EUROPEAN CONFERENCE ON COMPUTER VISION

Global Counterfactual Directions

Bartlomiej Sobieski [1], Przemyslaw Biecek [1, 2]

[1]Warsaw University of Technology [2]University of Warsaw

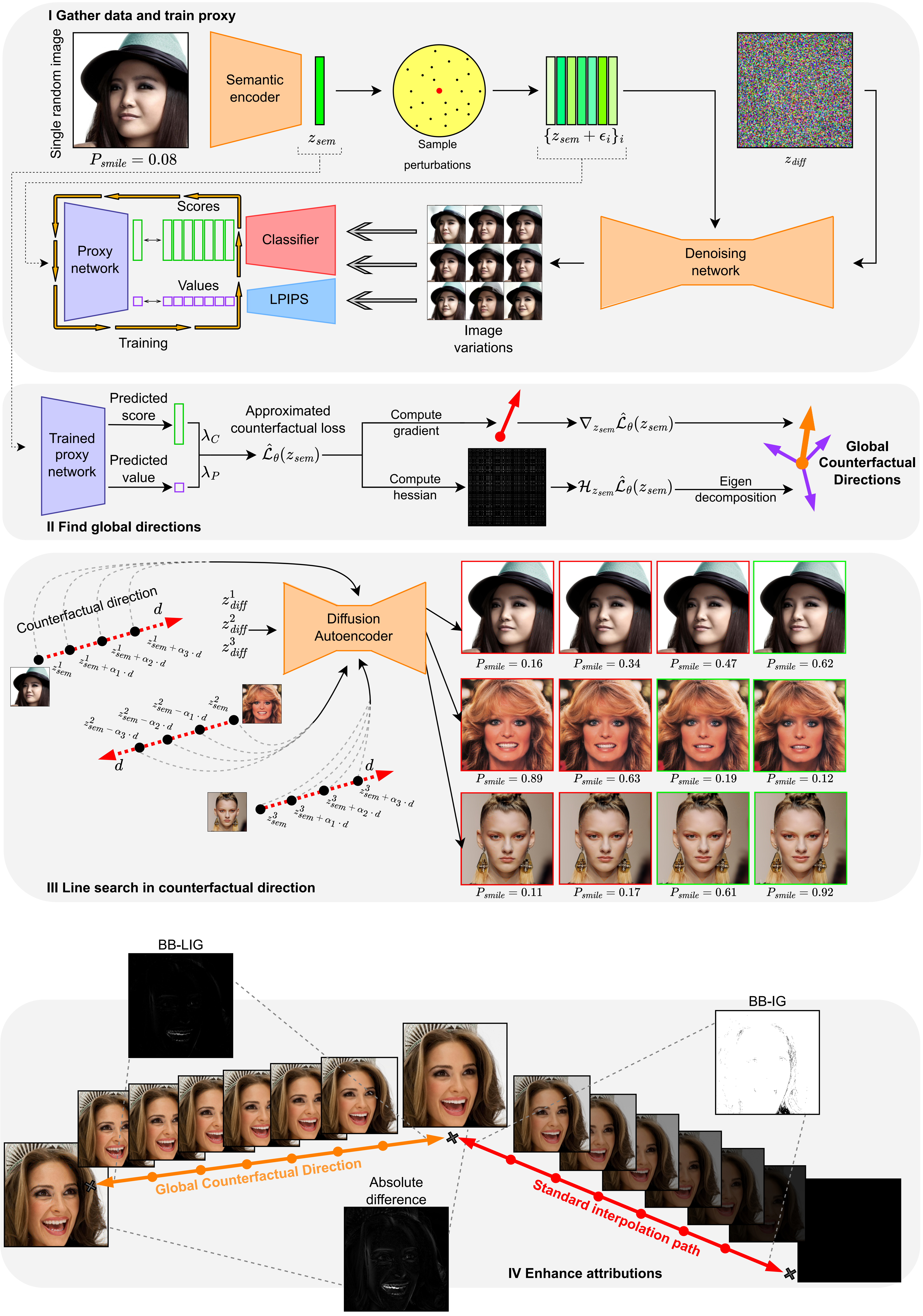
b.sobieski@student.uw.edu.pl

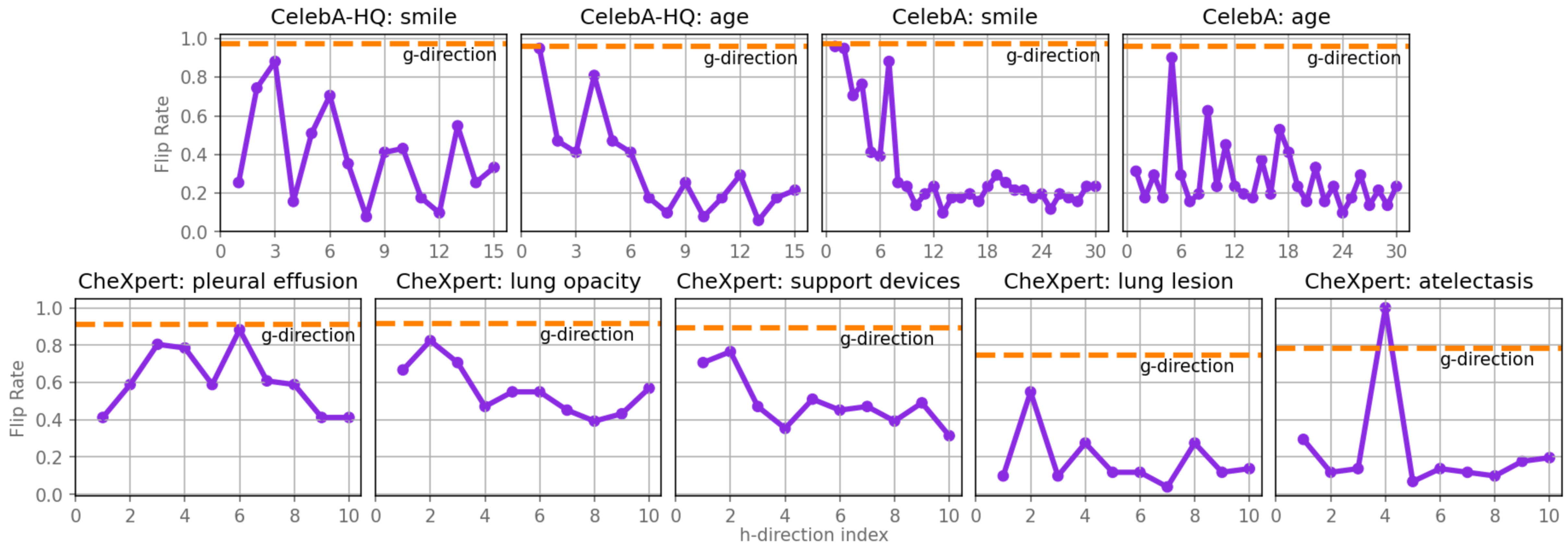
LinkedIn

Research group

Paper on arXiv

Let's talk about: explainable computer vision, counterfactuals, adversarial attacks, diffusion models





What are the problems?

Current state-of-the-art (SOTA) methods for generating visual counterfactual explanations (VCEs) assume full white-box access to the model of interest. Hence, once cannot apply them when the model's internal structure is not available, for example when using it through a closed-source API. Assuming black-box access is much more difficult, as one can only query the model with some input and receive the respective output.

Evaluating VCEs poses a significant challenge, even in the white-box scenario, where current methods lack any dedicated mechanisms for this task. Evaluating them in a black-box setting is even more challenging, as one cannot directly backpropagate through the model to utilize the gradient signal in any way.

SOTA approaches generate VCEs independently for each image. Although this methodology has proven to be very effective, deep neural networks are well known for utilizing global patterns present in the training data. Therefore, it seems encouraging to look for global patterns in VCEs as well, so as to improve the computational efficiency and open new research directions.

How do we solve them?

We propose to search for directions in the semantic latent space of Diffusion Autoencoders (DiffAE) that globally, i.e. for the entire dataset, flip the classifier's decision. Our methodology is fully black-box and requires only a single source image to discover a whole set of such Global Counterfactual Directions (GCDs).

