University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Bartłomiej Sadlej
Student no. 429589

Title in English

Bachelor's thesis
in MACHINE LEARNING

Supervisor:
prof. Przemysław Biecek
Wydział Matematyki Informatyki i Mechaniki

Warsaw, May 2017

Abstract

This thesis explores the generation of invariant sets of images with respect to specific objective functions, particularly in the context of Computer Vision model interpretation. While numerous approaches exist for explaining model behavior through concepts discovery or interpretable approximation, we focus on a novel problem: generating sets of images that yield identical predictions under a given objective function, thereby establishing an equivalence relation.

Unlike adversarial example optimization, our research aims to develop an algorithm that effectively samples meaningful and diverse examples from these invariant sets, primarily leveraging Diffusion Models for image generation. Building upon recent advances in treating diffusion models as unsupervised priors for inverse problems, we extend existing frameworks to accommodate arbitrary objective functions defined in either image space or model neuron space.

Our work addresses a significant gap in current research, which has focused primarily on analyzing training data and model behavior on known inputs. By diving into the unexplored domain of alternative inputs yielding identical outcomes, we contribute to a deeper understanding of model behavior and interpretation. We demonstrate the effectiveness of our approach through extensive experiments and provide insights into the practical applications of invariant set generation for model explanation.

Keywords

blabaliza różnicowa, fetory σ-ρ, fooizm, blarbarucja, blaba, fetoryka, baleronik

Thesis domain (Socrates-Erasmus subject area codes)

11.4 Sztuczna inteligencja

Subject classification

D. Software
D.127. Blabalgorithms
D.127.6. Numerical blabalysis

Tytuł pracy w języku polskim

Tytuł po polsku

Contents

Wprowadzenie

Explaining Computer Vision models is an active area of research in machine learning due to many practical benefits that can be derived from understanding the model's behavior. Over time, diverse approaches have been proposed to find particular features responsible for the model's predictions or to approximate the model's behavior in a more interpretable way . In this work we focus on the problem of finding invariant sets of images that given any objective function result in the same prediction and constitute to a equivalence relation defined by this function and its range. Our goal is not to optimize for adversarial examples , rather to establish an algorithm which can effectivly sample meaningful and diverse examples from the invariant set.

citations

citations

Chapter 1

Related work

1.1. Diverse posterior sampling for Inverse Problem

In th past, a lot of work has been done on treating diffusion models as unsupervised priors to solve inverse problems for image restoration such as denoising, inpainting, super-resolution, and deblurring. Those methods were commonly designed to find a single best solution for the given image. More recently the focus has shifted towards producing a range of diverse and meaningful valid solutions for evey image Cohen et al. [2024]. Chung et al. [2024] demonstrate that score-based diffusion models (SGMs) Song et al. [2021] can efficiently handle noisy nonlinear inverse problems. In our setting, the mesurement almost never lives in the dimensionality of data, which is explored in Batzolis et al. [2021] and provides systematic comparison of best ways of estimating the conditional score.

add citations

1.2. Concepts, Spurious features and Clever Hans detections

Lapuschkin et al. [2019] develops an automatic pipeline to for exploring shortcuts and biases learned by the model which are commonly referred to as Clever Hans Pfungst [1911]. Neuhaus et al. [2023] investigates how to automatically find spurious features in the training data. Recently Dreyer et al. [2025] answers the question of what concepts were learned by the model and where in the training data they were present. While those works are great at analyzing the training data and model's behavior on it, we shift our focus on the unexplored area of other inputs which can lead to the same outcomes.

1.3. Classifier guidance for conditional image generation

Diffusion models allows for incorporating an additional guidance signal into the denoising process. Most common approaches are either to use classifier-free guidance Ho and Salimans [2022] and guide the generation with class label or utilise text prompt Rombach et al. [2022]. Both approaches doesn't allow for incorporating new classes or text conditioning models at inference time either due to fixed network architecture or distributions mismatch. One way to include any classifier as a guidance signal is to map the noised image to its clean version at each step through the reverse processJeanneret et al. [2022] which is very computationally expensive or use Tweedie Formula [Robbins, 1992, Chung et al., 2022, Weng et al., 2024] to estimate the final solution which although introducing some bias works well in practise but limits the number of optimization steps to the number of diffusion steps. Augustin et al. [2024] uses classifier guidance to optimize initial noise fed to the diffusion model and makes it computationally feasible due to adoption of Latent Diffusion Models and moving from pixel space to latent space. We set out to further extend this framework to arbitrary objective function defined either in the image space or in the model neurons space.

Chapter 2

LOREM IPSUM

Bibliography

Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. Dig-in: Diffusion guidance for investigating networks – uncovering classifier differences neuron visualisations and visual counterfactual explanations, 2024. URL https://arxiv.org/abs/2311.17833.

Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models, 2021. URL https://arxiv.org/abs/2111.13606.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems*, 2022.

Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems, 2024. URL https://arxiv.org/abs/2209.14687.

Noa Cohen, Hila Manor, Yuval Bahat, and Tomer Michaeli. From posterior sampling to meaningful diversity in image restoration, 2024. URL https://arxiv.org/abs/2310.16047.

Maximilian Dreyer, Jim Berend, Tobias Labarta, Johanna Vielhaben, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Mechanistic understanding and validation of large ai models with semanticlens, 2025. URL https://arxiv.org/abs/2501.05398.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations, 2022. URL https://arxiv.org/abs/2203.15636.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), March 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08987-4. URL http://dx.doi.org/10.1038/s41467-019-08987-4.

Yannic Neuhaus, Maximilian Augustin, Valentyn Boreiko, and Matthias Hein. Spurious features everywhere – large-scale detection of harmful spurious features in imagenet, 2023. URL https://arxiv.org/abs/2212.04871.

Oskar Pfungst. *Clever Hans (the Horse of Mr. Von Osten): A Contribution to Experimental Animal and Human Psychology*, volume 8. Holt, Rinehart and Winston, 1911.

Herbert E Robbins. An empirical Bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, 1992.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Nina Weng, Paraskevas Pegios, Aasa Feragen, Eike Petersen, and Siavash Bigdeli. Fast diffusion-based counterfactuals for shortcut removal and generation. In *European Conference on Computer Vision*, 2024.