Utilization of GAN for automatic evaluation of counterfactuals: Challenges & Opportunities

Itisha Kothiyal[1], Anish Patil[1], Vitor A.C. Horta[2] and Alessandra Mileo[2]
[1] Dublin City University, Dublin, Ireland
[2] Insight Centre for Data Analytics, Dublin City University, Dublin, Ireland

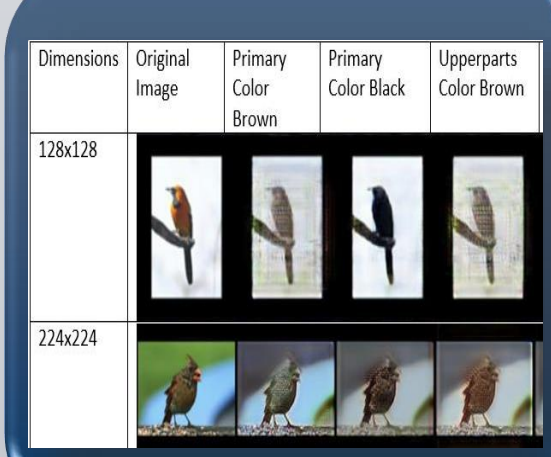
Insight SFI RESEARCH CENTRE FOR DATA ANALYTICS

A World Leading SFI Research Centre

Science Foundation Ireland SFI For what's next

1 Demand for **accountability and transparency** in AI models.

Semantic Counterfactual Explanations: **Pathway to Explainability**

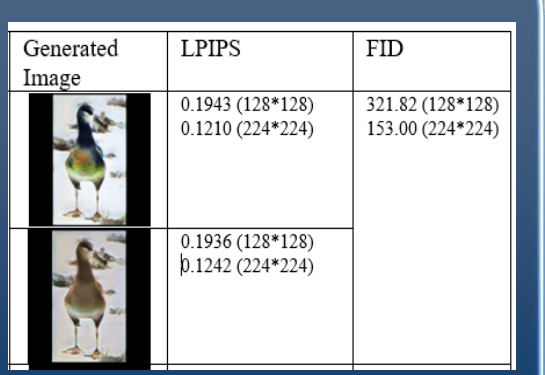


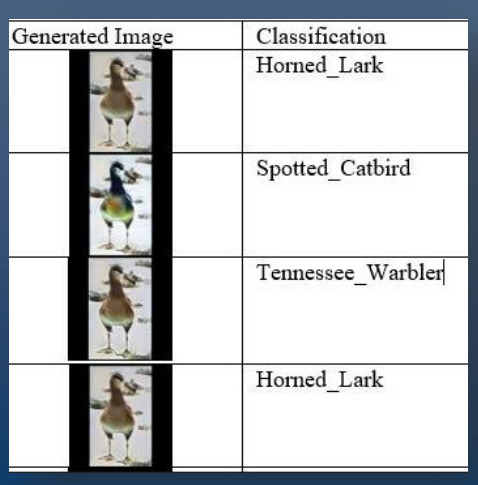
2 Systematic evaluation of **counterfactuals** is challenging.

Implemented StarGAN on **birds classification (CUB-200-2011 [2] dataset).**

4 Image Classifier trained on **higher resolution of 224*224** could possibly bring better results.

Specific attributes tend to have **more influence** in a counterfactual than others.





3 **Improvement in Prediction score** of the expected class for some attributes.

LPIPS and FID scores were found to be better for higher resolution of 224*224.

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

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

HOST INSTITUTIONS

PARTNER INSTITUTIONS

FUNDED BY: