Reproducibility Study of "Counterfactual Generative Networks"

Piyush Bagad Paul Hilders Jesse Maas Danilo de Goede

University of Amsterdam

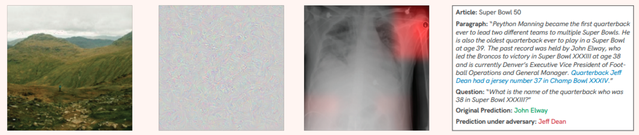
NEURAL INFORMATION PROCESSING SYSTEMS

Best Paper Award (MLRC 2021)

UNIVERSITY OF AMSTERDAM

Context and Scope

- Despite the considerable popularity of deep learning models within the field of artificial intelligence, recent literature suggests that these networks have a tendency of learning simple correlations that perform well on a benchmark dataset, instead of more complex relations that generalize better [1, 3, 4].
- This phenomenon, which is referred to as shortcut learning by [2], makes these models more sensitive to input perturbation and unseen input contexts.



Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

Figure 1. **Shortcut learning.** Deep neural networks have a tendency to solve problems by taking shortcuts instead of learning the intended solution, leading to a lack of generalisation and unintuitive failures [2].

In order to enhance the robustness and interpretability of classifiers, Sauer and Geiger [5] introduce the idea of a *Counterfactual Generative Network* (CGN). Using appropriate inductive biases to disentangle separate components within the input images, this model is capable of augmenting training data with generated counterfactual images.

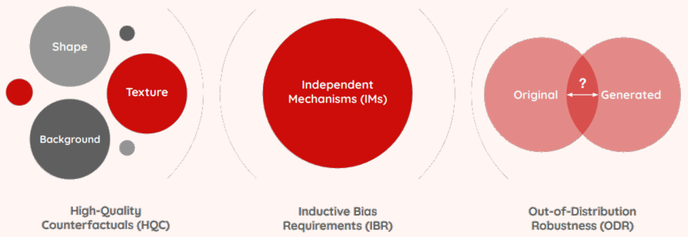


Figure 2. **Scope of Reproducibility.** In our reproducibility study, our main goal is to verify the three main claims of the original paper.

Counterfactual Generative Network

The counterfactual generative network (CGN) decomposes the image generation process into four indepdent mechanisms (IMs) whose losses are jointly optimized in an end-to-end matter: the shape mechanism, the texture mechanism, the background mechanism, and the composer.

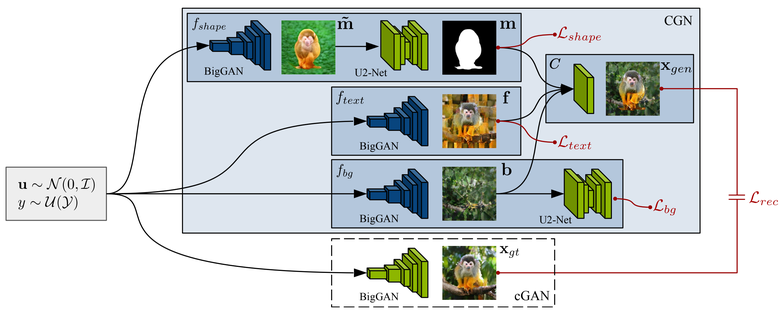


Figure 3. **CGN architecture.** Components with trainable parameters are blue, components with fixed parameters are green [5]. The dotted lines indicate that the cGAN is only used for training [5].

Reproducibility Results

1. Evaluating Counterfactual Samples

- To verify claim HQC, we qualitatively evaluate counterfactuals generated using CGN models on MNIST and ImageNet.

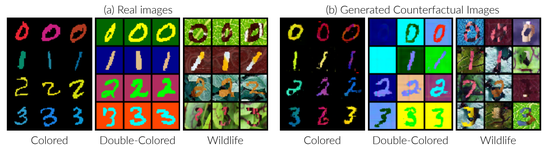


Figure 4. **Qualitative Analysis MNIST.** Left: Samples from the different MNIST variations. Right: Counterfactuals generated by the CGN on MNIST variants. Notice that the CGN generates varying shapes, colors, and textures.

Shape	Racer	Trench coat	Turtle	Vase	Malinois	Barrel
Texture	Clock	Cab	Cauliflower	Elephant	Viper	Piggy bank
Background	Toucan	Coral reef	Mushroom	Alp	Spider	Ibex

Figure 5. **Qualitative Analysis ImageNet.** Counterfactuals generated by the CGN on ImageNet.

2. Evaluating Loss Ablation

- Our loss ablation study results follow similar trends to those reported in the original paper.
- However, when disabling the texture loss, we found μ_{mask} to be 0.4, whereas the original paper reported a value of 0.9, which is an essential to support claim IBR.
- Nonetheless, we were able to support this claim by performing an additional qualitative experiment.

