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@mail.dcu.ie
anish.patil4@mail.dcu.ie
[2] Insight Centre for Data Analytics, Dublin City University, Dublin, Ireland
vitor.horta@insight-centre.org
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 aimns to enable better interpretability of the judgments/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 updates could be done in a particular image to attain the correct classification [16]. However, systematic evaluation of counterfactuals is challenging. This paper reports on the challenges of using GANs (Generative Adversarial Networks) to assess the quality of counterfactuals, using the CUB-200-2011 [9] birds dataset as a case study.

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

1 Introduction

Due to the demand for accountability and transparency when using AI models for making critical decisions, explainable artificial intelligence (XAI) has emerged as an active research area. Modern Convolutional Neural Networks (CNNs) have excelled in Computer Vision, but it is still difficult to explicit their decision-making process and validate the same, especially when errors are made.

Counterfactual explanations (or counterfactuals) are widely used in the XAI community. A counterfactual, given an input question, is a version of the input with minimal (or sparse) but meaningful (easily interpreted by humans) alterations that can correct (or even invert) [2] the model's output [3]. An example of counterfactual generated from the model in [1] is as follows: "If the attribute primary color of input image had the value brown and underparts color had the value brown and upper-parts color had the value brown, input image would more likely be classified as a Black Footed Albatross instead of a Blue Jay."

Counterfactuals have proven to be very effective for explainability, however there is no benchmark for their systematic evaluation. Hence, there is a strong

dependency on human-intervention to validate their efficacy as a semantic explanation. In this paper we explore the possibility of using GANs to constructively and systematically validate semantically generated counterfactuals.

GAN (Generative Adversarial Network) is a type of deep neural network that uses a zero-sum game between a discriminator and a generator. In GANs, a discriminator is a network which trains to be able to distinguish and classify real and fake images. It returns a metric which is a probability of how sure it is that an input image is a real image. A high score from the discriminator indicates it to be a real image. The generator, on the other hand, has the task of fooling the discriminator by generating fake images. As the GAN trains, it is able to generate images that are more and more realistic to the point that they are almost indistinguishable from real images. Our goal in this paper is to assess counterfactuals produced by the model in [1] by training a GAN on the CUB-200-2011 [9] birds' dataset on a the following attributes: primaryColor, underpartsColor and upperpartsColor. We selected 15 attributes for primaryColor, underpartsColor and upperpartsColor, out of the total 312 attributes for each bird.

Our objective was to train a GAN to produce perceptually good results on test samples for the CUB-200-2011 dataset that represent the counterfactuals generated in [1], and use such results to validate the counterfactuals. In order to assess the GAN-generated images were perceptually good, we used metrics like LPIPS (Learned Perceptual Image Patch Similarity) [11] and FID (Fréchet Inception Distance) [10] scores. While FID takes distance of feature vectors into account and works well for diverse datasets, LPIPS helps in calculating perceptual similarity which is similar to the way a human would perceive an image. Validation of the counterfactual was intended as an inverted classification for the GAN-generated image of the counterfactual from [1], when fed into a VGG-16 model pre-trained over ImageNet and fine-tuned on CUB200.

To this aim, we first compared how different GAN (namely DCGAN, StarGAN and AttGAN) would perform on the CelebA dataset. We then mobed to the pre-trained Style-AttnGAN model on CUB-200-2011 dataset. In doing so, we encounter several challenges that demonstrated the limitation of using GAN architectures for a systematic evaluation of counterfactuals. In this paper we want to report on such challenges and how they got manifested as we performed our evaluation, as well as discuss what potentially we could have done differently to get closer to our target.

These challenges include i) limited reproducibility for GAN approaches on the CelebA dataset, specifically in terms of compatibility of libraries; ii) insufficient resolution of GAN-generated images compared to the resolution of images used to train the original classifier; iii) suitability and robustness of metrics for validating the counterfactual quantitatively. The remainder of the paper is as follows: Section 2 reviews related works on vaildation of counterfactuals, including using GAN; Section 3 provides details about the datasets and data pre-processing before training GAN; Section 4 presents our methodology and experiment setup; Section 5 discusses our results; Section 6 presents improvements and challenges, ending with conclusions and future works in Section 7.

