Combining Counterfactuals With Shapley Values To Explain Image Models

Aditya Lahiri[1], Kamran Alipour[1], Ehsan Adeli[2] and Babak Salimi[1]

[1] University of California San Diego, [2] Stanford University,

[1] adlahiri@ucsd.edu,kalipour@eng.ucsd.edu, bsalimi@ucsd.edu, [2] eadeli@stanford.edu

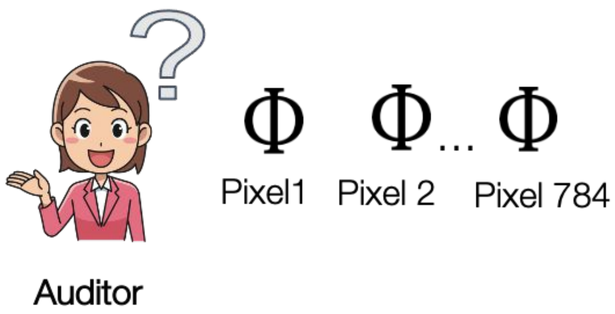
Algorithmic Decision Making

Individual Harms	Collective Social Harms
Hiring, Housing	**Loss Of Opportunity**
Credit, Goods' Pricing	**Economic Loss**
Liberty Loss, Surveillance	**Social Stigmatization**

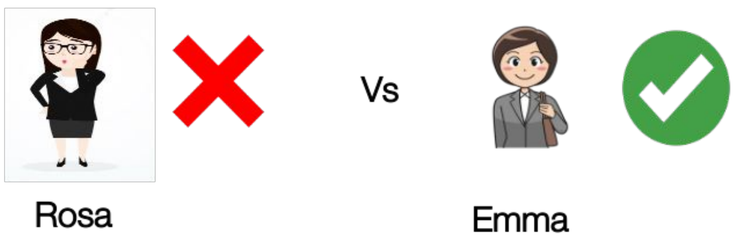
Source : Megan Smith, Former CTO of United States.