$\mathcal{L}_{shape}$	$\mathcal{L}_{text}$	$\mathcal{L}_{bg}$	$\mathcal{L}_{rec}$	IS ⇑		μ_{mask}	
✗	✓	✓	✓	100.8	85.9	0.3	0.2
✓	✗	✓	✓	186.5	198.4	0.4	0.9
✓	✓	✗	✓	200.9	195.6	0.1	0.1
✓	✓	✓	✗	19.3	38.4	0.4	0.3
✓	✓	✓	✓	156.1	130.2	0.3	0.3
BigGAN (Upper Bound)				202.9		-	

Figure 6. **Loss Ablation Study.** We turn off one loss at the time.

3. Evaluating Invariant Classifiers

- To evaluate the invariance in classifier heads on IN-mini, we reproduce the experiment regarding shape bias from the original paper.
- Additionally, we replicate the experiment regarding the evaluation of background robustness.
- For both experiments, we get different numbers than those reported in the original paper. Nonetheless the overall trend does support claim ODR.

Trained on	Shape Bias	top-1 ⇑	top-5 ⇑
IN + GCN/Shape	54.8		
IN + GCN/Text	16.7	74.0	91.7
IN + GCN/Bg	22.9		
IN-mini + GCN/Shape	49.1		
IN-mini + GCN/Text	20.5	56.2	79.1
IN-mini + GCN/Bg	25.7		

(a) **Shape vs. texture.** Evaluation of shape biases of independent classifiers.

Trained on	IN-9 ⇑	Mixed-Same ⇑	Mixed-Rand ⇑	BG-Gap ⇓
IN	95.6	86.2	78.9	7.3
SIN	89.2	73.1	63.7	9.4
IN + SIN	94.7	85.9	78.5	7.4
Mixed-Rand	73.3	71.5	71.3	0.2
IN + CGN	94.2	83.4	80.1	3.3
IN-mini + CGN	86.8	73.2	68.3	4.9

(b) **Backgrounds Challenge.** Evaluation of robustness against adversarially chosen backgrounds.

Results Beyond Original Paper

I. Improving CGN Training

- We predict digit masks collapse to erronous state during CGN training.
- We propose an edge-loss regularizer over predicted masks that improves training.

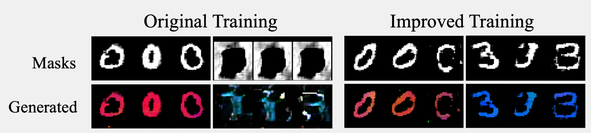


Figure 8. Adding the edge loss significantly improves CGN training on colored MNIST.

II. Explanability Analysis

- We visualize feature space using t-SNE & also visualize class activations using GradCAM.

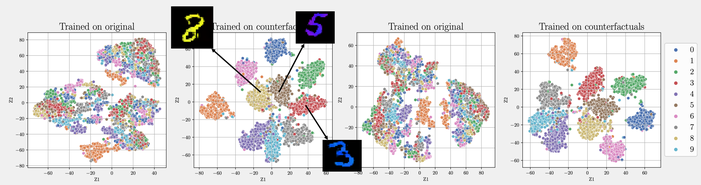


Figure 9. Feature space visualization of a CNN classifier trained on on colored MNIST variants

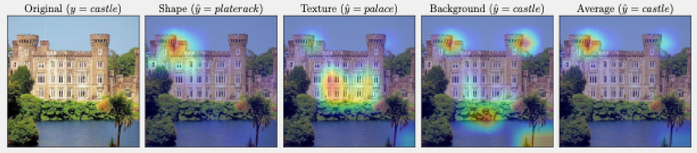


Figure 10. Visualization of GradCAM heatmaps for an example from IN-mini dataset.

III. Out-of-distribution Generalization

- We probe classifiers trained on counterfactuals to evaluate out-of-distribution robustness.

Model	Pretrained	Finetuned	IN-mini ⇑	IN-A ⇑	IN-Sketch ⇑	IN-Stylized ⇑
ResNet-50	IN-1k	-	75.580	3.400	24.092	19.218
CGN Ensemble	IN-1k	IN-mini + CF	56.793	1.387	11.775	17.188

Figure 11. Comparison of top-1 accuracy of invariant classifier with pretrained ResNet on OOD benchmarks

Key Takeaways and Limitations

- We largely succeeded in reproducing qualitative & quantiative results from CGN.
- Our additional experiments dig deeper into the utility of counterfactuals to train robust classifiers.

Limitations

- Since our experiments are on IN-Mini, it is not possible to reproduce the exact numbers.
- Some experimental details are unclear in CGN forcing us to use the default configurations.

References, Code and Paper

[1] M. A. Alcorn, Q. Li, Z. Gong, C. Wang, L. Mai, W.-S. Ku, and A. Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects, 2019.

[2] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[3] Y. Ming, H. Yin, and Y. Li. On the impact of spurious correlation for out-of-distribution detection, 2021.

[4] A. Rosenfeld, R. Zemel, and J. K. Tsotsos. The elephant in the room, 2018.

[5] A. Sauer and A. Geiger. Counterfactual generative networks. In *International Conference on Learning Representations (ICLR)*, 2021.