2   Related work

The framework described in[2], generates visual counterfactuals for the classifier while using a conditional GAN for transparent decision-making process in healthcare applications. The methodology revolves around perturbation of the original input image such that an explanation function is designed using cGAN keeping in mind the three properties of valid transformation namely: data consistency, classification model consistency and context-aware self-consistency. They used metrics like FID (Fréchet Inception Distance), CV (Counterfactual Validity) [12] and FOP (Foreign Object Preservation) to support the above three properties for valid transformations. FOP is a metric that they devised which helps in measuring if the patient-specific properties (foreign objects) are retained in the image. The intent of the task is similar to our experiment with a difference that in their method, counterfactuals are already images. However, in our experiment the counterfactuals are a human-readable text, that is easier for humans to interpret but difficult for testing and validating. Additionally, their work is specifically for healthcare service line and has been tested on celebA, MNIST and simulated data.

STEEX (STEering counterfactual EXplanations using semantics) model implemented in [3] makes use of the latest advancements made in the area of semantic-to-real image synthesis in order to achieve "region-targeted counterfactual explanations" (a concept introduced by the authors), which is the highlight of the paper. Their prime attention is to spotlight the how content-based image classification is vital than only region-based classification which is absolutely true when considering scenarios where safety is of utmost importance like self-driving cars. The metrics used for evaluation in their paper include: FID, Face Verification Accuracy (FVA) and Mean Number of Attributes Changed (MNAC), and their model has been trained for CelebA, CelebAMask-HQ and BDD100k datasets. Similarly to [2], this model updates the query image to produce the counterfactual image. Their method involves no textual explanations for producing a counterfactual image.

The approach in [4] relies on Causal Concept Effect (CaCE) measure for reducing errors arising from confounding. Their model relies on Variational AutoEncoder (VAE) and the VAE-CaCE metric, proposed to estimate the true concept causal effect. In order to showcase the effectiveness and generality of the CaCE metric, authors have tested their model on different datasets including MNIST, COCO Miniplaces and CelebA. However, reproducibility is very limited as no link to code is provided. Similarly to our approach, [4] uses StarGAN for producing the counterfactual images, but it focuses on the CaCE metric and does not provide any detail about the original counterfactual explanation from which the image is generated.

PIECE (PlausIble Exceptionality-based Contrastive Explanations) combines a GAN to create counterfactual and semi-factual explanatory images, with a CNN that makes predictions [13]. Similarly to our approach, PIECE uses semantic explanations as a base to generate the counterfactuals, but it has been tested only for CIFAR-10 and MNIST, producing very low resolution images that

are not likely to invert the classification for attribute driven counterfactuals in more complex datasets such as CUB-200-2011.

Based on the above research targeting the issule of counterfactuals' validation, current approaches are either limited to a specific domain like healthcare or tested on simple datasets like MNIST and CIFAR. We risked by taking the research in this area a step further when considering a dataset such as CUB-200-2011, which contains 312 attributes for each bird with a total collection of 11,788 images belonging to 200 class labels. Additionally, we needed looked at recent metrics which could measure the feature vector distance in the images and also indicate their perceptual similarity like FID and LPIPS scores.

3   Datasets, data preparation and training

Our experiments are conducted on two open source datasets.

The **CelebA** dataset [14] contains 200,000 celebrities face images annotated with 40 binary attributes and 5 landmark locations. We used this dataset to test attribute-based image editing and generation using GANs.

The **CUB-200-2011** dataset is an extended version of CUB-200 and has 11,788 images of 200 different birds. Each image is annotated with 15 part locations, 312 binary attributes and 1 bounding box. We randomly select 2,000 images as a test set and use all remaining images for training StarGAN. The main challenge in this dataset is that there is a huge variation and confounding features in background information compared to subtle inter-class differences among birds. We used this dataset in order to (i) generate and modify images by changing the attributes based on GAN pipeline, and (ii) validate the counterfactuals generated in [1].

Data Preprocessing. Images in the CelebA dataset have been cropped to 64x64 dimension and resized to further reduce the size by removing background clutter. As done in the original implementation of DCGAN, training images were thus scaled to the range of [-1, 1] of the tanh activation.

