CS 726 - Advanced Machine Learning

Using Diffusion Models to Generate Counterfactual Objects

Prof. Sunita Sarawagi

Spring 2022-23

TEAM TAVA :

Veda Pranav P - 190050094

Akash Reddy G - 190050038

Aniket Gudipaty - 190050041

Thivesh Chandra M - 190050124

May 3, 2023

Contents

1 Task Description

There are three stages to generating a counterfactual object:

1. Abduction
2. Action
3. Prediction

For generation of counterfactual objects using diffusion models, the most difficult step would be the abduction step, because of the non-deterministic irreversible nature of the generation process. Hence, the main task would be to come up with an effective abduction algorithm/heuristic. The other two steps will be based on that.

2 Related Works

- **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.

3 Work Splitup

We all searched and read the papers regarding this topic.