# Utilization of GAN for Automatic Evaluation of Counterfactuals: Challenges and Opportunities*

Itisha Kothiyal[1], Anish Patil[1], Vitor Horta[2], and Alessandra Mileo[2]

[1] Dublin City University, Dublin, Ireland
{itisha.kothiyal2,anish.patil4}@mail.dcu.ie
[2] Insight Centre for Data Analytics, Dublin City University, Dublin, Ireland
{vitor.horta,alessandra.mileo}@insight-centre.org

**Abstract.** Over the past few years, Explainable Artificial Intelligence (XAI) has grown significantly due to the fact that successful deep learning models are still difficult to understand and interpret. XAI aims to enable better interpretability of the classifications made by the neural networks for humans. In XAI research, counterfactual explanations are proven to be very effective in explaining the model's mistakes, describing what changes should be applied to a particular input (image in our case) to attain the correct classification [1]. However, systematic evaluation of counterfactuals is challenging and requires substantial human input.

In this work we focus on evaluating semantic textual explanations (expressed as attribute-value pairs) on birds classification (CUB-200-2011 [2] dataset). Being textual and not visual, the explanations are hard to systematically validate through the CNN. To tackle this problem, we experimented on the use of Generative Adversarial Networks (GANs) to modify a misclassified image based on a textual counterfactual. The resulting counterfactual image generated affected the original misclassification outcome but not as strongly as we expected. We believe this could be due to the known problem of poor resolution in GAN-generated images, and the challenge of multiple attribute modifications.

This paper reports on the challenges of using GANs to systematically assess the quality of textual counterfactuals via counterfactual image generation.

**Keywords:** Explainable AI (XAI) · Counterfactual explanations · Generative Adversarial Networks (GAN) · Computer Vision.

## References

1. R. M. J. Byrne, "Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning," www.ijcai.org, pp. 6276–6282, 2019, [Online]. Available: https://www.ijcai.org/proceedings/2019/876
2. "Perona Lab - CUB-200-2011," www.vision.caltech.edu. https://www.vision.caltech.edu/datasets/.
3. V. A. C. Horta, A. Mileo: Generating Local Textual Explanations for CNNs: A Semantic Approach Based on Knowledge Graphs. AI*IA 2021: 532-549

* Supported by Science Foundation Ireland Grant no. SFI/12/RC/2289_P2.