In the CUB-200-2011 dataset, each image has different dimensions and resolution, making it a complex dataset for training GAN, especially for larger images. In fact, the maximum resolution of generated images with GANs is limited to the resolution of the images used for training [14]. Also, the images in the dataset are not aligned and cropped. An image size of 224x224 dimensions was taken so as to get maximum resolution for the training purpose and a crop size of 512x512 was taken so that each image in the dataset is clearly visible to the generator and discriminator networks in the GAN, in order to increase the overall accuracy of generated images.

Training. For DCGAN, the model was trained with the L2 loss and the Adam optimizer with learning rate of 0.0002. It consists of 9 convolutional layers which were separated by batch normalization and Leaky ReLU and followed by one

fully connected layer. Training took about a day as it was done only using the CPU. When training the autoencoders, the images were split into a 70/15/15 ratio for training, testing and hyper-parameter tuning respectively.

For StarGAN, the model was trained using Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The images were flipped horizontally with a probability of 0.5 for the purposes of data augmentation. One generator update after five discriminator updates was performed. The batch size was set to 10 for all the experiments. Learning rate of 0.0001 was taken for the first 100,000 iterations and linearly decay the learning rate to 0 over the next 100,000 iterations of training. The attributes used for training StarGAN on CUB-200-2011 dataset were: primary_color (brown, black), underparts_color (brown, black) and upperparts_color (brown).

We also reproduced the tensorflow implementation of AttGAN [7] for CelebA dataset initially and thereafter on the custom cartoon dataset. The model uses 128x128 images for training and is trained on the Adam Optimizer ($\beta_1 = 0.5$,$\beta_2 = 0.999$) with 2e-4 learning rate. The model was trained for 60 epochs on a batch size of 32 for CelebA dataset, 500 epochs on a batch size of 32 for cartoon dataset and 20 epochs on a batch size of 35 for CUB-200-2011 dataset. The attributes used for training AttGAN on CUB-200-2011 dataset were: primary_color (blue, brown, black, grey and white), underparts_color (white, grey, blue, black and brown) and upperparts_color (white, black, grey, blue and brown).

4   Methodology and Experimental Setup

Our approach uses GANs to modify images by altering their attributes and use them to represent textual counterfactuals in [1] as visual counterfactuals that can be validated when passed through the pre-trained model in [1].

For validation, we implemented and compared DCGAN-encoder network [5], AttGAN [7], pre-trained Style-AttnGAN [8] and StarGAN [6]. DCGAN is a deep convolutional GAN with encoder networks for the generator and works by taking an input image and generating a similar image with specific characteristics by manipulating the vectors in the latent space [5]. StarGAN is an image to image translation model for multiple domains [6] which can change only one attribute at a time, while AttGAN allows changing more than one attribute of a single image at a time [7]. A pre-trained Style-AttnGAN [8] is a model which takes text captions as an input and generates images based on those captions.

Our GAN pipeline to generate visual counterfactuals starts by considering the celebA dataset and the above-mentioned GANs. The focus here was to manipulate attributes to generate a resulting image with changed attributes values.

DCGAN-Encoder networks. We started by using the encoder-decoder approach of DCGAN as in [5] for generating visual counterfactuals on the CelebA dataset, training the model for 25 epochs with learning rate of 0.0002 was used by using the Adam optimizer until the loss curve converged.

The trained generator and discriminator of the DCGAN was then used to build the encoder architecture. The GAN auto-encoder is used in such a way that it takes an input image and produces a z vector in the latent space, which produced an image close to the original one. CelebA dataset was divided into training, validation and test sets. To train the encoder model, training set of the CelebA was used by giving the input of an image from the dataset and measuring the loss as the result of any given similarity metric. Finally, we sampled the images from the validation set and run them through the encoder and decoder architecture to get the qualitative results.

The attributes in the CelebA dataset were represented as vectors in the latent space and each image in the dataset is labeled with one of the 40 binary attributes where 1 denotes the attribute is present and -1 denotes the attribute is absent for that corresponding image. Z vectors representing these attributes were calculated first by subtracting the average z vectors of all the images which do not have the specific attributes from the average z vectors of all the images which have the specific attributes in the training data [5]. The images were manipulated by first encoding the image in the latent space by using the encoder network and then adding the attribute z vector to the encoded image vector. This was done for manipulating the images with different attributes.