Shapley For Images: Challenges

High Dimensionality

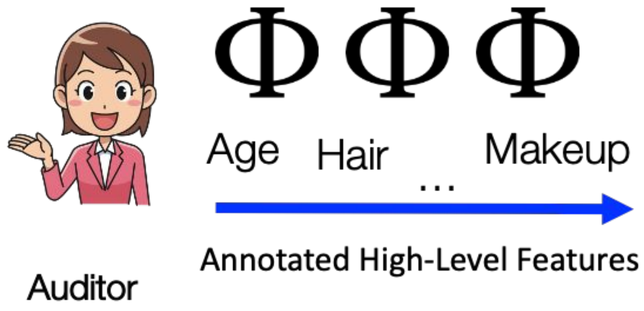


No Contrast

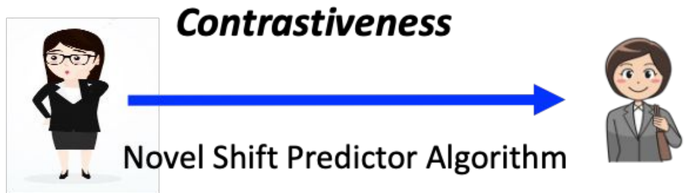


Proposed Modifications

Reducing Dimensionality



Contrastiveness

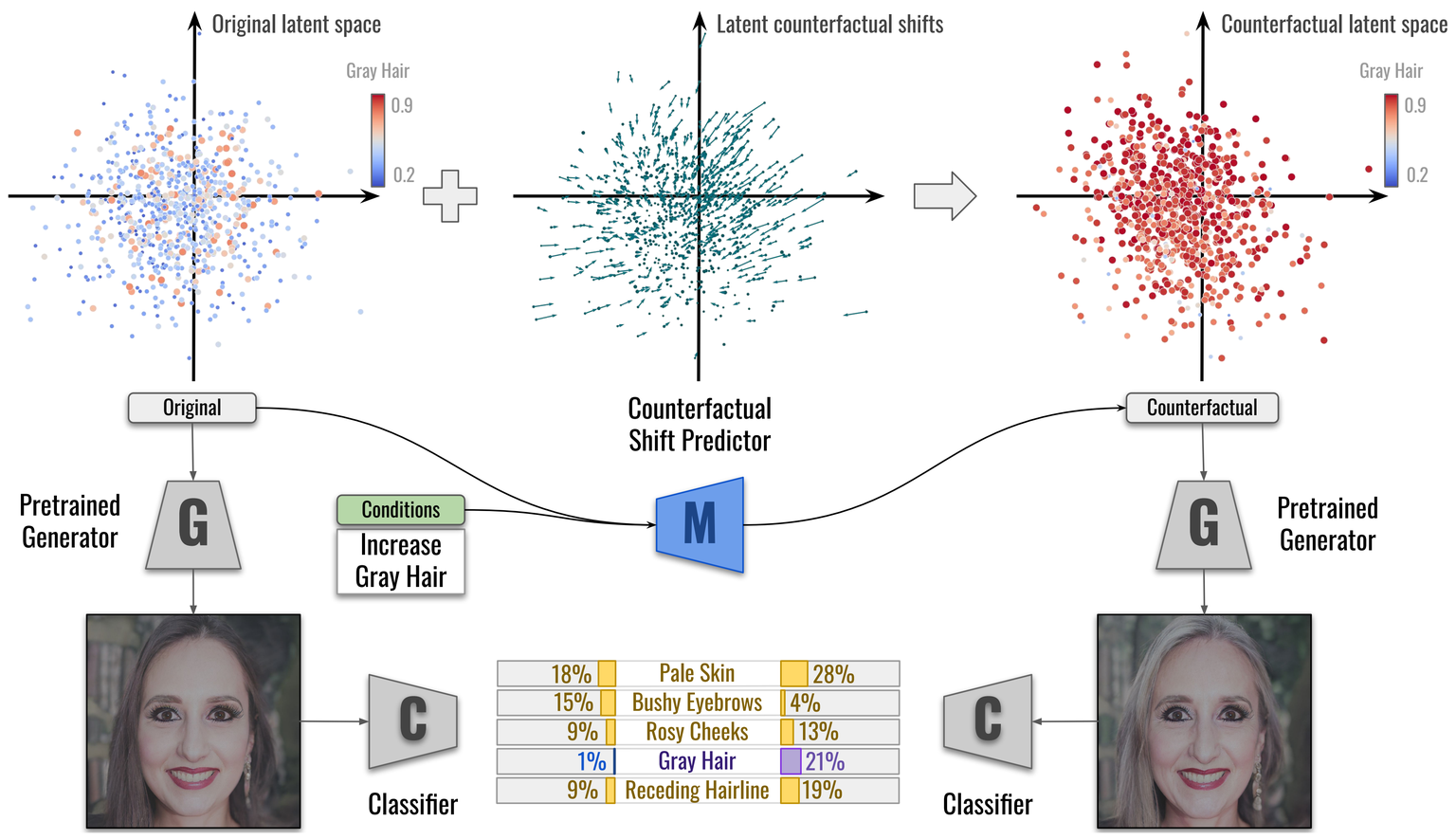


Shapley Values

3 desirable axioms :

- **Efficiency** Feature attributions add up the to difference between the individual's outcome and the average outcome.
- **Null** Features that do not affect model outcome in any way obtain a shapley value of 0.
- **Symmetry** Pair of features that interact with the model in a similar way, get same attributions.

Shift Predictor Algorithm



Counterfactuals in Shapley Value computation

<u>Challenge</u> → Find v(S) for any coalition S

<u>Setup</u>

Players S : high level attributes

Players can take 3 values → 0 (Same), +1 (Inc.), -1 (Dec.)

