RESCIENCE C

Replication / ML Reproducibility Challenge 2021

[Re] Counterfactual Generative Networks

Ankit[1], Sameer Ambekar[1 *], Baradwaj Varadharajan[1], and Mark Alence[1 *]
[1] Hogwarts University, Hogsmeade, UK

Edited by
Koustuv Sinha

Reviewed by
Anonymous Reviewers

Received
04 February 2022

Published
15 May 2022

DOI
10.0000/zenodo.0000000

Reproducibility Summary

Scope of Reproducibility

In this paper, we attempt to verify the claims that the paper [1] makes about their proposed CGN framework that decomposes the image generation process into independent causal mechanisms. Further, the author claims that these counterfactual images improves the out-of-distribution robustness of the classifier. We use the code provided by the authors to replicate several experiments in the original paper and draw conclusions based on these results.

Methodology

We use the same hyperparameters and architecture as mentioned in CGN [1]. We use the PyTorch code from the authors' publicly available repository. We make several changes to their code for the MNIST datasets since it gives spurious results/errors. Since we use ImageNet 1000 as a replacement for the ImageNet dataset, we modify the code accordingly. We reproduce tables 1-6 from CGN [1] paper, excluding results for models from other papers.

Results

We validated each of the author's claim through the experiments given in the original paper and few additional experiments of our own. Overall, we found many experiments yielding identical results while some deviations were observed with both the Counterfactual Generative Network and the subsequent classification task. We were able to explain most of these deviations through our additional experiments while some couldn't be validated due to computational limitations.

What was easy

Overall, clear environment setup instructions, well working code and availability of pretrained CGN models for both datasets proved valuable to validate the authors' claim.

Correspondence should be addressed to Baradwaj Varadharajan (ankitnitt1721@gmail.com)
The authors have declared that no competing interests exist.
Code is available at https://github.com/ – DOI 00.0000/zenodo.0000000. – SWH swh:1:dir:aaaaaaaaaaaaaaaaaaaaaaaaaaaaaa.
Open peer review is available at https://openreview.net/forum?id=forum_id.

What was difficult

Some experimental details were not reported in the original paper which made validations time consuming. ImageNet based experiments were replaced with ImageNet-1k(mini) due to the computational limitation which made it difficult to validate the author's original claims. Pre-trained classification models could have proven helpful in this case, but were unavailable, which meant we had to train the classifier from scratch. Code changes were required to obtain baseline results which was tedious considering different code architecture was implemented for MNIST & ImageNet.

Communication with original authors

We emailed the authors regarding inception score, MNIST dataset hyperparameters and ImageNet hyperparameters. We are awaiting a response from their end.
Code available at https://anonymous.4open.science/r/re_counterfactual_generative-E18F

1 Introduction

Neural Networks (NNs) have become ubiquitous in machine learning due to their predictive power. However, a shortcoming of NNs is their tendency to learn simple correlations that lead to good performance on test data rather than more complex correlations that generalise better. This shortcoming is apparent in the task of image classification, where NNs tend to overfit to factors like background or texture. To address this shortcoming, [1] proposes a method of generating counterfactual images that prevent classifiers from learning spurious relationships.

The authors take a causal approach to image generation by splitting the generation task into independent causal mechanisms. The authors considered three separately learned Independent Mechanisms (IMs) to generate shapes, textures and backgrounds for an image. For the MNIST setting, all IM specific losses are optimized end-to-end from scratch, while in the ImageNet setting, each IM is initialized with weights from pre-trained BigGAN-deep-256[2]. The counterfactual image is then generated by passing the result of each IM to a deterministic composer function.

In this report, we use the publicly available code provided by the authors to reproduce the results of the paper and validate the authors' claims. In this endeavour, we made modifications to the code to determine the efficacy of their generative model and validate its impact on improving the out of distribution robustness of a classifier.

2 Scope of reproducibility

In this report, we investigate the following claims from the original paper:

1. Generating high-quality counterfactual images that decompose into independent causal inductive biases, these mechanisms disentangle object shape, object texture and background
2. Using counterfactual images improves the shape vs texture bias which is an inherent problem of deep classifiers
3. Using counterfactual images improve the out-of-distribution robustness for the classifier during the classification task
4. The Generative model can be trained efficiently on a single GPU with the help of powerful pre-trained models

We attempt to reproduce the experiments from the paper [1] and perform exploratory analysis on the above mentioned claims. We propose using an extra loss function to mitigate some of the shortcomings during counterfactual generation process and generate heatmap plots to study the classifier behaviour.

3 Methodology

Alex et al. [1] propose a Counterfactual Generative Network (CGN) framework to generate high-quality counterfactual images, which can be used to train invariant classifiers. The architecture of a CGN is composed of three IMs that are trained to generate backgrounds, shapes, and textures. Each IM is provided with a label. The task of the invariant classifier is to predict the label of a specific IM, regardless of the labels of the others. In conjunction with the composer function, the use of counterfactual images generated by the three IMs prevents the classifier from learning spurious relationships that arise from training on a natural dataset only.