AttGAN and Style-AttnGAN. Next steps of the experiment focused on using AttGAN [7] for CelebA dataset to check the clarity of the results during sampling.
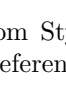

Fig. 1: Training sample at epoch 19 iteration 215 for AttGAN Implementation on CUB-200-2011 dataset.

We further checked the AttGAN implementation for another custom dataset i.e., cartoon dataset (which has 18 attributes and 10,000 cartoon face images). The results obtained for both training and testing samples were fairly adequate.

Next step for using AttGAN was to make it work for the CUB-200-2011 dataset to get similar or better results in comparison to StarGAN for changing the semantic attributes of the birds. However, as it is evident from Fig. 1, which highlights the training sample at epoch 19, the networks failed at manipulating the attributes of the images and generated same image for all the image samples. Even though the images generated by AttGAN were close to the real images, it could not generate the images with different attributes. One possible reason for this could be the less number of epochs and iterations used for the training.

We also executed the pre-trained model of Style-AttnGAN [8] on CUB-200-2011 dataset, which transforms the text captions provided to the model into

Reference Bird Name	Counterfactual Explanation	Converted Caption for Style-AttnGAN	Original Image from CUB-200 Dataset	Generated Image from Style-AttnGAN
Black Footed Albatross	If the attribute primary color of input img001 assumed the value brown and underparts color assumed the value brown and upperparts color assumed the value brown, input img001 would more likely be classified as a Black Footed Albatross instead of Blue Jay	bird with primary color as brown and underparts color as brown and upperparts color as brown		

(a)

Bird Name	Original Image	Image generated	Image Caption Reference	LPIPS score (AlexNet)	FID
American Crow			American_Crow_0031_25433	0.4694	226.78
Yellow billed Cuckoo			Yellow_Billed_Cuckoo_0006_265 78	0.6598	233.39
Purple Finch			Purple_Finch_0011_27633	0.5880	173.20

(b)

Fig. 2: (a) Images generated from Style-AttnGAN [8] from the text captions. (b) FID and LPIPS scores for reference image and resultant image from Style-AttnGAN [8].

images. Since our goal through this experiment is to validate the textual counterfactual explanations, so this approach of Style-AttnGAN [8] would have proved advantageous if it could produce attribute changes as mentioned in the text captions to the same image of which the counterfactual explanation was produced. But Style-AttnGAN [8], produces random images of the birds based on the text captions provided for test-samples. The resultant bird images fetched post testing are described in Fig. 2, along with their variation from the reference image. The text captions provided here were based on the counterfactual explanation generated from the VGG-16 [1]. As illustrated in Fig. 2, the mentioned captions generated random images of birds which could not satisfy the preliminary requirement for the validation of the counterfactual as illustrated in the example provided Section 1.

StarGAN. Our next approach was to test StarGAN [6] on the CelebA dataset to check the feasibility, code reproducibility and the clarity of the results obtained. StarGAN, proposed in [6] is a scalable image-to-image translation model meaning it aims on changing a particular aspect of a given image to another image by using a single generator and a discriminator of the generative adversarial networks. Fig. 4 shows the facial attribute transfer results on CelebA dataset for the facial attributes 'Black Hair', 'Blond Hair', 'Brown Hair', 'Male' and 'Young'. The method proposed in [6] provides considerable higher visual quality on test data compared to the experiment using DCGAN-encoder networks on CelebA dataset. This could be due to regularisation effect of StarGAN using a multi-task learning framework [6]. An interesting approach followed by using StarGAN is to train the model to flexibly translate images according to the labels of the target domain rather than training the same model to perform a static and fixed translation which could lead to overfitting of the model.

Validation of counterfactuals with StarGAN. The primary goal of this paper is to test if images generated by GANs by changing image attributes are suitable for validating textual counterfactuals on the CUB-200-2011 dataset.

Out of all the GANs we implemented for generating and manipulating the images of the CelebA dataset, StarGAN generated more visually appealing images. Hence, StarGAN was selected for assessing the counterfactual explanations in [1]. Our next steps to validate the counterfactual explanations, involved replicating the StarGAN for the CUB-200-2011 dataset.