<u>Example</u> : GC={Age(A), Makeup(M), Young(Y)}, S=(A, M)

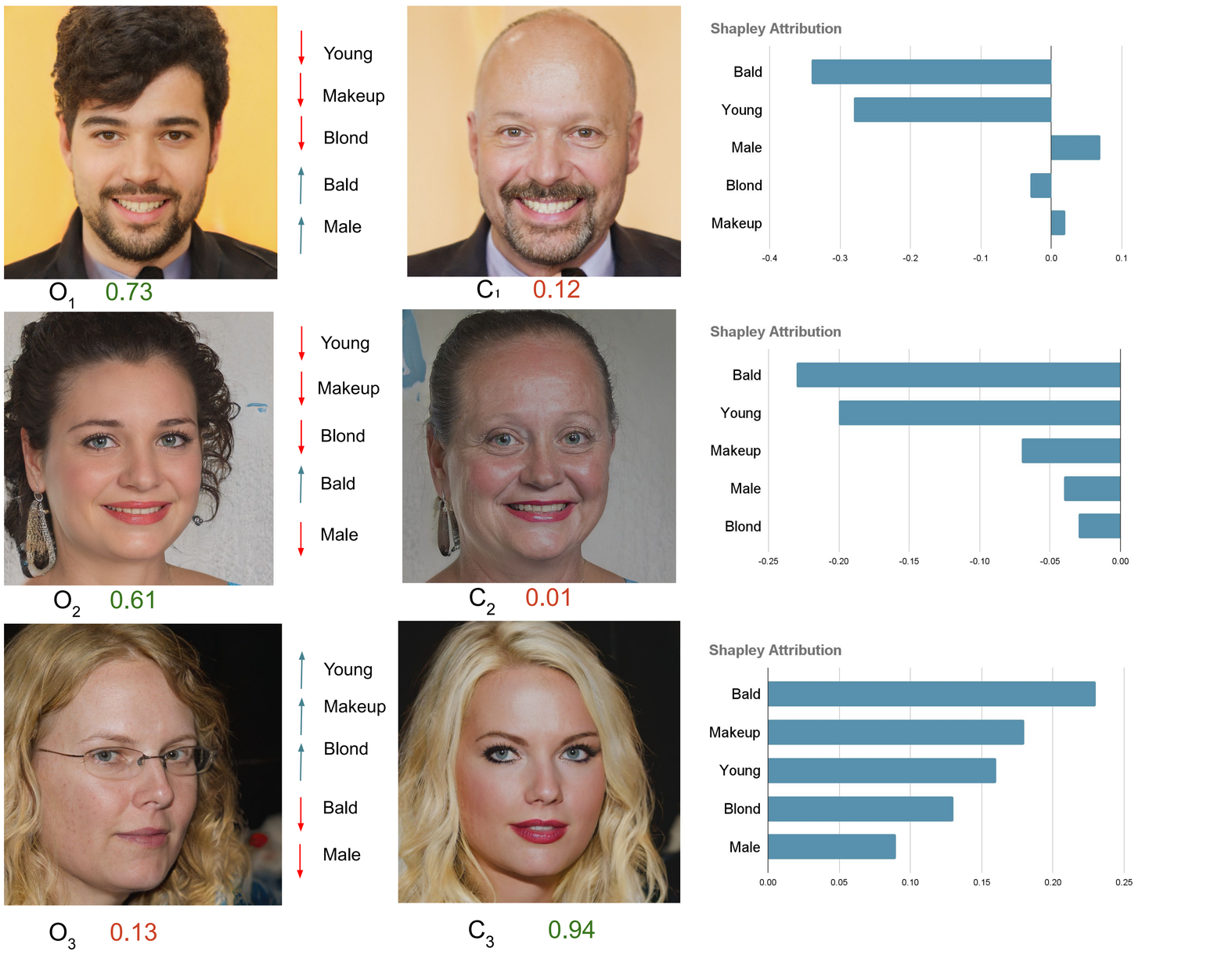
Original Image = (A = 0, M = 0, Y = 0)

Counterfactual = (A= 1, M = -1, Y = 0)

Out of coalition feature is 0(same) for both

User defined

Experiments And Results



Understanding behavior of a model trained on VGG backbone on CelebA dataset.