By combining GCDs with Latent Integrated Gradients through a simple finite-difference approximation of the classifier's gradient, we obtain Black-Box Latent Integrated Gradients (BB-LIG) method – an effective mechanism for filtering out irrelevant changes and assigning high attributions to edits that actually influence the classifier's decision.

Some math maybe?

We approximate the local relationship between the predictions of a classifier f, the semantic similarity metric LPIPS s and the semantic latent representation $\mathbf{z}_{sem}$ of a given source image $\mathbf{x}$ with a small MLP proxy network p_ψ. By fixing $\mathbf{z}_{diff}$, the high-dimensional part of the representation in DiffAE, we ensure that only $\mathbf{z}_{sem}$ influences the image, which is desirable because this compact representation controls the semantics of the image. After training the proxy, we find the *g-direction* $\mathbf{d}_g$ with

$$\mathbf{d}_g = \nabla_{\mathbf{z}_{sem}}(p^f_\psi(\mathbf{z}_{sem}) + \lambda p^s_\psi(\mathbf{z}_{sem}))$$

and the *h-directions* by first computing the hessian

$$\mathbf{H} = \nabla^2_{\mathbf{z}_{sem}}(p^f_\psi(\mathbf{z}_{sem}) + \lambda p^s_\psi(\mathbf{z}_{sem}))$$

and then eigendecomposing it to get the most relevant eigenvectors. By moving along these directions in the semantic latent space, we are able to obtain a counterfactual explanation $\mathbf{x}'$, which functions as the *baseline* for the BB-LIG method:

$$BB\text{-}LIG_i(\mathbf{x}) = \frac{1}{m-1}(\mathbf{x}_i - \mathbf{x}'_i)\sum_{k=1}^{m-1}\frac{f(\tilde{\mathbf{x}}^{k+1}) - f(\tilde{\mathbf{x}}^k)}{\tilde{\mathbf{x}}^{k+1}_i - \tilde{\mathbf{x}}^k_i}.$$