We trained the model with different setups: with image sizes of 128x128 and 224x224 and crop size of 512x512. After generating the modified images, we passed them through the CNN to validate whether they would invert the classification from the wrong class to the correct class as per our hypothesis.

5   Results

Results are presented considering the two key objectives of this paper, namely (i) quality of counterfactuals generated using GAN and (ii) results of counterfactuals validation using GANs. In what follows we present the former by comparing different GANs and the latter by focusing on StarGan.
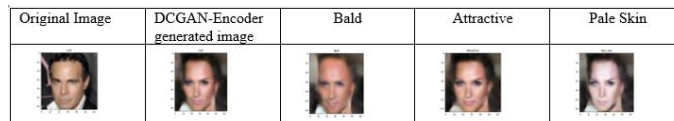

Original Image	DCGAN-Encoder generated image	Bald	Attractive	Pale Skin

Fig. 3: DCGAN [5] results for CelebA with a model trained on 40 attributes

Counterfactuals generation with DCGAN-Encoder. The qualitative results obtained from the trained DCGAN-Encoder model which were generated by first generating z encoding vectors and later decoded by the generator for obtaining the images. The generated images were not satisfactory as the images obtained from the sample-training and testing dataset were not clear enough and suffered from blur and generated an image with more feminine attributes due to bias in the dataset to detect the face attribute changes as depicted in the Fig. 3. Upon using the Encoder networks on top of the trained discriminator, to generate an image with changed attributes were not satisfactory as well. This could be possibly due to the blur and non-realistic generated image from DCGAN and getting an image with the desired attributes by using the same image. This could be seen evidently from Fig. 3.

Counterfactuals generation with StarGAN. The qualitative results obtained for StarGAN on CelebA dataset using the training parameters mentioned

Original Image	Black Hair	Blond Hair	Brown Hair	Male	Young

(a)

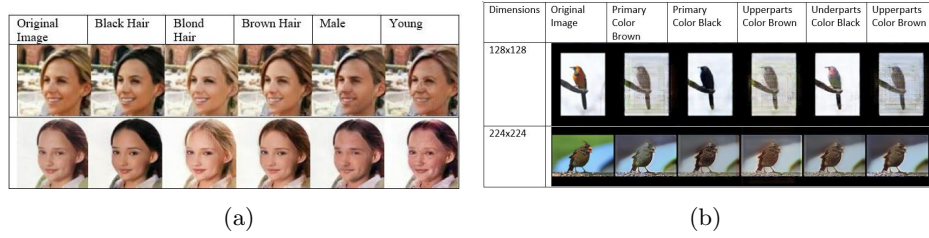| Dimensions	Original Image	Primary Color Brown	Primary Color Black	Upperparts Color Brown	Underparts Color Black	Upperparts Color Brown
128x128						
224x224						

(b)

Fig. 4: (a) StarGAN [6] results for CelebA with a model trained on 5 attributes. (b) StarGAN [6] results for CUB-200-2011 with a model trained on 5 attributes in two different dimensions.

earlier were satisfactory as depicted in Fig. 4a. StarGAN generated images by attribute manipulation that are less blurry and where different attributes can be visually identified for an image size 128x128. The qualitative results obtained for StarGAN on the CUB-200-2011 dataset were also visually satisfactory, producing better quality 128x128 and 224x224 images as in Fig. 4b.

We assessed perceptual similarity by looking at quantitative results based on LPIPS and FID score. The lower values of LPIPS and FID as represented in Fig. 5a indicates better perceptual similarity of the images generated by the StarGAN by manipulating the attributes of primary color black and primary color brown for both Black Footed Albatross and Black Tern. It is interesting to note here that LPIPS and FID values for the both the images for the attributes given in Fig. 5a was lower for the resolution 224x224 than for resolution 128x128 meaning better perceptual similarity for the images generated with higher resolution.

Counterfatuals validation. Given its superior performance, we used images generated with StarGAN for validation of visual counterfactuals generated from [1]. The StarGAN higher resolution (224x224) images were provided as inputs to the VGG-16 model to see if the wrong classification for Black Footed Albatross is flipped. Results can be seen in Fig. 5b.