The architecture of the CGN consists of a GAN as the backbone of each IM. Each IM samples random noise $\mu \sim$ N(0,1), along with an independently sampled label to generate samples. The output x_{gen} is generated using an analytical function from the Composer 'C',

$$x_{gen} = C(m, f, b) = m \otimes f + (1 - m) \otimes b$$

where 'm' is the mask (alpha map), f is foreground and b is background. $\otimes$ denotes the element wise multiplication.
The losses $\mathcal{L}_{rec}$ (x_{gt}, x_{gen}), $\mathcal{L}_1$ reconstruction loss, $\mathcal{L}_{perceptual}$ as shown in Fig. ?? are used to improve the quality of generated images. Once the CGN is trained, u and y are randomized per mechanism such that new counterfactual x_{gen} are generated. Furthermore, hyperparameters such as CF ratio (the ratio indicates how many counterfactuals are generated per sampled noise) can be used to control the number of samples that are being generated. These samples are then used to train the classifier and evaluated on the corresponding test set.

3.1 Model descriptions

The ImageNet variant follows the architecture that is illustrated in Fig. ??. The MNIST variant applies a simpler architecture by applying a second texture mechanism rather than a background mechanism.

3.2 Experimental setup and code

We use the datasets mentioned in [1], excluding ImageNet [3] due to limited resources and computational constraints.

Dataset	Description
Colored MNIST	Consists of digits in red or green.
Double Colored MNIST	Consists of more varied backgrounds and digits than Colored MNIST.
Wildlife MNIST	An attempt to build MNIST [4] closer to the ImageNet[3], texture was added as a bias to the data. The ten digits of the striped texture class encode the foreground lables and the background is labelled with the with the texture class 'veiny'.
ImageNet-1k(mini)	Subset of the ImageNet-1k[ImageNet-1k],available here[1] that contains 34745 images in train set and 3923 for validation set, each split among 1000 classes individually.

Table 1. Datasets used

For all the experiments, we make use of standard dataset splits akin to the CGN paper [1]. Considering the computational constraint to train a classifier on ImageNet[5], we used the pre-trained CGN to generate counterfactual images and trained a classifier on ImageNet-1k(mini) and mini-imagenet datasets.

3.3 Hyperparameter search

We found that the hyperparameters provided by the authors were stable, and so we did not conduct a hyperparameter search in this report.

[01] https://kaggle.com/ifigotin/imagenetmini-1000

3.4 Computational requirements

All models are run on Nvdia GTX1080Ti GPUs (11Gb VRAM). For the MNIST datasets, training a CGN and a classifier each took approximately one hour.

4 Results

A lack of compute power prevented us from replicating the experiments on ImageNet. As a workaround, we limit ourselves to verifying the results using the ImageNet-1k(mini) dataset. This is beneficial because it extends the results of the paper and evaluates the method on a new dataset, and ensures that results can be reproduced with limited resources by referring to our report/code and the CGN paper.

4.1 Results reproducing original paper

Can Image generation process be decomposed into independent causal inductive biases effectively? — We begin the experiment by training a CGN on the three variants of the MNIST dataset. We observe in Fig. ?? that the digits in case of colored MNIST dataset lose their shape when reconstructed, whereas for double colored and wildlife MNIST, the digits look much better. Since we do not clearly understand why the shape in Colored MNIST is poor, we generated a mask timeline to verify any patterns. Fig. ?? details the same. Further analysis on this was conducted and recorded in ??. We also propose an additional loss function to help mitigate this problem.
Quality of Counterfactual Images on ImageNet-1k
To quantify the quality of the composite images produced by the CGN, the authors calculate the inception score (IS). The details of the IS calculations (inception model used, number of images used) were not mentioned in the paper. In an attempt to recreate the results regarding IS, we use the OpenAI implementation [1]. We plot the results of IS vs the number images using 10 splits in Fig. ??. We observe the IS converges to an IS of 198.
We made use of the pre-trained CGN trained on ImageNet-1k that was present as part of the codebase to generate counterfactual images. Since there is no quantitative way to measure the quality of counterfactual images, we reproduced the images given in the original paper. We achieved a similar quality of counterfactual images but also noted deviations. Fig. ?? shows all the images that were given in the original paper. A deviation in the mask is observed for the class 'Agaric' and 'Cauliflower'. The difference in the images to the original paper prompted us to collect the classes with poorer counterfactual images to observe any patterns.
Fig. ?? is generated from the pre-trained CGN that have a low quality of images picked from random classes. Since the analysis is qualitative, we relied on the realism of the counterfactual compared to original images from that class. Images under the classes 'Cliff dwelling' 'American Chameleon' suffer from Texture-background entanglement resulting in the counterfactual with no subject. On the other hand, the images under the class 'Goldfinch', 'Junco' suffer from reduced realism due to linear constraints applied on the composer.

References

1. A. Sauer and A. Geiger. "Counterfactual generative networks." In: *arXiv preprint arXiv:2101.06046* (2021).
2. A. Brock, J. Donahue, and K. Simonyan. *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. 2019. arXiv:1809.11096 [cs.LG].

[1] https://github.com/nnUyi/Inception-Score

3. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "Imagenet: A large-scale hierarchical image database." In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
4. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.
5. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. "ImageNet: a Large-Scale Hierarchical Image Database." In: June 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.