CS 726 Final Project Presentation

Team TAVA

Problem Statement

Using diffusion models to efficiently generate counterfactuals

Motivation

Diffusion models have gained popularity as a powerful tool for analyzing complex systems and predicting their behavior over time. By simulating the spread of a particular phenomenon or variable, these models can help us understand how different factors may interact and affect outcomes. So we plan to use diffusion models to generate meaningful counterfactuals.

Solution Approach

The generation of counterfactuals involves 3 steps

1) Abduction
2) Intervention
3) Prediction

We plan to perform these steps using the diffusion model.

Related literature

- **Deep Structural Causal Models for Tractable Counterfactual Inference** - This paper explores the use of Normalising Flows to ensure reversibility of the generation process, which makes the abduction step realisable.
- **Diffusion Causal Models for Counterfactual Estimation** - Here, the authors perform abduction of the noise by utilising a relation between Denoising Diffusion Implicit Models and neural ODEs, which leads to deterministic inference of the noise. They generate counterfactuals using an 'anti-causal predictor', which essentially scores the counterfactual object while generation.
- **Diffusion Models for Counterfactual Explanations** - This paper uses guided diffusion model for generation, and modifies the loss/score function to generate counterfactual objects in a fairly intuitive manner.

Related literature

Proceedings of Machine Learning Research vol 140:1–21, 2022 1st Conference on Causal Learning and Reasoning

Diffusion Causal Models for Counterfactual Estimation

Pedro Sanchez PEDRO.SANCHEZ@ED.AC.UK
Sotirios A. Tsaftaris
The University of Edinburgh

Editors: Bernhard Schölkopf, Caroline Uhler and Kun Zhang

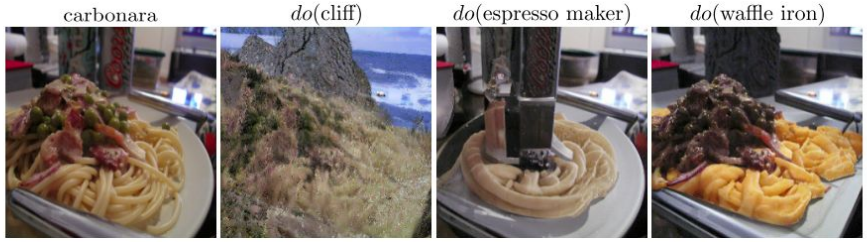

Figure 1: Counterfactuals on ImageNet 256x256 generated by Diff-SCM. *From left to right*: a random image sampled from the data distribution and its counterfactuals *do*(class), corresponding to "how the image should change in order to be classified as another class?".