Bird Class	Original Image	Generated Image	Attribute Change	LPIPS	FID
Black Footed Albatross			has_primary_color_black	• 0.1943 (128*128) • 0.1210 (224*224)	• 321.82 (128*128) • 153.00 (224*224)
			has_primary_color_brown	• 0.1936 (128*128) • 0.1242 (224*224)	

(a)

Original Image	Attribute selected	Generated Image	Classification
Black Footed Albatross	has_primary_color_brown		Horned_Lark
	has_underparts_color_black		Spotted_Catbird
	has_underparts_color_brown		Tennessee_Warbler
	has_upperparts_color_brown		Horned_Lark

(b)

Fig. 5: (a) Metric values for generated images using StarGAN on CUB-200-2011 with 128 x 128 and 224 x 224-pixel resolutions. (b) Classifier Decision for the StarGAN generated images (224 x 224)

The images generated by StarGAN are 128x128 in size, and the major drawback with the CUB-200-2011 dataset is that the images have different sizes. As a result, for the first training model on StarGAN we used image size = 128 and crop size = 512 pixels, which did not yield to good resolution images [14] for the classifier in [1] to produce accurate results. We retrained StarGAN to get images of size 224x224. Even at this resolution, the classifier [1] experiences a lot of noise and classification results look random as shown in Fig. 5b.

6 Discussion

Based on our experiments on AttGAN, StarGAN and DCGAN, we realized that the output images obtained are of relatively lower resolution, introducing a lot of noise when it comes to validating classifier's predictions. This appears to be a limitation of current GANs and a possible reason why counterfactuals do not produce the expected inverted classification.

Despite this somehow discouraging result, GAN-generated images representing counterfactual does improve the probability of the expected class in some cases. In addition, attributes for which we obtained better probabilities could be ranked as the top predictors of getting the expected class in some cases.

For example, the prediction score of the expected class improves for the image of a Black footed albatross used in [1] generated by StarGAN for the attributes $'PrimaryColor - Brown'$ and $'UpperpartsColor - Brown'$ of resolution 224x224 when compared to the same image generated by StarGAN for same attributes at 128x128 resolution. However, this was not the case for the attribute $'UnderpartsColor - Brown'$.

In case of generated image of a Black Tern used in [1] by StarGAN, the prediction score of the expected class improves for the attribute $'PrimaryColor - Black'$ of resolution 224x224 when compared to the same image generated by StarGAN for the same attribute at 128x128 resolution. However, this was not the case for the attribute $'UnderpartsColor - Black'$

Considering the obtained results and our considerations, we argue that if the image classifier was trained on lower resolution images from CUB-200-2011 (128X128 or 224X224) we might have obtained better probabilities of the expected classes and possibly invert the classification based on the counterfactual explanation. It is also interesting to note that StarGAN trained on CUB-200-2011, does not recognise the attribute $'UnderparsColor'$ very well and hence the prediction score of the expected class does not improve even when increasing the resolution of the StarGAN generated image. This means specific attributes might have more influence in a counterfactual than others.

It is worth mentioning that training GANs was very time consuming, especially without a GPU. In terms of technical challenges, when implementing GAN models on tensorflow we realized that most of the models rely on tensorflow 1.15 which is compatible with Python version 3.6 and lower. If one uses higher versions of tensorflow and Python, there are many import errors with respect to various packages like trace, absl-py etc. that require to update the code and do

a substantial amount of debugging. Additionally, working with the tensorflow implementation of AttGAN for version 1.15 installs tensorflow-estimator version 2.0 or higher by default, which creates additional version compatibility errors.

7 Conclusion and Future Work

In this paper, we evaluated the use of different GAN to generate plausible visual counterfactuals that are good enough to be used to validate textual counterfactual explanations generated in [1]. The target dataset was the CUB-200-2011. Our hypothesis was that the images generated by changing semantic attributes according to a textual counterfactual could be used to validate such counterfactual [1]. Since there is little to no research in the area of using GAN to generate visual counterfactuals from texts for their validation, we used the CelebA dataset to create a baseline pipeline for generating and modifying images based on their attributes. After creating a data-parser for the CUB-200-2011 dataset, we trained different GAN models to obtain the modified images with respect to the attributes specified in the counterfactual.

