Principal Counterfactual Directions

for black–box predictive models in computer vision

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

[1]MI2.AI, Warsaw University of Technology [2]MI2.AI, University of Warsaw

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

Explainable AI

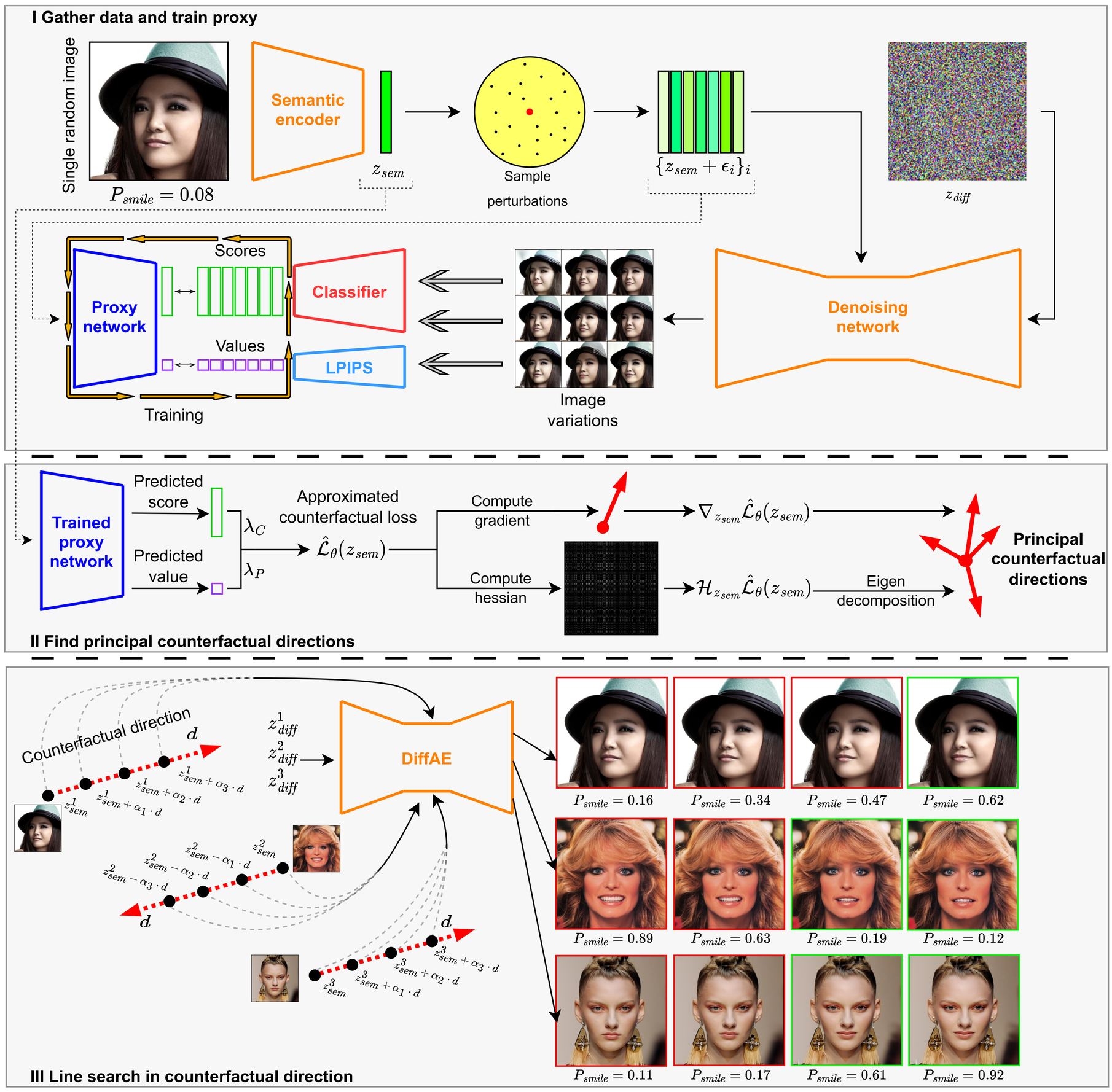
In the ever-evolving landscape of **artificial intelligence** (AI), the remarkable capabilities of machine learning models have revolutionized the way we interact with and depend on technology. From personalized recommendations to autonomous vehicles, AI systems are now deeply woven into the fabric of our daily lives. As these systems continue to advance and permeate diverse domains, so too does the need to understand and trust their **decision-making** processes. The premise of the area of **Explainable Artificial Intelligence** (XAI) is to develop tools and methods that allow for a deeper understanding of these complex models and make their decision-making process understandable to humans.

