #1
("Eyebrow Thickness")

Attribute #2
("Facial hair")

= Counterfactual

p = Probability of being male

Results



□ = **Counterfactual**

p = **Probability of being healthy**

User study (n=54)

➢ **Classification study (coherence)**

- Users are shown two random examples of the same transformation *x*;
- Given two examples of transformation *x* and *y*, classify which is which.

➢ **Verbal Description Study (distinctness)**

- Users are shown 4 random animated images;
- Describe in 1-4 words the most prominent changing attribute.

User study (n=54)

➢ **Classification study (coherence)**

- Attributes are recognizable, but less so than in the original paper;
- Smaller image size? Training procedure subtleties?

➢ **Verbal Description Study (distinctness)**

- Common descriptor between the descriptions, one or two common words.

Dataset	Wu *et al.*	Lang *et al.*	Ours
FFHQ - Perceived Gender	0.783 (±0.186)	0.96 (±0.047)	Model 1: 0.52 (±0.2081) Model 2: 0.79 (±0.1599)
Plant Village - Perceived Health	0.91 (±0.081)	0.916 (±0.081)	0.66 (±0.323)

Table 2. User study results. Partial reproduction of Table 2 of the original paper, on a subset of the datasets.

UNIVERSITY OF AMSTERDAM

NEURAL INFORMATION PROCESSING SYSTEMS

Sufficiency

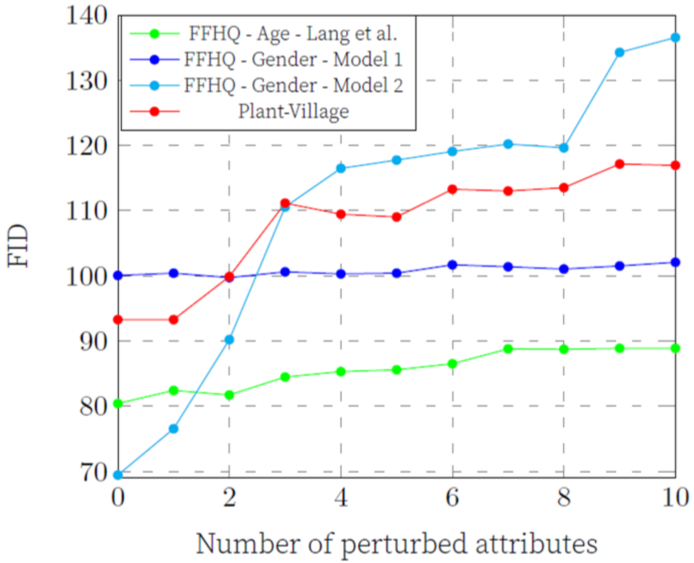
- Change top-10 attributes for image of class x, count images which flip to class y.
- The pretrained model has sufficiency within 1% of the reported value in original paper.
- Our models show significantly lower sufficiency.

Dataset	Ours
FFHQ - Perceived Age	94.8%
FFHQ - Perceived Gender (Model 1, $s = 2$)	51%
FFHQ - Perceived Gender (Model 2, $s = 1$)	21%
Plant Village - Perceived Health ($s = 2$)	30%

Table 1. Percentage of flipped classifications on different datasets. Row in *italics* shows our experiment on the original authors' model. s represents the shift size used to generate the results. The shift sizes have been chosen by qualitatively looking at the produced images.

Going beyond the original work

- We explore the effect of perturbing attributes on the quality of encoded images
- We find a steady increase in FID score over both datasets
- Suggests perturbing attributes results in unlikely combinations that are not seen in the original dataset (i.e. young boy with lipstick)



Conclusion & Discussion

➢ Numerical results not fully comparable

- Experimental results support claim 1&2 in the paper of our own models, albeit not as strong

➢ Does this fully refute the claims made? **No!**

- Computational limitations;
- Hyperparameter tuning;
- Training procedure.

Future directions

➢ Use more computational resources to clearly verify the posed claims

➢ Explore the effect of different classifiers on the detected attributes

Bibliography

[Lang *et al.*, 2021] https://arxiv.org/abs/2104.13369

[Lang *et al.*, 2022] https://ai.googleblog.com/2022/01/introducing-stylex-new-approach-for.html

UNIVERSITY OF AMSTERDAM

NEURAL INFORMATION PROCESSING SYSTEMS

Get in touch!

Noah van der Vleuten noahvdvleuten@gmail.com

Tadija Radusinović radusinovictadija@gmail.com

Rick Akkerman itsrickakkerman@gmail.com

Meilina Reksoprodjo meilinareksoprodjo12@gmail.com