The whole experimental process was focused at achieving good resolution generated images with semantic attribute changes as per the counterfactual explanations obtained from [1] so as to assess the same by feeding the outputs back to the VGG-16 model in [1]. StarGAN produced the best results in terms of generation of the visual counterfactual, but when it comes to validation on a complex dataset of images such as CUB-200-2011, we were able to highlight some limitations of GANs for systematic validation of counterfactuals.

Specifically, beyond the technical issues of reproducibility when it comes to deploying and training GANs, the low resolution of the generated images did not allow to fully validate textual counterfactuals by obtaining an inverted classification as expected. However, we identified some opportunities due to the fact that the probability of some classes was affected, and we were also able to observe the different impact of certain attributes in the classification result. This is a first evidence that the textual counterfactuals generated by [1] captured the semantic attributes that affected the misclassification, even though the generated visual counterfactuals were not sufficient to invert the outcome correctly.

This also leads to promising avenue for future investigation. For example, training StarGAN on more attributes would provide insights on their effect on Counterfactual Validity (CV) scores [12]. CV as an additional metric would provide more evidence to the counterfactual evaluation for the model in [1]. Another observation is that using link prediction scores from [1] in the validation process would produce different counterfactual images from the trained GAN and validate the link prediction approach in [1]. Intuitively, visual counterfactuals produced by changes in semantic attributes can be a starting point for a deeper investigation of possible existing biases towards such attributes.

References

1. V. A. C. Horta, A. Mileo: Generating Local Textual Explanations for CNNs: A Semantic Approach Based on Knowledge Graphs. AI*IA 2021: 532-549
2. Singla, S., Pollack, B., Wallace, S. and Batmanghelich, K. (2021) 'Explaining the Black-box Smoothly- A Counterfactual Approach, Available at: https://arxiv.org/pdf/2101.04230.pdf
3. Jacob, P., Zablocki, E., Ben-Younes, H., Chen, M., Perez, P., Cord, M., Valeo.ai and Universite, S. (2021) 'STEEX: Steering Counterfactual Explanations with Semantics', Available at: https://arxiv.org/pdf/2111.09094.pdf
4. Goyal, Y., Feder, A., Shalit, U. and Kim, B. (2020) 'Explaining Classifiers with Causal Concept Effect (CaCE)', Available at: https://arxiv.org/pdf/1907.07165.pdf
5. R. Homero, B. Roman, M. Yang, and Zhang, "Photoshop 2.0: Generative Adversarial Networks for Photo Editing." Available at: http://cs231n.stanford.edu/reports/2017/pdfs/305.pdf
6. Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation," arxiv.org, Nov. 2017, [Online]. Available: https://arxiv.org/abs/1711.09020
7. Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial Attribute Editing by Only Changing What You Want," IEEE Transactions on Image Processing, vol. 28, no. 11, pp. 5464–5478, Nov. 2019, doi: 10.1109/tip.2019.2916751.
8. T. Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," arXiv.org, 2017. https://arxiv.org/abs/1711.10485
9. "Perona Lab - CUB-200-2011," www.vision.caltech.edu. https://www.vision.caltech.edu/datasets/.
10. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," arXiv:1706.08500 [cs, stat], Jan. 2018, [Online]. Available: https://arxiv.org/abs/1706.08500
11. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," arXiv:1801.03924 [cs], Apr. 2018, [Online]. Available: https://arxiv.org/abs/1801.03924
12. R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Jan. 2020, doi: 10.1145/3351095.3372850.
13. E. Kenny and M. Keane, "On Generating Plausible Counterfactual and Semi-Factual Explanations for Deep Learning." Accessed: Aug. 05, 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17377/17184
14. Park, M.; Lee, M.; Yu, S. HRGAN: A Generative Adversarial Network Producing Higher-Resolution Images than Training Sets. Sensors 2022, 22, 1435. https://doi.org/10.3390/s22041435
15. "Image Editing using GAN," Image-Editing-using-GAN. https://tandon-a.github.io/Image-Editing-using-GAN/ (Accessed 05 January, 2022).
16. 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