Counterfactual explanations

Counterfactual explanations help users understand why a specific decision was made by the model by exploring alternative scenarios. From a mathematical point of view, finding a counterfactual explanation for a classifier C, binary class y and an observation x is equivalent to solving the following optimization problem:

$$\min_{\tilde{x}} \mathcal{L}(\tilde{x}) = \min_{\tilde{x}} \mathcal{L}_C^y(\tilde{x}|x) + \mathcal{L}_S(x, \tilde{x}), \qquad (1)$$

where $\mathcal{L}_C^y(\tilde{x}|x)$ denotes the score assigned by classifier C to an observation $\tilde{x}$ that it represents an instance of the same class y as x, and $\mathcal{L}_S(x, \tilde{x})$ is a measure of similarity between the original observation x and its modification $\tilde{x}$. When considering machine learning models that operate on images, this problem can be understood as finding a minimal semantic change in the original image x to flip the model's decision, i.e. change its prediction to the opposite class.



Diffusion Autoencoders

Modifying an image in a semantically meaningful way is a necessity when trying to produce a **visual counterfactual explanation**. It is also a long-standing problem in the field of generative models and representation learning - **how to find a rich, compact representation of an image and properly manipulate it to change image attributes?** We propose to adapt one of the latest breakthroughs in these domains - **Diffusion Autoencoders** (DiffAE) - to generate counterfactual explanations for black-box models, i.e. predictive models for which only pairs of input data and predictions are available. DiffAE are an extension of classical diffusion models - a type of generative model that generates images by sequential denoising - that possess an additional semantic encoder. In this way, DiffAE allow for decomposing the traditional latent space of diffusion models into a two-part code. More specifically, for an image x, its latent representation z is composed of

- z_{diff} – the most noised, high-dimensional version of image x which is responsible for small, stochastic details of the final image,
- z_{sem} – a flat vector which controls the majority of the semantic attributes of the final image.

Principal Counterfactual Directions

In DiffAE, generating a modified version of an original image x is done by manipulating its semantic representation z_{sem} while keeping z_{diff} constant. Thus, there exists a direct relationship between z_{sem}, a modification $\tilde{x}$ of an image x that results from changing its z_{sem}, and classifier's prediction on $\tilde{x}$. We propose to approximate this relationship using a lightweight **proxy network**. By perturbing z_{sem}, thus generating variations of the original image, we can query the classifier of interest to obtain its predictions on new observations that closely resemble the original. Making use of a trained proxy, we propose two approaches of moving in the latent space of z_{sem} that result in flipping the classifier's prediction:

- *Gradient–based* – we input the original z_{sem} into our proxy and compute the gradient of a weighted sum of its outputs with respect to z_{sem},
- *Hessian–based* – we input the original z_{sem} into the proxy, compute the hessian $\mathcal{H}$ of a weighted sum of its outputs with respect to z_{sem} and perform an eigendecomposition of $\mathcal{H}$.

We empirically verify that, in both cases, the discovered directions not only allow for flipping the prediction for the original image, but are also fully transferable to the whole dataset of other images.

High-stakes decision making

Current evaluation benchmarks in the task of generating counterfactual explanations are typically based on solving simple problems such as identifying face attributes on the *CelebA* dataset or detecting the type of an object from *ImageNet*. We argue that such methods should be evaluated on more **challenging** tasks in which their use might be found practically helpful. With an increasing capacity of computer vision models in medicine, specifically in the detection of diseases connected with **X-ray images**, the *CheXpert* dataset has become a valuable source of data for creating models that perform on-par with human experts. We show that our approach can be adapted even to medical images and is able to generate meaningful counterfactual explanations.